THÈSE
présentée par

François Bérard

pour obtenir le titre de
DOCTEUR de L'UNIVERSITE JOSEPH-FOURIER-GRENOBLE I
(arrêtés ministériels du 5 juillet 1984 et du 30 mars 1992)
Spécialité : Informatique

Vision par ordinateur pour l'interaction homme-machine fortement couplée

Composition du Jury :

Directeur de thèse :	Mme. Joëlle Coutaz
Co-Directeur de thèse :	Mr. James L. Crowley
Rapporteurs :	Mme. Monique Thonnat Mr. Michel Beaudouin-Lafon
Jury :	Mr. Jean-Pierre Verjus Mr. Michael J. Black Mr. Giorgio Faconti

Thèse préparée au sein du laboratoire de Communication Langagière et
Interaction Personne-Système
Fédération IMAG
Université Joseph Fourier - Grenoble I

à mes parents...

Remerciements

L'environnement que Joëlle a créé (communément appelé "équipe IHM"), est une sorte de paradis du thésard. On y travaille beaucoup, mais sans avoir l'impression de travailler : ce qu'on y fait est tellement chouette. On s'y sent bien, sans doute grâce à l'intérêt que porte Joëlle à tout ce qu'on dit et ce qu'on fait. Merci Joëlle pour le respect que tu nous portes et ton sens des responsabilités à notre égard. Merci pour ton optimisme et tes idées, pour les moyens dont tu nous fait bénéficier. Merci, enfin, pour l'aide que tu m'as apportée : quand il le fallait, et sans compter.

Merci Jim d'avoir été à l'écoute, d'avoir fait l'effort de m'expliquer et me convaincre, merci pour tes cours particuliers, pour les opportunités dont tu m'a fait profiter.

Merci à Alex Pentland pour m'avoir accueilli dans un laboratoire et une équipe "tip top", pour m'avoir permis de découvrir l'efficacité et le pragmatisme "à l'américaine".

Merci à Michael J. Black, pour son accueille dans un laboratoire non moins "tip top", pour l'exemple que représente son approche sans concession de la recherche : sobre et rigoureuse. Merci d'avoir accepté de participer à mon Jury de thèse et pour l'intérêt porté à mon travail.

Merci à Michel Beaudouin-Lafon et Monique Thonnat d'avoir accepté la charge de rapporter cette thèse, merci pour l'intérêt qu'ils y ont porté et pour leurs remarques.

Merci à Giorgio Faconti d'avoir participé à mon Jury de thèse et pour l'intérêt porté à mon travail.

Merci à Jean-Pierre Verjus de m'avoir miraculeusement fait l'honneur de présider mon Jury de thèse.

Merci aux étudiants qui ont du me supporter dans leur projet, et ont malgré tout réussi un super travail : Franck et Fabrice, et Seb, Benny et Michel. Les travaux présentés ici sont aussi le résultat de leurs efforts.

Je tiens également à remercier Jacques Davy, responsable au C.N.E.T. du projet COMEDI, pour son intérêt et pour le soutien financier dont ce travail a bénéficié.

Merci à tous les amis, au labo et ailleurs, pour avoir fait que ces années de thèse soient des années géniales. Merci à Jef, Jean-Pascal, Karine, Arnaud, Franck (Freu), Françoise, Franck & Ziz, Philou, Sandrine, Daniel, Laurence, Seb, Steph, Éric, Céline, Deb, Nuria, Fred (Freu), Fred, Hubert, Mickaële, Anne, Dave, Jean-Luc, Stella, Céline (une autre), Manu, William, Laure, Raph, Sandrine, Yann, Leon, Nick, Gaëlle, Alex, Ninie, Chris, Philippe, Céline (encore une !!), et les autres...

Merci à ma famille, Mireille, Marie, Elisabeth et Allain (avec 2 "l"). Merci d'avoir supporté mon indisponibilité et mon ingratitude sans le moindre reproche. Merci d'être toujours là, et surtout quand ça va mal.

et merci à ma Marie...

Table des matières

TABLE DES MATIÈRES VII

TABLE DES FIGURES XI

INTRODUCTION 1

1.Contexte : vision par ordinateur et interaction fortement couplée 2
1.1.Vision par ordinateur : une opportunité pour de nouvelles formes d'interaction 2
1.2.Concept d'interaction fortement couplée 3

2.Sujet et approche de recherche 3
2.1.Approche en interaction homme-machine 4
2.2.Approche en vision par ordinateur 4
2.3.Notre approche 4

3.Structure du mémoire 6

CHAPITRE I INTERACTION FORTEMENT COUPLÉE : DÉFINITION ET EXEMPLES 7

1.Définition et modélisation 9
1.1.Définition 9
1.2.Modélisation : principe 9
1.3.Modèle affiné 10

2.Systèmes immersifs 11
2.1.Venez tel quel : VideoPlace et ALIVE 12
2.2.Perception de l'espace en trois dimensions 17
2.3.Synthèse sur les systèmes immersifs 24

3.Systèmes de Réalité Augmentée 25
3.1.Interfaces saisissables 26

3.2.Interfaces digitales 33
4.Résumé du chapitre 36

CHAPITRE II REQUIS DE L'INTERACTION FORTEMENT COUPLÉE 39

1.Requis fonctionnels 39
1.1.Espace des services 40
1.2.Détection 41
1.3.Identification 41
1.4.Suivi 41
2.Requis non fonctionnels 41
2.1.Système en boucle fermée et latence 42
2.2.latence de l'utilisateur 43
2.3.latence du système 47
2.4.Synthèse : requis pour la latence du dispositif 53
2.5.Qualité des informations rendues par le dispositif 55
3.Résumé du chapitre 59

CHAPITRE III VISION PAR ORDINATEUR : PROBLÉMATIQUE ET APPROCHES 61

1.Problématique de la vision par ordinateur 61
1.1.Le domaine 61
1.2.Difficultés 63
1.3.Constat 65
2.Approches en vision par ordinateur 66
2.1.Vision orientée modèle 67
2.2.Vision par apparence 69
2.3.Notre approche : vision par apparence centrée sur la tâche utilisateur 71
3.Résumé du chapitre 73

CHAPITRE IV TECHNIQUES DE SUIVI EN VISION PAR ORDINATEUR 75

1.Suivi d'entité 75
1.1.Principe 76
1.2.Mesure 76
1.3.Estimation de la position 77
1.4.Validation 80
1.5.Prédiction 81
2.Suivi par différence d'images 83
2.1.Principe 83
2.2.Réalisation 84
2.3.Performances 86
2.4.Discussion 87
3.Suivi par modèle de couleur 88
3.1.Principe 88
3.2.Réalisation 89
3.3.Performances 93
3.4.Discussion 94
4.Suivi par corrélation 95
4.1.Principe 95

4.2.Réalisation 96
4.3.Performances 100
4.4.Discussion 101

5.Coopération de techniques 103
5.1.Architectures pour la coopération de techniques de suivi 103
5.2.Notre suivi de visage par coopération de techniques 107

6.Résumé du chapitre 112

CHAPITRE V LE TABLEAU MAGIQUE 113

1.Motivations 114
1.1.Adéquation du tableau blanc 114
1.2.Insuffisances 117

2.Le système 120
2.1.Appareillage 121
2.2.Transformation entre repères 122
2.3.Capture des inscriptions 124
2.4.Suivi du doigt 130
2.5.Interaction 136

3.Évaluation 138
3.1.Remarques générales 139
3.2.Interaction fortement couplée 141

4.Résumé du chapitre 142

CHAPITRE VI LA FENÊTRE PERCEPTUELLE 145

1.Motivations 146
1.1.Parallélisation des actions et minimisation des mouvements 146
1.2.Aspect cognitif 148
1.3.Suppression des intermédiaires de l'interaction 150

2.Le système 151
2.1.Dispositif 152
2.2.Suivi du visage 152
2.3.Interaction 156

3.Performances utilisateur pour une tâche exploratoire 162
3.1.Motivations 163
3.2.Protocole expérimental 163
3.3.Résultats 165
3.4.Discussion 166

4.Performances utilisateur pour une tâche de glisser-déposer 167
4.1.Motivations 167
4.2.Protocole expérimental 168
4.3.Résultats 171
4.4.Discussion 171

5.Résumé du chapitre 173
5.1.Faisabilité 173
5.2.Avantages 173

CONCLUSION 175

1.Résumé de la contribution 175

2.Originalité et points forts 176
2.1.Approche scientifique 176

3.Limites et Perspectives 178
3.1.Limites et perspectives à court terme 178
3.2.Perspectives à moyen terme 179

ANNEXE A TECHNIQUES DE VISION PAR ORDINATEUR À USAGE GÉNÉRAL 181

1.Seuillage 181
1.1.Principe 181
1.2.Choix du seuil 182

2.Analyse en composantes connexes 182
2.1.Définition 182
2.2.Implémentation efficace 183

3.Calcul de validité 183

ANNEXE B CONSIDÉRATIONS D'IMPLÉMENTATION 185

1.Bibliothèque de services 185
1.1.Apport 186
1.2.Constat 186
1.3.Notre approche 187

2.Acquisition du flux vidéo 188
2.1.Indépendance vis-à-vis du matériel 188
2.2.Format du flux vidéo 189

BIBLIOGRAPHIE 191

Table des figures

TABLE DES MATIÈRES VII

TABLE DES FIGURES XI

INTRODUCTION 1

CHAPITRE I INTERACTION FORTEMENT COUPLÉE : DÉFINITION ET EXEMPLES 7

1 Interaction fortement couplée modélisée par un système en boucle fermée 9
2 Boucle de l'interaction fortement couplée (d'après [Ware 94]) 10
3 Le dispositif de "**VideoPlace**" (d'après [Krueger 90]) 12
4 Une image générée dans "**VideoPlace**" (extrait de [Krueger 90]) 13
5 L'interaction "Critter" de **VideoPlace** (extrait de [Krueger 90]) 13
6 Le système **ALIVE** (extrait de [Maes 94]) 15
7 Parallaxe par mouvement (extrait de [Voorhorst 98]) 18
8 Dispositif de type "fenêtre virtuelle" (extrait de [Voorhorst 98]) 21
9 Réalité Virtuelle de type "aquarium" (extrait de [Ware 93]) 22
10 Principe d'une interface saisissable (extrait de [Fitzmaurice 95]) 26
11 Aspect direct de l'interaction fondée sur les interfaces saisissables (d'après [Fitzmaurice 96]) 27
12 Multiplexage spatial extrême des fonctions : une table de mixage 28
13 Interaction à deux mains (extrait de [Fitzmaurice 95]) 28
14 L'application **GraspDraw** (extrait de [Fitzmaurice 96]) 30
15 Le système **MetaDESK** (extrait de [Ullmer 97]) 31
16 Aspect direct de l'interaction fondée sur les interfaces digitales 33
17 Sélection au doigt sur le **Bureau Digital** (extrait d'une vidéo non publiée de Xerox EuroPARC) 34

CHAPITRE II REQUIS DE L'INTERACTION FORTEMENT COUPLÉE 39

1 Boucle de l'interaction fortement couplée (d'après [Ware 94]) ... 42
2 Représentation schématique du modèle du processeur humain (d'après [Card 83]) ... 44
3 Acquisition d'une cible de largeur L à une distance D du pointeur ... 46
4 Latence utilisateur et latence dispositif ... 49

CHAPITRE III VISION PAR ORDINATEUR : PROBLÉMATIQUE ET APPROCHES 61

1 Image hors-contexte d'un objet ... 64
2 Reconstruction 3D (extraite de [Socher 95]) ... 67
3 Identification des visages par la techniques des eigenfaces ... 70
4 L'objet de la figure 1 page 64 replacé dans son contexte ... 74

CHAPITRE IV TECHNIQUES DE SUIVI EN VISION PAR ORDINATEUR 75

1 Statistiques du second ordre de la distribution spatiale des pixels ... 79
2 Différence entre images successives ... 84
3 Différence par rapport à une image de référence ... 85
4 Frontière de la cible (un doigt) ... 86
5 Constitution du modèle de couleur par l'exemple ... 90
6 Modèle histogramme (a, c) et modèle gaussien (b, e) ... 91
7 Suivi du visage par modèle de couleur de peau (modèle gaussien) ... 93
8 Conservation de l'apparence d'une entité durant une translation parallèle au plan de la caméra ... 95
9 Suivi par corrélation ... 96
10 Déplacement maximal de l'entité en fonction de la taille de la zone de recherche ... 98
11 Vitesse maximale de l'entité (modèle et mesure) et fréquence de fonctionnement en fonction de la taille (t) de la zone de recherche ... 99
12 Architecture pour le Focus d'Attention Incrémental (FAI, d'après [Toyama 96]) ... 104
13 Architecture SERP (d'après [Crowley 94a]) ... 105
14 Détection des yeux par différence d'images pour l'initialisation du modèle de couleur de peau ... 108
15 Automate de contrôle de la coopération des techniques de suivi ... 109
16 Fréquence de fonctionnement des différentes techniques de suivi coopérantes ... 110

CHAPITRE V LE TABLEAU MAGIQUE 113

1 L'interface de **Tivoli** (extrait de [Pedersen 93]) ... 116
2 Capture de la trajectoire de l'outil de dessin : problèmes de résolution spatiale et temporelle 119
3 L'appareillage du **tableau magique** ... 121
4 Image de mire projetée (a) et capturée (b) ... 123
5 Analyse en composantes connexes de l'image de différence de la mire ... 124
6 Image "brute" capturée par la caméra ... 125
7 Seuillage adaptatif ... 125
8 Principe de l'assemblage de la mosaïque ... 127
9 Exemple de capture du **tableau magique** ... 129
10 Rapport des tailles de cible, motif, et zone de recherche ... 131
11 Variation d'apparence de l'index en fonction de l'orientation du bras ... 132
12 Énergie de la zone sensible lors d'une activation du suivi ... 135

CHAPITRE VI LA FENÊTRE PERCEPTUELLE 145

1 Tâches des expériences de Buxton et Myers ([Buxton 86]) ... 147
2 Illustration de l'aspect direct de l'interaction avec la **fenêtre perceptuelle** ... 151
3 Le dispositif de la **fenêtre perceptuelle** ... 152
4 Translation du motif du suivi ... 153
5 Résolution du système de suivi ... 154
6 Courbure de la fonction de transfert en fonction du paramètre d'accélération *a* ... 159
7 Forme de la fonction de transfert verticale ... 160
8 Tâche exploratoire au moyen de la **fenêtre perceptuelle** avec contrôle de la vitesse de défilement ... 164
9 Moyennes des temps d'accomplissement de la tâche et leurs écarts-type en fonction des conditions expérimentales ... 165
10 Fin de navigation de la tâche glisser-déposer ... 168
11 Vue radar active ... 169
12 Moyennes des temps d'accomplissement de la tâche et leurs écarts-type en fonction des conditions expérimentales ... 170

CONCLUSION 175

ANNEXE A TECHNIQUES DE VISION PAR ORDINATEUR À USAGE GÉNÉRAL 181

1 Analyse en composantes connexes ... 182

ANNEXE B CONSIDÉRATIONS D'IMPLÉMENTATION 185

1 Entrelacement dans un flux vidéo analogique ... 189

BIBLIOGRAPHIE 191

Introduction

Nos travaux de recherche doctorale s'inscrivent dans un mouvement de convergence entre l'interaction homme-machine et la vision par ordinateur.

L'interaction homme-machine a longtemps confiné ses recherches au développement de techniques fondées sur l'usage du triplet écran-clavier-souris. Aujourd'hui, elle s'oriente vers de nouveaux paradigmes : l'utilisateur doit pouvoir évoluer sans entraves dans son milieu naturel ; les doigts, la main, le visage ou les objets familiers sont envisagés comme autant de dispositifs d'entrée / sortie ; la frontière entre les mondes électronique et physique tend à s'estomper. Ces nouvelles formes d'interaction nécessitent le plus souvent la capture du comportement observable de l'utilisateur et de son environnement. Elles s'appuient pour cela sur des techniques de perception artificielle, et notamment de vision par ordinateur.

Pour sa part, la vision par ordinateur tire profit de la croissance en puissance de calcul des processeurs, de la miniaturisation et de la banalisation des dispositifs d'acquisition. On l'associe le plus souvent aux travaux en robotique avec la mise en œuvre de systèmes d'inspection de produits et le pilotage d'engins autonomes. Mais depuis cinq ans environ, la recherche en vision par ordinateur voit dans l'interaction homme-machine une source fertile d'expérimentations. Cependant, on vise l'élaboration de modèles génériques à fondements mathématiques sans pour autant se préoccuper des critères d'utilisabilité identifiés en interaction homme-machine. Il en résulte des techniques valides sur le plan théorique mais qui ne remplissent pas nécessairement les conditions d'exécution en milieu naturel pour un usage courant. Notre objectif vise à combler cette insuffisance.

En somme, la vision par ordinateur comme technique d'interaction offre de nouvelles opportunités, mais ces possibilités d'ouverture doivent répondre à des requis bien identifiés. Dans le cadre de cette thèse, nous considérons les requis que doit satisfaire la vision par ordinateur au service de l'interaction fortement couplée. Nous développons ci-dessous ces deux points qui constituent le contexte motivant de nos travaux.

1. Contexte : vision par ordinateur et interaction fortement couplée

1.1. VISION PAR ORDINATEUR : UNE OPPORTUNITÉ POUR DE NOUVELLES FORMES D'INTERACTION

Techniquement, la vision par ordinateur nécessite l'utilisation de caméras reliées à une unité de calcul et de mémoire. Alors que les dispositifs usuels de l'interaction, clavier et souris, nécessitent une action humaine explicite, la caméra assure un échantillonnage sans contact. Cette complémentarité avec les dispositifs actionnables ouvre, comme le microphone et la reconnaissance de la parole, de nouvelles formes d'interaction.

En particulier, la vision par ordinateur permet :

- d'étendre nos capacités visuelles par délégation à la machine de tâches d'observation hors de notre champ de vision ou difficiles pour notre système visuel et cognitif : l'analyse de scène et la télésurveillance sont des exemples où la vision par ordinateur intervient comme prothèse visuelle ;
- d'interagir avec le système de manière non intrusive, c'est-à-dire sans "fil à la patte", ouvrant ainsi la voie à la disparition de la station de travail telle que nous la connaissons aujourd'hui ;
- de servir au plus près la notion d'engagement dans l'action en éliminant les dispositifs intermédiaires lorsqu'ils s'avèrent inutiles. Par exemple, le doigt, dont la trajectoire est suivie par un système de vision, peut agir en direct sur la représentation d'un concept. Dans les interfaces à manipulation directe usuelles, la souris est un intermédiaire imposé. Avec un système de suivi par vision artificielle, le doigt *est* le dispositif d'action. L'intermédiaire physique, surplus encombrant, est éliminé.

C'est ce dernier point, *l'engagement direct de l'utilisateur dans l'action*, qui justifie mes travaux de recherche en relation avec *la vision par ordinateur*. Mais pour fonctionner convenablement, l'engagement direct suppose un couplage étroit entre les actions de l'utilisateur et les réactions du système. C'est ce que Shneiderman désigne sommairement par "retour d'information immédiat"[1] [Shneiderman 87]. À la notion de retour

d'information qui ne traduit que le côté système des phénomènes observables, nous substituons celle d'*interaction fortement couplé*e.

1.2. CONCEPT D'INTERACTION FORTEMENT COUPLÉE

L'interaction entre l'utilisateur et son système se modélise le plus souvent sous forme d'une décomposition hiérarchique de tâches en sous-tâches et ceci de manière itérative jusqu'aux tâches élémentaires ([Card 83][1], [Scapin 89], [Hartson 90]). Une tâche se définit par le couple "but, procédure" où le but désigne l'état cible recherché et la procédure les moyens pour l'atteindre. Une procédure s'exprime en termes de sous-tâches organisées selon des relations temporelles et structurelles. Une sous-tâche est élémentaire lorsqu'elle recouvre un but terminal et que sa procédure est constituée d'actions.

Une action est une opération terminale pour un niveau de granularité de description donné. Au niveau de granularité le plus fin, une action s'appelle "action physique". Une *action physique* est une opération indivisible observable par un dispositif artificiel ou naturel et qui provoque une transition d'état de ce dispositif. Par exemple, enfoncer et relâcher une touche du clavier sont deux actions physiques utilisateur qui affectent l'état du système.

Dans le cadre de cette thèse, nous considérons les relations de dépendance entre les actions physiques humaines observées par le système et les actions physiques produites en retour par le système et observées par le dispositif naturel qu'est l'œil humain. C'est sur le *couplage homme-machine au niveau des actions physiques* que nous concentrons notre recherche en vision par ordinateur. Le concept d'interaction fortement couplée désigne une phase de l'interaction entre l'homme et la machine durant laquelle la dépendance mutuelle est maintenue pendant toute la durée d'exécution d'une tâche élémentaire. C'est le cas par exemple lorsque l'utilisateur contrôle la souris afin de déplacer son pointeur à un emplacement précis.

2. *Sujet et approche de recherche*

Le sujet de nos travaux de recherche est la conception et le développement de techniques de vision par ordinateur qui répondent aux requis de l'interaction fortement couplée.

1. "immediate feedback"

1. "The GOMS Model of Manuscript Editing"

2.1. APPROCHE EN INTERACTION HOMME-MACHINE

L'approche de conception classique en interaction homme-machine consiste, en une première étape, à cerner les besoins par une étude des utilisateurs et de leur activité. En particulier, cette étude fait abstraction de tout problème d'implémentation. Une fois que les besoins sont établis, on passe à l'étape de réalisation du système. C'est donc une approche de type "top-down" au sens des besoins. L'intérêt de cet approche est d'assurer que la conception est dirigée en fonction des besoins des utilisateurs, et non pas en fonction des services disponibles a priori sur le système.

Toutefois, il est fréquent de découvrir dans l'étape de réalisation que les besoins identifiés dans la première étape se révèlent techniquement irréalisables ou nécessitant un effort de développement trop important. Ce fut le cas par exemple pour les système de **bureau digital** ([Wellner 93b]) et de **fenêtre virtuelle** ([Gaver 95]). Ces deux systèmes s'appuient sur une interaction fortement couplée fondée sur la vision par ordinateur. Ne disposant pas des services de vision par ordinateur adéquats, Wellner et Gaver développent tout deux une solution ad-hoc qui ne répond pas réellement aux besoins. De fait, leurs systèmes restent à l'état de prototype et leur apport ne peut être confirmé par l'usage.

2.2. APPROCHE EN VISION PAR ORDINATEUR

De son côté, le domaine de vision par ordinateur, poussé par les progrès scientifiques et technologiques récents, s'oriente vers l'analyse en temps réel des scènes comportant des sujets humains. Ces travaux trouvent naturellement leur application en interaction homme-machine. Néanmoins, les applications imaginées ne sont, bien souvent, que de simples faire-valoir des techniques de vision par ordinateur sous-jacentes. Le défaut d'une telle approche, de type "bottom-up", est de concevoir des techniques qui ne répondent pas aux besoins réels des systèmes interactifs. Par exemple, les travaux en vision par ordinateur mettent l'accent sur l'exactitude des techniques alors que les besoins prioritaires concernent un faible temps de réponse et la stabilité des données. De fait, les systèmes interactifs développés ne sont pas utilisables et, une fois encore, leur apport ne peut être confirmé par l'usage.

2.3. NOTRE APPROCHE

Notre approche est dirigée par le principe de conception centrée sur l'utilisateur. En cela, notre approche est profondément ancrée dans le domaine de l'interaction homme-machine. Cependant, lorsque nous sommes confronté au besoin d'un service de vision par ordinateur qui n'est pas immédiatement disponible, nous poursuivons l'analyse dans le domaine de la vision par ordinateur. Cette analyse a pour but de concevoir des techniques de vision adaptées aux besoins afin de réaliser des systèmes réellement utilisables. Ces systèmes sont le support d'une validation expérimentale par confrontation avec les utilisateurs.

Identification des besoins

La première étape de notre approche consiste à cerner les besoins. Nous basons notre analyse sur une revue de littérature concernant un ensemble de systèmes interactifs fondés sur de nouveaux paradigmes d'interaction. Notre revue se limite aux systèmes qui mettent en évidence des situations d'interaction fortement couplée où la vision par ordinateur est susceptible d'apporter une solution adéquate. Cette revue de littérature nous permet d'exhiber les fonctions que doivent réaliser les dispositifs d'entrée de ces systèmes. Ces fonctions constituent les requis fonctionnels de l'interaction fortement couplée.

Nous proposons alors une modélisation de l'interaction fortement couplée en nous basant sur certains résultats de psychologie expérimentale. Cette modélisation nous permet d'exhiber un ensemble de requis quantitatif sur les propriétés des dispositifs d'entrée. Cet ensemble constitue les requis non fonctionnels de l'interaction fortement couplée.

Conception centrée sur la tâche

Ayant établi l'ensemble de requis nécessaire à la réalisation d'une interaction fortement couplée, la deuxième étape de notre approche consiste à concevoir et développer des techniques de vision par ordinateur qui satisfont ces requis.

L'étude des deux approches classiques en vision par ordinateur, vision orientée modèle et vision par apparence, nous amène à choisir la vision par apparence car ses traitements sont en règle générale de complexité moindre et permettent d'envisager une exécution en temps réel. Nous affinons l'approche de vision par apparence par le principe de conceptions centrée sur la tâche. En particulier, nous centrons notre recherche sur l'extraction de l'information minimale nécessaire à la satisfaction du besoin exprimé par la tâche. De plus, nous autorisons l'introduction de contraintes sur la scène traitée afin de faciliter la résolution des problèmes de vision. Seules les contraintes qui n'ont pas d'incidence sur la tâche sont introduites.

Validation expérimentale

La dernière étape de notre approche est une validation expérimentale. Le sujet de nos travaux concernant l'interaction homme-machine, la validation expérimentale met en jeux la confrontation de nos systèmes face à des utilisateurs "neutres", c'est à dire des utilisateurs n'ayant pas participé au développement du système. La "neutralité" des utilisateurs est essentielle, car elle permet l'observation de comportements non biaisés par la connaissance des forces et faiblesses du système.

Les résultats qualitatifs et quantitatifs de l'expérimentation permettent de confirmer ou de remettre en question l'approche, les raisonnements et les techniques qui ont conduit à la réalisation des systèmes.

3. Structure du mémoire

Au **chapitre I** nous définissons de manière précise notre concept d'interaction fortement couplée et décrivons le modèle retenu pour raisonner sur les requis et les propriétés de l'interaction fortement couplée. Le concept est illustré par un état de l'art sur les systèmes interactifs qui bénéficient, ou sont susceptibles de bénéficier, d'une interaction fortement couplée fondée sur la vision par ordinateur. L'état de l'art sert à l'établissement des requis fonctionnels proposé au **chapitre II**. Ce chapitre se poursuit par une modélisation précise de l'interaction fortement couplée destinée à permettre l'établissement des requis non fonctionnels.

Au **chapitre III**, nous étudions les deux approches classiques de vision par ordinateur : orientée modèle et orientée apparence. Nous justifions notre choix de l'approche orientée apparence et affinons cette approche par une conception centrée sur la tâche. Cette approche est mise en œuvre au **chapitre IV** pour la conception de techniques de suivi d'objet en vision par ordinateur. Nous y rapportons nos travaux, fondés sur l'existant, pour le développement de techniques satisfaisant les requis de l'interaction fortement couplée.

Le **chapitre V** rapporte les travaux concernant notre premier cas d'étude : le **tableau magique**. Ce prototype nous permet d'expérimenter le suivi d'objet pour l'interaction au doigt. L'expérience nous permet en particulier de confirmer la validité des requis exprimés au chapitre II. Notre deuxième cas d'étude, la **fenêtre perceptuelle**, est détaillée au chapitre VI. Ce prototype est fondé sur un suivi des mouvements du visage pour le contrôle d'une interface graphique classique. Il sert de support à deux expérimentations dont les résultats illustrent l'apport quantitatif d'une interaction fortement couplée fondée sur la vision par ordinateur.

Chapitre I *Interaction fortement couplée : définition et exemples*

En Interaction Homme-Machine, la notion de *couplage étroit*[1] est utilisée de manière informelle pour désigner différents types de dépendance entre un système et ses utilisateurs. Le concept de “retour d’information immédiat et informatif”, introduit avec la manipulation directe, exprime le besoin d’une réaction système conforme à l’attente des utilisateurs tant du point de vue informationnel que temporel. Dans [Ahlberg 94], le couplage étroit couvre plusieurs aspects de l’interaction :

- Interactions rapides, incrémentales et réversibles entre les composants de l’interface graphique (par exemple, en recherche d’information, l’effet de l’ajustement d’un filtre doit être immédiatement perceptible).
- Satisfaction des invariants de l’interface graphique (par exemple, un retour d’information proactif).
- Affichage continu montrant de manière permanente l’espace informationnel pertinent.
- Affordances compréhensibles et cohérentes servant de guidage intuitif (mots en surbrillance, poignées explicites, etc.).
- Utilisation des sorties du système comme données d’entrée.

Dans cette définition de la notion de couplage, le niveau des actions physiques est ignoré. Pour Ahlberg et Shneiderman, les conditions de succès de ce type de couplage sont supposées remplies. Il est vrai que dans le cas général, l’utilisateur imprimant un mouvement continu à la souris, le curseur, retour d’information physique du système, permet à tout instant à l’utilisateur de contrôler son geste. Le pointeur de la souris semble se déplacer sans délai vers la cible, les opérations de glisser-déposer ou de sélection de composants graphiques se font instantanément.

1. “tight coupling”

Si le couplage étroit au niveau des actions physiques est techniquement résolu pour les interfaces usuelles, il n'en va pas de même pour des dispositifs consommateurs de ressources matérielles tels que la vision par ordinateur, ou des dispositifs nouveaux comme les **Briques** de Fitzmaurice [Fitzmaurice 95]. Dans son espace d'aide à la conception d'interfaces saisissables (on y reviendra au paragraphe "Interfaces saisissables" page 26), Fitzmaurice introduit le couplage étroit comme un type de "lien entre les couches physiques et virtuelles" d'un système. On relève ceci :

> *"Dans les systèmes fortement couplés, les représentations virtuelles et physiques sont parfaitement synchronisées, les objets physiques sont suivis de manière continue en temps réel"*[1].

Nous adoptons ce point de vue mais il convient d'en pousser l'analyse de façon à préciser les termes qualitatifs tels que "synchronisation parfaite", "suivi continu", et "temps réel".

Dans ce chapitre, nous définissons de manière précise notre concept d'interaction fortement couplée et décrivons le modèle retenu pour raisonner sur les requis et les propriétés de cette notion. Nous effectuons ensuite une revue des systèmes qui bénéficient, ou sont susceptibles de bénéficier, d'une interaction fortement couplée fondée sur la vision par ordinateur. Ces systèmes serviront d'illustration aux analyses présentées aux chapitres suivants. Par souci de simplification, nous les classons en deux familles :

- les systèmes *immersifs* dont l'idée directrice est de plonger l'utilisateur dans le monde électronique,
- les systèmes de *Réalité Augmentée* qui adoptent l'approche opposée en immergeant les services électroniques dans le monde réel.

Comme l'observent Dubois et Nigay dans [Dubois 99], un continuum de possibilités est envisageable entre ces deux extrêmes.

1. "Tightly coupled systems have the physical and virtual representations perfectly synchronized, the physical objects are tracked continuously in real time"

1. Définition et modélisation

1.1. DÉFINITION

Une interaction est fortement couplée sur un intervalle de temps donné lorsque les systèmes humain et artificiel sont engagés de manière continue dans l'accomplissement d'actions physiques mutuellement perceptibles et dépendantes sur cet intervalle.

Le déplacement de la souris vers une cible est un exemple d'interaction fortement couplée. Le système humain (l'utilisateur) accomplit une action physique (le mouvement) perçue par le système artificiel (la souris). Le système produit en retour une action physique : l'affichage du curseur en relation avec la nouvelle position de la souris. La position du curseur est alors perçue par l'utilisateur qui contrôle ainsi son geste de manière incrémentale. Les actions des systèmes humain et artificiel sont mutuellement perceptibles et dépendantes sur l'intervalle de temps qui marque le déplacement de la souris sur une cible. De même, le contrôle du pointeur avec le doigt, dont la trajectoire est observée par une caméra, relève de l'interaction fortement couplée.

Du côté humain, l'interaction fortement couplée concerne le niveau des "comportements basés sur les habiletés"[1] du modèle de Rasmunssen [Rasmussen 86]. Elle ne fait pas intervenir les niveaux cognitifs supérieurs. Du côté système, l'interaction fortement couplée fait intervenir des traitements de complexité bornée afin d'assurer un temps de réponse aussi réduit que possible.

1.2. MODÉLISATION : PRINCIPE

L'interaction fortement couplée peut être modélisée sous forme d'un *système en boucle fermée* ([Card 83], [MacKenzie 93], [Ware 94]). Selon ce modèle, les données en sortie du système sont ré-introduites en entrée du système à un instant ultérieur. Appliqué à l'interaction fortement couplée, le système comprend ici un sujet humain et un système artificiel (ou ordinateur). Le modèle du système en boucle fermée traduit une dépendance continue sur un intervalle de temps donné entre les actions humaines et les réactions de l'ordinateur, et réciproquement. Typiquement, un mouvement de l'utilisateur est conditionné par le retour d'information de l'ordinateur résultant du mouvement précédent. La

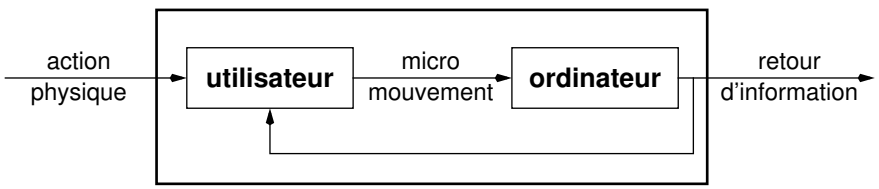


Figure 1
Interaction fortement couplée modélisée par un système en boucle fermée

1. "skill level behavior"

figure 1 montre un exemple d'interaction fortement couplée initiée par l'utilisateur via une action physique. La sortie du système est le retour d'information de l'ordinateur. Au sein du système, on assiste à une boucle de rétro-action continue fonctionnant comme suit : l'utilisateur produit un mouvement échantillonné par l'ordinateur. Cet échantillon constitue un micro-mouvement à l'origine d'un retour d'information de l'ordinateur. Ce nouveau retour d'information est nécessaire à la génération du micro-mouvement suivant. L'interaction se termine lorsque le dernier micro-mouvement a permis d'atteindre l'état cible souhaité.

1.3. MODÈLE AFFINÉ [Ware 94]

Ware et Balakrishnam affinent le modèle du système en boucle fermée sous forme d'un cycle comprenant quatre étapes [Ware 94]. Ce cycle, représenté sur la figure 2, débute par la perception de l'état courant de l'ordinateur (étape 1). L'utilisateur effectue un mouvement destiné à rapprocher l'état courant de l'état cible (étape 2). Le mouvement est capté par l'ordinateur (étape 3) qui calcule le nouvel état et produit le retour d'information correspondant (étape 4). L'utilisateur perçoit le nouvel état (étape 1) et débute un nouveau cycle. La succession des cycles se termine lorsque l'utilisateur perçoit un état correspondant à l'état cible.

Comme le montre la figure 2, chaque étape se traduit par un délai. Nous verrons dans les exemples qui suivent, et au chapitre II, les implications des effets temporels sur la qualité d'une interaction fortement couplée.

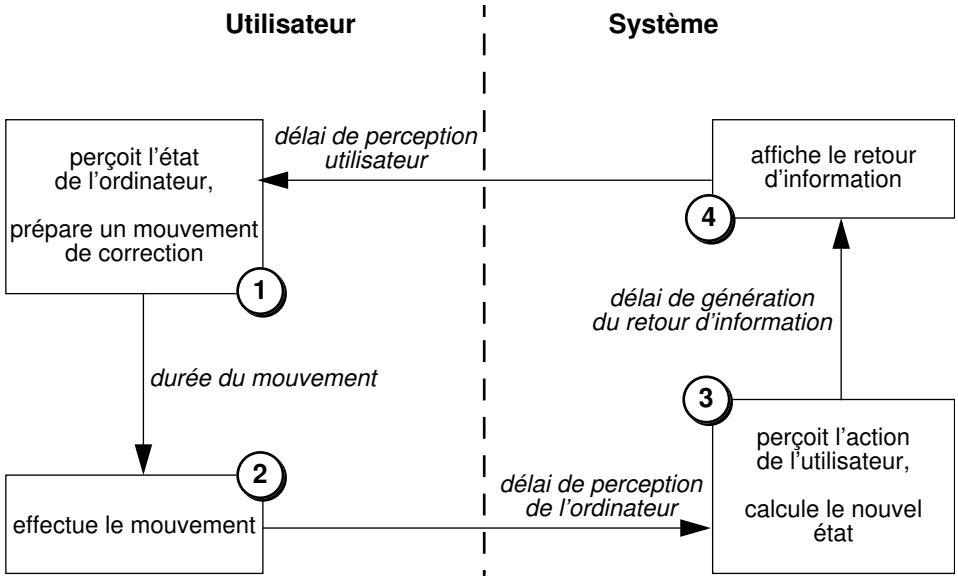


Figure 2
Boucle de l'interaction fortement couplée (d'après [Ware 94])

2. Systèmes immersifs

L'immersion a pour effet de placer une ou plusieurs personnes dans un monde simulé dont l'aspect et les événements sont si convaincants qu'ils sont perçus comme réels. De notre point de vue, la sensation d'immersion implique la satisfaction de trois requis :

1. Alimenter les sens du participant par des informations numériques qui se *substituent* aux informations originaires du monde physique.
2. Reproduire dans le monde simulé, les lois du monde physique pertinentes pour la tâche envisagée. Les informations qui traduisent ces lois doivent être cohérentes avec l'expérience sensorielle acquise dans le monde physique. Par exemple, lorsque l'utilisateur tourne la tête vers la gauche, le "paysage" doit défiler vers la droite.
3. Assurer une interaction fortement couplée. En reprenant l'exemple ci-dessus, le défilement du paysage doit être asservi au mouvement de la tête. À défaut, le participant serait simple spectateur.

Lorsque le réalisme est convaincant, l'utilisateur est amené à se comporter dans l'environnement simulé comme il le ferait dans le monde réel. L'apprentissage du système en est facilité par la réutilisation des connaissances acquises sur, et entretenues dans, le monde réel.

L'immersion totale, telle que l'envisage la plupart des systèmes de Réalité Virtuelle, implique le port de dispositifs d'interaction spécifiques :

- casques stéréoscopiques pour fournir une visualisation en relief du monde simulé,
- capteurs de position magnétique pour informer le système de la position et de l'orientation de la tête nécessaires à la génération de l'image correspondante dans le casque,
- gant numérique dont la capture du mouvement des doigts permet de synthétiser la position et la forme de la main dans le monde simulé. Le participant peut ainsi "tendre" la main et "attraper" des objets virtuels.

L'approche "casque et gant" trouve des applications et un public prêt à accepter de "porter tout ce qui est nécessaire" pour accéder à la sensation d'immersion ([Krueger 90]). C'est le cas pour les jeux, ou pour des professionnels, comme les pilotes de chasse, habitués à supporter un équipement dans leur travail. Toutefois, il est peu probable que cette approche se généralise pour des tâches routinières : hormis le coût du matériel spécifique, les dispositifs d'interaction casques et gants sont encombrants. Le faible succès de films en trois dimensions ne nécessitant pourtant que le port de lunettes polarisantes légères est un témoignage éloquent de la faible tolérance de l'individu vis-à-vis d'équipements portés.

À cette approche "portez tout ce qui est nécessaire", Krueger oppose une approche plus naturelle, qu'il résume par l'expression "venez tel quel"[1].

Une immersion sans préparation préalable en est le principe directeur. Dans ce chapitre, nous nous limiterons à ce type de systèmes. **VideoPlace** de Krueger et ses Réalités Virtuelles en sont les exemples pionniers [Krueger 90]. **ALIVE**, développé au MIT, lui fait suite. Nous les présentons maintenant. En liaison étroite avec l'immersion, la perception de l'espace en trois dimensions sera présentée en page 17.

2.1. VENEZ TEL QUEL : VIDEOPLACE ET ALIVE

VideoPlace et **ALIVE** sont parmi les rares systèmes interactifs fondés sur la vision par ordinateur ayant fait l'objet d'une utilisation publique en dehors du milieu contrôlé de laboratoire. Sachant que la majorité des systèmes interactifs fondés sur la vision restent à l'état de démonstrateur, le succès de la confrontation au public de **VideoPlace** et de **ALIVE** peut être considéré comme une validation de l'approche "venez tel quel" fondée sur la vision par ordinateur.

VideoPlace
[Krueger 90]

VideoPlace que schématise la figure 3, est constitué d'un projecteur vidéo qui restitue l'image d'un monde virtuel sur un écran de grande taille placé face à l'utilisateur. Sous l'écran, une caméra de surveillance capture les évolutions de l'utilisateur. Derrière lui, une plaque de plastique blanc partiellement transparente et rétro-éclairée par des tubes fluorescents vise à simplifier l'analyse d'image. Le système de vision par ordinateur a pour rôle d'extraire "en temps réel" la silhouette du participant dans l'image acquise par la caméra. La silhouette, une surface de couleur uniforme, est utilisée pour représenter le participant dans le monde virtuel. Un exemple de monde virtuel contenant la silhouette du participant est visible sur la figure 4.

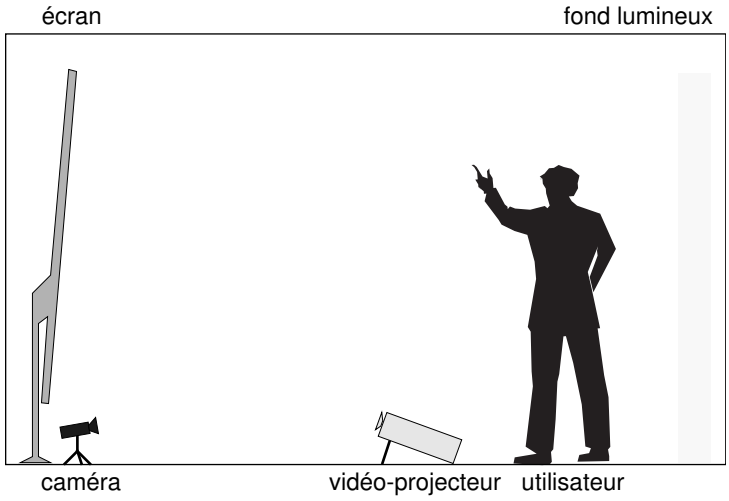


Figure 3
Le dispositif de "VideoPlace" (d'après [Krueger 90])

1. Les citations originales de Krueger sont "wear whatever is necessary" et "come as you are" ([Krueger 90] page 103).

Figure 4
Une image générée dans "VideoPlace" (extrait de [Krueger 90])
La silhouette du participant est segmentée et intégrée en temps réel dans l'image du monde virtuel.

Au-delà de son rôle de représentation, la silhouette sert de moyen d'interaction entre le participant et les objets du monde virtuel : **VideoPlace** est le support d'une série "d'interactions". Une interaction, telle que l'entend Krueger, se définit par les objets qui peuplent le monde et les lois qui les gouvernent. La figure 5 représente l'interaction "Critter" où un petit animal virtuel, le Critter, soumis à une faible gravité, interagit avec le participant par contact avec sa silhouette.

À la différence des systèmes immersifs qui utilisent un casque de visualisation, **VideoPlace** ne vise pas l'isolation totale du monde réel mais cherche à inventer une nouvelle forme de réalité. À l'évidence, la majeure partie du champ visuel est alimentée par l'information projetée sur l'écran, mais en périphérie, le participant peut aussi percevoir l'environnement physique dans lequel il évolue. De plus, le participant se voit "de

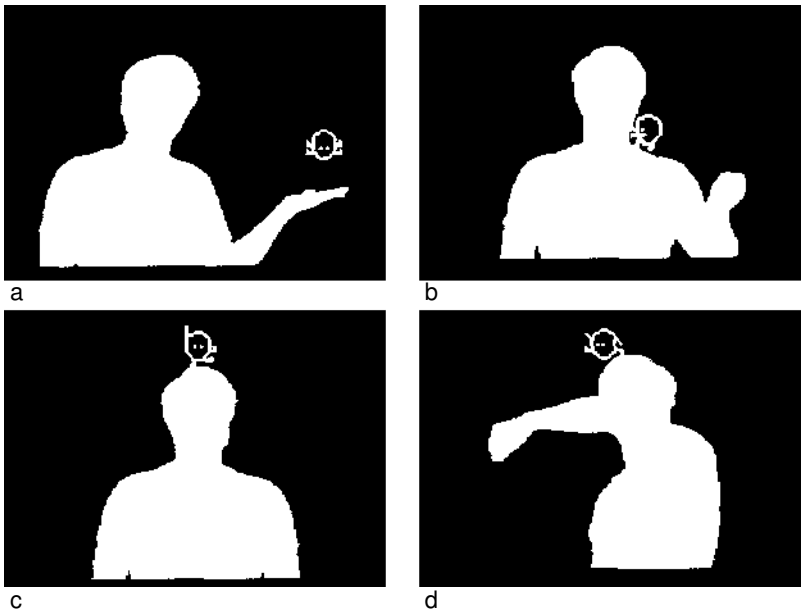


Figure 5
L'interaction "Critter" de VideoPlace (extrait de [Krueger 90])
Le Critter chute jusqu'à entrer en contact avec la silhouette du participant (a). Il tente alors l'escalade de la silhouette (b). Une fois au sommet, il manifeste sa satisfaction par une danse (c).

l'extérieur" : c'est la silhouette, et non le participant lui-même, qui est intégrée au monde virtuel.

Si l'utilisateur est immergé de manière partielle dans le monde virtuel, Krueger constate que la plupart des utilisateurs ont un fort sentiment de possession de leur image. Ce phénomène de conscience de soi à travers la silhouette tient à la mise en œuvre d'un effet miroir parfait, autre exemple d'interaction fortement couplée. Pour que l'effet miroir fonctionne, le temps de réaction du système doit être perçu comme nul. Dans **VideoPlace**, le délai entre la perception par le système d'un mouvement du participant et la génération de l'image associée est de l'ordre de 1 / 30ème de seconde (cf. les étapes 3 et 4 du modèle de la figure 2). D'après Krueger, ce délai ne doit pas être dépassé :

> "*Je pense que 30 réponses par secondes est la fréquence minimale pour fournir une sensation d'interaction en temps-réel ; elle est trop lente, et de loin, pour de nombreuses tâches*"[1] ([Krueger 90], page 115).

Cette idée intuitive de temps de réaction minimal sera étudiée de manière rigoureuse au chapitre suivant. Krueger, pour sa part, l'a considérée comme essentielle dès les années 1975 pour une mise en œuvre effective de ses divers systèmes de Réalités Artificielles. Il développa pour cela des circuits intégrés dédiés à la tâche de segmentation du profil du participant. Un seul circuit ne pouvant traiter toute l'image, plusieurs circuits en traitent chacun un sous-ensemble en parallèle. La technique utilisée par Krueger pour extraire la silhouette du participant de l'image est présentée à la section "Suivi par différence d'images" du chapitre IV.

Pionniers dans le domaine des systèmes immersifs sans équipement porté, les travaux de Krueger sont référencés en tant que source d'inspiration pour de nombreux travaux ultérieurs. Le système **ALIVE** [Maes 94] en est un exemple.

ALIVE

[Maes 94]
[Wren 97]

À l'instar de **VideoPlace** et comme le montre la figure 6a, le système **ALIVE** utilise un grand écran sur lequel est projeté le monde virtuel et, comme unique dispositif d'entrée, une caméra vidéo placée au-dessus de l'écran. Pour **ALIVE**, le choix d'une solution de type "venir tel quel" trouve plusieurs justifications au regard de l'approche "porter tout ce qui est nécessaire" :

- Une sécurité accrue dans les déplacements : les participants voient où ils marchent sans risque de collision avec des objets ou d'entrave due aux câbles.

1. "I think that 30 responses per second is the minimum rate required to provide a sens of real-time interaction; it is far too slow for many purposes".

- Une plus grande liberté de comportement et d'expression : Dans [Maes 94], on relève que les participants exécutent des figures acrobatiques (roues, sauts).
- Une meilleure implication dans l'environnement : les utilisateurs n'ont pas à se préoccuper d'un équipement complexe qui ne leur est pas familier (casque et gants).

Si **ALIVE** et **VideoPlace** relèvent de principes d'interaction communs, ils diffèrent par la nature des traitements de vision par ordinateur.

Analyse comparative

L'analyse de **ALIVE** et de **VideoPlace** met en évidence quatre points techniques distinctifs : le matériel de traitement, les contraintes sur le fond de la scène, la nature du monde virtuel et, liée à ce monde, les informations extraites de l'image.

Matériel de traitement. Dans **ALIVE**, le système de vision par ordinateur est exécuté sur un matériel standard (en 1996, une station Silicon Graphics Indy équipée d'un processeur R5500 à 150 Mhz). Par rapport à **VideoPlace**, qui utilise des processeurs dédiés, cette différence permet de réduire les coûts de développement et de diffuser le système ou ses algorithmes de traitement.

Fond de scène. En vision par ordinateur, la nature du fond de scène a une incidence sur la complexité des algorithmes de traitement. Alors que

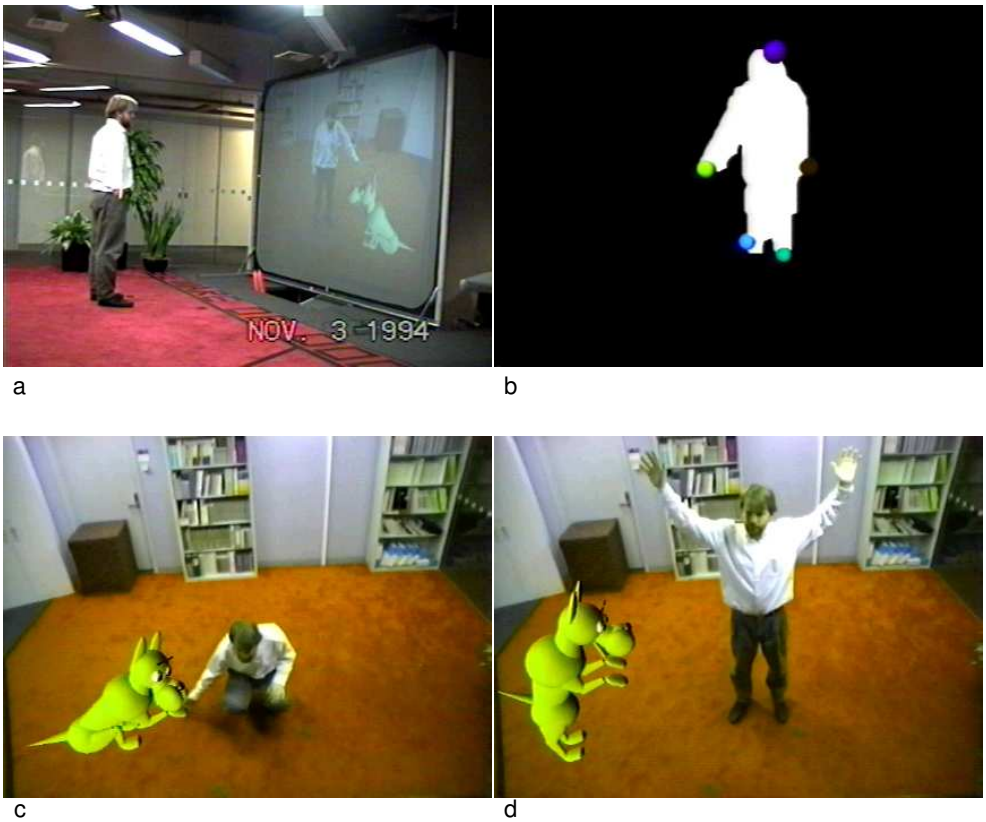


Figure 6
Le système ALIVE (extrait de [Maes 94])
Le participant est face à l'écran sur lequel est projeté le monde virtuel. Ce monde comprend des agents et l'image vidéo de l'utilisateur (a). De la silhouette de l'utilisateur, sont extraites les positions de la tête, des mains et des pieds (b). Les agents intelligents réagissent aux évolutions du participant (c) et (d).

VideoPlace impose un fond blanc lumineux uniforme, **ALIVE** relâche partiellement la contrainte : le fond peut être quelconque mais doit être statique. Le système ne fonctionnerait pas si la caméra captait, par exemple, le mouvement des pales d'un ventilateur ou l'image d'un écran.

Monde virtuel. VideoPlace part d'une image virtuelle dans laquelle est intégrée une expression du monde réel : la silhouette du participant. Il s'agit d'un monde planaire. Comme le montre la figure 6, **ALIVE** pratique l'inverse : le système part de l'image du monde réel et y insère des agents virtuels. L'image résultante projetée traduit un monde 3D impliquant la nécessité de gérer la profondeur. Par exemple, lorsqu'un agent virtuel se déplace derrière le participant par rapport à la caméra, il est occulté par l'image du participant. La gestion de l'occultation est une condition nécessaire au bon fonctionnement de l'interaction dans **ALIVE**. Dans le cas contraire, il y aurait violation de lois perceptuelles acquises dans le monde réel : si les agents virtuels, même en fond de scène et derrière le participant, étaient affichés en permanence par-dessus l'image du participant, il y aurait conflits visuels. Notre second requis sur les conditions d'immersion, exprimé en introduction de cette section page 11, ne serait pas satisfait.

Information extraite. ALIVE va plus loin que **VideoPlace** dans la nature de l'information extraite de l'image. Alors que **VideoPlace** se limite à la détection de la silhouette, **ALIVE** analyse la silhouette pour en extraire les positions de la tête, des mains et des pieds (représentés par des boules de couleur sur la figure 6b). Parce que le monde de **ALIVE** est tridimensionnel, ces positions doivent être calculées dans un espace 3D incluant notamment l'information de profondeur.

La profondeur est déduite de la manière suivante : considérant que le participant a ses pieds au sol, la position verticale des pieds dans l'image (résultat de l'analyse de la silhouette) est une estimation de la distance du participant à l'écran (la profondeur). La profondeur des autres membres (tête, mains) est assimilée à celle des pieds.

Cette hypothèse simplificatrice fournit une approximation grossière de la profondeur mais convient généralement au type d'application envisagée. Il est vrai que certains participants, nous l'avons vu, font des figures acrobatiques pendant lesquelles leurs pieds ne reposent plus sur le sol. Alors, l'estimation de profondeur est incorrecte : elle place le participant à une distance de l'écran plus grande que la distance réelle. Aussi, lorsque le participant tend les bras en avant ou se penche, la position de la tête et des mains ne correspond pas à celle des pieds.

Nous verrons au paragraphe "Notre approche : vision par apparence centrée sur la tâche utilisateur" (chapitre III page 71) comment la prise en compte des conditions spécifiques à l'application permet de simplifier les

algorithmes de vision et de satisfaire ainsi aux requis de l'interaction fortement couplée.

Synthèse

VideoPlace et **ALIVE** sont essentiellement des systèmes prospectifs destinés à explorer l'interactivité en situation d'immersion non intrusive ("en venant tel quel"). En l'absence d'étude formelle sur l'utilité et l'utilisabilité de ces systèmes, nous ne nous prononcerons pas sur leurs qualités interactives. Dans le cadre de cette thèse, nous retenons que ces systèmes démontrent la faisabilité technique de l'interaction fortement couplée au moyen de la vision par ordinateur. Nous précisons au chapitre suivant les conditions et hypothèses qui la rendent possible.

Notre étude ne concerne pas seulement la faisabilité d'une interaction fortement couplée au moyen de la vision par ordinateur, mais également l'évaluation de ses bénéfices du point de vue de l'interaction homme-machine. À la différence de **VideoPlace** et **ALIVE**, les systèmes que nous étudions dans la suite de ce chapitre répondent à un besoin utilisateur bien cerné. Il est alors possible d'évaluer l'apport au regard du besoin. Concernant les systèmes immersifs, nous étudions le besoin de perception de l'espace en trois dimensions.

2.2. PERCEPTION DE L'ESPACE EN TROIS DIMENSIONS

La capacité de percevoir l'espace en trois dimensions permet à l'Homme d'appréhender la forme des objets, d'apprécier leurs relations, d'évaluer les distances, etc. Les applications informatiques sont immédiates : analyse d'objets virtuels intrinsèquement tridimensionnels comme les molécules ou un crâne dont la topologie a été mesurée par scanner, maquettage industriel, visualisation scientifique, etc. Dans les systèmes immersifs où l'espace est une composante du processus d'interaction, le requis de conformité avec l'expérience visuelle dans le monde réel nécessite un rendu 3D aussi réaliste que possible.

Nous présentons brièvement les approches à la perception de l'espace en trois dimensions. Parmi ces éléments, nous retenons la parallaxe par mouvement dont l'efficacité a été démontrée empiriquement et dont la mise en œuvre peut tirer profit de la vision par ordinateur. Nous illustrons l'utilisation de la parallaxe par deux systèmes interactifs immersifs : la fenêtre virtuelle de type "venez tel quel" et l'aquarium de Deering et Ware de type "portez tout ce qui est nécessaire".

Approches à la perception en 3D

On relève deux approches complémentaires au rendu de l'espace en trois dimensions :

- La stéréoscopie, qui exploite la disparité binoculaire,
- L'extraction d'indices en vision monoculaire tels que la perspective, les occlusions, les ombres et la parallaxe par mouvement.

Historiquement, la stéréoscopie est l'approche privilégiée pour le rendu de la perception en trois dimensions. Elle s'appuie sur la disparité binoculaire : chaque œil capte une image du monde selon un point de vue différent. La mise en correspondance des disparités de ces deux images permet d'estimer la distance entre les objets observés et le point d'observation.

Contrairement à une idée répandue, la disparité binoculaire n'est pas le seul phénomène intervenant dans la perception de l'espace en trois dimensions. Voorhorst ([Voorhorst 98] page 13) recense de nombreux autres indices accessibles par vision monoculaire :

- Certains indices visuels sont accessibles sur image fixe, par exemple la perspective (le fait que les lignes parallèles se coupent à l'infini), l'occultation que nous avons évoquée avec **ALIVE**, les ombres et bien d'autres.
- D'autres indices sont accessibles sur une séquence temporelle d'images : la parallaxe par mouvement.

Dans la mise en œuvre de ses systèmes, Voorhorst produit l'effet 3D au moyen de deux indices visuels monoculaires particuliers : les ombres et la parallaxe par mouvement. Nous rappelons ci-dessous l'effet parallaxe avant de rapporter une analyse empirique de son efficacité.

Parallaxe par mouvement

De manière formelle, la parallaxe est l'angle formé par deux droites menées de l'objet observé à deux points d'observation. En pratique, la parallaxe dénote le déplacement de la position apparente d'un objet dû à un changement de position de l'observateur. La figure 7 illustre le phénomène. On y relève trois points de vue successifs correspondant à un observateur qui se déplace vers la gauche. Le cylindre central représente le point de fixation du regard de l'observateur. Les objets situés devant le point de fixation semblent se déplacer vers la droite, c'est-à-dire dans la direction opposée au mouvement de l'observateur. Inversement, les objets situés derrière le point de fixation semblent se déplacer dans la même direction que l'observateur. Enfin, la vitesse de déplacement apparente des objets croît avec leur éloignement du point de fixation.

La parallaxe par mouvement peut être considérée comme le résultat d'une forme de disparité, au même titre que la stéréoscopie est le résultat de la disparité binoculaire. Au lieu de considérer les disparités entre deux

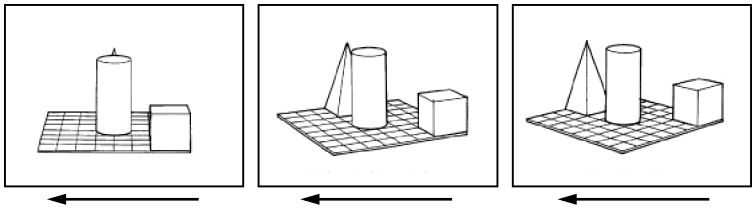


Figure 7
Parallaxe par mouvement (extrait de [Voorhorst 98])
Trois points de vue sur la même scène lorsque l'observateur se déplace latéralement vers la gauche.
Les flèches dénotent le sens de déplacement du point d'observation.

images captées de deux points de vues différents à un instant donné (stéréoscopie), on considère la disparité entre les différentes images captées par un point de vue mobile. Ainsi présenté, le phénomène de parallaxe offre une information plus riche que la stéréoscopie pour deux raisons :

1. La multiplication des points de vue au cours du temps (alors que la stéréoscopie ne considère que deux points de vue) ;
2. La possibilité de considérer des images plus dissemblables parce que captées de points de vue plus éloignés que l'écart qui sépare les deux yeux.

Ware démontre expérimentalement la supériorité prédite par notre analyse intuitive des faits [Ware 93].

Supériorité de la parallaxe par mouvement
[Ware 93]

Ware et ses collaborateurs rapportent les résultats d'une étude empirique sur l'efficacité relative des différents indices visuels dans la perception de l'espace en trois dimensions [Ware 93]. Leur système expérimental met en œuvre trois types d'indices : les indices statiques (effet de perspective, ombres), la stéréoscopie et la parallaxe par mouvement. Quatre conditions expérimentales sont considérées :

1. présence des indices statiques uniquement,
2. indices statiques + stéréo,
3. indices statiques + parallaxe[1],
4. indices statiques + stéréo + parallaxe.

Deux expériences complémentaires ont été effectuées. L'une, dite "subjective", consiste en une série de comparaisons des conditions expérimentales deux à deux : pour chaque combinaison, les sujets doivent identifier laquelle des deux conditions est la plus convaincante pour la perception d'une scène en trois dimensions. Deux scènes sont utilisées : l'une représente une sphère, l'autre un tube coudé. Les résultats obtenus pour les deux scènes ne sont pas différents de façon significative. Tous révèlent une forte préférence pour la condition 3 (parallaxe). En particulier, la parallaxe seule est jugée plus convaincante en comparaison de la stéréo seule pour environ 90% des sujets.

Une deuxième expérience, dite "objective", mesure les performances utilisateur pour une tâche de perception en trois dimensions. La tâche consiste à reconstruire une structure arborescente ternaire d'une trentaine de feuilles observable sur un écran selon les conditions de la table 1. Les

1. La condition "parallaxe" est en fait affinée en deux conditions : parallaxe monoculaire (il est demandé aux sujets de fermer leur œil gauche) et parallaxe binoculaire. Au vu des résultats expérimentaux, cette distinction n'est pas significative.

Indices visuels	Taux d'erreur
Statiques	21.8 %
Statiques + stéréo	14.7 %
Statiques + parallaxe	3.7 %
Statique + stéréo + parallaxe	1.3 %

Table 1 : Résultat de l'expérience "objective" de Ware [Ware 93] : taux d'erreur en fonction des indices visuels de perception en trois dimensions.

différences de taux d'erreur sont toutes significatives deux à deux et sont rapportées dans la table 1.

Cette expérience met en évidence l'efficacité de la parallaxe par mouvement au regard de la réduction des taux d'erreurs et sa supériorité sur la stéréoscopie pour la perception de l'espace en trois dimensions. Cependant, le couplage parallaxe-stéréoscopie permet d'améliorer les performances par un facteur 16 sur les indices statiques seuls. Si l'on note la nette préférence subjective pour la parallaxe, il convient de retenir qu'un bon rendu de la perception en 3D doit passer en priorité par la mise en œuvre d'un système de parallaxe par mouvement. Nous illustrons la mise en application de la parallaxe avec des systèmes de type "fenêtre virtuelle" et "aquarium".

Exemples de système fondé sur la parallaxe

Dispositifs de type "Fenêtre Virtuelle" . Le premier dispositif de type "fenêtre virtuelle" a été conçu à l'université technique de Delft [Smets 88]. Son principe directeur est le support à la perception de l'espace en trois dimensions par reproduction de l'effet parallaxe. À la différence de la plupart des systèmes de réalité virtuelle qui génèrent des images de synthèse, les images produites par la fenêtre virtuelle sont captées par une caméra vidéo distante. Le dispositif est illustré sur la figure 8. L'utilisateur perçoit les images d'une scène distante au travers d'un moniteur vidéo. Les déplacements de la tête sont captés par un dispositif. Ces déplacements sont reproduits par le support de la caméra qui se déplace en temps réel à l'identique des déplacements du visage de l'utilisateur. On obtient ainsi l'effet parallaxe.

Le principe de la fenêtre virtuelle a été appliqué à différents domaines. Gaver en a étudié l'utilisation en communication interpersonnelle médiatisée [Gaver 95] : l'utilisateur est en situation de communication avec une personne distante par l'intermédiaire de liaisons audio et vidéo. La caméra fournit les images de la liaison vidéo. Elle est montée sur un support mobile qui se déplace latéralement. Les déplacements latéraux du visage de l'utilisateur sont captés par un système de vision par ordinateur et transmis au support mobile de la caméra distante. Ce montage permet ainsi à un utilisateur d'explorer une scène distante par translation du visage.

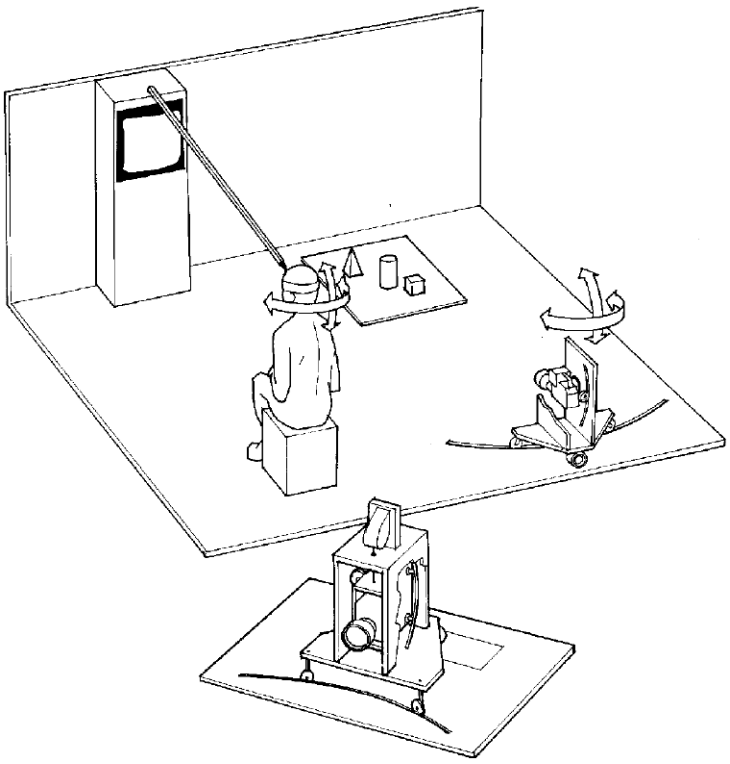


Figure 8
Dispositif de type "fenêtre virtuelle" (extrait de [Voorhorst 98])
L'utilisateur observe un moniteur affichant l'image d'une caméra vidéo. La caméra est montée sur un support mobile contrôlé pour refléter les déplacements du visage.

Voorhorst, quant à lui, montre l'utilité du principe de la fenêtre virtuelle dans le domaine de l'endoscopie [Voorhorst 98] : il met au point un prototype d'endoscope dont l'extrémité peut se déplacer latéralement. L'extrémité de l'endoscope est couplée aux translations du visage du chirurgien. Grâce à ce système, le chirurgien a une perception en trois dimensions des organes de la cavité abdominale sur laquelle il opère.

La principale difficulté technique de la mise en œuvre d'un système de fenêtre virtuelle est le déplacement de la caméra. En règle générale, elle implique la construction d'un support mobile complexe. Dans le cas des systèmes de réalité virtuelle, ce problème disparaît puisque l'image fournie à l'utilisateur est générée de façon logicielle. Le déplacement du point de vue se traduit par un simple changement de paramètre dans l'algorithme de génération.

Dispositifs de réalité virtuelle de type "aquarium". Le terme "aquarium"[1] fait référence aux dispositifs dont le monde est maintenu à l'intérieur d'un écran graphique (à l'instar des poissons dont le monde est limité à l'aquarium). Il est employé pour la première fois dans [Ware 93] qui fait référence aux travaux antérieurs de Deering [Deering 92] sur les requis d'un tel système.

Comme le montre la figure 9, un "aquarium" comprend un écran graphique observé au moyen de lunettes stéréoscopiques. Un mécanisme

1. "Fish Tank Virtual Reality".

Figure 9
Réalité Virtuelle de type "aquarium" (extrait de [Ware 93])
La position d'observation est captée par un bras mécanique attaché à la tête de l'utilisateur. Une scène en trois dimensions (nécessitant le port de lunettes stéréoscopiques) est générée sur l'écran en fonction de la position d'observation.

de suivi de la tête informe le système de la position du point d'observation. L'image générée est calculée en permanence pour correspondre au point de vue courant. Un tel système met en œuvre à la fois la stéréoscopie (port de lunettes) et la parallaxe par mouvement. Sur le plan interactionnel, on assiste à un fort couplage entre la position d'observation et l'image générée.

Alors que la Réalité Virtuelle "classique" génère les images du monde virtuel sur deux écrans miniature installés dans un casque, l'approche aquarium utilise un seul moniteur graphique standard. L'intérêt de l'approche classique est la possibilité de générer les images du monde virtuel quel que soit le point de vue du participant. Ware et Deering justifient l'approche aquarium par la possibilité d'obtenir des images de résolution bien supérieure avec un écran standard comparés aux écrans miniature des casques actuels. Ware ajoute que l'approche aquarium n'obstrue pas complètement le champ visuel : l'utilisateur perçoit, en vision périphérique, le monde physique qui l'entoure réduisant ainsi les risques de collision ou de chute, de même que la sensation de mal de mer. Si cette semi-immersion peut être perçue comme une limitation (l'étendue du monde virtuel créé est de fait limitée), elle rend plus crédible l'utilisation de ce type de système pour des durées prolongées.

Ces exemples de systèmes, pour lesquels la parallaxe par mouvement constitue un facteur essentiel du rendu en trois dimensions, mettent en évidence un couplage étroit entre la position d'observation et l'image générée. Voorhorst [Voorhorst 98] insiste sur l'importance d'un lien direct entre les déplacements de la tête et du point vue pour produire une impression convaincante d'espace en trois dimensions. Nous passons maintenant en revue les moyens techniques de mise en œuvre de ce couplage et concluons en synthèse sur le rôle de la vision par ordinateur en la matière.

Couplage point d'observation / image générée

Les systèmes que nous venons de décrire utilisent trois types de dispositifs pour la mise en œuvre du couplage entre le point d'observation et l'image générée : les dispositifs magnétiques (les plus employés), les bras physiques et la vision par ordinateur.

Dispositifs magnétiques. Les systèmes magnétiques, tels le **Polhemus Isotrack** ([Polhemus 99]) et l'Ascension **Flock of Birds** ([Ascension 99]), comprennent un générateur de champ magnétique contrôlé et des capteurs magnétiques fixés aux entités à suivre (dans le cas qui nous intéresse, à la tête de l'utilisateur). Les capteurs renseignent le système sur leur position dans l'espace selon six degrés de liberté (trois coordonnées pour la position et trois angles pour l'orientation).

Les systèmes magnétiques présentent trois limitations fâcheuses :

1. L'utilisateur doit être équipé de capteurs qui nécessitent une connexion par câbles. Bien que les capteurs soient de petite taille, la présence de câbles est une source de gêne, voire un facteur de risque.
2. Le délai entre la capture de la position et la transmission de l'information au système est perceptible par l'humain (cf. étape 3 du modèle de la figure 2 page 10). Il en résulte un retard sensible de l'affichage de la scène en relation avec les mouvements de la tête.
3. L'estimation de la position est bruitée. Lorsqu'un capteur est immobile, les informations de position transmises sont instables, oscillant autour de la position réelle du capteur.

On trouvera dans [Liang 91] des mesures précises sur le délai et l'oscillation des estimations pour le cas du **Polhemus**. Les auteurs proposent des techniques de prédiction pour réduire le délai et de filtrage du bruit pour stabiliser les estimations. Mais ces mesures correctives ne sont que partielles. Deering [Deering 92] utilise également la prédiction pour réduire le délai, mais reconnaît les limitations de cette technique. Ware et Balakrishnam effectuent une étude approfondie de l'effet du délai sur les performances utilisateur [Ware 94] et montrent que les performances se dégradent rapidement lorsque le délai croît. Cette dégradation de performances peut avoir des conséquences importantes sur l'interaction :

> *"(...) ce type de dégradation de performance peut facilement faire la différence entre un système qui est perçu comme utile et un qui ne l'est pas."*[1] ([Ware 94], p. 343)

Bras physique. Smets [Smets 88], Ware [Ware 93] et Voorhorst [Voorhorst 98] utilisent un bras physique fixé à la tête des sujets pour mesurer leurs déplacements (un bras physique est visible sur la figure 9

1. "(...) this kind of performance degradation may easily make the difference between a system that is perceived as useful and one that is not".

page 22). Le bras est soit directement relié à la caméra générant les images de la scène observée ([Smets 88]), soit il est connecté à des dispositifs convertissant les déplacements du bras en informations numériques ([Ware 93], [Voorhorst 98]).

L'utilisation d'un bras physique permet de réaliser un excellent couplage entre la position de la tête et l'image générée car les délais de mesure sont imperceptibles et l'information de position mesurée est précise et stable. La qualité de l'information de position produite par un bras physique permet de faire abstraction des effets du dispositif sur les performances utilisateurs. C'est pourquoi Ware choisit d'utiliser un bras physique dans ses expériences. Par contre, son utilisation est peu réaliste pour un usage courant ([Ware 93]) : il est encombrant et les déplacements autorisés sont limités.

Système de vision par ordinateur. Avec la fenêtre virtuelle, Gaver [Gaver 95] introduit l'utilisation de la vision par ordinateur pour capter la position de la tête de l'utilisateur. Une caméra fixe, située sur le moniteur en face de l'utilisateur, fournit un flux vidéo au système. L'estimation de la position de la tête est calculée par la technique de différence d'images et de seuillage. Les détails de cette technique seront donnés au paragraphe "Suivi par différence d'images" (chapitre IV page 83).

Un tel dispositif fondé sur la vision par ordinateur présente deux avantages essentiels sur ses concurrents :

- coût d'achat moindre,
- intrusion acceptable : une caméra de petite taille est discrète. De plus, elle entre *de facto* dans le modèle "venez tel quel" : il suffit à l'utilisateur de se présenter devant le moniteur pour que le système fonctionne.

Par contre, dans l'implémentation de Gaver, les informations de position extraites des images sont grossières et instables. Gaver reconnaît les défauts de sa technique et admet que le système est, au final, inutilisable. Il suggère l'étude d'autres techniques de suivi du visage en vision par ordinateur. C'est ce à quoi nous nous sommes employés dans nos travaux de recherche.

2.3. SYNTHÈSE SUR LES SYSTÈMES IMMERSIFS

Les systèmes immersifs s'appuient, nous l'avons vu, sur la réutilisation des habiletés acquises dans le monde réel pour la conduite de tâches dans des mondes virtuels. La difficulté fondamentale tient à la conception et à la réalisation de mécanismes artificiels d'immersion réaliste. Parmi ces conditions et pour certaines applications, nous relevons la nécessité du rendu de la perception de l'espace en trois dimensions. Parmi les techniques qui participent à l'effet 3D, la parallaxe par mouvement est la plus performante. Mais son efficacité tient à la qualité du couplage entre le point d'observation et l'image produite. Parmi les dispositifs capables

de mettre en œuvre ce couplage, seule la vision par ordinateur élimine les équipements particuliers à fixer sur l'utilisateur. Cette propriété en fait un candidat de choix pour réaliser la fonction de couplage entre le point d'observation et la génération d'image pour les applications nécessitant une approche "venez tel quel" tels que **VideoPlace** et **ALIVE**.

Les systèmes de Réalité Augmentée que nous présentons dans la section suivante s'appuient, comme les systèmes immersifs, sur la réutilisation d'habiletés acquises dans le monde réel. En revanche, ces systèmes visent la conduite de tâches non pas dans des mondes simulés mais dans le monde réel. Si les systèmes de Réalité Augmentée n'impliquent pas la difficile mise en œuvre de mécanismes d'immersion, le couplage qu'ils imposent entre les objets physiques et l'information électronique présente des difficultés techniques importantes.

3. Systèmes de Réalité Augmentée

La notion de Réalité Augmentée (RA) couvre plusieurs définitions traduisant différents courants de recherche menés en parallèle dès les années 90. En Graphique, la RA dérive de la Réalité Virtuelle : par exemple, de "vrais" pixels d'origine vidéo sont fusionnés à des pixels de synthèse [Azuma 93]. En interaction homme-machine, la RA, en réaction à l'immersion virtuelle, s'appuie résolument sur la conservation et l'amplification du réel : par exemple, la feuille de papier physique, objet familier, est "augmentée" de capacités de traitement de l'information ([Mackay 93], [Mackay 96]). Dans le cadre de cette thèse, nous entendons la RA comme l'amplification d'objets familiers par le traitement numérique.

Comme pour les systèmes immersifs de la section 2, nous illustrons les principes de la RA au moyen de systèmes représentatifs pour lesquels la vision par ordinateur a un rôle à jouer. Nous retenons deux catégories de systèmes : les systèmes fondés sur le principe des *interfaces saisissables* et les *interfaces digitales* qui, comme leur nom l'indique, utilisent le doigt comme dispositif privilégié d'interaction.

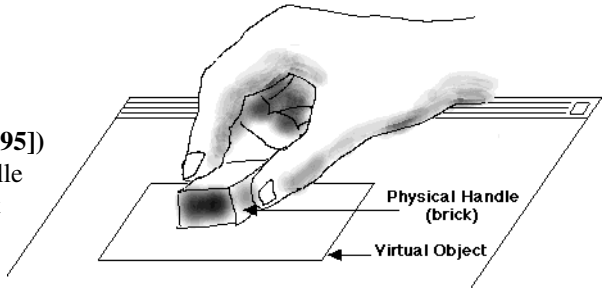


Figure 10
Principe d'une interface saisissable (extrait de [Fitzmaurice 95])
Un objet physique (ici une brique) est associé à une entité virtuelle en le posant sur la représentation de l'entité. L'association est rompue dès que l'objet est soulevé.

3.1. INTERFACES SAISISSABLES

La notion d'interface saisissable[1], due à Fitzmaurice ([Fitzmaurice 95], [Fitzmaurice 96]), dénote une famille d'interfaces homme-machine dans lesquelles des objets physiques saisissables servent de dispositif d'entrée. La position et l'orientation de ces objets sont à tout instant connues du système. L'utilisateur interagit avec le système en associant un ou plusieurs de ces objets à des *entités virtuelles* : objets graphiques (fenêtre graphique, icône de fichier) ou fonctions (opération sur l'espace de travail tels que déplacement, effet de zoom, rotation). Comme le montre la figure 10, l'association entre un objet physique et une entité virtuelle se crée en posant l'objet sur la représentation de l'entité. Une fois l'association établie, l'état de l'entité est modifié en déplaçant l'objet. Elle est rompue dès que l'objet est soulevé. En général, la surface de restitution des systèmes à interface saisissable est horizontale afin d'y déposer les composants physiques. On peut toutefois imaginer une interface saisissable sur une surface verticale, par exemple un tableau, avec des objets munis d'aimants.

Les "interfaces manipulatoires"[2] ou "interfaces matérialisées"[3] de Fishkin et Harrison s'appuient également sur le principe de l'analogie avec les gestes produits dans le monde réel [Harrison 98]. Mais la nature des artefacts démontrés ne saurait tirer profit a priori des techniques de vision.

Ayant présenté le principe interactionnel des interfaces saisissables, nous allons en énoncer les caractéristiques puis les apports. Nous utilisons comme élément de comparaison la notion de multiplexage.

Multiplexage spatial et temporel des fonctions

Le multiplexage est un procédé employé pour répartir l'activation des fonctions d'un système. Fitzmaurice observe que les interfaces graphiques usuelles appliquent, avec la souris, un *multiplexage temporel* alors que les interfaces saisissables pratiquent le *multiplexage spatial* ([Fitzmaurice 96]).

1. "Graspable user interface"
2. "Manipulative interface"
3. "Embodied interface"

Le multiplexage temporel désigne une activation séquentielle des fonctions du système. Le pointeur de la souris, qui ne désigne, à un instant donné, qu'un seul emplacement de l'écran, ne permet d'activer qu'une fonction à la fois. L'activation d'une fonction avec la souris (par exemple, le déplacement d'une fenêtre) est marquée par l'association du curseur à la représentation graphique de la fonction. Cette association s'exprime par la position du curseur conjointe à l'enfoncement du bouton de la souris. Elle cesse avec le relâchement du bouton. La souris est alors disponible pour l'activation d'une autre fonction.

Le multiplexage spatial désigne le partage, à un instant donné, de l'espace entre plusieurs entités. Dans notre contexte, le multiplexage spatial se traduit par la présence simultanée de plusieurs représentations graphiques à l'écran : les boutons, barres de titre des fenêtres, barres de menu et poignées de contrôle se répartissent l'espace d'affichage. Ces représentations offertes simultanément à l'activation sont, avec la souris, utilisées en séquence. Le potentiel de simultanéité n'est pas exploité. Autrement dit, le multiplexage spatial du rendu graphique est bridé par le multiplexage temporel de la souris.

Les interfaces saisissables, dont le principe est l'association d'un objet physique distinct à chaque fonction, offrent une correspondance directe entre les dispositifs d'entrée et les fonctions ou concepts sous-jacents. Comme le montre la figure 11, l'activité d'association due au multiplexage temporel de la souris est éliminée. Mais, jusqu'où faut-il pousser le principe de la correspondance bi-univoque? N'y a-t-il pas un bon équilibre à trouver entre le multiplexage temporel extrême et le multiplexage spatial extrême ?

Si l'on écarte les raccourcis clavier, les interfaces graphiques standard pratiquent le multiplexage temporel extrême : l'accès à l'ensemble des

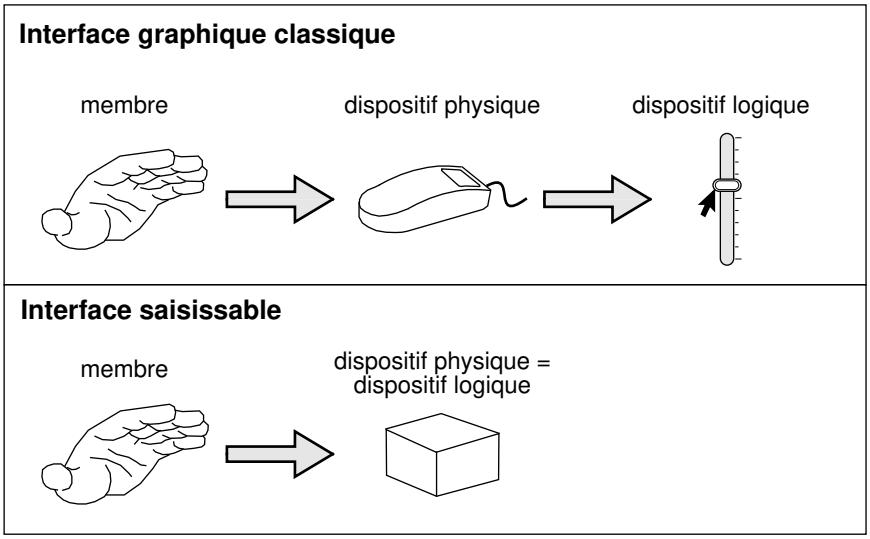


Figure 11
Aspect direct de l'interaction fondée sur les interfaces saisissables (d'après [Fitzmaurice 96])
Avec une interface graphique classique, l'utilisateur doit acquérir la souris avec la main, puis acquérir un widget avec le pointeur de la souris.
Avec une interface saisissable, une seule acquisition est nécessaire car l'objet saisi contrôle directement le système.

Figure 12
Multiplexage spatial extrême des fonctions : une table de mixage
La fonction de chaque potentiomètre est fixe et ne peut être modifiée.

fonctions du système passe par la souris. La table de mixage de la figure 12 est un cas de multiplexage spatial extrême : chaque potentiomètre de la table est lié, en dur, à une et une seule fonction. Les limites du multiplexage spatial sont claires : le nombre de fonctions accessibles est limité par l'espace (la table de mixage est un bon exemple de démesure) et le dispositif d'entrée n'offre pas la capacité d'adaptation nécessaire au contrôle des logiciels.

Fitzmaurice considère en pratique que les interfaces saisissables doivent se situer entre les deux extrêmes. Ainsi, à un fin niveau de granularité temporelle, les objets physiques sont multiplexés spatialement : chaque objet a une fonction bien définie. À un niveau de granularité temporel plus large, il est possible d'attacher et de détacher les objets physiques à leur fonction (tel que présenté sur la figure 10). On multiplexe ainsi leur fonction dans le temps. Si, en pratique, le bon équilibre entre multiplexage temporel et spatial semble un principe raisonnable, Fitzmaurice ne fournit aucune heuristique explicite sur les conditions d'un dosage adapté.

Apports

On trouvera dans [Fitzmaurice 96] un recensement des apports des interfaces saisissables. Nous les résumons ci-dessous :

- l'interaction à deux mains est favorisée, voire encouragée. L'utilisateur peut envoyer des ordres au système par l'intermédiaire de deux flux parallèles, accroissant ainsi le débit d'information à destination du système. La figure 13 montre un exemple d'interaction à deux mains. Dans une expérience empirique comparant les deux types de

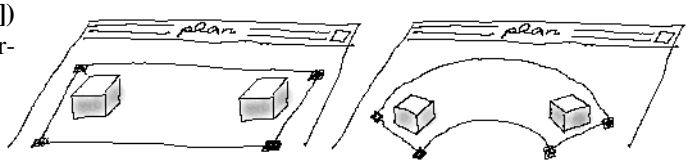


Figure 13
Interaction à deux mains (extrait de [Fitzmaurice 95])
Deux objets sont manipulés pour spécifier une transformation complexe sur un rectangle : les paramètres de position, échelle, orientation et courbure sont exprimés en un seul geste.

multiplexage, spatial et temporel, Fitzmaurice montre que le multiplexage spatial aboutit à de meilleures performances pour des tâches comportant des transformations d'objets graphiques (translations, changements de taille, et rotations). Le gain de performance est sensible pour des tâches nécessitant plusieurs transformations simultanées.

- Une interface saisissable met en jeu les habiletés de préhension de la vie courante. La souris, qui ne sollicite qu'une main limitée à des mouvements de translation, est, de ce point de vue, un dispositif réducteur. Les objets physiques des interfaces saisissables impliquent des manipulations plus riches pour le contrôle fin du positionnement et de l'orientation.
- Une interface saisissable encourage l'interaction à plusieurs. On constate que, dans la vie courante, un groupe d'individus parvient à se synchroniser efficacement dans la réalisation de tâches de manipulation : c'est le cas des chefs cuisiniers autour du fourneau, des contrôleurs aériens en salle de contrôle ou du personnel hospitalier dans le bloc opératoire. La configuration classique "clavier-écran-souris" n'encourage pas l'interaction de groupe. Elle est destinée à un usage mono-utilisateur. Inversement, avec une interface saisissable, les utilisateurs peuvent se répartir et s'échanger le contrôle des objets physiques.
- L'externalisation des représentations, jusqu'à présent confinées au virtuel, favorise l'utilisabilité. Leur matérialisation les rend plus saillantes facilitant la manipulation.

Les apports des interfaces saisissables ont motivé la réalisation de plusieurs prototypes que nous présentons maintenant : les **Briques**, le **MetaDesk** et **Built-It**

Exemples d'interfaces saisissables

Le système "Briques". Fitzmaurice concrétise le principe des interfaces saisissables avec le prototype "**Briques**"[1] [Fitzmaurice 95]. L'interface en entrée est constituée de petits objets parallélépipédiques à l'image des briques "lego" des jouets d'enfant (une brique lego est représentée dans la marge). "**GraspDraw**", fondé sur **Briques**, permet de dessiner des formes géométriques colorées simples (voir figure 14). La sélection d'une couleur se fait en posant une brique dans un support qui représente des pots de peinture. Plusieurs formes d'interaction à deux mains sont mises en jeu (et étudiées) pour la création, la déformation et le placement de formes géométriques.

Sur le plan technique, les briques sont réalisées au moyen d'un Ascension **Flock of Birds** ([Ascension 99]). Par sa forme, ce dispositif est adapté à la réalisation de briques puisque les capteurs sont de petits cubes d'environ 2,5 centimètres de côté. Chaque capteur renseigne le système sur sa

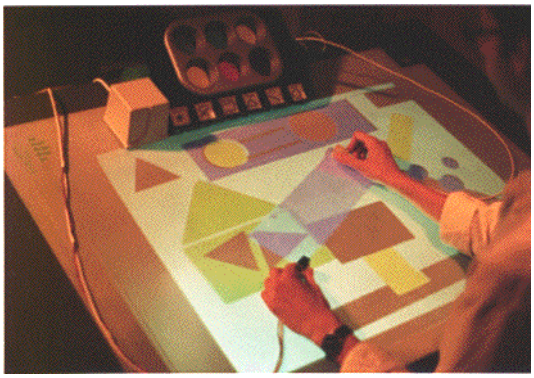


Figure 14
L'application GraspDraw (extrait de [Fitzmaurice 96])
Deux "briques" permettent de réaliser les tâches suivantes : choix des couleurs et des formes géométriques, création, déplacement et déformation d'entités graphiques.

position et son orientation dans l'espace. Cependant, le **Flock of Birds** est sujet aux limitations des dispositifs magnétiques déjà évoqués ("Dispositifs magnétiques" page 23) :

- Les valeurs de position et d'orientation sont instables. Fitzmaurice rapporte que, pour la manipulation de formes géométriques de petite taille, les utilisateurs de **Briques** sont gênés par les oscillations, pourtant faibles, du curseur. Notons qu'une tablette graphique est utilisée pour les mesures de performances utilisateur évoquées au paragraphe "Apports" page 28 : la tablette graphique rend des valeurs de position stables dont la précision est bien supérieure à celle des dispositifs magnétiques.
- Pour ces mêmes expériences, il n'est fait aucun commentaire sur les contraintes de délai de capture des informations de position et d'orientation. Ces problèmes sont éludés, encore une fois, par l'usage de la tablette graphique.
- Le câble qui relie chaque capteur au système est une source de gêne. Bien que le **Flock of Birds** soit capable de gérer un grand nombre de capteurs, **GraspDraw** ne comprend que deux briques : Fitzmaurice estime qu'au-delà, les fils deviennent trop encombrants sur le plan de travail. L'expérience avec **Graspdraw** montre qu'en général, les utilisateurs placent les câbles de part et d'autre du plan de travail de façon à laisser un triangle dégagé entre le corps et les deux mains. La figure 14 illustre la situation. On note aussi que la limitation à deux briques n'empêche pas les utilisateurs d'exprimer leur gêne vis-à-vis des câbles.

Les recherches de Fitzmaurice ont inspiré d'autres travaux dont nous allons étudier les réalisations.

1. "Bricks"

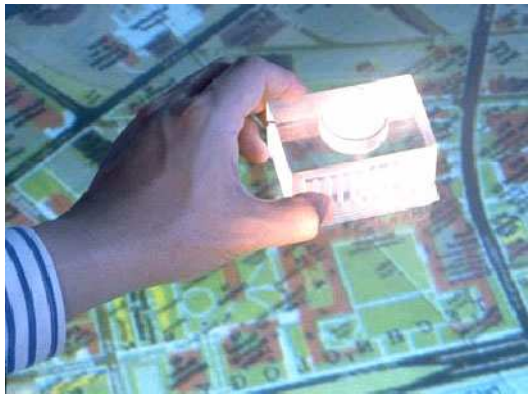


Figure 15
Le système MetaDESK (extrait de [Ullmer 97])
Le fait de poser le Dôme miniature du Massachusetts Institute of Technology (M.I.T.) sur le **MetaDESK** fait apparaître le plan du M.I.T. Le Dôme sert ensuite de "poignée" pour déplacer et faire pivoter le plan.

MetaDESK. La première application du **MetaDESK** ([Ullmer 97] [Ishii 97]) est un système de navigation sur une carte (voir la figure 15). À la différence des briques banalisées de Fitzmaurice, les objets physiques du **MetaDESK** sont spécialisés, telle la réplique miniature du bâtiment principal (le Dôme) du Massachusetts Institute of Technology (M.I.T.). Poser cet objet sur le **MetaDESK** fait apparaître la carte du campus du M.I.T centrée sur la position du Dôme miniature. La carte peut ensuite être déplacée et pivotée par translation et rotation respectives de la miniature. En plaçant une autre miniature sur la carte (le bâtiment du MediaLab dans l'exemple du **MetaDESK**), l'utilisateur contrôle le facteur d'échelle de la carte : les emplacements des deux bâtiments, Dôme et MediaLab, sont maintenus en permanence par le système sous leur miniature respective. En éloignant les deux miniatures, l'utilisateur effectue un effet de zoom sur la carte (et inversement).

Les miniatures du Dôme et du MediaLab sont identifiables par leur aspect. Ces objets, qui fonctionnent par analogie de forme avec les concepts qu'ils permettent de manipuler, sont appelés des "phicons" (pour "PHysical ICON" ou "icône physique") [Ishii 97]. Les phicons peuvent également jouer le rôle de récipient enregistreur : la miniature du Dôme pourrait être prêtée à une autre personne qui pourrait alors consulter la carte du M.I.T. sur un système différent.

La réalisation du **MetaDESK** fait intervenir à la fois un dispositif magnétique **Flock of Birds** et un système de vision par ordinateur. Le dispositif magnétique est dans certains cas préféré au système de vision qui fonctionne seulement a une fréquence de sept images par seconde et utilise une technique de différence d'images connue pour son imprécision (les détails sur cette technique sont présentés au paragraphe "Suivi par différence d'images" page 83). Les capteurs magnétiques sont installés sur deux objets particuliers du **MetaDESK** :

1 la loupe, qui permet d'obtenir une information spécifique (par exemple, une photo aérienne) sur la surface du **MetaDESK**,

2 la fenêtre 3D, qui offre une visualisation en trois dimensions des bâtiments présents sur la carte.

Malgré les faibles performances du système de vision, les concepteurs du **MetaDESK** préfèrent utiliser ce système pour le suivi des phicons sur le bureau plutôt que le dispositif magnétique : l'objectif visé est l'utilisation de n'importe quel objet. Il est donc nécessaire d'imposer un minimum de contraintes sur leur nature. Une approche similaire est utilisée dans "**Illuminating Light**" pour identifier et repérer des briques représentant, cette fois-ci, des composants optiques. Chaque brique est habillée de pastilles de couleurs qui facilitent les traitements d'identification ([Underkoffler 98]).

Built-it. Les interfaces saisissables suggèrent, nous l'avons vu, la manipulation à deux mains et à plusieurs. Ces avantages sont exploités dans le système **Built-It** ([Rauterberg 98] [Fjeld 98]). Le prototype est appliqué à la conception de plans d'usine mais le paradigme d'interaction est envisageable pour d'autres domaines tels que l'architecture d'intérieur et la planification urbaine.

Les utilisateurs sont assis autour d'une table sur laquelle est projeté le plan de l'usine en deux dimensions. Sur un mur proche de la table, est projetée une représentation en trois dimensions de l'usine. Toute l'interaction entre les utilisateurs et le système se fait à l'aide d'une brique. Les principales manipulations consistent à :

- attacher une brique à un concept du domaine (par exemple, une machine de l'usine) en posant la brique sur sa représentation graphique,
- déplacer et orienter la machine sur le plan en faisant glisser la brique qui lui est associée (les représentations 3D et 2D sont asservies à la brique),
- détacher la brique de la machine en plaçant la main au-dessus de la brique, puis en la ramassant.

Le suivi des briques est assuré par un système de vision par ordinateur. Nous n'avons pas trouvé de publication détaillant la réalisation et les performances de ce système.

Synthèse sur les interfaces saisissables

Les interfaces saisissables mettent à profit nos capacités expertes de préhension utilisées dans le monde réel. Au niveau cognitif, les phicons, représentations analogiques et tangibles de concepts, visent à favoriser la compréhension du fonctionnement du système. Comparée aux IHM traditionnelles, l'association d'objets physiques à des concepts électroniques améliore le caractère direct de l'interaction : l'intermédiaire logique qu'est le widget, et son acquisition par le pointeur de la souris, sont éliminés (voir l'illustration de la figure 11 page 27).

Les **Briques** de Fitzmaurice, le **MetaDESK** et **Built-It** relèvent des interfaces saisissables et montrent l'intérêt de la vision par ordinateur pour leur mise en œuvre : élimination des fils et relâchement des contraintes sur la nature et le nombre de briques.

Les interfaces digitales suppriment aussi un intermédiaire mais de nature différente.

3.2. INTERFACES DIGITALES
[Wellner 93b]

Nous allons successivement présenter le principe des interfaces digitales, puis, au regard d'un exemple représentatif, les points saillants de l'interaction ayant un impact sur la vision par ordinateur.

Le principe

Comme leur nom l'indique, les interfaces digitales utilisent le doigt comme base d'interaction. Dans une interface digitale, le doigt participe à un geste sémiotique : sa trajectoire et sa forme font sens [Cadoz 96]. Dans les interfaces saisissables, le doigt a un rôle ergotique : il applique une énergie à un instrument, par exemple à une brique, dont il modifie l'état. C'est l'interprétation de ce changement d'état qui fait sens. Alors que les interfaces saisissables éliminent le dispositif logique (le widget), les interfaces digitales conservent le dispositif logique mais suppriment le dispositif physique (l'instrument) : comme le montre la figure 16, le *doigt est le dispositif*. Il agit en direct sur l'information électronique. Nul besoin d'acquérir la souris ou la brique : le doigt est sur soi ! À l'instar de la souris, le doigt est multiplexé temporellement mais on dispose de plusieurs doigts répartis sur deux mains. Les interfaces digitales ont donc la possibilité d'exploiter le multiplexage spatial. Et comme les interfaces saisissables, les interfaces digitales réutilisent les habiletés motrices naturelles.

Le "**Bureau digital**" est l'exemple pionnier de ce type d'interfaces.

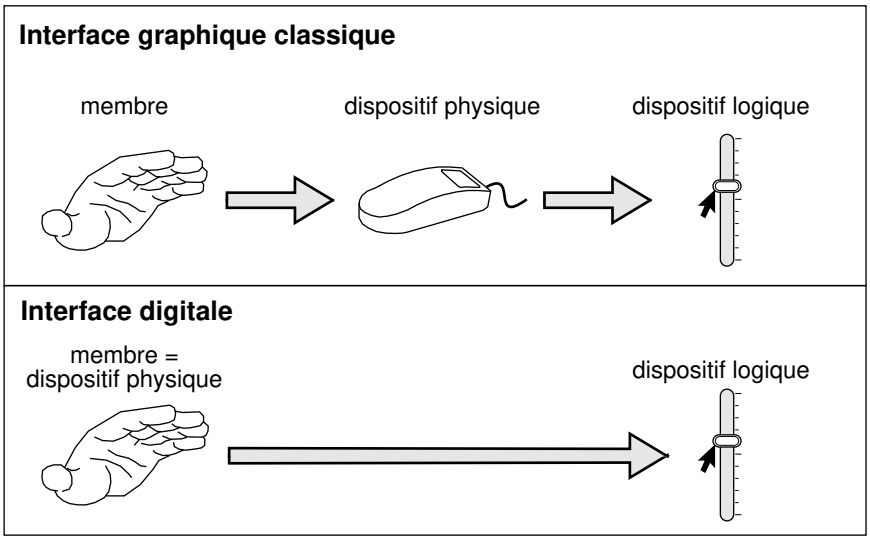


Figure 16
Aspect direct de l'interaction fondée sur les interfaces digitales
La main de l'utilisateur contrôle directement le dispositif logique (par exemple un widget). Le dispositif physique intermédiaire est supprimé.

Le bureau digital

Le **Bureau Digital**, défini par Wellner ([Wellner 91], [Wellner 93b]), est avant tout un bureau physique. Il en possède tous les attributs : on peut y entreposer ses livres, stylos, gommes, et tasse à café. Le **Bureau Digital** est équipé d'une caméra et d'un projecteur vidéo disposés au-dessus du plan de travail. La caméra est reliée à un système de vision par ordinateur destiné à interpréter les gestes de l'utilisateur, tandis que le projecteur restitue sur le plan de travail l'information numérique en provenance du système. Ce dispositif permet *d'augmenter* le bureau physique qui nous est familier de services électroniques.

L'idée est de conserver ce qui fonctionne bien dans le monde physique mais de l'amplifier par des services électroniques dont l'équivalent dans le monde physique est difficile, voire impossible, à réaliser. Dans le cas du **Bureau Digital**, l'utilisateur conserve ses feuilles de papier, crayons et gomme et la duplication de texte, opération complexe dans le monde physique, devient une opération triviale sous forme électronique. Par exemple, l'utilisateur sélectionne le paragraphe d'un livre posé sur le bureau. Il se sert de ses deux index pour désigner les extrémités opposées d'un rectangle de sélection. Le système projette un retour d'information immédiat sur le livre : le rectangle de sélection s'affiche sur le livre en suivant les déplacements des doigts. Cette manipulation est illustrée sur la figure 17. Une fois le texte sélectionné, le lecteur peut dupliquer et déplacer le paragraphe en faisant glisser le rectangle de sélection au moyen du doigt.

Le prototype du **Bureau Digital**, imaginé par Wellner et présenté sous forme d'une vidéo n'est en fait qu'une simulation destinée à démontrer le paradigme d'interaction. Au moment de sa conception, Wellner ne dispose pas des connaissances et techniques nécessaires à la réalisation d'un prototype entièrement fonctionnel. Seul un suivi a été mis en œuvre pour expérimenter la désignation au doigt.

Figure 17
Sélection au doigt sur le Bureau Digital (extrait d'une vidéo non publiée de Xerox EuroPARC)
L'utilisateur désigne la zone à sélectionner avec les doigts de ses deux mains. Le système projette un rectangle de sélection asservi aux déplacements des doigts.

Désignation au doigt

Le **Bureau Digital** met en oeuvre un système de vision par ordinateur pour localiser, à tout instant, le doigt sur le bureau. La technique employée par Wellner est basée sur la différence d'images que nous présentons en détail au paragraphe "Suivi par différence d'images" (chapitre IV page 83).

Du fait de sa taille (un doigt couvre environ une vingtaine de pixels), le doigt est inadapté aux pointages de précision. Cette limitation tient à la taille du doigt et non pas au manque de précision de ses mouvements. Une solution à ce problème est d'afficher un pointeur, similaire à celui de la souris, à proximité du doigt : les déplacements du pointeur reflètent exactement ceux du doigt. Pour désigner un emplacement, l'utilisateur ne considère pas la position du doigt mais celle du pointeur. Cette forme d'interaction correspond à la partie "interface digitale" de la figure 16 page 33 : la main est directement associée au pointeur qui permet de contrôler les dispositifs logiques (barres de défilement, boutons) de l'interface.

Du point de vue technique, l'association d'un pointeur au doigt a une conséquence subtile, mais importante. Si le doigt pouvait être utilisé "nu", c'est-à-dire sans pointeur associé, l'utilisateur n'aurait pas besoin de retour d'information système pour amener son doigt à l'emplacement désiré. Le système pourrait se contenter de localiser le doigt une fois que celui-ci marque une pause. Par contre, si l'utilisateur doit positionner le pointeur sur un emplacement cible, le système doit être capable de localiser le doigt et de mettre à jour la position du pointeur en temps réel. On assiste ici à un cas d'interaction fortement couplée qui ne souffre donc pas de délai.

Instants d'intérêt

Les déplacements du pointeur ne suffisent pas à eux seuls à exprimer une désignation : le pointeur est le plus souvent en transit entre deux positions "d'intérêt". Dans les interfaces usuelles, ce problème est résolu avec le clic d'un bouton de la souris. Quel en est l'équivalent pour une interaction digitale ? Le prototype du **Bureau Digital** utilise une tape sur le bureau. Le contact du doigt sur le bureau est détecté au moyen d'un microphone fixé sous le bureau. Lorsque l'utilisateur effectue une tape sur le bureau, l'intensité du signal sonore capté par le microphone subit un pic que le système détecte. Cette solution a des limites : le contact n'est pas détecté si l'utilisateur tape sur un livre. Inversement, un contact est détecté alors qu'il s'agit de la chute involontaire d'un objet sur la surface de travail. De plus, cette solution nécessite un dispositif audio et la mise en œuvre de traitements du signal audio. Nous verrons au paragraphe "Détection des pauses" (chapitre V page 136) notre solution à ce problème en nous appuyant sur une analyse spatio-temporelle de la trajectoire du doigt.

4. Résumé du chapitre

Dans ce chapitre, nous avons défini le concept d'*interaction fortement couplée*. Cette notion traduit l'interdépendance de l'action et de la perception humaine et artificielle. Le couplage qui nous intéresse se situe à bas niveau d'abstraction : celui des habiletés humaines et, du côté machine, l'acquisition et le rendu d'information. Nous modélisons la situation d'interaction fortement couplée sous forme d'un système en boucle fermée qui permet d'exprimer les relations entre action et perception ainsi que le rôle du temps dans la qualité du couplage.

L'interaction fortement couplée intervient de manière fondamentale dans les nouvelles formes d'interaction. Celles-ci s'appuient toutes sur l'exploitation du monde physique qui nous est familier. Nous les avons regroupées en deux familles qui traitent "le familier" de manière duale :

- Les *systèmes immersifs*, de type *venir tel quel* ou de type *porter tout ce qui est nécessaire*, qui visent à reproduire électroniquement l'environnement physique familier de l'utilisateur. Cette famille de systèmes joue le jeu de la virtualité mais utilise la physicalité comme modèle ;
- Les *systèmes de Réalité Augmentée* offrant soit des *interfaces saisissables* soit des *interfaces digitales* qui introduisent l'électronique dans le monde physique familier. Ces systèmes conservent la physicalité et y introduisent la virtualité.

Comme effet de bord à la mise en œuvre de situations familières, nous constatons la suppression des intermédiaires dans l'interaction mais aussi l'ouverture vers l'exploitation du parallélisme. L'interaction directe devient plus directe. Elle devient aussi bimanuelle et multi-utilisateur. Concernant l'aspect direct, les exemples étudiés révèlent plusieurs manières de réduire les intermédiaires : élimination du dispositif logique dans le cas des interfaces saisissables (voir la figure 11 page 27) ou bien élimination du dispositif physique dans le cas des interfaces digitales (voir la figure 16 page 33). Notre concept de fenêtre perceptuelle présenté au chapitre VI supprime à la fois les dispositifs logique et physique : il va en ce sens un pas plus loin.

Tous les exemples d'interaction fortement couplée présentés dans ce chapitre démontrent l'intérêt de la vision par ordinateur comme technique d'acquisition d'information sur le comportement humain. Ils imposent aussi des requis de performance. En particulier, nous avons relevé la nécessité de :

- calculer la position d'une entité (un objet physique ou un membre de l'utilisateur),
- produire le retour d'information correspondant à la position de l'entité suivie.

Nous précisons au chapitre suivant la nature des requis pour la mise en œuvre de l'interaction fortement couplée au moyen de la vision par ordinateur.

Chapitre II *Requis de l'interaction fortement couplée*

Les familles de systèmes présentées au chapitre précédent, expriment en filigrane, les requis de l'interaction fortement couplée. Si les conditions du couplage ne sont pas (ou ne peuvent pas être) satisfaites, il convient d'en identifier les raisons. Si les causes ne peuvent être résolues sur le plan technique, ou si la solution implique l'introduction de contraintes supplémentaires inacceptables pour l'utilisateur dans son contexte d'usage, le concepteur d'interface se doit d'envisager une autre modalité d'interaction. L'objet de ce chapitre est de préciser les requis de l'interaction fortement couplée pour raisonner sur le bien fondé du choix ou du rejet d'un dispositif.

Ce chapitre est organisé en deux parties : les requis fonctionnels de l'interaction fortement couplée puis les requis non fonctionnels en précisant autant que possible des métriques de fonctionnement.

1. Requis fonctionnels

Les requis fonctionnels d'un système désignent l'ensemble des services attendus de ce système. Lorsqu'il s'agit d'un système logiciel, ces services peuvent couvrir plusieurs niveaux d'abstraction. Nous considérons les services de base que l'on regrouperait dans une boîte à outils de niveau équivalent à celui de la bibliothèque Xlib ([Nye 88]). Nous présentons un espace de services suivi d'une présentation succincte de chacun d'eux.

1.1. ESPACE DES SERVICES

Nous considérons uniquement les services dont l'utilité est démontrée par la revue des systèmes présentés au chapitre précédent :

- la détection,
- l'identification,
- le suivi.

Avant de définir la couverture de la détection, de l'identification et du suivi, nous en précisons les paramètres communs. Ces services peuvent être appliqués à des *entités* de nature distincte imposant chacune des requis : certaines entités sont rigides, d'autres déformables ; certaines sont inanimées, d'autres sont mobiles. On distingue :

- les objets de la vie courante comme les phicons du **MetaDesk**, les briques de Fitzmaurice, une gomme, etc.
- tout ou partie du corps humain : le doigt dans le cas du prototype opérationnel du **Bureau Digital**, la main et les pieds dans **ALIVE**, le visage pour la Fenêtre Virtuelle, le corps tout en entier dans **VideoPlace**.
- une activité d'agent vivant (une personne, un animal, etc.): le fait qu'une personne monte un escalier, change d'orientation, etc.

Les services de détection, d'identification et de suivi dépendent non seulement de la nature des entités sur lesquelles elles opèrent mais aussi sur la *cardinalité* : à un instant donné, le service considéré est-il capable de traiter une ou plusieurs entités ? Si plusieurs entités sont permises, quel en est le nombre maximal ? Sont-elles de même nature ou sont-elles hétérogènes ? Peuvent-elles se recouvrir et donc se confondre dans l'image ?

Le multiplexage spatial de **Briques** ou du **MetaDesk**, qui permet à un instant donné l'association de plusieurs objets mobiles, nécessite des services de :

- détection : y-a-t-il une ou plusieurs brique(s) ?
- identification : dans l'ensemble des phicons possibles, lesquelles sont présentes ?
- de suivi : quelle est la trajectoire de la (ou des) brique(s) ?

Le **Bureau Digital** et son successeur **Ariel** ([Mackay 95]), vont au-delà de ces requis : la scène observée peut contenir des entités de classes distinctes : doigt, main, gomme, crayon, phicons, etc.

Enfin, Il convient de s'interroger sur *l'espace* : les services fonctionnent-ils pour un univers planaire (comme dans **VideoPlace** et le **Bureau Digital**) ou bien sont-ils capables de gérer un espace à trois dimensions ?

En synthèse, pour une situation d'interaction donnée, la nature des entités, leur cardinalité et leur espace définissent les paramètres des services de base. Nous présentons maintenant chacun des services par ordre de complexité croissante.

1.2. DÉTECTION

La détection détermine l'occurrence d'une classe d'entité : pour une classe d'entité donnée, la sortie de cette fonction exprime l'existence (ou l'absence) d'une entité relevant de cette classe, par exemple, la présence d'une brique ou d'un visage. Dans le cas de **Built-It**, la présence de la brique indique au système l'activation des composants logiciels gérant la brique : le service de suivi doit être enclenché. L'occultation de la brique par la main, signale la désactivation du suivi et l'activation du service de détection. S'il y a plusieurs briques possibles, le service d'identification entre en jeu.

1.3. IDENTIFICATION

L'identification vise à désigner parmi les exemplaires d'une classe, le (ou les) exemplaire(s) présent(s) dans la scène. Dans le cas du **MetaDesk**, ce service permet de distinguer les différentes phicons, porteuses chacune de sens. Un autre exemple serait le visage *V* qui correspondrait, au niveau applicatif, à une personne autorisée à utiliser le système. Dans une interaction multi-utilisateur à plusieurs mains, le service d'identification devrait être en mesure d'extraire de la scène, les mains qui appartiennent à chaque utilisateur. La présence de deux mains peut correspondre à deux situations de sémantique distinctes. Il peut s'agir des mains d'une même personne, ou bien de mains appartenant à des personnes différentes.

1.4. SUIVI

Le suivi concerne les entités mobiles. Il s'agit de localiser en permanence une entité sans que cette entité soit munie de dispositif de repérage "artificiel". L'entité doit être, si possible, "telle quelle". L'information de position est plus ou moins complexe si l'entité évolue dans un espace à deux dimensions comme les briques ou le doigt du **Bureau Digital** ou dans un espace à trois dimensions tel que **ALIVE**.

Le suivi s'appuie sur les services de détection et d'identification s'il doit gérer respectivement plusieurs classes d'entités et plusieurs entités d'une même classe. Puisqu'il suit des entités, il constitue le service de base indispensable à la réalisation de l'interaction fortement couplée. Au chapitre IV, nous montrerons sa mise en œuvre technique au moyen de la vision par ordinateur. Dans la section qui suit, nous en précisons les requis de fonctionnement.

2. Requis non fonctionnels

Les requis non fonctionnels désignent un ensemble de propriétés permettant d'évaluer la qualité, et par suite, l'adéquation d'un service à une situation donnée. Les propriétés peuvent être qualitatives ou quanti-

tatives. Dans notre analyse, nous privilégions les propriétés quantitatives comme éléments d'évolution de l'Interaction Homme-Machine vers une Science de l'Ingénieur.

Les requis non fonctionnels que nous avons retenus au regard de l'interaction fortement couplée se répartissent comme suit :
- la *latence* qui caractérise le comportement du couplage,
- la *résolution* et la *stabilité statique* qui mesurent la qualité des informations rendues par le système de suivi.

Nous rappelons en 2.1 le modèle qui sert de fondement à notre analyse du temps de latence. Ce modèle, qui distingue l'utilisateur et le dispositif, nous conduit à raisonner sur le temps de latence de l'utilisateur en 2.2. et sur celui du dispositif en 2.3. En 2.4 nous présentons une synthèse pour le requis de latence. En 2.5, nous abordons les propriétés des informations rendues par le dispositif.

2.1. SYSTÈME EN BOUCLE FERMÉE ET LATENCE

L'interaction fortement couplée, nous l'avons vu, forme un système en boucle fermée (voir chapitre I page 9). Cette boucle d'asservissement traduit l'engagement des systèmes humain et artificiel dans l'accomplissement d'actions physiques mutuellement perceptibles et dépendantes sur un intervalle de temps donné. Nous recopions sur la figure 1 le schéma illustratif du modèle introduit initialement au chapitre I page 10. Dans le cadre de notre analyse, le système artificiel est un dispositif de suivi.

Le cycle de la figure 1 peut aussi se voir comme le couplage de deux systèmes de type stimulus / réponse :
1. l'utilisateur pour lequel le retour d'information du dispositif de suivi constitue un stimulus qui engendre une réponse utilisateur,
2. le dispositif de suivi pour lequel la réponse de l'utilisateur constitue un stimulus auquel il réagit par un nouveau retour d'information.

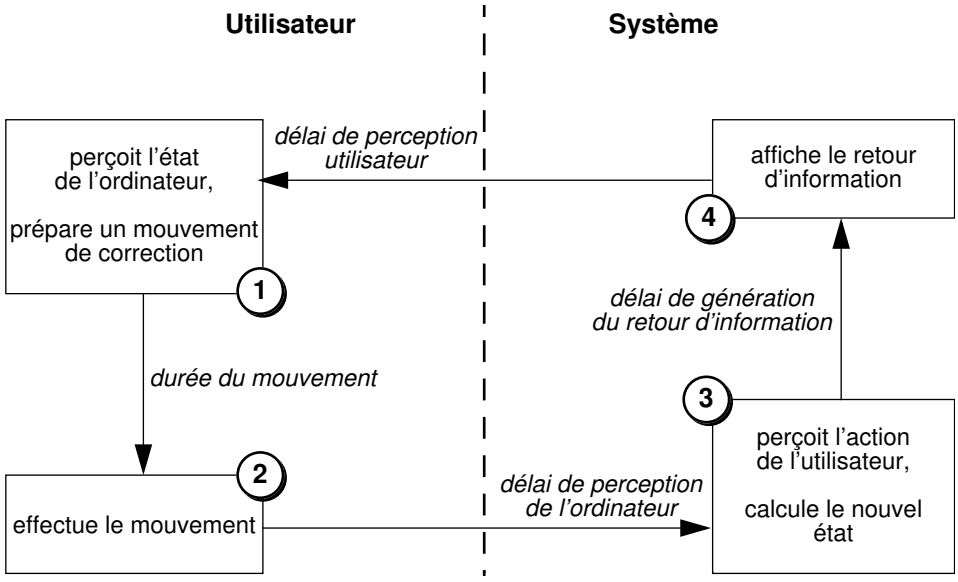


Figure 1
Boucle de l'interaction fortement couplée (d'après [Ware 94])

Le temps de latence est utilisé en psychologie et en physique comme un indicateur du fonctionnement interne d'un système de type stimulus / réponse. Nous l'exploitons à notre tour comme expression de requis.

Le *temps de latence* L (ou plus simplement, *latence*) d'un système est le temps écoulé entre la présentation d'un stimulus à ce système et le début de la réponse correspondante. Pour une granularité de temps donnée, la latence est en règle générale constante, ou peu variable. Dans le cas d'un processus synchrone, la fréquence de fonctionnement F du système est égale au nombre de couples "stimulus réponse" ayant lieu dans une seconde, on peut donc associer la latence à l'inverse de la fréquence de fonctionnement du processus, soit $L = 1/F$.

Dans le modèle de la figure 1, la latence du dispositif de suivi est la somme des délais nécessaires au passage de l'état 2 à l'état 4. Il s'agit des délais de :
- perception du mouvement de l'utilisateur,
- calcul du nouvel état,
- génération du retour d'information.

Ces délais peuvent être mesurés par une instrumentation du système de suivi en des points idoines.

La latence du système humain est la somme des délais nécessaires au passage de l'état 4 à l'état 2 de la figure 1. Il s'agit des délais de :
- perception du retour d'information du système,
- préparation du mouvement,
- exécution du mouvement.

L'estimation des délais humains est plus complexe car il n'est pas possible (ou alors extrêmement difficile) d'instrumenter les mécanismes de perception, de cognition et d'action de l'Homme. L'alternative est l'utilisation de modèle prédictif.

2.2. LATENCE DE L'UTILISATEUR

Nous fondons notre analyse de la latence humaine sur un modèle prédictif quantitatif issu de la psychologie cognitive : le Modèle du Processeur Humain [Card 83]. Nous le présentons brièvement dans la section qui suit et montrons comment nous l'exploitons pour mesurer la latence du système humain engagé dans une interaction fortement couplée. Cette latence, conjointement à la Loi de Fitts, servira de base à la détermination de la latence souhaitée pour le système de suivi.

Modèle du Processeur Humain
[Card 83]

Le "Modèle du Processeur Humain" de Card, Moran et Newell représente l'individu comme un système de traitement de l'information [Card 83]. Comme le montre la figure 2, le processeur humain comprend trois sous-systèmes indépendants : les systèmes sensoriel, moteur et cognitif. La mise en correspondance du Modèle du Processeur Humain avec le modèle

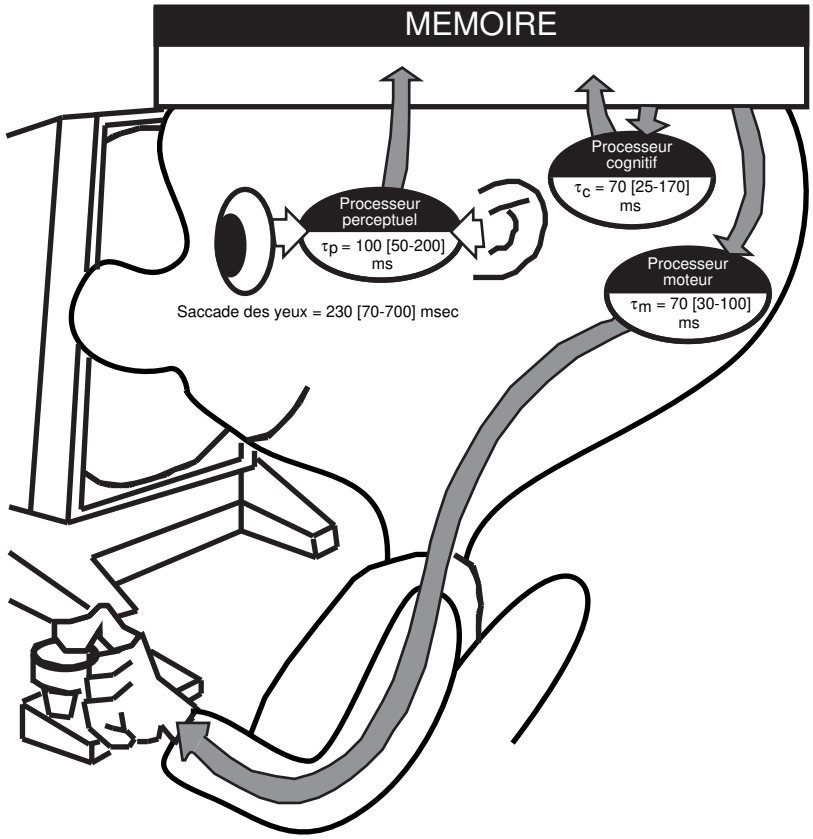


Figure 2
Représentation schématique du modèle du processeur humain (d'après [Card 83])
L'utilisateur est modélisé par un ensemble de trois processeurs et une mémoire.

de Ware et Balakrishnan indique que le processeur sensoriel est la source du "délai de perception utilisateur" (passage de l'état 4 à l'état 1), que le processeur cognitif est responsable de la "préparation du mouvement de correction" (état 1), et que le processeur moteur est la cause de la "durée du mouvement" (passage de l'état 1 à l'état 2).

Chaque sous-système du Modèle du Processeur Humain dispose d'une mémoire locale et d'un processeur dont les performances sont caractérisées par des métriques. Parmi ces métriques, nous retenons, pour les besoins de notre analyse, le *temps de cycle* d'un processeur. Card et ses co-auteurs ne fournissent pas de définition précise de cette notion ([Card 83], p. 25). Mais leur utilisation dans l'analyse des performances présentées dans ([Card 83], pages 25 à 86), nous permet d'assimiler le temps de cycle d'un processeur à sa latence.

S'appuyant sur de nombreux résultats expérimentaux, Card et ses co-auteurs relèvent que les temps de cycle sont constants pour une personne donnée mais varient en fonction des conditions (stress, fatigue, vivacité du stimulus) et en fonction des personnes. Les bornes inférieures et supérieures des temps de cycle de chaque processeur, de même que les valeurs nominales sont représentées sur la figure 2. Ces valeurs sont également rapportées sur la table 1. Un temps de cycle minimal correspond aux meilleures performances (sujets les plus rapides dans les

Processeur	Notation	Temps de cycle en milliseconde (ms.)		
		Minimal	Nominal	Maximal
Perceptuel	τ_p	50	100	200
Cognitif	τ_c	25	70	170
Moteur	τ_m	30	70	100

Table 1 : Temps de cycle des différents processeurs du modèle du processeur humain ([Card 83]).

meilleures conditions dit "Fastman"). Inversement, un temps de cycle maximal, correspond aux personnes les plus lentes dans les pires conditions de stress et de fatigue ("Slowman"). Les valeurs nominales ("Middleman") désignent des valeurs typiques entre les deux bornes extrêmes Fastman et Slowman.

Card, Moran et Newell utilisent les temps de cycle pour analyser ou prédire l'adéquation de certains systèmes aux utilisateurs. Par exemple, le principe du cinéma est analysé ainsi : l'objectif est de projeter des images fixes à une fréquence suffisamment élevée pour que les spectateurs aient une impression d'image animée continue. La perception du flux visuel ne fait intervenir que le processeur sensoriel visuel. Celui-ci ne peut distinguer la succession d'images fixes si elles sont présentées à une fréquence supérieure à sa fréquence de fonctionnement. Pour le "Fastman", la fréquence de fonctionnement du processeur sensoriel est $1/\tau_p = 1/0.05$ soit 20 Hertz (Hz). Le modèle du processeur humain permet ainsi d'estimer que les images d'un film doivent être présentées à une fréquence supérieure à 20 Hz pour être perçues comme une séquence animée continue.

Le modèle ne prétend pas fournir des valeurs précises mais des seuils (Fastman et Slowman) et des valeurs indicatives (Middleman). Ces valeurs permettent de calculer des ordres de grandeur des performances requises d'un système artificiel. Reprenant l'exemple du processeur visuel, le modèle permet de prédire que la fréquence ne doit pas être inférieure à 20 Hz et qu'elle doit être de l'ordre de 20 Hz. Reprenant l'exemple du cinéma, à 20Hz des oscillations de luminosité sont perceptibles. Pour cette raison, le cinéma utilise traditionnellement une fréquence de 24 Hz. Inversement, un rafraîchissement de 200 images par seconde serait inutile : le processeur visuel, même dans les meilleures conditions, serait incapable de discerner toutes les images. Il convient cependant de noter que le temps de cycle du processeur visuel varie inversement avec l'intensité du stimulus. Les moniteurs vidéo, qui produisent une image au moyen d'éléments phosphorescents, sont plus lumineux que les écrans de projection du cinéma. Parce que le stimulus produit est plus intense, le temps de cycle du processeur visuel est plus

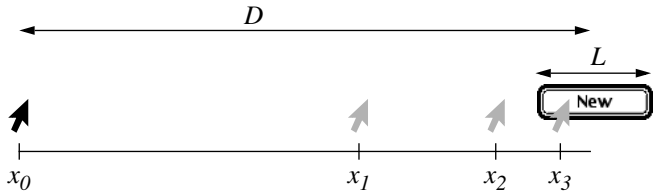


Figure 3
Acquisition d'une cible de largeur L à une distance D du pointeur
Le mouvement du pointeur est une suite de micro-mouvements, la précision relative de chaque mouvement est supposée constante ($x_1/x_0 = x_2/x_1$...).

rapide. Il est alors nécessaire d'utiliser des fréquences élevées (de l'ordre de 60 Hz) pour rendre imperceptibles les oscillations de luminosité.

La Loi de Fitts, présentée ci-dessous, concerne le processeur moteur impliqué, comme le processeur visuel, dans l'interaction fortement couplée.

Loi de Fitts

[Fitts 53]

La loi de Fitts permet de prédire le temps moyen d'acquisition d'une cible en fonction de la taille de la cible et de la distance initiale entre l'effecteur (par exemple le pointeur de la souris) et la cible. Cette action est représentée sur la figure 3. L'acquisition de cible est modélisée comme une séquence de micro-mouvements destinés à approcher le pointeur de la souris de la cible. Chaque micro-mouvement correspond à un cycle du système en boucle fermée :

- l'utilisateur perçoit la position courante du pointeur et celle de la cible identifiant ainsi la trajectoire à parcourir pour atteindre la cible,
- l'utilisateur exécute un micro-mouvement destiné à approcher le pointeur de la cible,
- le micro-mouvement est pris en compte pas le système qui met à jour le pointeur,
- la nouvelle position du pointeur, sortie du système, est nécessaire à la génération du micro-mouvement suivant : elle est utilisée en entrée du système au cycle suivant.

Ce modèle permet de montrer que le temps moyen d'acquisition d'une cible est donné par la relation suivante :

$$MT = \frac{1}{IP} \cdot ID \qquad ID = \log_2\left(\frac{2D}{L}\right) \qquad (1)$$

où *MT* est le temps moyen d'acquisition de la cible, *IP* est l'*indice de performance* du dispositif de pointage étudié, *ID* est l'*indice de difficulté* correspondant à l'action d'acquisition, *D* est la distance initiale du pointeur à la cible et *L* est la taille de la cible. L'indice de performance (IP) est une constante liée au dispositif de pointage utilisé. Il est déterminé de façon empirique. L'indice de difficulté (ID) est proportionnel au nombre de micro-mouvements nécessaires. Il augmente lorsque la tâche devient plus difficile, c'est-à-dire lorsque la distance augmente ou lorsque la taille de la cible diminue.

La loi de Fitts a été régulièrement vérifiée par de nombreuses expérimentations[1] ([MacKenzie 92], [Balakrishnan 97], [Accot 97], [Douglas 99]). Nous considérons ces vérifications expérimentales comme une validation du modèle sous-jacent de système en boucle fermée.

Ayant introduit les modèles quantitatifs prédictifs nécessaires à notre analyse, nous sommes en mesure de calculer la latence de l'utilisateur dans une situation d'interaction fortement couplée.

Calcul de la latence utilisateur

L'interaction fait intervenir les trois processeurs : le processeur sensoriel pour percevoir l'état courant du système, le processeur cognitif pour calculer le micro-mouvement suivant, et le processeur moteur pour effectuer le micro-mouvement proprement dit. Le Modèle du Processeur Humain fait l'hypothèse que ces processeurs peuvent fonctionner en parallèle. Dans le cas de la situation stimulus / réponse, le parallélisme n'est pas possible : chaque processeur a besoin des données du processeur précédent pour exécuter sa fonction.

Si les processeurs s'exécutent en séquence de manière synchrone, la latence utilisateur L_u est la somme des temps de cycle des trois processeurs, soit :

$$L_u = \tau_p + \tau_c + \tau_m \qquad (2)$$

Dans le cas des sujets les plus rapides placés dans des conditions minimales de stress et de fatigue, on estime la latence L_{ur} (utilisateur rapide) du mécanisme stimulus / réponse humain à :

$$L_{ur} = \tau_p + \tau_c + \tau_m = 50 + 30 + 25 = 105 \text{ ms} \qquad (3)$$

Parce que nous raisonnons sur des ordres de grandeur, nous ramenons la valeur indicative de 105 ms. à une valeur plus exigeante du même ordre, soit :100 ms.

2.3. LATENCE DU SYSTÈME

Jusqu'ici, on ne s'est pas préoccupé de la notion de latence des systèmes à interface graphique classique parce que les dispositifs d'entrée impliqués convenaient aux exigences de l'interaction fortement couplée. Pour la souris par exemple, le délai de perception du système est dû à la numérisation d'un simple signal électrique et l'acheminement de deux valeurs dans le système : les positions en abscisse et ordonnée de la souris. Ces opérations ne nécessitent que quelques millisecondes et sont négligeables par rapport à la latence humaine. Le temps de génération du retour d'information est également faible : il s'agit en général de déplacer le

1. Scott MacKenzie maintient une bibliographie de travaux de recherches sur la loi de Fitts regroupant plus de 270 entrées à l'adresse suivante : http://www.uoguelph.ca/~imackenz/Fitts_bib.html

pointeur de la souris sur l'écran. Cette opération se traduit par le déplacement d'une faible quantité de données, le motif du pointeur de la souris, dans la mémoire d'écran graphique. Au final, la latence d'un système à interface graphique classique est de l'ordre de 10 ms., délai presque imperceptible au regard des 100 ms. de latence des utilisateurs les plus rapides.

Dans le cas des systèmes immersifs et de réalité augmentée, nous l'avons vu au chapitre précédent, les nouveaux dispositifs d'interaction (capteurs magnétiques, vision par ordinateur) et la richesse des retours d'information de synthèse graphique, mettent en jeu des coûts de calcul bien supérieurs à ceux des interfaces usuelles. Notre revue de la littérature indique implicitement que, pour ces systèmes, le temps de latence devient une mesure fondamentale à l'utilisabilité. Ware fait état d'une hypothèse intuitive utilisée en informatique graphique selon laquelle la perception de mouvements fluides à l'écran nécessite une fréquence de fonctionnement de 10 Hz (équivalent à une latence de 100 ms pour des dispositifs synchrones) [Ware 94]. Selon les notes de Ware, les spécialistes du domaine s'accordent à dire que cette valeur n'est pas satisfaisante. Nous affinons cette hypothèse par déduction analytique et prédisons son effet sur les performances utilisateur. Nous confortons ensuite notre raisonnement par plusieurs expériences empiriques relevées dans la littérature.

Déduction analytique du temps de latence du dispositif

La latence idéale du dispositif est de toute évidence une latence nulle. Le fonctionnement résultant est représenté par le cas 1 de la figure 4 : dès que l'utilisateur a fini son mouvement, il est capté par le dispositif et provoque l'apparition immédiate d'un nouveau retour d'information. La phase de perception de l'utilisateur, au cycle suivant, s'effectue sur un retour d'information "à jour" par rapport au mouvement qui a immédiatement précédé la phase de perception. En pratique il est impossible de réaliser une latence de dispositif nulle (ou même négligeable). Dans ces conditions, il est impossible de fournir un retour d'information à jour d'un cycle utilisateur à l'autre. On va chercher, au mieux, à fournir un retour d'information ayant un seul cycle utilisateur de retard.

Supposons que la latence du dispositif soit supérieure à la latence utilisateur. Supposons que, idéalement, le dispositif débute son cycle de stimulus / réponse au moment où l'utilisateur vient de finir son premier mouvement. La latence utilisateur étant plus courte que celle du dispositif, l'utilisateur débute la phase de perception du cycle suivant alors que le dispositif n'a pas encore mis à jour son retour d'information. Ce cas est illustré par le cas 2 de la figure 4 : l'utilisateur perçoit un retour d'information correspondant à 2 cycles de retard. Évidemment, plus la latence du dispositif est grande, plus le nombre de cycles de retard augmente.

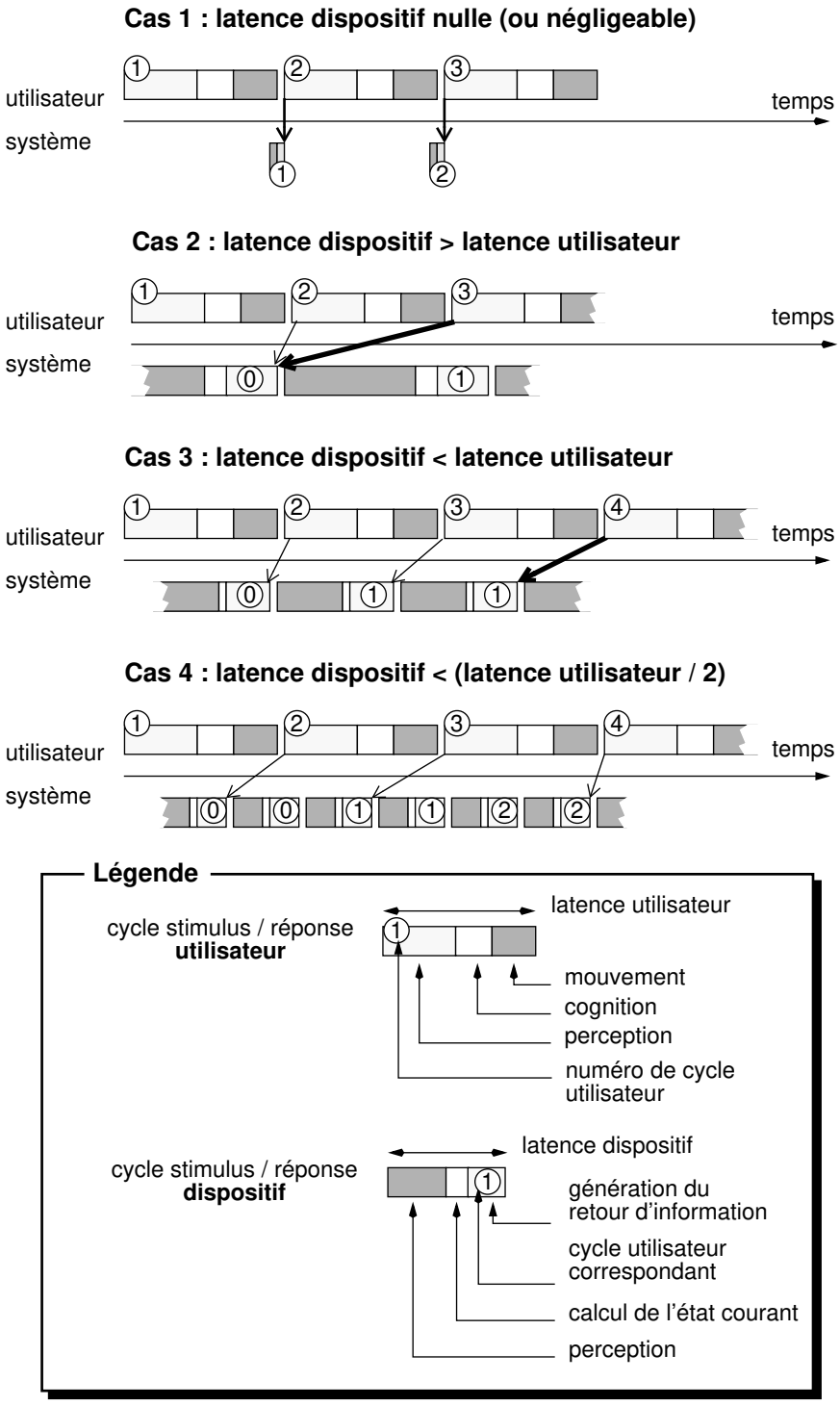


Figure 4
Latence utilisateur et latence dispositif
Cas 1 : cas idéal : l'utilisateur perçoit toujours le retour d'information correspondant au cycle précédent.
Cas 2 : le dispositif n'a pas encore généré le retour d'information du cycle 1 lorsque l'utilisateur entre dans le cycle 3. La phase de perception du cycle 3 considère le retour d'information correspondant au cycle 0 (2 cycles de retard).
Cas 3 : au cycle 4, l'utilisateur perçoit le retour d'information correspondant au cycle 1 (2 cycles de retard).
Cas 4 : cas idéal pour une latence dispositif non nulle : l'utilisateur perçoit toujours le retour d'information avec un seul cycle de retard.

Supposons maintenant que la latence du dispositif soit légèrement inférieure à celle de l'utilisateur. Lorsque le dispositif débute son cycle de stimulus / réponse immédiatement après un mouvement, le retour d'information apparaît avant la phase de perception utilisateur du cycle suivant. Par contre, le dispositif débute un nouveau cycle alors qu'il n'est plus en phase avec l'utilisateur : au cycle suivant, le retour d'information perçu par l'utilisateur ne correspond pas au dernier mouvement. Le cas 3

de la figure 4 illustre cette situation : certains cycles utilisateur s'exécutent sur un retour d'information qui a deux cycles de retard.

Notre raisonnement indique que la fréquence de fonctionnement du dispositif de suivi doit être au moins supérieure au double de la fréquence du cycle utilisateur. Avec l'hypothèse que l'utilisateur et le dispositif de suivi sont des systèmes synchrones, la latence du dispositif doit être au moins deux fois moindre à celle de l'utilisateur, soit :

$$L_{dispositif} \leq (L_{ur}/2 = 50 \text{ ms.}) \quad (4)$$

Au-delà du seuil de 50 ms environ, certains cycles stimulus / réponse consécutifs de l'utilisateur sont exécutés, on l'a vu, sur le même retour d'information du dispositif (cas 2 et 3 de la figure 4). Ce phénomène est d'autant plus sensible lorsque la latence du dispositif est largement supérieure à celle de l'utilisateur, provoquant l'exécution d'un grand nombre de cycles utilisateurs sur le même retour d'information du système. À partir de là, quelles prédictions peut-on énoncer sur le comportement de l'utilisateur ?

Conséquences prévisibles de la latence du dispositif

Les modèles qui nous servent de fondement nous permettent de prédire deux conséquences sur le comportement humain : redondance et oscillation

Redondance. Les cycles stimulus / réponse effectués par l'utilisateur sur un même retour d'information du dispositif sont *redondants* : en l'absence de mise à jour du retour d'information du dispositif, aucune correction de mouvement ne peut être évaluée par le processeur cognitif. Ces cycles inutiles se traduisent par une perte de temps.

Le temps perdu par redondance correspond à *n-1* fois la latence utilisateur si *n* est le nombre de cycles utilisateur exécutés sur le même retour d'information. Cette conséquence serait négligeable si la perte de temps n'avait lieu qu'une seule fois, ce qui n'est pas le cas : la nature itérative de l'interaction fortement couplée implique que le mouvement de l'utilisateur est décomposé en micro-mouvements, chaque micro-mouvement nécessitant un cycle stimulus / réponse *effectif* (par opposition aux cycles redondants). Il s'en suit que la perte de temps est à multiplier par le nombre de cycles effectifs nécessaires à l'accomplissement de l'action élémentaire. Ce nombre est proportionnel à l'indice de difficulté introduit au paragraphe "Loi de Fitts" page 46. En d'autres termes, plus la tâche à exécuter est difficile (petite cible située loin de la position de départ), plus la latence du dispositif aura d'effet sur le temps d'accomplissement de la tâche.

Oscillation. En apparence, différents stimulus de l'utilisateur provoquent la même réponse du dispositif. Certains mouvements de l'utilisateur sont donc perçus comme n'ayant pas d'effet sur le système. Un tel phénomène

incite le système cognitif à amplifier les mouvements pour "forcer" l'apparition d'un retour d'information. Les mouvements trop amples sont ensuite rapportés (avec du retard) par le retour d'information du système et doivent être compensés. La conséquence générale est une oscillation de comportement qui peut aboutir à l'erreur.

Les études empiriques rapportées ci-dessous confirment le seuil de latence d'un dispositif de suivi que nous avons établi de manière déductive à 50 ms. Les résultats de ces études sont également cohérents avec nos prédictions de l'effet de la latence sur le comportement humain.

Etudes empiriques du temps de latence du dispositif

[Liang 91]
[Deering 92]
[MacKenzie 93]
[Ware 94]

Avec un dispositif expérimental ingénieux basé sur un pendule, Liang [Liang 91] mesure avec précision la latence d'un capteur magnétique, un Polhemus **Isotrack**. En ajustant les paramètres du dispositif, la latence minimum observée est de 85 ms. Il propose d'appliquer un filtre prédictif sur les données du dispositif magnétique (filtre de Kalman que nous présentons en page 82) afin de minimiser les effets de la latence mais reconnaît les limites de cette approche. Deering établit un ensemble de requis pour les systèmes de réalité virtuelle visant la "haute résolution" [Deering 92]. Il reconnaît le problème du délai et estime, en conformité avec notre résultat analytique ci-dessus, que "le délai perçu ne devrait pas dépasser 50-100 ms."[1] ([Deering 92], page 198).

D'autres travaux plus récents ont cherché à mesurer et à modéliser la dégradation de performance due à la latence système. MacKenzie et Ware ([MacKenzie 93]) réalisent une expérimentation de type "Loi de Fitts" en faisant varier l'indice de difficulté pour des tâches d'acquisition de cible au moyen de la souris. Faisant varier la latence du dispositif (à 8.3, 25, 75, et 225 ms.), ils constatent que la latence du système a un effet significatif sur le temps d'accomplissement de la tâche et sur le taux d'erreur. Les performances sont :

- similaires pour les latences de 8.3 ms. et 25 ms,
- sensiblement inférieures pour une latence de 75 ms : le temps d'accomplissement de la tâche est supérieur de 16% pour une latence de 75 ms comparé à la latence de 8.3 ms,
- largement dégradées pour une latence de 225 ms : une dégradation de 63.9% comparée à la latence de 8.3 ms.

MacKenzie et Ware proposent de modéliser l'effet de la latence sur le temps moyen (TM) d'accomplissement de tâche en étudiant la représentativité de différentes équations par rapport aux données utilisateur mesurées. Ces équations sont rapportées dans la table 2. Le symbole ID_e apparaissant dans les équations dénote l'indice de difficulté effectif, une

1. "perceived lags should be no more than 50-100 ms."

#	Modèle pour TM (ms.)	Corrélation	Quantité de variance exprimée
1	$TM = 435 + 190 \cdot ID_e$	$r = 0,560$	$31,3\%$
2	$TM = 894 + 46 \cdot Latence$	$r = 0,630$	$39,8\%$
3	$TM = -42 + 246 \cdot ID_e + 3,4 \cdot Latence$	$r = 0,948$	$89,8\%$
4	$TM = 230 + (169 + 1,03 \cdot Latence) \cdot ID_e$	$r = 0,967$	$93,5\%$

Table 2 : Modèles du temps moyen d'accomplissement d'une tâche de type Fitts faisant intervenir (ou non) la latence (extrait de [MacKenzie 93]).

variante de l'indice de difficulté *ID* présenté en page 46 (se référer à [MacKenzie 92] pour le détail du calcul de ID_e).

D'après les statistiques calculées, les modèles qui correspondent le mieux aux données sont les modèles 3 et 4 de la table 2. Ces modèles expliquent respectivement 89,8% et 93,5% de la variance des données mesurées. Le modèle numéro 3 fait apparaître un facteur multiplicatif de la latence de 3,4. Ce coefficient, fortement supérieur à 1, indique que la latence a un effet très négatif sur les performances. Le modèle peut être affiné en tenant compte de l'interaction entre la latence et l'indice de difficulté : c'est le cas du modèle numéro 4 de la table 2 où la latence apparaît en coefficient multiplicatif de l'indice de difficulté. Notons que l'analyse présentée au point 1 du paragraphe "Redondance" page 50 permet d'anticiper ce phénomène de dégradation plus importante des performances pour des indices de difficulté élevés.

Dans cette même expérimentation, MacKenzie et Ware constatent également que le taux d'erreur est influencé par la latence de manière significative : comparé à la latence de 8.3 ms., le taux est identique pour la latence de 25 ms. Il est supérieur de 36% pour la latence de 75 ms, et supérieur de 214% pour la latence de 225 ms. Cette étude confirme notre analyse prédictive sur le phénomène d'oscillation introduit en page 50.

L'expérimentation que nous venons de décrire a été conduite pour un dispositif classique "écran graphique et souris". Ware et Balakrishnan [Ware 94] réalisent une expérience visant les mêmes objectifs, mesurer et modéliser l'effet de la latence sur les performances utilisateur, mais pour des systèmes de réalités virtuelles. Ici, la tâche de sélection s'exécute dans un espace en trois dimensions. Les utilisateurs portent un casque de visualisation stéréoscopique, la tête et la main sont suivies par des dispositifs magnétiques afin, respectivement, de produire l'effet de parallaxe par mouvement (voir page 18), et de permettre la désignation dans l'espace. Les résultats confirment la validité du modèle 4 de la table

2 : la latence intervient comme coefficient multiplicatif de l'indice de difficulté selon la formule :

$$TM = C_1 + 1,59 \cdot (LatenceUtilisateur + LatenceSysteme) \cdot ID \qquad (5)$$

D'après les différentes expériences rapportées dans [Ware 94], le coefficient de corrélation des données mesurées par rapport à ce modèle est compris entre 0.90 et 0.99. La constante C_1 dépend de la tâche. Elle représente le temps d'initiation et de terminaison (clic souris par exemple). Les valeurs de latence utilisateur rapportées sont comprises entre 100 et 250 ms, en cohérence avec la valeur de 100 ms estimée précédemment de manière analytique.

2.4. SYNTHÈSE : REQUIS POUR LA LATENCE DU DISPOSITIF

En résumé, les études à la fois analytique et empirique montrent que la latence d'un dispositif est un facteur essentiel de l'interaction fortement couplée. Comme le notent Ware et Balakrishnan [Ware 94] :

> "*(...) pour de petites cibles (indice de difficulté = 5,0), une simple sélection nécessitera 1,5 s. de plus pour une latence de 200 ms qu'en l'absence de latence. Pour de nombreux systèmes hautement interactifs, la sélection de cible est un service de base fondamental de l'interface, et ce type de dégradation de performance peut facilement faire la différence entre un système qui est perçu comme utile et un qui ne l'est pas*"[1].

Une latence de 50 ms constitue une valeur indicative de seuil à ne pas franchir pour le choix et / ou la mise en œuvre d'un dispositif d'interaction fortement couplée. Pour un système synchrone, une latence de 50 ms est équivalente à une fréquence de fonctionnement de 20 Hz.

La latence du dispositif comprend le cycle entier de stimulus / réponse, soit la somme des délais de perception, de calcul du nouvel état et de génération du retour d'information par le dispositif (ces étapes sont illustrées sur la figure 1 page 42). Notre travail concernant l'introduction d'un nouveau type de dispositif d'entrée, nous concentrons notre recherche sur la première phase du cycle : la perception des mouvements de l'utilisateur.

Nous étudierons aux chapitres suivants les techniques permettant de minimiser le délai dû à cette phase. Nous prenons note que le délai de perception ne doit pas dépasser Δ_p milliseconde., où $\Delta_p = 50 - \Delta_c - \Delta_g$, et

1. "(...) for selection of small targets (ID=5.0) a lag of 200 msec will cause a simple selection to take 1.5 seconds longer than it would without lag. In many highly interactive systems target selection is a fundamental building block of the interface, and this kind of performance degradation may easily make the difference between a system that is perceived as useful and one that is not."

Δ_c et Δ_g sont respectivement les délais de calcul du nouvel état et de génération du retour d'information.

La satisfaction du requis de latence minimale est de notre point de vue, le premier objectif à atteindre pour réaliser une interaction fortement couplée réellement utilisable. Or ce requis a été le plus souvent éludé dans les précédentes tentatives d'introduction de la vision par ordinateur comme support d'une interaction fortement couplée : la table 3 résume notre revue de la littérature concernant les latences et fréquences de fonctionnement. La fréquence est en règle générale inférieure ou égale à 10 Hz, soit inférieure ou égale à la moitié de la fréquence seuil estimée. Les exceptions notables sont le système de Toyama ([Toyama 98]) et le système **VideoPlace** de Krueger ([Krueger 90]).

Référence	Latence (ms)	Fréquence (Hz)
[Krueger 90]	33.3	30
[Wellner 91]	143-167	6-7
[Azarbayejani 93]	100	10
[Ullmer 97]	143	7
[Wren 97]	100	10
[Toyama 98]	33,3	30
[Kang 98]	250	4
[Rauterberg 98]	~ 100 (estimée sur vidéo)	~ 10 (estimée sur vidéo)

Table 3 : Latence et fréquence de fonctionnement, rapportées dans la littérature, des systèmes de vision par ordinateur intervenant dans une interaction fortement couplée.

Nous considérons que la non satisfaction de ce requis est responsable pour une grande part de l'absence, à l'heure actuelle, de systèmes de vision par ordinateur dans la mise en œuvre d'interaction fortement couplée. En cela, nous partageons l'opinion de MacKenzie [MacKenzie 93], exprimée à propos des systèmes de réalité virtuelle mais applicable à toute forme d'interaction fortement couplée :

> "*La latence doit être considérée sérieusement et reconnue comme un goulot d'étranglement majeur de l'utilisabilité. L'attitude courante de reconnaître l'existence de la latence mais "d'apprendre à vivre avec" deviendra de plus en plus inacceptable tandis que les systèmes de réalité virtuelle passent du statut de curiosité de recherche à celui d'outil.*
> (...)
> *Un aspect majeur* (pour rendre acceptable cette technologie par les utilisateurs) *est de réaliser une réponse quasi-immédiate du retour d'information graphique aux stimuli d'entrée*"[1]

Une latence satisfaisant au requis des 50 ms est une condition nécessaire mais non suffisante à l'utilisabilité du dispositif en situation d'interaction fortement couplée. La qualité des informations fournies en sortie du dispositif intervient également dans l'utilisabilité de ce dispositif.

2.5. QUALITÉ DES INFORMATIONS RENDUES PAR LE DISPOSITIF

Les dispositifs intervenant dans une interaction fortement couplée ont pour rôle de renseigner le système sur le mouvement de l'entité suivie : doigt, main, corps de l'utilisateur, objets de la vie courante, etc. (voir page 40). À cette fin, ces dispositifs captent en permanence la position de l'entité suivie. L'information générée est en règle générale une position et une orientation en deux dimensions (x, y, θ), ou une position et une orientation en trois dimensions (on parle alors des six degrés de liberté suivants : la position de l'entité dans l'espace x, y, z et les trois angles définissant son orientation α, β, γ).

Quelle que soit la nature des données capturées, il existe toujours une différence entre la valeur effectivement mesurée et la position réelle de l'entité suivie. Cette différence, appelée *erreur de mesure*, est le résultat de plusieurs phénomènes liés à la réalisation du dispositif. Les données générées par la souris ont des caractéristiques telles que l'erreur de mesure est imperceptible. Il n'en est pas de même pour des dispositifs de suivi destinés à remplacer la souris, là où son usage est inapplicable.

L'erreur de mesure peut être *caractérisée* par plusieurs facteurs. Nous considérons ici la *résolution* et la *stabilité statique*. Alors que pour la latence, nous avons pu identifier une valeur universelle de seuil, les requis pour les facteurs caractérisant une erreur de mesure dépendent du contexte d'utilisation du dispositif.

Dans cette section, nous définissons les caractéristiques de résolution et de stabilité des données rendues par le dispositif d'entrée. Nous analysons les requis de ces caractéristiques au regard des applications étudiées au premier chapitre.

Résolution

De manière générale, *la limite de résolution* ou plus simplement, *la résolution*, est la plus petite variation perceptible de la grandeur à mesurer dans des conditions de mesure données. Dans le cas de l'interaction fortement couplée, la grandeur considérée est le déplacement de l'entité suivie par un dispositif d'entrée. La résolution du dispositif de suivi est

1. "Lag must be taken seriously and recognized as a major bottleneck for usability. The current attitude of acknowledging lag but "learning to live with it" will be increasingly unacceptable as VR systems shift from research curiosities to application tools. (...) A major component of (garnering user acceptance) is the delivery of near-immediate response of graphic output to input stimuli"

donc le plus petit mouvement qui peut être détecté par ce dispositif dans des conditions précises de capture.

Le mouvement peut représenter une translation. Dans ce cas, la résolution est exprimée par une distance. Il peut également représenter une rotation. La résolution est alors exprimée par un angle. Par exemple, la souris standard des ordinateurs Apple Macinstosh a une résolution de 0.005 pouces ([Apple 93a]), soit 0,127 mm. Autrement dit, les mouvements de l'utilisateur inférieurs à 0,127 mm. ne sont pas détectés par la souris. En pratique, cette résolution est suffisante pour les tâches de désignation, même précises, sur une interface graphique. De plus, 0,127 mm. est de l'ordre du plus petit mouvement que peut contrôler une personne dans l'espace.

La résolution est liée au mécanisme d'*échantillonnage* des dispositifs d'entrée. L'échantillonnage est l'étape incontournable précédant le traitement numérique d'un phénomène continu. Le monde physique est continu. Par exemple, le déplacement d'une main est un phénomène continu dans l'espace : pour aller d'un point A à un point B, la main passe par une infinité de points entre A et B. Malheureusement, la continuité (de même que l'infini) ne peut pas être représentée dans le monde numérique. Ainsi, lorsque la main déplace une souris d'un point A à un point B, un nombre fini de positions de souris, les *échantillons*, est envoyé au système. De même, lorsqu'une caméra capture une image contenant les deux points A et B, l'espace entre A et B est échantillonné en un nombre fini de points (les *pixels*). Il est important de faire la distinction entre les concepts de *résolution* et de *définition*. La définition désigne le nombre d'échantillons différents que peut mesurer un dispositif. Par exemple, une caméra vidéo au format PAL a une définition de 768 x 576 pixels. La définition est un paramètre intrinsèque du dispositif d'entrée, alors que la résolution est liée aux conditions de mesure : pour une même caméra, c'est à dire à définition constante, la résolution de l'image d'un objet est d'autant plus fine que l'objet est proche de la caméra.

Ayant défini le concept de résolution, nous étudions maintenant l'adéquation d'un dispositif à la tâche qu'il supporte, en fonction de sa résolution.

VideoPlace et ALIVE. (voir page 12) L'objectif de ces deux applications est l'immersion des participants dans un monde virtuel. L'application "Critter" par exemple, illustrée sur la figure 5 page 13, cherche à donner l'illusion au participant qu'un agent électronique se promène sur lui. Le participant aura naturellement tendance à faire des mouvements amples pour faire réagir l'agent. Il en est de même pour **ALIVE**. Ces deux applications, qui n'exigent pas de gestes fins, n'imposent pas de contraintes fortes sur la résolution. Une résolution de l'ordre d'un échantillon par centimètre semble raisonnable.

Parallaxe par mouvement. (voir page 18) Dans la mise en œuvre de la parallaxe par mouvement, on cherche à effectuer un couplage entre la position d'observation et la scène présentée. Il est difficile d'estimer a priori la résolution requise. Les prototypes existants mettent en œuvre des dispositifs magnétiques dont la résolution est assez élevée (un dispositif de type **Flock of Birds** a une résolution de l'ordre de 0,5 mm.) et ne rapporte pas de problème lié à la résolution. Il est possible qu'il ne soit pas nécessaire d'obtenir une telle résolution pour que l'interaction "fonctionne". Une étude empirique reste à faire pour cerner plus précisément ce requis.

Interfaces saisissables et Bureau Digital. (voir page 26 et page 33) La résolution requise pour ces deux types de système est similaire à la résolution requise pour les interfaces graphiques classiques : ces systèmes mettent tous en œuvre des manipulations précises sur des représentations graphiques. Que la main contrôle la souris, des objets saisissables, ou directement une représentation graphique, dans tous les cas, sa précision est du même ordre. Il est raisonnable de s'inspirer de la résolution de la souris et d'estimer que l'ordre de grandeur de la résolution requise est de 0,1 mm.

Stabilité statique

Un dispositif de suivi est dit stable si la mesure de position ne varie pas lorsque l'entité suivie est immobile. Si la souris est un dispositif stable, il n'en va pas de même pour les dispositifs permettant une interaction à distance : capteurs magnétiques et systèmes de vision par ordinateur. Ces dispositifs sont en permanence soumis à des perturbations qui varient au cours du temps et qui ont pour conséquence de faire osciller les mesures de position d'un objet immobile : perturbation du champ magnétique, ou du capteur visuel des caméras.

La stabilité statique peut être mesurée en calculant l'écart type des données fournies par le dispositif pendant un court moment. L'écart type augmente rapidement au cours du temps puis se stabilise. La stabilité statique est la valeur de l'écart type lorsqu'elle se stabilise. Les spécifications techniques du dispositif **Isotrack** ([Polhemus 99]) annoncent un écart type de 2,54 mm. en position et 0,75 degrés en orientation. Pour le **Flock of Birds** ([Ascension 99], l'écart type est de 1,77 mm. en position et 0,5 degrés en orientation.

Aux paragraphes suivants, nous tentons de déterminer les requis de stabilité au regard des applications présentées au chapitre I : réalité virtuelle, tâche de désignation, interfaces saisissables.

Réalité virtuelle avec casque. La conséquence de l'instabilité des mesures est perceptible dans l'utilisation des systèmes de réalité virtuelle avec casque de visualisation. La position du casque est en règle générale suivie par un dispositif magnétique afin de générer les images corres-

pondant à la position d'observation (voir page 17). L'information du capteur magnétique étant instable, le participant perçoit un monde virtuel qui oscille en permanence, même lorsque la tête est parfaitement fixe.

Il est difficile d'établir a priori la stabilité requise pour les dispositifs destinés à capturer le point d'observation de l'utilisateur. Jusqu'à présent, l'approche a été de constater le problème et "d'apprendre à vivre avec". Une étude empirique permettrait de cerner le requis avec précision.

Tâches de désignation. Concernant la désignation, le modèle de l'interaction fortement couplée nous permet d'estimer la stabilité requise selon le raisonnement suivant : d'après le modèle, l'interaction se termine lorsque l'utilisateur perçoit un état du système correspondant à l'état cible. Le problème de l'instabilité se manifeste à cet instant précis : l'utilisateur est immobile puisqu'il ne cherche pas à effectuer de mouvement supplémentaire (il est satisfait de l'état perçu). Par exemple, la désignation d'un bouton se termine lorsque l'utilisateur constate que le curseur de la souris est situé "à l'intérieur" du bouton. L'instabilité des mesures a pour conséquence de simuler un mouvement utilisateur même lorsque celui-ci est immobile. Il convient alors de considérer deux cas selon l'amplitude perçue de l'oscillation du retour d'information :

1. le changement d'état du système induit par l'instabilité est d'une amplitude suffisante pour que l'état courant ne soit plus considéré comme satisfaisant. Du point de vue de l'utilisateur, l'état du système oscille entre "satisfaisant" et "non satisfaisant". Dans le cas du bouton, le curseur oscille entre des positions à l'intérieur et à l'extérieur du bouton. Il est évident que cette situation est frustrante pour l'utilisateur, à l'origine d'une perte de temps (l'utilisateur doit attendre le bon moment pour valider sa désignation) et source d'erreur (le taux d'erreur sera d'autant plus élevé que la fréquence de changement d'état est élevée)[1].
2. le changement d'état du système induit par l'instabilité est d'une amplitude trop faible pour changer le caractère "satisfaisant" de l'état courant. Dans le cas d'une désignation, cela signifie que l'oscillation de position maintient le curseur à l'intérieur du bouton. L'utilisateur peut ainsi valider sa désignation (par exemple par un clic souris). L'instabilité n'a pas de conséquence sur la réalisation de la tâche, si ce n'est la gêne visuelle due à l'oscillation.

Ce raisonnement permet d'établir le requis de la stabilité statique pour des tâches de désignation : pour un système interactif donné, l'écart type caractérisant l'instabilité doit être inférieur à la taille de la plus petite cible que l'on peut désigner dans ce système. Ce requis est nécessaire mais pas

1. Notons que si l'utilisateur cherche à corriger l'oscillation, il peut créer une situation d'instabilité dynamique.

suffisant : il peut y avoir une gêne visuelle due à l'oscillation du retour d'information.

Interfaces saisissables. Dans le cas des interfaces saisissables (page 26) et du **Bureau Digital** (page 33), le retour d'information est similaire à celui d'une interface graphique classique. Les tâches de désignation les plus précises sont de l'ordre de la taille d'un pixel, c'est-à-dire de l'ordre de 0,1 mm., soit encore de l'ordre de la résolution requise. Requérir une stabilité de l'ordre de la résolution est similaire à requérir une stabilité parfaite : la donnée oscille dans un intervalle inférieur à la résolution du dispositif, l'oscillation est alors imperceptible. En conclusion, il est nécessaire, pour les systèmes de type interface saisissable et **Bureau Digital**, de réduire l'instabilité du dispositif d'entrée à un niveau non perceptible.

3. Résumé du chapitre

L'interaction fortement couplée nécessite la mise en œuvre de dispositifs capables d'assurer un service de suivi qui remplisse des requis de fonctionnement précis. Le suivi peut, selon le nombre et la nature des entités suivies, s'appuyer sur des fonctions de détection et d'identification. Détection, identification et suivi, fonctionnent dans un monde planaire et / ou tridimensionnel, en réponse aux besoins des applications visées.

Parmi les facteurs caractérisant le fonctionnement d'un système de suivi, nous avons retenu la latence, la résolution et la stabilité statique. Ces trois propriétés sont transparentes dans les interfaces graphiques usuelles mais déterminantes pour la conception ou l'utilisation de nouveaux dispositifs de suivi. Nous avons donc défini des métriques pour chacun des trois facteurs en étayant notre raisonnement sur les expériences empiriques rapportées dans la littérature et sur le modèle retenu de l'interaction fortement couplée : une boucle fermée de deux systèmes de type stimulus / réponse.

Pour la latence, nous aboutissons à une métrique indépendante de l'application. Nous recommandons un seuil de l'ordre de 50 ms. au-delà duquel les performances de l'utilisateur sont compromises. Cette valeur tient avec l'hypothèse faite implicitement par Card, Moran et Newell que les processeurs cognitif, moteur et sensoriel humains sont agencés en système synchrone.

Pour la résolution et la stabilité statique, les seuils critiques sont déductibles au cas par cas. Nous donnons des exemples d'estimation des

valeurs requises dans les cas étudiés au chapitre I. En particulier, l'écart type de la stabilité doit être inférieur à la taille de la plus petite cible que l'on peut désigner dans le système.

Ce chapitre clôture la partie "identification des besoins" de la thèse. Au chapitre suivant, nous considérons la vision par ordinateur en tant que candidat à la satisfaction de ces besoins. Nous étudions différentes approches de résolution des problèmes de vision par ordinateur afin d'identifier la plus adaptée.

Chapitre III *Vision par ordinateur : problématique et approches*

Les exemples de systèmes et les paradigmes d'interaction analysés au chapitre I, démontrent la potentialité de la vision par ordinateur comme technique de mise en œuvre de dispositifs d'interaction fortement couplée. Le chapitre II précise les requis de l'interaction fortement couplée, sans toutefois viser de dispositif particulier. Dans la suite de ce mémoire, nous envisageons la conception et la mise en œuvre de dispositifs fondés sur la vision par ordinateur pour des situations d'interaction fortement couplée.

Ce chapitre introduit la problématique de la vision par ordinateur : ses objectifs, les difficultés intrinsèques du domaine et les deux grandes familles d'approches : la vision "orientée modèle" et la vision "par apparence". Nous justifions notre choix de l'approche par apparence que nous affinons pour satisfaire les requis de l'interaction fortement couplée.

1. Problématique de la vision par ordinateur

1.1. LE DOMAINE

La vision par ordinateur désigne la partie de la *perception artificielle* concernée par le canal visuel. La perception artificielle a pour objectif de fournir à un système des informations sur le monde qui l'entoure par l'intermédiaire de *capteurs*. Un capteur, à la frontière du système et de son environnement, détecte les changements d'état du monde et en informe le système ([Russel 95], "Perception"). Ce peut être une simple bascule binaire détectant l'état d'ouverture ou de fermeture d'une porte,

ou un dispositif aussi complexe que la rétine humaine qui comprend des centaines de millions d'éléments photosensibles.

Un système de vision par ordinateur est chargé de renseigner le système sur le monde qui l'entoure avec, comme capteurs, des caméras vidéo. Une caméra, réplique artificielle de la rétine humaine, convertit le flux lumineux acquis par son objectif optique en une suite d'images ordonnées dans le temps : le *flux vidéo*. Chaque image est modélisée sous forme d'une matrice à deux dimensions de pixels. Un système de vision par ordinateur traite les informations de bas niveau d'abstraction du flux vidéo pour en extraire des informations de "plus haut niveau d'abstraction".

La discussion présentée au paragraphe "Espace des services" du chapitre II donne une idée générale de l'espace des informations "de plus haut niveau d'abstraction" : présence d'entités, classe d'entités (document, chaise, mur, homme, femme), identité (notamment, le numéro de référence d'un document, l'identité d'une personne), position, ou déplacement au cours du temps d'une entité. Dans le cas particulier d'une personne, on peut aussi chercher à connaître la direction du regard [Collet 99], l'expression du visage ([Essa 95], [Black 97]), ou la nature de son activité (en réunion, au téléphone, disponible, occupée, etc.) [Chomat 99].

Russel et Norvig ([Russel 95]) utilisent un formalisme mathématique pour schématiser le problème de la vision par ordinateur. Soient :

- W, le monde,
- f, la fonction qui décrit la façon dont un stimulus visuel est produit par un monde donné,
- S, le stimulus visuel.

Alors :

$$S = f(W) \qquad (1)$$

La fonction f représente la formation des images correspondant à un monde donné. L'étude de cette fonction est l'objet du domaine de recherche en synthèse d'image. D'un point de vue géométrique, la transformation que représente f s'appelle *projection perspective* ([Foley 82]). Les images de plus en plus réalistes que les infographistes sont capables de synthétiser témoignent d'une connaissance approfondie de f.

Le problème de la vision par ordinateur se pose de manière inverse à celui de la synthèse d'image : "étant donné la fonction f et un stimulus S, quel est le monde W qui a produit S ?". L'approche directe consiste à tenter d'inverser la fonction f :

$$W = f^{-1}(S) \qquad (2)$$

Ainsi présenté, le problème de la vision par ordinateur est assimilable au problème inverse de la synthèse d'image. Cependant, la fonction f, qui projette le monde 3D (W), en un monde 2D (l'image S), entraîne nécessairement une perte d'information. Autrement dit, la fonction f n'est pas réversible. Retrouver W à partir de S constitue un problème difficile. Nous analysons plus avant la nature de ces difficultés.

1.2. DIFFICULTÉS

Le flux vidéo, on l'a vu, constitue la "matière première" des traitements de vision par ordinateur. Ce flux présente des défauts intrinsèques : instabilité statique, ambiguïté, et grand débit imposant de facto des traitements efficaces alors qu'il s'agit de résoudre des problèmes complexes. Nous reprenons successivement ces trois points et identifions leurs conséquences sur l'approche à adopter pour la conception et la mise en œuvre de dispositifs d'interaction fondés sur la vision par ordinateur.

Instabilité statique (bruit)

Le pixel, élément constitutif d'une image, est, en vision par ordinateur, l'information élémentaire du niveau d'abstraction le plus bas. Malheureusement, cette information est statiquement instable : pour une caméra fixe, des conditions d'éclairage constantes et une scène statique (c'est-à-dire dans laquelle aucun objet du champ de la caméra ne bouge), la valeur des pixels varie de manière aléatoire. On dit que le flux vidéo est *bruité* et que l'oscillation de la valeur des pixels est due au *bruit de caméra*.

Lorsque le bruit n'est pas pris en compte et traité explicitement, les données extraites du flux vidéo sont elles-mêmes instables. Nous en verrons l'illustration au chapitre suivant ("Suivi par différence d'images" page 83 et "Suivi par modèle de couleur" page 88).

L'instabilité des valeurs des pixels est source d'ambiguïté au niveau d'abstraction le plus bas, phénomène aggravant pour les niveaux d'abstraction supérieurs qui, en raison de la nature de f, sont également ambigus.

Ambiguïté de l'information

La non réversibilité de f implique qu'à un stimulus visuel peuvent correspondre plusieurs mondes susceptibles d'en avoir été la source. Prenons, pour nous en convaincre, l'exemple suivant : soit un monde W constitué d'un gros cube et d'un petit cube, le petit cube étant masqué à la caméra par le gros. Le stimulus S calculé par l'équation 1 ne contient alors qu'une image du gros cube et aucune représentation du petit cube. À l'évidence, il est impossible, à partir de S uniquement, de calculer un monde correspondant à W. Il convient donc de faire appel à des informations qui complètent S. Comme le notent Russel et Norvig ([Russel 95], "Perception"), la difficulté est d'identifier l'information appropriée :

> *"Un point clé de l'étude de la perception est de comprendre quelle information additionnelle il convient de considérer pour lever l'ambiguïté."*[1]

Il n'existe pas de réponse générale à cette question, si ce n'est que les informations complémentaires nécessaires à la résolution des ambiguïtés relèvent de *connaissances* sur le monde. Ces connaissances couvrent différents niveaux d'abstraction dont la pertinence dépend du problème à résoudre. Les deux exemples qui suivent illustrent notre propos.

Considérons le suivi d'objet dans un flux vidéo. Le flux est ici le film d'un jongleur qui manipule des balles d'apparence identique. L'objectif est de connaître en permanence la position de l'une des balles. Supposons que dans l'une des images, la balle suivie *croise* la trajectoire d'une autre balle donnant l'impression que les deux balles ont *fusionné* dans l'image. Dans les images suivantes, les deux balles se séparent à nouveau. Laquelle des deux balles, issue de la fusion, correspond à celle que nous suivons ? L'information contenue dans le flux vidéo ne permet pas, à elle seule, de lever l'ambiguïté en raison de la similitude des apparences des balles. La réponse à notre question est cependant immédiate si l'on considère que les balles effectuent des trajectoires paraboliques continues dans l'espace : la balle suivie est celle qui poursuit son chemin sur la parabole amorcée avant la fusion. La connaissance à laquelle il est fait appel concerne ici le comportement de l'objet d'intérêt. Dans l'exemple qui suit, la connaissance nécessaire a trait au contexte spatial de l'objet, aux relations que les objets entretiennent dans cet espace, etc.

Nous invitons le lecteur à identifier l'objet de la figure 1 avant de poursuivre plus avant la lecture. Cette image a été présentée à une quinzaine de personnes. Une seule d'entre elles a pu reconnaître l'objet. Confrontées à la même image replacée dans son contexte (se référer à la figure 4 page 74), toutes les personnes, sauf une, ont réussi à identifier l'objet. Dans cet exemple, l'ambiguïté n'est pas seulement levée par la présentation d'une image de plus grande taille. La reconnaissance de

Figure 1
Image hors-contexte d'un objet
La même image est représentée à différentes résolutions (de gauche à droite : 150, 75, 30. et 15 points par pouce). Il est très difficile de reconnaître cet objet lorsqu'il est placé hors de son contexte.
La même image de cet objet replacée dans son contexte est représentée sur la figure 4 page 74.

1. "A key aspect of the study of perception is to understand what additional information can be brought to bear to resolve ambiguity."

l'objet fait aussi appel à un ensemble de connaissances générales sur le monde : relations entre objets qui permettent de reconnaître une scène de bureau, connaissances qui permettent d'estimer la position de l'objet à mi-hauteur de la porte alors que la porte n'est que partiellement visible, localisation qui permet d'inférer qu'il s'agit d'un interrupteur, assertion confirmée par l'apparence de l'objet.

La représentation des connaissances est un problème complexe qui fait l'objet de recherches actives. A l'heure actuelle, on ne sait pas représenter l'ensemble des connaissances acquises par un individu. Cette lacune explique l'incapacité des systèmes de vision par ordinateur à égaler les performances de perception visuelle de l'Homme. De fait, aucun système de vision n'est en mesure de reconnaître les objets d'une scène quelconque. Le fonctionnement d'un système de vision suppose toujours connu le type de scène traitée.

Le flux vidéo est non seulement statiquement instable et ambigu. Il représente aussi un très grand débit d'information.

Grand débit d'information

Un flux vidéo au format standard PAL, se traduit, une fois numérisé, par un débit de l'ordre de 40 Mo/s. Un tel débit d'information a pour conséquence de réduire l'ensemble des traitements qu'on peut lui appliquer en temps réel. Seuls les traitements ayant un coût limité en temps de calcul sont applicables, ce qui a pour effet d'aggraver les difficultés précédentes : la résolution des ambiguïtés implique en général des algorithmes coûteux en temps de calcul, de même que l'extraction d'indices stables à partir de données par nature instables.

Les difficultés que soulève l'interprétation du flux vidéo expliquent en partie notre constat sur le développement quelque peu restreint des systèmes interactifs fondés sur la vision par ordinateur.

1.3. CONSTAT

En introduction de ce mémoire, nous énoncions les potentialités de la vision par ordinateur pour l'interaction homme-machine : extension des capacités visuelles de l'Homme, interaction non intrusive (sans fil à la patte), dispositifs d'interaction fortement couplée, etc. Si l'apport de la vision semble compris, il convient de noter qu'à l'heure actuelle, cette technologie appliquée à l'interaction homme-machine, n'apparaît que dans des démonstrateurs de laboratoire. Si ceux-ci commencent à être nombreux ([Azarbayejani 93], [Wellner 93b], [Gaver 95], [Ullmer 97], [Kang 98], [Toyama 98], [Yang 98a]), ils ne satisfont pas pour autant les requis de l'interaction homme-machine. Peu ou pas utilisés, leur apport n'a pu être confirmé par l'usage.

Quelques exceptions méritent cependant d'être soulignées :

- **VideoPlace** et **ALIVE**, on l'a vu au chapitre I, ont été confrontés au public avec succès. Mais ces systèmes, conçus pour des applications

ludiques, ne démontrent pas l'apport de la vision par ordinateur au cas des tâches productives.

- Les systèmes **BrightBoard** ([Stafford-Fraser 96a]) et **ZombieBoard** ([Saund 96]) sont exploités quotidiennement dans un laboratoire de recherche par des utilisateurs qui n'en sont pas les concepteurs. Ce trait est important : les utilisateurs sont motivés par l'apport intrinsèque du système. Toutefois, **BrightBoard** et **ZombieBoard** ne sont pas fondés sur une interaction fortement couplée : la latence de leur cycle stimulus / réponse est de l'ordre de la minute. Par conséquent, ces systèmes ne démontrent pas la faisabilité, ni l'apport, d'une interaction fortement couplée fondée sur la vision par ordinateur.
- **Build-It**, que nous avons présenté au chapitre I est, quant à lui, en voie de commercialisation. Ce système, qui adopte le paradigme des interfaces saisissables, nécessite *de facto* une interaction fortement couplée. Cependant les performances de ce système, observées par nous-même, sur une vidéo ([Rauterberg 98]), montrent une latence de l'ordre de la seconde, soit 20 fois supérieure au requis estimé au chapitre II page 53. **Built-It** ne peut donc être considéré comme une preuve de la faisabilité et de l'apport d'une interaction fortement couplée fondée sur la vision par ordinateur.

Au vu des difficultés que nous venons d'énoncer, il est légitime de s'interroger sur l'opportunité d'introduire la vision par ordinateur comme support de l'interaction fortement couplée. En effet, l'interaction fortement couplée requiert précisément ce qui constitue des difficultés fondamentales en vision par ordinateur : stabilité des données et "temps réel" pour une latence conforme à celle du système humain. Nous pensons qu'il est possible d'utiliser la vision par ordinateur en interaction fortement couplée sous réserve d'adopter une approche résolument centrée sur la tâche. Nous présentons dans la section qui suit les approches possibles et précisons notre point de vue.

2. Approches en vision par ordinateur

La recherche en vision par ordinateur se répartit en deux courants de pensée qui, avec le temps, tendent à se rapprocher et à se compléter. On distingue :

- L'approche classique dite *vision orientée modèle* qui consiste à effectuer la reconstruction 3D d'entités observées, puis à comparer la reconstruction à un modèle 3D pré-acquis de ces entités en utilisant des techniques de rétro-projection et de mise en correspondance ;

- L'approche récente de la *vision par apparence* qui, comme son nom l'indique, s'appuie sur toutes les manifestations visuelles possibles des entités observées, c'est-à-dire, en théorie, sur toutes les images de ces entités prises de tous les points de vue et éclairages possibles.

Dans la première approche, les modèles se font dans le repère de la scène (ce sont donc des modèles 3D). En vision par apparence, les modèles sont construits dans le repère image (ce sont donc des modèles 2D).

2.1. VISION ORIENTÉE MODÈLE

Si l'on reprend la formulation de l'équation 2 page 62, l'approche orientée modèle vise à construire un modèle géométrique en trois dimensions d'un monde W correspondant à un stimulus S donné. Selon le cas, S est une image ou un flux vidéo. Puisque la fonction f^{-1} n'existe pas, l'approche orientée modèle consiste à construire un modèle du monde W, à calculer sa projection avec la fonction f puis à faire correspondre le stimulus calculé avec le stimulus mesuré par la caméra. Si les deux stimulus ne correspondent pas, les différences sont corrigées en modifiant le monde W et les paramètres de la caméra.

Les paramètres de la caméra définissent le point de vue de la caméra sur le monde. Ils comprennent la distance focale de la caméra et sa position selon les six degrés de liberté (trois coordonnées dans l'espace cartésien et trois angles définissant l'orientation). La modification du monde W correspond à un changement d'état dans un espace de très grandes dimensions. Parmi ces dimensions, on trouve la nature des entités présentes dans le monde, leur position, leur orientation, leur forme (qui peut être variable dans le cas d'entités déformables), leur texture, leur cardinalité, la nature et la position des sources lumineuses, etc.

Une recherche exhaustive des paramètres qui permettraient de reconstruire une image proche du stimulus source est vouée à l'échec en raison

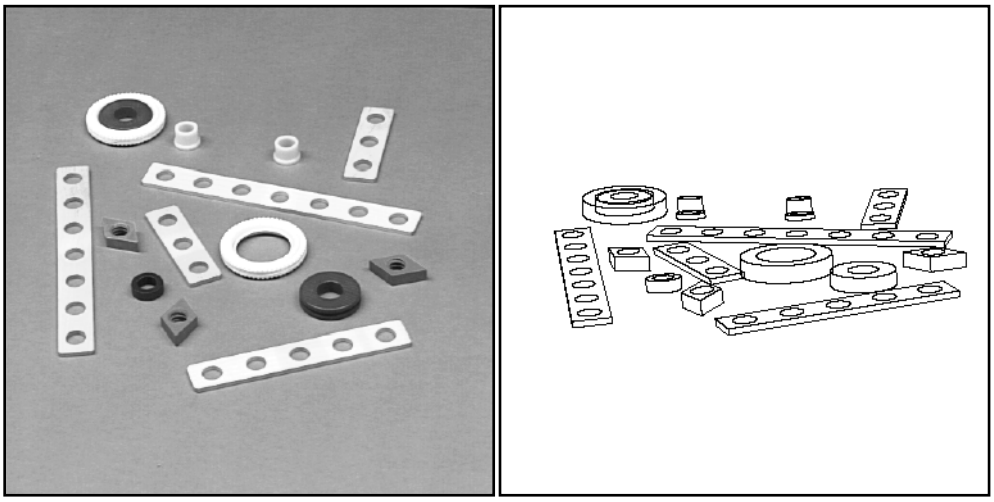


Figure 2
Reconstruction 3D (extraite de [Socher 95])
Scène photographiée (gauche) et modèle 3D reconstruit vu sous un angle différent (droite).

des coûts de calcul induits par le grand nombre de paramètres possibles. L'espace de recherche peut être réduit si l'on restreint l'analyse à certains indices *invariants* de l'image. Par exemple, l'effet des paramètres de luminosité (nombre, nature, position des sources de lumière) sur l'image source est fortement amoindri si l'on considère, non pas l'intensité lumineuse, mais les variations de l'intensité lumineuse. Le calcul des zones de fortes variations lumineuses est appelé *détection de contour* : c'est en général sur le contour des objets que se situent les plus grandes variations d'intensité lumineuse. D'autres indices invariants de nature géométrique sont également considérés. Ce sont par exemple les segments de droite car la linéarité est invariante par projection perspective. La figure 2, extraite de [Socher 95], illustre le résultat d'une extraction de modèles 3D d'objets rigides à partir d'images de ces objets.

L'approche orientée modèle présente plusieurs problèmes :

- La résolution du système nécessite, en règle générale, un processus itératif coûteux en temps de calcul. Toutefois, dès qu'un monde correspondant à l'image initiale a été modélisé, la mise à jour du monde reflétant l'évolution des images dans le flux vidéo peut être calculée de façon efficace. Harris [Harris 92] propose un algorithme qui permet de suivre un objet rigide dans le flux vidéo par mise à jour de son modèle 3D en fonction du flux. La mise en œuvre de cet algorithme fonctionne à 50 Hz sur une station de travail Sparc 2. Cependant, l'absence de solution au problème de l'état initial du monde réduit l'intérêt d'une telle approche pour le suivi.
- L'extraction des indices invariants est un processus complexe, coûteux en temps de calcul. L'extraction de contours est difficile dans le cas de scènes contenant des objets texturés ou présentant des combinaisons complexes d'ombres et de lumière. Ce traitement est de plus sensible au bruit de caméra (voir page 63). L'oscillation des valeurs de pixels entraîne celle des indices extraits de l'image.
- Malgré la réduction de l'espace de recherche par prise en compte d'invariants, le nombre d'inconnues du problème reste trop grand pour permettre sa résolution. Il est nécessaire de fixer un grand nombre d'inconnues, en introduisant des connaissances a priori sur la scène analysée. Par exemple, on fixe le nombre et la nature des objets présents dans la scène. Par une phase de calibrage initial, on détermine le point de vue de la caméra sur la scène et l'on fixe les paramètres de caméra en s'assurant que la caméra reste fixe. La nécessité d'introduire ces connaissances limite le champ d'application de cette approche à des scènes "contrôlées".

En synthèse, la vision orientée modèle tente de résoudre un problème trop difficile pour être traité dans toute sa généralité. Si l'objectif est valide sur le plan scientifique, un modèle 3D du monde peut s'avérer inutile pour certaines applications. La vision par apparence constitue l'alternative.

2.2. VISION PAR APPARENCE

> "*L'apparence est ce qu'on voit d'une personne ou d'une chose, la manière dont elle se présente à nos yeux*." ([Robert 67])

Par analogie, la vision par apparence se fonde sur les manifestations visuelles d'entités d'intérêt captées par une caméra. *L'espace des apparences* d'une entité est l'ensemble des manifestations visuelles de cette entité prises selon tous les points de vue et conditions d'éclairage possibles. L'espace d'apparence est un concept abstrait et idéal. Il est impossible de décrire parfaitement toutes les apparences d'une entité. Mais dans de nombreux cas, il est possible de se limiter aux manifestations d'apparences susceptibles d'être observées. En pratique, les modèles sont construits par l'exemple, c'est-à-dire par l'analyse d'un ensemble d'images représentatives de l'entité ou du phénomène d'intérêt.

Le principe de la perception par apparence est de projeter l'espace d'apparences dans un espace de dimensionnalité réduite choisi de façon à faciliter les tâches à accomplir (indexation, reconnaissance, estimation de position, etc.). Turk et Pentland sont parmi les premiers à avoir appliqué cette approche pour l'identification des visages ([Turk 91a], [Turk 91b]). Dans ce cas :

- L'espace d'apparences est un ensemble d'images de visage appelé "base de visages". La figure 3a montre un espace d'apparences composé de trois images. Chaque image de l'espace est un point dans un espace de grande dimension où chaque pixel représente une dimension de cet espace. L'intensité lumineuse d'un pixel, pour une image donnée, représente la coordonnée de cette image dans la dimension correspondant au pixel. Ainsi, pour des images de définition 256 x 256 pixels par exemple, chaque image est un point dans un espace à 65536 dimensions.
- Une analyse en composantes principales (ACP) est effectuée sur le nuage de points que constitue la base de visages. Le résultat de cette ACP est un sous-ensemble de points qui représente au mieux le nuage de points initial. Ce sous-ensemble, appelé "espace des visages" ou "eigenface", est centré sur la moyenne des images de la base. Ses axes sont définis par les "eigenfaces" résultats du calcul de l'ACP. Moyenne et eigenfaces sont illustrés sur la figure 3. L'intérêt de l'espace des visages est double : 1) sa dimension est considérablement inférieure à celle de l'espace des apparences initial, et 2) il constitue un modèle d'images de visage.
- L'identification d'une personne est illustrée sur la figure 3d. Le processus se résume à la projection de l'image du visage de la personne dans l'espace des visages (par un produit scalaire) et au calcul de deux distances euclidiennes. La projection de l'image est peu coûteuse en temps de calcul car l'espace des visages est de faible dimension. Le temps de calcul des distances euclidiennes est

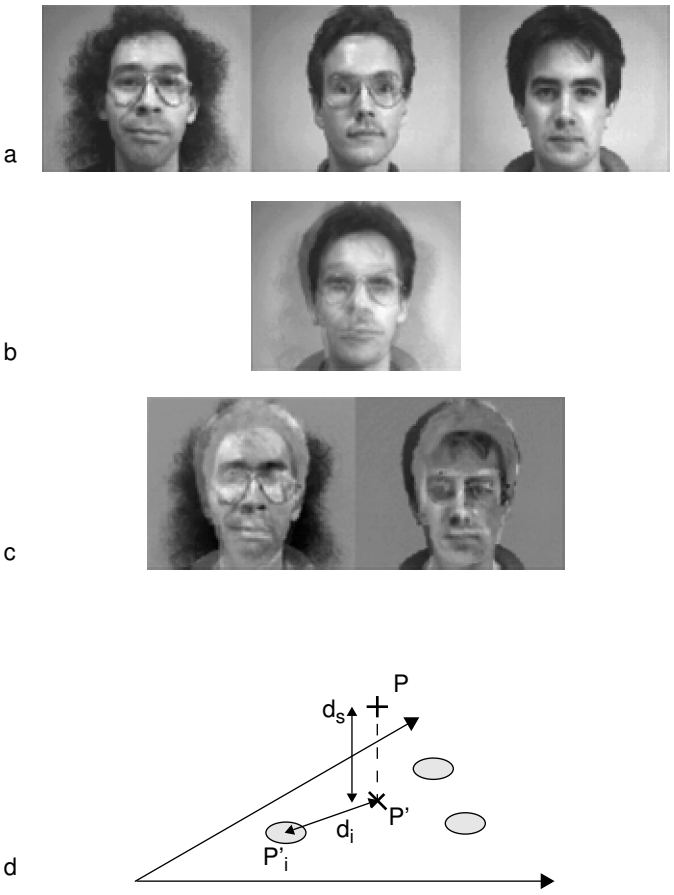


Figure 3
Identification des visages par la techniques des eigenfaces
La moyenne (b) et les eigenfaces (c) sont calculés à partir de la base des visages (a).
Le processus (d) consiste à projeter l'image à identifier P dans l'espace des visages dont l'origine est la moyenne et les axes sont les eigenfaces.
Si la distance d_s entre l'image P et sa projection P' est inférieure à un seuil donné, P est une image de visage.
Si la distance d_i entre la projection P' et la projection du visage P'$_i$ est inférieure à un seuil donné, P est l'image du visage de la personne i.

négligeable. La distance entre l'image du visage et sa projection dans l'espace des visages est considérée en premier lieu : si cette distance est supérieure à un seuil donné, on estime que l'image ne représente pas un visage. Dans le cas inverse, on calcule la distance minimale entre la projection de l'image dans l'espace des visages, et chacune des projections des visages de la base dans cet espace. Si cette distance minimale est inférieure à un seuil donné, on estime que l'image correspond au visage de la base le plus proche. Sinon, on estime que c'est l'image d'un visage inconnu.

La vision par apparence est parfois confondue avec une projection en espace de composantes principales. L'analyse en composantes principales n'est qu'un des outils possibles pour représenter un espace d'apparence. D'autres méthodes permettent de réduire la taille de cet espace : les vecteurs de dérivées gaussiennes, les ondelettes de Gabor ([Granlund 78]), et le "Local Jet" de Koenderinck ([Koenderink 87]). Ces filtres sont à plusieurs niveaux de résolution en utilisant une pyramide de gaussiennes ([Crowley 81]).

L'approche de Turk et Pentland est néanmoins représentative de la vision par apparence :

- Le modèle mis en œuvre (c'est-à-dire, l'espace des visages) est fondé sur l'apparence des visages (les images), non pas sur des propriétés intrinsèques des visages telles que la présence de deux yeux, leur symétrie, leur forme, etc.
- Le modèle est construit grâce à un apprentissage par l'exemple,
- Le modèle est valide uniquement pour une tâche ciblée : l'identification de visages présentés de face. L'analyse en composantes principales "fonctionne" dans le cas des images de visages présentés de face car ces images sont globalement similaires. L'approche de Turk et Pentland n'est pas applicable, par exemple, à l'identification d'objets disparates.

En synthèse, les modèles de la vision par apparence sont spécifiques aux tâches prévues. L'absence de généralité est toutefois compensée par l'existence d'algorithmes moins coûteux que ceux de la vision orientée modèle. Ce constat justifie notre approche.

2.3. NOTRE APPROCHE : VISION PAR APPARENCE CENTRÉE SUR LA TÂCHE UTILISATEUR

Le temps de latence, nous l'avons vu au chapitre 2, est un facteur déterminant dans la mise en œuvre de dispositifs de suivi pour l'interaction fortement couplée. Par conséquent, notre approche doit utiliser l'efficacité computationnelle comme argument directeur. Nous reprenons à notre compte les principes de la vision par apparence plus efficace, par essence, que la vision orientée modèle. Nous spécialisons les traitements selon deux facettes : l'*extraction d'information minimale* et l'*ajout contrôlé de contraintes*. L'extraction d'information et la pose de contraintes sont l'une et l'autre *centrées sur la tâche de l'utilisateur*. La première vise la simplification des techniques à mettre en œuvre, la seconde vise la simplification du domaine applicatif. Les systèmes **VideoPlace** et **ALIVE**, deux exemples de succès, étayent nos choix.

Extraction de l'information minimale

L'extraction d'information minimale consiste à définir l'information minimale requise pour un objectif donné et à n'extraire de l'image (ou du flux vidéo) que cette information. L'objectif fixé est centré sur un besoin lié à la qualité de l'interaction homme machine. L'extraction d'information minimale, qui réduit la complexité des algorithmes, permet d'améliorer les performances. Elle permet aussi de réduire l'espace des paramètres des algorithmes et par conséquent, de simplifier leur mise au point. Nous en illustrons le principe avec **VideoPlace** et **ALIVE**.

L'objectif interactionnel des "réalités virtuelles" de **VideoPlace** est de créer un sentiment d'immersion (se rapporter au chapitre I pour une description détaillée). La silhouette du participant est ici considérée comme suffisante pour créer l'effet interactionnel souhaité. En conséquence, Krueger se limite à l'extraction de la silhouette, et ne cherche pas à extraire la position des différents membres du participant.

La silhouette est l'information minimale requise pour cet objectif précis. L'extraction de la silhouette uniquement garantit la capacité réactive du système.

La gestion de la profondeur du système **ALIVE** illustre de façon encore plus flagrante le principe de l'extraction minimale. Ici, l'objectif est la gestion des occlusions dans un espace tridimensionnel. Une idée répandue veut que l'accès à la troisième dimension nécessite la multiplication des points de vue (avec deux caméras par exemple, ou avec un point de vue mobile). Par une astuce habile, les concepteurs de **ALIVE** parviennent à extraire l'information qui leur est nécessaire à l'aide d'une seule caméra à point de vue fixe (voir page 16). Cette information n'est pas parfaite, mais elle est suffisante pour gérer le problème d'occlusion pour les situations interactionnelles visées[1].

Dans le point suivant, nous observons que le domaine d'application de ces techniques, non pas les techniques elles-mêmes, peut également être simplifié pour atteindre l'objectif.

Ajout contrôlé de contraintes

L'ajout de contraintes sur la situation interactionnelle est aussi une source de simplification. Il est cependant essentiel de considérer les risques que présente l'introduction de toute nouvelle contrainte du point de vue de l'intérêt général du système. Pour cette raison, nous parlons d'ajout *contrôlé* de contraintes.

Une contrainte ne sera introduite que si elle satisfait les deux conditions suivantes :

1. elle simplifie un problème de vision pour lequel aucune solution n'est a priori disponible, contribuant à rendre possible la réalisation du système,
2. les conséquences de son introduction ne sont pas fatales au système final, autrement dit le système fait encore sens ; il est utilisable.

On retrouve ce principe dans **VideoPlace** et **ALIVE**. L'un et l'autre ont réduit le domaine d'application par ajout de contraintes sur l'environnement observé : **VideoPlace** nécessite un fond lumineux ; **ALIVE** tolère un fond quelconque, à condition qu'il soit statique. Les deux systèmes sont limités à un seul participant.

1. Notons que l'équipe de recherche VISMOD, dans laquelle a été développé **ALIVE**, a par la suite réalisé un système utilisant deux caméras pour extraire la position en trois dimensions des mains et du visage [Azarbayejani 96]. Il n'a cependant pas été jugé utile d'introduire cette technique dans **ALIVE** alors que le système est encore un support de recherche actif. Il semble que la technique initiale d'estimation de la profondeur soit suffisante pour les applications développées.

L'écran lumineux de **VideoPlace** augmente le coût du système et interdit son installation dans une salle commune où la lumière pourrait être gênante pour les autres personnes. La limitation à un seul participant prive **VideoPlace** et **ALIVE** de situations ou plusieurs participants pourraient interagir par l'intermédiaire d'agents ou d'objets virtuels (tel qu'un jeu de football où le ballon, virtuel, ne serait pas soumis à la gravité).

L'écran lumineux et la limitation du nombre de participants satisfont néanmoins nos deux conditions : les traitements fondés sur la vision sont simplifiés et les systèmes sont utilisables (et utilisés). D'autres contraintes auraient pu être introduites, par exemple la limitation de la vitesse de déplacement du participant, ou le port d'une combinaison facilement repérable. Mais ces contraintes auraient porté atteinte à la liberté de mouvement des participants et à l'approche "venez tel quel", motivations essentielles de **VideoPlace** et **ALIVE** au regard des systèmes immersifs utilisant casques et gants.

3. Résumé du chapitre

Le problème de la vision par ordinateur est assimilable au problème inverse de la synthèse d'image. Elle a pour objectif de retrouver le monde qui a produit un stimulus donné, capté par une caméra sous forme d'image ou de flux vidéo. La difficulté du problème (ambiguïté, instabilité et grand flux d'information) explique en partie le faible déploiement de la vision par ordinateur en tant que technique d'interaction. La vision par apparence, dont les modèles sont plus simples en terme de complexité de calcul que ceux de la vision orientée modèle, ouvre des perspectives encourageantes. Nous en exploitons les principes simplificateurs que nous enrichissons avec l'extraction de l'information minimale et l'ajout contrôlé de contraintes :

- Seule l'information minimale permettant d'atteindre le but identifié doit être extraite du flux vidéo. Ce principe a pour objectif d'améliorer les performances du système.
- L'environnement observé peut être contraint sous réserve que le système reste utilisable. Ce principe a pour objectif de réduire la généralité des problèmes de vision à traiter au profit de la faisabilité.

Ayant défini la portée de la vision par ordinateur et précisé notre approche, nous entrons, avec le chapitre suivant, dans le détail des

techniques que nous avons mises en œuvre et intégrées dans des prototypes interactifs fondés sur la vision par ordinateur.

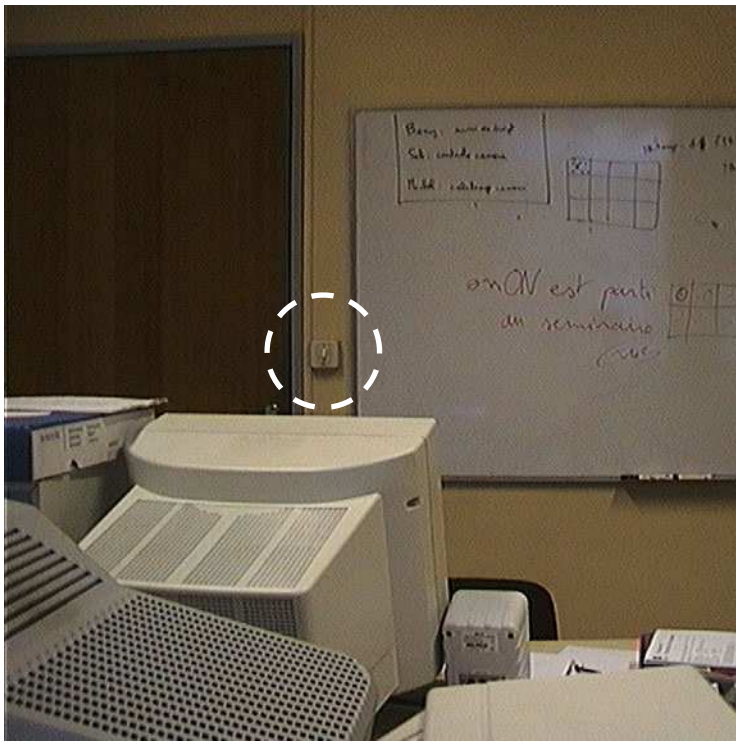


Figure 4
L'objet de la figure 1 page 64 replacé dans son contexte
L'image représente visiblement une scène de bureau. L'objet est situé à proximité de la porte à mi-hauteur, c'est un interrupteur.

Chapitre IV *Techniques de suivi en vision par ordinateur*

Le suivi d'entité, nous l'avons vu, est un service indispensable à l'interaction fortement couplée. Dans ce chapitre, nous nous intéressons à la réalisation de systèmes de suivi en vision par ordinateur. Ces systèmes, on le rappelle, doivent répondre aux requis de fonctionnement justifiés au chapitre II.

Le suivi d'entité fait l'objet d'une présentation générale en première partie de ce chapitre. Les trois sections qui suivent détaillent chacune une approche à sa mise en œuvre. Chaque technique ayant ses forces et ses faiblesses au regard des requis de l'interaction fortement couplée, nous présentons en dernière partie une architecture coopérative qui, par combinaison de techniques complémentaires, vise à assurer un suivi robuste et autonome.

1. Suivi d'entité

Le suivi d'entité par vision par ordinateur a pour objectif de déterminer la position d'une entité donnée dans chaque image du flux vidéo. Nous appelons "cible du suivi", ou plus simplement "cible", l'entité ou la partie d'une entité suivie par le système.

On relève de nombreuses techniques de suivi en vision par ordinateur, chacune adoptant une stratégie différente. Dans cette section, nous présentons les concepts communs à ces techniques et le principe des traitements communément rencontrés dans un suivi.

1.1. PRINCIPE

Le suivi d'entité est un processus cyclique à plusieurs étapes dont la nature et le nombre dépendent des mises en œuvre applicatives. Nous citerons : la mesure, l'observation, la validation, l'ajustement de l'estimation, la prédiction.

1. La *mesure* consiste, comme son nom l'indique, à *mesurer* une propriété donnée de l'image de façon à mettre la cible en évidence. Chaque technique de suivi est fondée sur une propriété particulière de l'image.
2. L'*observation* consiste à faire une hypothèse (appelée "observation") sur la position de la cible à partir de la mesure fournie par l'étape 1.
3. La *validation* détermine la validité de la position estimée à l'étape 2. Elle peut s'appuyer sur des connaissances externes à l'image mais caractéristiques de l'application ou bien sur la valeur attendue de la position à cet instant. Cette valeur de position est le résultat de l'étape de prédiction (étape 5) du cycle précédent.
4. L'*estimation* met à jour l'estimation de la position de la cible maintenue par le processus de suivi. La mise à jour s'effectue par prise en compte de l'observation (étape 2) si celle-ci est validée (étape 3).
5. La *prédiction* calcule la position la plus probable de la cible dans l'image suivante. Cette étape fait appel à des connaissances externes à l'image sur la cible et sur ses déplacements.

Les deux premières étapes, "mesure" et "observation", sont parfois regroupées en une seule étape appelée *observation* ([Crowley 94a]). Dans certains suivis, comme les nôtres, observation et estimation ne sont pas dissociés : l'observation, lorsqu'elle est validée, tient lieu d'estimation de position. Ainsi, dans ce qui suit, nous reprenons les étapes utilisées pour la mise en œuvre de nos techniques de suivi et, par *estimation,* nous entendons l'*étape 2* du cycle ci-dessus.

La suite d'étapes énoncée ci-dessus doit se voir comme un cadre analytique de décomposition des problèmes, non pas comme une solution assise de mise en œuvre systématique d'un suivi.

1.2. MESURE

En vision par ordinateur, la caméra sert de dispositif de mesure. En règle générale, l'information produite par la caméra est de trop bas niveau d'abstraction pour d'emblée localiser la cible. Des traitements d'image sont nécessaires à la mise en évidence de la cible.

Caméra

Une caméra "noir et blanc" fournit une mesure de l'intensité lumineuse reçue par chaque pixel. L'ensemble des pixels constitue la capture de la caméra. Une image représente la mesure, à un instant donné, de l'intensité lumineuse de tous les pixels. Dans le cas des caméras couleur, chaque pixel est composé de trois mesures de l'intensité lumineuse selon trois bandes de fréquence qui correspondent respectivement aux couleurs rouges, vertes, bleues.

Une caméra, comme tout dispositif de mesure d'un phénomène physique, fournit une mesure imparfaite dite bruitée. Ce bruit, responsable de l'oscillation des valeurs mesurées (voir à ce propos le paragraphe "Instabilité statique (bruit)" page 63), nécessite l'application de traitements particuliers.

Traitements d'image

Les techniques de traitement d'image ont pour objectif de mesurer une propriété particulière de l'image. Cette propriété facilite l'extraction de la position de la cible de l'image. Dans les sections qui suivent, nous étudions trois techniques de suivi fondées sur des traitements distincts :

- mesure de la *variation d'intensité lumineuse au cours du temps* (cas du suivi par différence d'images),
- mesure de la *ressemblance à une teinte* (cas du suivi par modèle de couleur),
- mesure de la *ressemblance à un motif* (cas du suivi par corrélation).

D'autres propriétés de l'image telles que la *variation d'intensité lumineuse dans l'espace* ([Lindeberg 96]), ou *le flux de déplacement dans l'image* ([Basu 96]) ont été largement étudiées. Mais la complexité des traitements nécessaires au calcul de ces propriétés ne permet pas d'envisager une exécution en temps réel.

Les traitements de calcul de propriété sont appliqués à la mesure d'intensité lumineuse fournie par la caméra. Cette mesure étant bruitée, la propriété mesurée l'est également. Cependant, on peut atténuer voire s'affranchir du bruit, en appliquant un seuillage. On trouvera en annexe A page 181 la présentation d'une telle technique.

1.3. ESTIMATION DE LA POSITION

L'étape d'estimation consiste à calculer la position de la cible à partir des mesures effectuées à l'étape de mesure. L'expression de la position peut prendre plusieurs formes.

Forme de l'estimation de la position

Une position peut être exprimée dans le repère de la scène ou dans celui des images. On retrouve ici la distinction entre l'approche orientée modèle (voir page 67) et l'approche par apparence (voir page 69) : en règle générale, les techniques de vision orientées modèle expriment la position d'une entité par sa structure géométrique dans le repère scène. Lorsque l'objet n'est pas déformable, sa position est parfaitement déterminée par un sextuplet qui regroupe les six degrés de liberté de la cible dans l'espace : trois coordonnées dans l'espace cartésien de la scène et trois angles définissant l'orientation ([Harris 92]). En vision par apparence, la position de l'entité prend l'une des formes suivantes :

- coordonnées en deux dimensions dans l'image ([Martin 95]),
- rectangle englobant la cible dans l'image ([Gaver 95], [Oliver 97]),

- statistiques du second ordre de la distribution spatiale des pixels représentant la cible ([Oliver 97]),
- six paramètres d'une transformation affine projetant une image de la cible dans l'image de la caméra ([Shi 94], [Black 97]),
- ensemble de points correspondant au contour de la cible ([Kass 87]).

Dans le cadre de notre travail, le choix de la forme de la position est dirigé par les besoins applicatifs. Nos deux cas d'étude, le **tableau magique** (chapitre V) et la **fenêtre perceptuelle** (chapitre VI), nécessitent seulement le couple de coordonnées cartésiennes de la cible dans l'image. Le suivi du visage, étudié à la section "Coopération de techniques" page 103, nécessite le calcul de la boîte englobante. Appliquant l'approche énoncée au paragraphe "Extraction de l'information minimale" page 71, notre étude met l'accent sur l'extraction de ces seules informations.

Dans certains cas, la propriété de l'image calculée à l'étape de mesure, est *discriminante* : les pixels qui satisfont la propriété sont les pixels qui représentent la cible dans l'image. Nous présentons ici deux techniques de calcul de la position fondée sur une image présentant une propriété discriminante. La première a pour but d'extraire les statistiques de la distribution spatiale des pixels qui représentent la cible, la seconde calcule la boîte englobante de ces pixels.

Statistiques de la distribution spatiale. Soit I l'image résultat de l'étape de mesure. Le centre de gravité de la distribution spatiale des pixels définit les coordonnées en deux dimensions de la cible :

$$\hat{x} = \frac{\sum_{\forall x, \forall y} x \cdot I(x, y)}{N} \qquad \hat{y} = \frac{\sum_{\forall x, \forall y} y \cdot I(x, y)}{N} \qquad N = \sum_{\forall x, \forall y} I(x, y) \qquad (1)$$

Notons que les formules ci-dessus sont valides aussi bien pour une image seuillée, dont les pixels sont binaires, que pour une image dont les pixels sont valorisés. Dans ce dernier cas, la position correspond au barycentre des pixels pondéré par leurs valeurs. Cette pondération est justifiée, par exemple, lorsque la valeur d'un pixel correspond à sa probabilité de représenter la cible.

Les besoins applicatifs peuvent justifier le calcul de statistiques d'ordre supérieur. Par exemple, il peut être nécessaire de connaître, en plus de sa position, l'orientation de la cible dans l'image. Dans [Oliver 97], nous calculons la matrice de covariance du nuage de points afin d'en extraire les vecteurs propres et les valeurs propres associées. Les vecteurs propres définissent les axes principaux du nuage de points qui représente la cible. Dans le cas d'un visage, la distribution de points est allongée. Nous pouvons donc nous fonder sur l'orientation des axes principaux pour

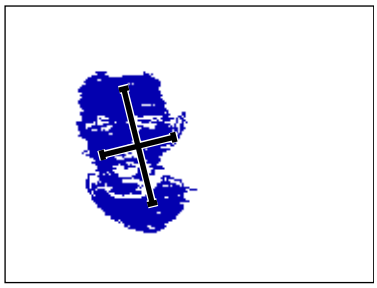


Figure 1
Statistiques du second ordre de la distribution spatiale des pixels
Le centre et les axes principaux du nuage de points (représentés par une croix sur l'image) définissent respectivement la position et l'orientation du visage.

l'estimation de l'orientation du visage dans l'image. La figure 1 illustre le résultat de ce calcul.

La matrice de covariance est calculée comme suit :

$$C = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix} \tag{2}$$

avec

$$\sigma_{xx} = \frac{\sum_{\forall x, \forall y} (x - \hat{x})^2 \cdot I(x, y)}{N} \qquad \sigma_{xy} = \frac{\sum_{\forall x, \forall y} (x - \hat{x}) \cdot (y - \hat{y}) \cdot I(x, y)}{N} \tag{3}$$

où $\hat{x}$, $\hat{y}$, et N sont définis par l'équation 1. La matrice C étant symétrique, nous pouvons la diagonaliser par une analyse en composantes principales :

$$C = \Phi L \Phi^t \tag{4}$$

où Φ est orthogonale et L est diagonale. Les valeurs de la diagonale de L représentent la taille des segments sur la figure 1, alors que Φ représente la matrice de rotation entre le repère de l'image et le repère défini par la croix sur la figure 1.

Boîte englobante. La boîte englobante est une autre forme d'expression de la position de la cible. Elle se définit comme le rectangle dont les côtés correspondent aux abscisses et aux ordonnées minimales et maximales des pixels correspondant à la cible dans l'image. Le calcul des coordonnées maximales et minimales implique que chaque pixel soit étiqueté comme appartenant ou non à la cible. L'image résultat de l'étape de mesure doit donc être seuillée afin de contenir des pixels à valeur binaire.

Le bruit de caméra a un effet extrêmement néfaste sur le calcul de la boîte englobante. L'apparition d'un seul pixel dû au bruit a pour effet d'accroître la surface de la boîte englobante pour contenir ce pixel. Le seuillage sert à réduire le nombre de pixels dus au bruit mais ne les élimine pas tous. L'analyse en composantes connexes, définie en annexe A page 182, permet de ne considérer que les surfaces de l'image

présentant une forte densité de pixels. Les pixels dus au bruit de caméra, dispersés dans l'image, sont exclus du calcul de la boîte englobante. À des fins d'optimisation, l'analyse en composantes connexes et le calcul des coordonnées minimales et maximales sont effectués en un seul parcours de l'image. Le détail est donné en annexe A. La figure 7d page 93 illustre le résultat du calcul, pour le cas d'un visage, de la boîte englobante sur la plus grande classe de pixels connexes.

Résolution spatiale

En vision par apparence, la position de la cible est exprimée dans le repère image. Il paraît donc naturel d'exprimer la résolution d'une technique de suivi dans le repère image. Toutefois, du point de vue de l'utilisateur, la résolution "utile" est la valeur exprimée dans le repère de la scène. La relation entre la résolution dans les repères image et scène dépend de la distance de la cible à la caméra. Par exemple, supposons que la résolution soit de trois pixels dans le repère image, la résolution dans le repère scène est égale à la taille de ce que représentent les trois pixels. Cette taille est d'autant plus petite que l'entité que représentent les trois pixels est proche de la caméra. Dans la suite de ce chapitre, nous exprimerons la résolution des techniques de suivi dans le repère image, mais nous gardons à l'esprit que la résolution "utile", du point de vue de l'utilisateur, est la résolution exprimée dans le repère scène, et que cette résolution doit être estimée en fonction des conditions de mesure. Ces estimations sont documentées dans le cas du **tableau magique** (page 131) et de la **fenêtre perceptuelle** (page 154).

La position de la cible dans l'image une fois estimée, l'étape suivante consiste à valider cette estimation et préparer la recherche de la cible dans la prochaine image.

1.4. VALIDATION

En vision par ordinateur, l'échec est fréquent. Dans le cas du suivi d'entité, ce phénomène doit être détecté de façon à pouvoir le gérer au niveau du client du suivi. Les causes possibles de l'échec d'un suivi sont multiples, parmi lesquelles:

- L'absence de solution : la scène évolue de telle façon que le problème de vision ne peut avoir de solution. Dans le cas du suivi, une cible occultée, c'est-à-dire hors du champ de la caméra, ne peut être localisée.
- Le non-respect de contraintes : le paragraphe "Ajout contrôlé de contraintes" page 72 justifie la mise en place de contraintes sur la scène. Le non-respect de l'une de ces contraintes, même momentané, peut faire échouer le suivi. Par exemple, le suivi de personne du système **ALIVE** (voir page 14) suppose la présence d'une seule personne dans le champ de la caméra. Si une deuxième personne se présente, le suivi échoue.

- Le changement de conditions par rapport au calibrage : de nombreuses techniques de vision par ordinateur, par exemple la technique de seuillage présentée en annexe, nécessitent une phase préalable de calibrage qui mesure une propriété supposée constante de la scène pendant cette phase. Si, en phase opérationnelle, les conditions changent, la mesure de la propriété peut devenir invalide et provoquer l'échec du suivi.

L'estimation de la validité est utile au contrôle de haut niveau du processus de suivi. Si la position est estimée "valide", elle peut être utilisée pour prédire la prochaine position de la cible. Si la position est estimée "invalide", la recherche de la cible doit être élargie.

Certaines techniques de suivi incluent de manière intrinsèque le calcul de la validité de l'estimation. Par exemple, le suivi par corrélation présenté en page 95, effectue une mesure de similarité entre l'image courante de la cible et une image de la cible mémorisée pendant la phase de calibrage. Cette mesure de similarité tient lieu de mesure de validité.

D'autres techniques de suivi n'incluent pas de calcul de validité de l'estimation de la position. Dans ce cas, il faut faire appel à un mécanisme ad hoc d'évaluation de la validité de l'estimation. Dans [Coutaz 96], nous proposons une technique d'estimation de validité généralisable à toute technique de suivi fournissant une représentation de la position de la cible sous forme d'un vecteur de paramètres d'observation. Notre technique est valide dans la limite de l'hypothèse d'une distribution gaussienne des paramètres valides. Le détail de cette technique est fourni en annexe A page 183.

La validité de l'estimation peut être calculée de façon explicite, comme nous venons de le voir. Elle peut également être calculée de façon implicite lors de la phase de prédiction.

1.5. PRÉDICTION

L'objectif de la prédiction est d'estimer la position de la cible dans l'image suivante. Pour un cycle donné, la prédiction effectuée au cycle précédent permet d'estimer la validité de l'estimation de la position courante mais aussi d'optimiser l'espace de recherche. C'est ce dernier usage que nous privilégions dans nos techniques de suivi.

Optimisation de l'espace de recherche. Reprenons l'exemple du jongleur cité en page 64 du chapitre III. La cible du suivi est l'une des balles du jongleur. Nous savons que, une fois lancée, la balle est soumise à la pesanteur et suit une trajectoire parabolique. Lorsque le suivi a correctement localisé les positions de la balle dans les premiers instants du lancé, il est possible d'estimer la forme de la parabole, et donc de prédire la position future de la balle à tout instant, jusqu'à ce qu'elle soit rattrapée par le jongleur. Ainsi, la recherche de la balle, dans une nouvelle

image peut être réduite à un voisinage restreint de sa position prédite. L'exemple de la balle du jongleur est extrême car la trajectoire spatio-temporelle de la balle est déterminée par la loi de la gravitation. Une connaissance même partielle des paramètres de mouvement de la cible est suffisante pour réduire l'espace de recherche par prédiction.

Supposons par exemple que la norme de l'accélération maximale γ_m d'une cible soit connue. La vitesse $\vec{v}$ de la cible peut être estimée à l'instant t en fonction de ses positions à l'instant précédent t. Lorsqu'il s'agit de localiser la cible dans une nouvelle image à $t + \Delta t$, la norme de la vitesse de la cible n'a pu varier de plus de $\gamma_m \cdot \Delta t$. Il est donc raisonnable de restreindre la recherche de la cible dans un disque de rayon $\gamma_m \cdot \Delta t^2/2$, centré sur la position de la cible à t translatée de $\vec{v} \cdot \Delta t$. La taille du disque est d'autant plus faible que la norme de l'accélération maximale est faible. Ce cas est représentatif de la plupart des véhicules dont la forte masse limite la capacité d'accélération.

Validation de l'estimation de position. Reprenons le cas précédent où l'accélération maximale de la cible est connue. Supposons que l'étape d'estimation résulte en une position de la cible à $t + \Delta t$ largement en dehors de la prédiction du disque de recherche. Il est clair que la cible n'a pu se déplacer à cet endroit, et que l'estimation est invalide. C'est pourquoi validation et prédiction sont fréquemment intégrées en un traitement unique. Le filtre de Kalman ([Kalman 60], [Maybeck 79]) est largement utilisé pour implémenter l'étape de validation et prédiction du suivi d'entité car il intègre ces deux phases de façon efficace grâce à une formulation récursive.

Dans le cadre d'un suivi pour l'interaction fortement couplée, nous relevons deux problèmes à la mise en œuvre d'une phase de prédiction :

- Le comportement humain est imprévisible. La prédiction, nous l'avons vu, est fondée sur une régularité de comportement de la cible. Or, il est difficile d'identifier des régularités dans le comportement d'une personne. Par exemple, les personnes sont capables d'accélération de grande amplitude de leurs membres (bras, main, tête) au regard de leur vitesse de déplacement maximale. Ce phénomène réduit l'intérêt de la prédiction fondée sur la connaissance de l'accélération maximale : en pratique, le rayon du disque de recherche est supérieur à la taille de l'image. Il ne permet donc pas de réduire l'espace de recherche. Il conviendrait de s'appuyer sur des études du comportement moteur humain pour étayer plus avant notre argumentation.
- Le calcul de la prédiction à l'instant $t + \Delta t$ nécessite une estimation précise des paramètres de mouvement à l'instant t. En effet, toute erreur d'estimation est amplifiée lorsque la trajectoire de la cible est "projetée" dans le futur. Le filtre de Kalman apporte une solution à ce problème en "lissant" les mesures au cours du temps : l'effet des

erreurs de mesure individuelles est atténué en calculant les paramètres du mouvement en fonction de plusieurs mesures réalisées dans le passé proche. Le problème est que cette approche introduit un délai : l'estimation n'est plus uniquement fondée sur la mesure courante, mais aussi sur les mesures du passé. L'estimation ne reflète donc pas la position courante mais plutôt une position "récente" de la cible. Il convient de faire un compromis entre délai de calcul et précision de l'estimation. Dans le cadre d'une interaction fortement couplée, la contrainte de latence définie au chapitre II page 53 milite en faveur d'une réduction des délais, limitant l'intérêt du filtrage et de la prédiction réalisée par l'algorithme de Kalman.

Dans une expérience décrite dans [Oliver 97] sur le suivi du visage, nous constatons que le filtre de Kalman introduit des délais incompatibles avec le requis de latence. Dans cette expérience, nous nous limitons à un filtre de Kalman d'ordre 0 (c'est-à-dire sans étape prédictive). Le filtre est uniquement utilisé à des fins de validation et de lissage de l'estimation de la position. Pour le contexte précis de l'interaction fortement couplée, nous abandonnons l'usage d'un filtre de Kalman.

Ayant introduit le principe général du suivi d'entité, nous présentons trois techniques de mise en œuvre : par différences d'image, par modèle de couleur et par corrélation. Pour chacune d'elles, notre analyse comprend quatre volets : principe, réalisation, performances et discussion. La réalisation sera analysée selon les étapes "mesure, estimation, validation-prédiction" du principe général de suivi.

2. *Suivi par différence d'images*

2.1. PRINCIPE

Le suivi par différence d'images s'appuie sur la stabilité statique des pixels. Nous avons vu au paragraphe "Instabilité statique (bruit)" page 63 que le flux généré par une caméra vidéo est statiquement instable. La technique de seuillage, présentée en annexe A page 181, permet de se ramener dans des conditions de stabilité statique. La stabilité statique des pixels signifie que la valeur des pixels du flux est constante si le point de vue de la caméra est statique, si toutes les entités de la scène sont fixes, et si les conditions lumineuses de la scène ne varient pas. Dans ces conditions, si certains pixels ne sont pas constants entre les différentes images du flux, c'est qu'ils représentent une (ou des) entité(s) en mouvement. Suivant notre approche d'ajout contrôlé de contraintes sur l'environnement (page 72), nous faisons l'hypothèse qu'il ne peut exister,

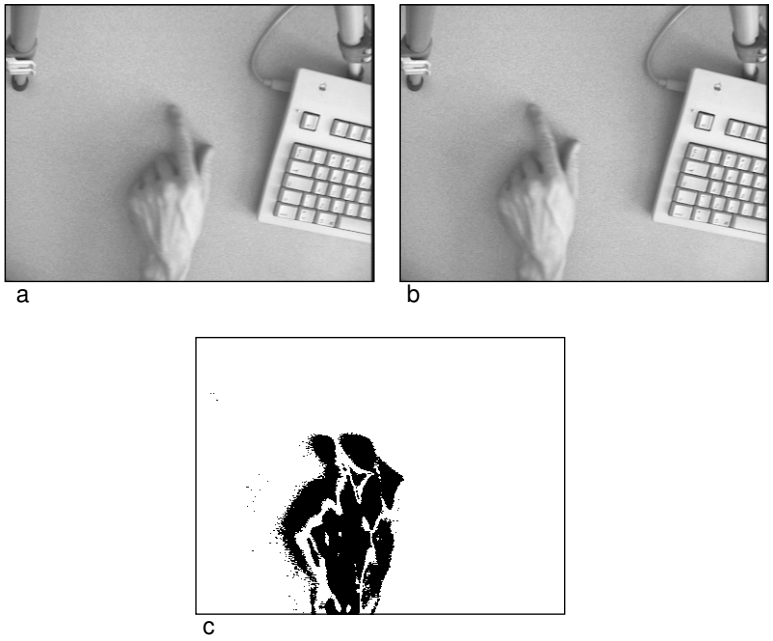


Figure 2
Différence entre images successives
Deux images successives (a et b) d'un flux vidéo, et leur différence (c).
Dans l'image de différence, les points sont d'autant plus sombres que l'intensité lumineuse a varié entre les deux images.

à un instant donné, qu'une seule entité en mouvement dans la scène. Celle-ci est localisée lorsqu'elle se déplace, à l'emplacement des pixels dont la valeur a varié.

2.2. RÉALISATION

Mesure

Les pixels qui correspondent au mouvement d'entité sont mis en évidence par le calcul de la différence entre deux images du flux vidéo. Soit $I_t(x, y)$ l'intensité lumineuse du pixel de coordonnées (x, y) dans l'image I du flux vidéo au temps t. L'*image de différence* D s'obtient par :

$$\forall x, \forall y, D_t(x, y) = \left| I_t(x, y) - I_{t-1}(x, y) \right| \qquad (5)$$

Dans l'image D, les pixels situés à un emplacement où il n'y a pas eu de mouvement entre les instants $t-1$ et t ont la valeur 0, les autres pixels, ont une valeur supérieure à 0.

Le bruit de caméra (voir page 63) est responsable de l'oscillation des valeurs de tous les pixels. S'il n'est pas traité, la majeure partie des pixels de l'image D ont une valeur supérieure à 0. Il est donc nécessaire d'appliquer une technique de seuillage (voir en annexe A page 181). Le seuillage permet de considérer uniquement les différences de valeurs *significatives*, c'est-à-dire celles qui dépassent le seuil. La valeur du seuil est déterminée par une phase initiale de calibrage selon la technique de Stafford-Fraser (voir page 182).

Comme l'illustre la figure 2c, l'image de différence met en évidence les zones de l'image où quelque chose a bougé. Cela concerne aussi bien les

zones d'où l'entité est partie que les zones où elle est arrivée. Il est difficile de déterminer, sur la figure 2c, lequel des deux doigts apparents correspond à la position la plus récente du doigt. Une façon de résoudre ce problème est de calculer une différence d'images, non pas entre les images successives du flux, mais entre les images du flux et une *image de référence*. L'image de référence est mémorisée à l'initialisation du suivi alors que l'entité à suivre n'est pas dans le champ de vue de la caméra. L'image de différence entre l'image de référence et une image du flux vidéo met en évidence les pixels qui ont changé par rapport à l'image de référence, c'est-à-dire les pixels représentant la position courante de l'entité. La figure 3 en illustre le principe. Krueger ([Krueger 90], "VideoPlace" page 12) réalise une variante de cette technique : il installe un écran lumineux comme fond de la scène. Ainsi, la capture de l'image de référence n'est pas nécessaire : tout pixel dont l'intensité est nettement plus faible que l'intensité du fond lumineux correspond au participant.

Estimation de la position

L'estimation de la position de la cible est réalisée soit par un calcul des statistiques de distribution spatiale des pixels de mouvement (voir page 78), soit par calcul d'une boîte englobante après analyse en composantes connexes (voir page 79). Une boîte englobante est représentée sur la figure 3c.

Krueger ([Krueger 90]) utilise l'ensemble des pixels de l'image D en tant que silhouette du participant. Quatre images seuillées contenant la silhouette du participant sont représentées sur la figure 5 page 13.

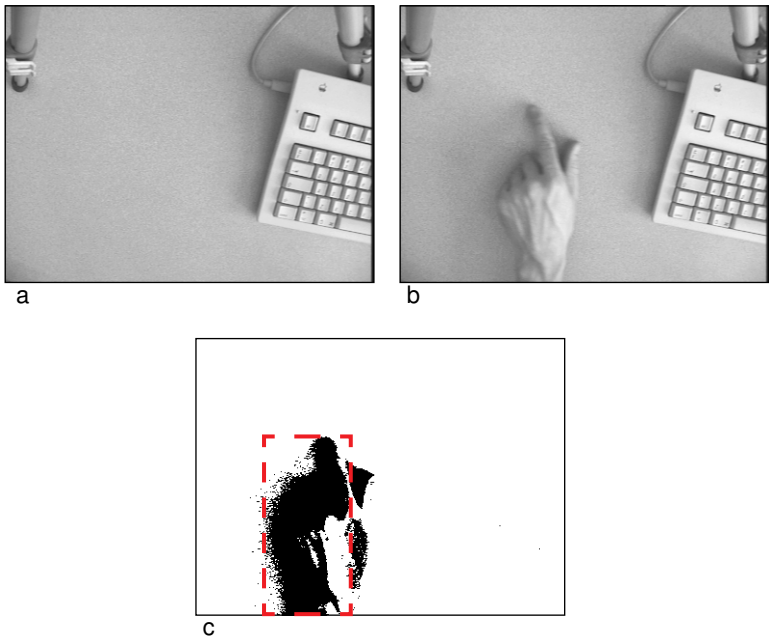


Figure 3
Différence par rapport à une image de référence
Image de référence (a), image courante (b), et image de différence seuillée (c). Le cadre pointillé sur l'image de différence représente le rectangle englobant de la plus grande classe de pixels connexes.

Validation et prédiction

La validité de l'estimation de position est déduite des paramètres de position en appliquant la technique de calcul de validité décrite dans l'annexe A page 183.

Lorsque la cible n'est pas en mouvement, elle n'apparaît pas dans l'image de différence. L'image de différence ne contient aucun pixel. La cible étant immobile, sa position se définit comme la dernière position où son déplacement a été détecté.

Du fait de sa faible complexité de calcul, le suivi par différence d'images autorise une recherche de la cible sur toute la surface de chaque image du flux vidéo. Il n'est donc pas nécessaire de concevoir une phase de prédiction pour optimiser la recherche.

2.3. PERFORMANCES

La faible complexité des traitements mis en jeux autorise un fonctionnement à fréquence élevée. Pour un flux vidéo dont les images sont de taille 384 x 288 (quart d'une image au format PAL), notre réalisation sur processeur PowerPC 604 à 350 MHz fonctionne à 41 Hz dans le cas de différence entre images successives et 52 Hz dans le cas de différence avec une image de référence. La fréquence moins élevée du premier cas s'explique par la nécessité de mémoriser l'image courante à chaque cycle alors que cette mémorisation n'a pas lieu d'être dans le second cas.

Exprimée dans le repère image, la résolution du suivi par différence d'images est de un pixel. Par contre, l'estimation de position n'est pas stable statiquement. L'instabilité statique est un effet de la discrétisation lors de la numérisation des images. La frontière entre la zone qui représente la cible dans l'image et le fond de l'image n'est pas "franche". Lorsque cette frontière traverse un pixel en son milieu, le pixel est affecté d'une intensité lumineuse mitoyenne entre celle de la cible et celle du fond. Ce phénomène est illustré sur la figure 4. De nombreux pixels, situés à la frontière de la cible, se voient affectés d'une intensité lumineuse proche du seuil de binarisation de l'image de différence (page 84). Ces pixels sont donc susceptibles d'êtres classés en tant que pixel de mouvement, ou non, même lors de très faibles variations de leur intensité lumineuse. En pratique, l'oscillation de valeur due au bruit de

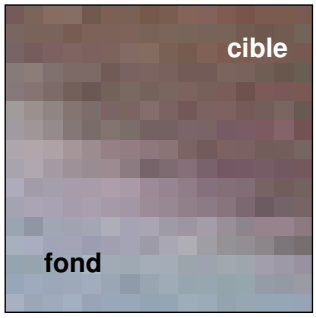


Figure 4
Frontière de la cible (un doigt)
Les pixels situés à la frontière entre la cible et le fond de l'image ont une apparence qui représente à la fois la cible et le fond.

caméra a effectivement pour conséquence d'attribuer ou non à la cible les pixels situés sur sa frontière. Ce phénomène explique l'instabilité statique de l'estimation de position.

2.4. DISCUSSION

Le suivi par différence d'images est une technique simple dont la réalisation correspond à des algorithmes de faible complexité. L'avantage est double : la réalisation d'un suivi demande peu d'effort de développement, et les performances en terme de fréquence de fonctionnement sont élevées même sur architecture matérielle modeste. Ces avantages font de cette technique le choix privilégié des chercheurs du domaine de l'interaction homme-machine : Stafford-Fraser pour le suivi des personnes ([Stafford-Fraser 96a]), Krueger pour VideoPlace ([Krueger 90], voir page 12), Gaver pour le suivi du visage dans son prototype de fenêtre virtuelle ([Gaver 95], voir page 20), et Weller pour le suivi du doigt dans le bureau digital ([Wellner 93b], voir page 33).

Le suivi par différence d'images nécessite cependant l'ajout contrôlé de contraintes sur l'environnement (voir page 72) :

1. La caméra doit être fixe. Un mouvement de caméra, même minime, est perçu comme un déplacement de l'ensemble des pixels de la scène.
2. Il ne peut y avoir qu'une seule entité en mouvement dans la scène. Lorsque plusieurs entités se déplacent, chacune génère un ensemble de pixels de mouvement dans l'image de différence. Il est difficile d'attribuer ces pixels à chaque entité, notamment lorsque deux entités se rapprochent et que les pixels de mouvement correspondants sont connectés.
3. Les conditions lumineuses de la scène ne doivent pas varier. Une variation de condition lumineuse de la scène, qu'elle soit locale (on allume une lampe, le passage d'une personne crée une ombre) ou globale (la lumière du jour s'atténue), entraîne une variation de la valeur d'un grand nombre de pixels de la scène. Ces pixels sont détectés comme des pixels de mouvement dans l'image de différence. Cette contrainte est moins sensible si les différences sont calculées sur les images successives du flux vidéo et que la variation d'intensité est lente. Dans ce cas, l'écart d'intensité lumineuse entre deux images reste faible et n'est pas détecté comme un mouvement.

La contrainte de la caméra fixe ne présente pas, en règle générale, de difficulté. Les applications décrites au premier chapitre font toutes l'hypothèse d'une caméra fixe. Les deux autres contraintes sont plus gênantes : dans un environnement non contrôlé, il peut arriver que plusieurs entités se déplacent dans le champ de la caméra et la majorité des environnements de travail est sujette à des variations d'intensité lumineuse.

En pratique, cette technique de suivi n'est applicable que dans le cas où une interaction en environnement contrôlé est acceptable (par exemple, VideoPlace et ALIVE). Gaver ([Gaver 95]) et Wellner ([Wellner 93b]) utilisent la différence d'images pour réaliser un démonstrateur de concept. Gaver reconnaît que les faiblesses du système de suivi de la **fenêtre virtuelle** expliquent pour une large part l'inutilisabilité du système.

En résumé, nous retenons la faible complexité de la technique de suivi par différence d'images, qui autorise un traitement peu coûteux sur l'ensemble du flux vidéo. Nous recommandons de l'utiliser comme *détecteur de mouvement* dans le flux vidéo, puis, s'il y a mouvement, faire appel à d'autres techniques mieux adaptées au suivi.

3. Suivi par modèle de couleur

3.1. PRINCIPE

Le principe du suivi par modèle de couleur consiste à détecter dans l'image les pixels dont la couleur est proche de celle de la cible. Une couleur est modélisée par un triplet de valeurs rvb qui représente l'intensité lumineuse de la couleur dans les trois bandes de fréquence : rouge, verte et bleue. Connaissant la valeur rvb de la couleur de la cible, les pixels de l'image dont la valeur est égale à ce triplet sont considérés comme appartenant à la cible.

Cette technique suppose que la couleur de la cible est uniforme et *discriminante* (en d'autres termes, les entités non pertinentes de la scène sont supposées de couleur différente de celle de la cible). Nous utilisons cette technique pour le suivi du visage ou des mains qui sont de couleur de peau. Cette couleur est en général discriminante dans un environnement d'intérieur tel qu'un bureau.

La représentation rvb permet de représenter toutes les apparences d'un pixel dans une image, et notamment une même teinte selon différents niveaux de luminosité. Par exemple, les deux triplets rvb (50 %, 0 %, 0 %) et (100 %, 0 %, 0 %) représentent la teinte rouge saturée, mais à deux niveaux de luminosité (respectivement, luminosité médiane et luminosité maximale). Il résulte que la représentation rvb associe une couleur différente à une entité donnée pour chaque condition d'illumination. En conséquence, un triplet rvb choisi sur une zone "illuminée" de l'entité affecte une faible valeur de ressemblance à cette même entité lorsqu'elle est placée dans des conditions plus sombres. Schiele et Waibel ([Schiele 95]) proposent de s'affranchir de ce problème en *normalisant* la couleur des pixels par leur luminosité. Ils considèrent la *teinte*, c'est-à-

dire la couleur normalisée par la luminosité, plutôt que la couleur. Les pixels représentant la même entité sous des éclairages différents ont la même teinte. Considérer la teinte, plutôt que la couleur, permet par exemple de suivre la cible alors qu'elle passe d'une zone éclairée de la scène à une zone plus sombre. Les variations de luminosité globales de la scène sont également tolérées.

En outre, un simple triplet rvb normalisé par la couleur est un modèle inadapté aux scènes réelles. Les entités d'une scène courante ne sont pas représentées par des pixels de teinte unie, mais par un ensemble de pixels dont la teinte varie (la peau d'un individu présente des variations de teinte). Il est donc nécessaire de modéliser la teinte par une représentation statistique plus complexe qu'un simple triplet rvb normalisé. Nous étudierons deux modèles statistiques fondés sur l'histogramme et sur la fonction gaussienne.

3.2. RÉALISATION

Normalisation de la luminosité

Une approximation peu complexe de la luminosité d'un pixel p est obtenue en additionnant les trois composantes du pixel :

$$p_l = p_r + p_v + p_b \tag{6}$$

Les composantes normalisées d'un pixel sont calculées par :

$$p_R = \frac{p_r}{p_l} \qquad p_V = \frac{p_v}{p_l} \qquad p_B = \frac{p_b}{p_l} \tag{7}$$

On dit que les composantes normalisées représentent la *teinte* du pixel alors que ses composantes "brutes" représentent sa *couleur*. On note que les trois composantes normalisées sont linéairement dépendantes :

$$p_R + p_V + p_B = \frac{p_r + p_v + p_b}{p_l} = 1 \tag{8}$$

Il est donc inutile d'effectuer des traitements sur les trois composantes puisque deux suffisent à représenter entièrement l'information de teinte.

Histogramme de couleur

Swain et Ballard ([Swain 91]) montrent que l'*histogramme de couleurs* est un modèle fiable pour la reconnaissance d'entités colorées. Un histogramme de couleur est défini pour un ensemble de pixels E. Il représente, pour chaque couple de composantes normalisées (R, V), le nombre de pixels de l'ensemble E dont la teinte est égale au couple. Un histogramme est construit *par l'exemple*, c'est-à-dire par calcul de l'histogramme sur un ensemble de pixels prédéterminés (l'ensemble E). Nous utilisons une interface graphique pour sélectionner un rectangle de

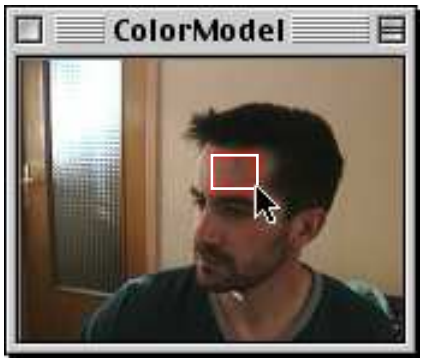


Figure 5
Constitution du modèle de couleur par l'exemple
Un cadre, défini à la souris, délimite l'ensemble des pixels qui serviront d'exemple pour la constitution du modèle.

l'image correspondant à la couleur de peau. Cette manipulation est illustrée sur la figure 5.

L'histogramme h d'un ensemble E de pixels est calculé de la façon suivante. L'histogramme est d'abord initialisé :

$$\forall R, \forall V \qquad h(R, V) \leftarrow 0 \tag{9}$$

On incrémente ensuite la valeur de l'histogramme correspondant à la teinte de chaque pixel de l'ensemble E (l'équation 7 exprime le calcul des composantes normalisées d'un pixel).

$$\forall p \in E, \qquad h(p_R, p_V) \leftarrow h(p_R, p_V) + 1 \tag{10}$$

Il est souvent commode de considérer l'histogramme comme la loi de probabilité de l'appartenance d'une teinte au modèle. La somme des valeurs d'une fonction représentant une loi de probabilité doit être égale à 1. Nous calculons donc l'histogramme normalisé H par :

$$\forall R, \forall V \qquad H(R, V) = \frac{h(R, V)}{N} \tag{11}$$

où N est le nombre de pixels de l'image.

Modèle gaussien

Pour que l'histogramme soit représentatif de la couleur de l'entité à suivre, il est souhaitable que l'ensemble de pixels sur lequel il est calculé soit assez vaste pour représenter toutes les teintes possibles de l'entité. En pratique, il est difficile de déterminer un tel ensemble. Par exemple, concernant la couleur de peau, il faudrait avoir le moyen de sélectionner tous les pixels du visage. Une simple sélection rectangulaire, telle que celle de la figure 5, ne permet pas une désignation aussi fine. De plus, il faudrait prendre en compte la variation de teinte au cours du temps du fait du bruit de caméra.

Nous proposons une autre approche qui consiste à utiliser un modèle statistique permettant de prédire la distribution des teintes de la cible en fonction d'un nombre réduit d'échantillons. Nous observons sur de nombreux histogrammes construits par l'exemple que les teintes de couleur de peau se répartissent dans l'espace des teintes selon une

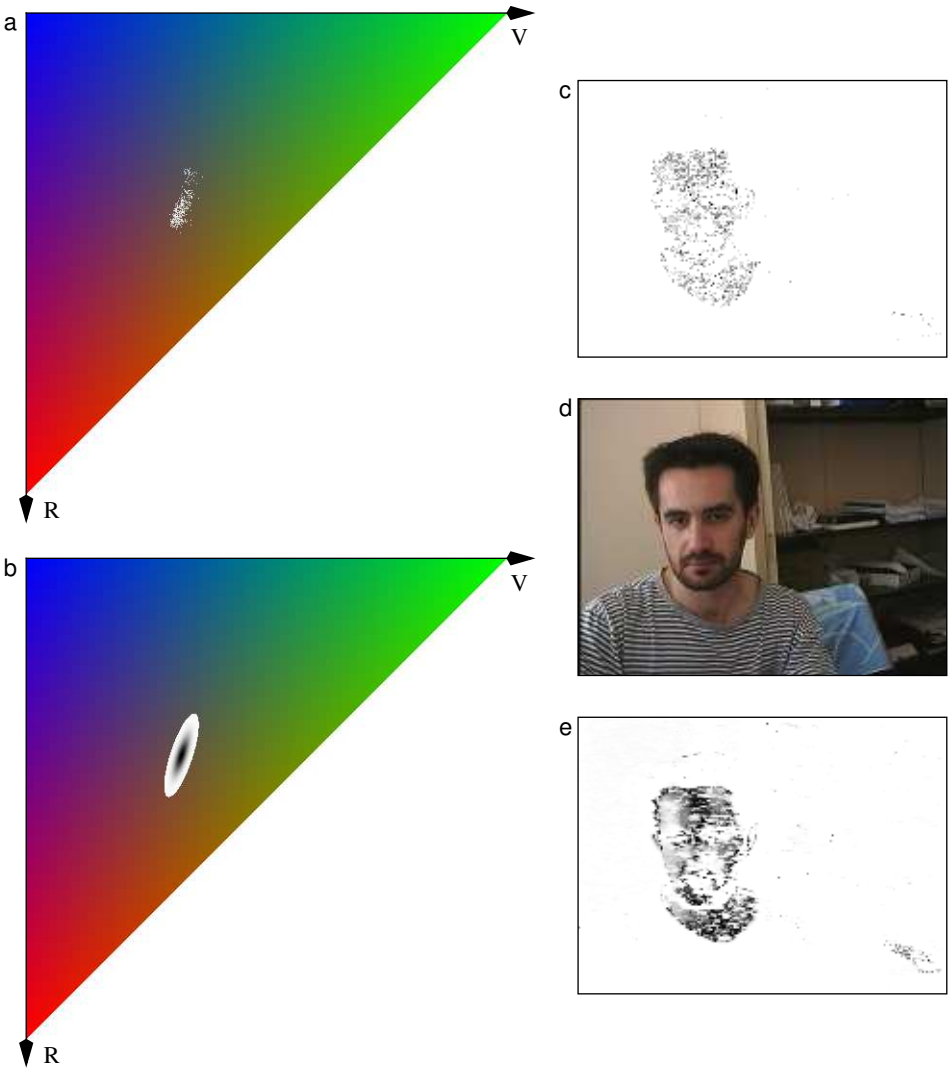


Figure 6
Modèle histogramme (a, c) et modèle gaussien (b, e)
Dans les espaces de couleurs normalisées (a) et (b) et les images de probabilité (c) et (e), les points sont d'autant plus sombres que la probabilité d'appartenance à la couleur de peau est forte.

fonction gaussienne à deux dimensions. Nous choisissons de représenter la couleur de peau par une courbe gaussienne à deux dimensions. Les paramètres de la gaussienne sont calculés sur l'ensemble E de pixels de couleur de peau. La figure 6 illustre l'intérêt du modèle gaussien : certaines teintes, qui ne sont pas présentes dans l'ensemble E, sont proches du centre de la gaussienne et sont intégrées au modèle. Ceci résulte, sur l'image de probabilité, en une plus forte densité des pixels de couleur de peau et facilite le calcul de la position du visage.

Le modèle gaussien est paramétré par la moyenne μ et la matrice de covariance C des composantes normalisées des pixels de E :

$$\mu = \begin{bmatrix} \mu_R \\ \mu_V \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{p \in E} p_R}{\mathrm{card}(E)} \\ \dfrac{\sum_{p \in E} p_V}{\mathrm{card}(E)} \end{bmatrix} \qquad C = \begin{bmatrix} \sigma_{RR} & \sigma_{RV} \\ \sigma_{RV} & \sigma_{VV} \end{bmatrix} \tag{12}$$

avec

$$\sigma_{xx} = \frac{\sum_{p \in E} (p_x - \mu_x)^2}{\text{card}(E)} \qquad \sigma_{xy} = \frac{\sum_{p \in E} (p_x - \mu_x) \cdot (p_y - \mu_y)}{\text{card}(E)} \tag{13}$$

Les équations 13 nécessitent deux parcours de l'ensemble E pour calculer la matrice de covariance, le premier parcours servant au calcul de la moyenne. Ces équations sont transformées pour permettre le calcul de tous les paramètres de la gaussienne en un seul parcours de l'ensemble E :

$$\sigma_{xx} = \frac{\sum_{p \in E} p_x^2 - 2p_x\mu_x + \mu_x^2}{\text{card}(E)} = \frac{\sum_{p \in E} p_x^2}{\text{card}(E)} - 2\mu_x \frac{\sum_{p \in E} p_x}{\text{card}(E)} + \mu_x^2 \tag{14}$$

$$\sigma_{xx} = \frac{\sum_{p \in E} p_x^2}{\text{card}(E)} - \mu_x^2 \tag{15}$$

Le calcul de σ_{xy} se transforme de façon similaire. Sous cette forme, la somme des carrés est calculée en même temps que la moyenne. Un seul parcours des pixels de E est donc nécessaire au calcul de tous les paramètres du modèle.

La probabilité qu'un pixel appartienne au modèle est naturellement définie par la loi normale à deux dimensions. Pour des raisons d'efficacité, cette probabilité est précalculée dans une table. Cette table étant de même nature que l'histogramme, nous conservons la notation H :

$$\forall t = \begin{bmatrix} t_R \\ t_V \end{bmatrix} \qquad H(t_R, t_V) = \frac{1}{2\pi\sqrt{\det(C)}} \cdot e^{-\frac{1}{2} \cdot (t-\mu)^t \cdot C^{-1} \cdot (t-\mu)} \tag{16}$$

Mesure

Une fois constituée, la table H est un modèle (de type histogramme ou gaussien) de la teinte de la cible. L'étape de mesure d'une image I consiste à créer une *image de probabilité* R. La valeur d'un pixel de R est égale à la probabilité d'appartenance au modèle de la teinte du pixel correspondant dans I. Le calcul de R se fait par consultation de la table H :

$$\forall x, \forall y, \qquad R(x, y) = H(I_R(x, y), I_V(x, y)) \tag{17}$$

La figure 7b montre une image de probabilité de couleur de peau.

Estimation de la position

L'estimation de position peut être calculée sous forme de statistique de la distribution spatiale des pixels ayant la teinte du modèle (le résultat est illustré sur la figure 1 page 79), ou bien, après seuillage et analyse en composantes connexes, sous la forme d'une boîte englobante. Les différentes étapes de calcul de la boîte englobante sont représentées dans la figure 7.

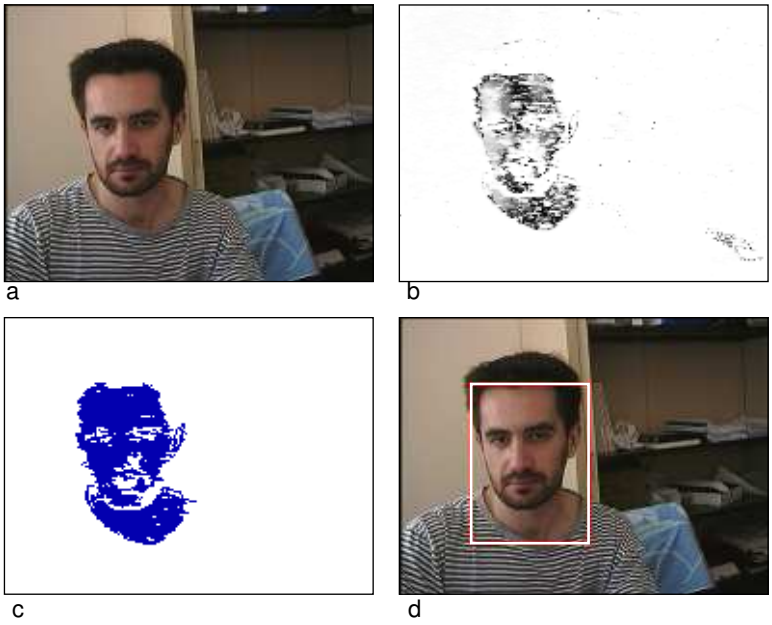


Figure 7
Suivi du visage par modèle de couleur de peau (modèle gaussien)
Le modèle de couleur est appliqué sur une image du flux vidéo (a) afin de calculer l'image de probabilité de couleur de peau (b). En (b), les points sont d'autant plus sombres qu'ils ressemblent à la peau.
La position du visage est extraite par d'autres traitements sur l'image de probabilité : nous donnons l'exemple du seuillage et recherche de la plus grande zone connexe (c) à partir de laquelle est calculée la boîte englobante (d).

Validation et prédiction

La validation de l'estimation se fait en appliquant la technique de calcul de validité détaillée en annexe A page 183. Cette technique permet de rejeter les estimations correspondant à une forme de cible aberrante, telle qu'une boîte englobante plus large que haute dans le cas du suivi de visage.

Comme pour le suivi par différence d'images, la faible complexité de calcul du suivi par modèle de couleur permet de rechercher la cible sur toute la surface de chaque image du flux vidéo. Il n'est donc pas nécessaire de concevoir une phase de prédiction pour optimiser la recherche.

3.3. PERFORMANCES

Le modèle, de type histogramme ou gaussien, est précalculé dans une *table*. Le calcul de l'image de probabilité consiste à parcourir l'image en consultant la valeur de la table pour chaque pixel. Les traitements se résument donc à :

1. une somme pour calculer la luminosité du pixel (équation 6 page 89),
2. deux divisions pour normaliser les composantes du pixel (équation 7 page 89),
3. la consultation de la valeur de la table correspondante (équation 17 page 92).

En raison de la faible complexité de calcul de cette technique, notre réalisation sur processeur PowerPC 604 à 350 MHz fonctionne à la fréquence de 15 Hz sur des images de taille 384 x 288 et 42 Hz sur des images de taille 192 x 144.

Comme pour le suivi par différence d'images, la résolution spatiale de l'estimation de position du modèle de couleur (exprimée dans le repère de

l'image) est de un pixel. L'estimation de position est également statiquement instable du fait de la discrétisation à la frontière de la cible (voir notre analyse en page 86).

3.4. DISCUSSION

Le suivi par modèle de couleur présente des avantages sur le suivi par différence d'images. Ces avantages se traduisent par une réduction des contraintes sur l'environnement :

- La caméra n'est pas nécessairement fixe.
- Plusieurs entités en mouvement peuvent être présentes dans le champ de la caméra à condition qu'une seule d'entre elles présente la couleur du modèle. Par exemple, les pales d'un ventilateur en mouvement dans le champ de la caméra ne perturbent pas le suivi mais l'apparition d'une entité de même teinte peut être cause d'oscillations.
- La considération de la teinte, plutôt que la couleur, permet de s'affranchir des problèmes de luminosité : par exemple le suivi fonctionne lorsque la luminosité générale diminue ou que l'entité suivie passe d'une zone d'ombre à une zone de lumière.

Le suivi par modèle de couleur présente aussi des limitations :

- L'information de position est statiquement instable.
- Il est nécessaire de construire le modèle de couleur par l'exemple. Une désignation manuelle de l'ensemble exemple, telle qu'illustrée sur la figure 5 page 90, serait acceptable si le modèle construit était totalement général. Or l'expérience montre qu'il est nécessaire d'adapter le modèle de couleur à différents environnements lumineux : la teinte de la peau est très différente selon qu'elle est illuminée en lumière artificielle (incandescente, néon) ou en lumière naturelle. Il est donc souhaitable de pouvoir construire un nouveau modèle lorsque cela est nécessaire, sans intervention de l'utilisateur.
  Yang et ses collaborateurs ([Yang 98b]) proposent un algorithme d'adaptation des paramètres du modèle gaussien au cours du temps. Mais cet algorithme ne tolère que des modifications progressives. Il n'est donc pas applicable aux changements brutaux de l'éclairage. Nous proposons, page 107, l'initialisation automatique du modèle de couleur.

Les deux techniques de suivi étudiées jusqu'ici, différence d'image et modèle de couleur, ne permettent de localiser qu'une position grossière de l'entité suivie. Elles ne sont pas applicables, par exemple, à l'extraction de mouvements fin du visage. De plus, ces techniques sont sensibles au bruit de caméra et fournissent une information de position statiquement instable. Elles conviennent à des applications qui ne reposent pas sur une position précise et stable de l'entité suivie. Par exemple, Black et Jepson ([Black 98b]) utilisent notre système de suivi par modèle de couleur pour

créer un corpus de trajectoires destinées à un système de reconnaissance de geste. Un système de reconnaissance de geste doit, par nature, reconnaître un geste précis malgré les variations importantes de trajectoires entre les différentes instances du geste effectuées par l'utilisateur. Il n'est donc pas nécessaire, dans ce cas, d'extraire une position précise et stable.

Un de nos cas d'études pour l'interaction fortement couplée, la **fenêtre perceptuelle** (voir le chapitre VI), nécessite le suivi d'une partie du visage, non pas du visage entier. Aucune des deux techniques de suivi présentées jusqu'ici n'est apte à fournir une telle information. De plus, le requis de stabilité identifié au chapitre II nous amène à considérer une technique de suivi produisant une estimation stable de la position de la cible : le suivi par corrélation.

4. Suivi par corrélation

4.1. PRINCIPE

Lorsqu'une entité se déplace dans le champ de la caméra, son apparence ne change pas si le déplacement est une translation dans un plan parallèle à celui de l'image et si l'éclairage est constant et diffus (sans ombres). Ces conditions sont rarement satisfaites dans un univers non contrôlé. Cependant, si le mouvement de l'entité est proche d'une translation dans un plan parallèle à celui de l'image et si les ombres de la scène ne sont pas prononcées, on constate que l'apparence de l'entité varie peu. La figure 8 illustre le phénomène.

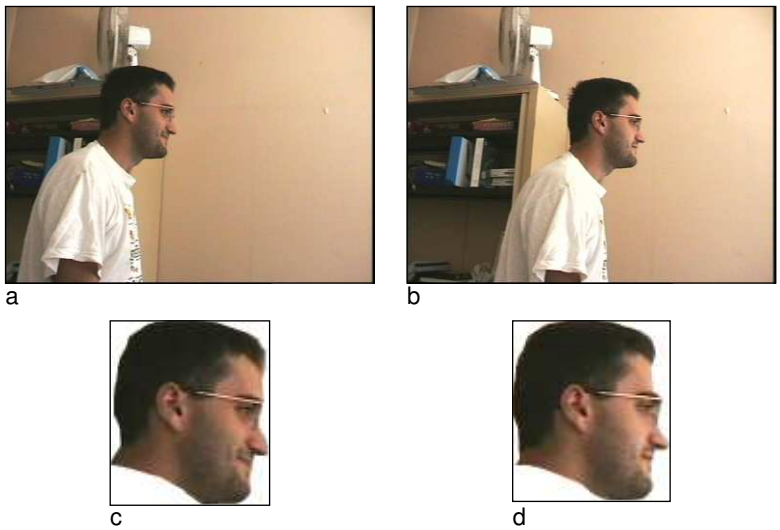


Figure 8
Conservation de l'apparence d'une entité durant une translation parallèle au plan de la caméra
Dans les deux images (a) et (b), les apparences du visage en (c) et en (d) sont très similaires : leur mouvement est proche d'une translation parallèle au plan de l'image et l'éclairage ambiant est diffus.

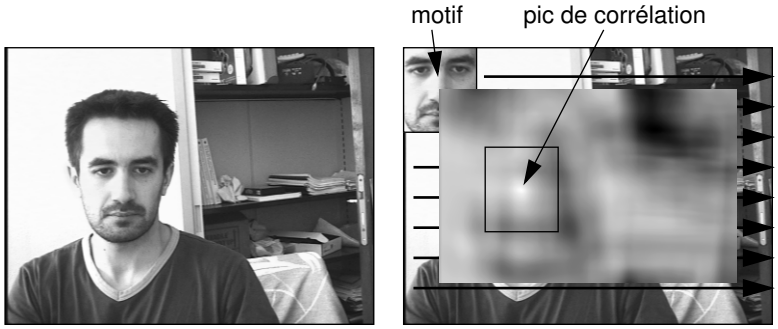


Figure 9
Suivi par corrélation
Le motif est comparé à l'image en tout lieu par un parcours systématique. Les flèches indiquent le sens de ce parcours. Les points de la carte de corrélation (à droite) sont d'autant plus clairs que le motif est similaire à l'image en ce lieu. Le pic de corrélation (valeur de corrélation maximale) est élu comme position de la cible dans l'image.

Le principe du suivi par corrélation consiste à mémoriser, dans une phase d'initialisation, l'apparence de la cible. Dans le cas du suivi par corrélation, l'ensemble de pixels ainsi mémorisés est appelé *motif*. En phase opérationnelle, le suivi recherche la partie de l'image la plus ressemblante au motif.

La localisation du motif dans une nouvelle image s'effectue par un parcours de toutes les sous-parties de l'image de même taille que le motif. À chaque étape du parcours, le motif et une des parties de l'image sont comparés. Le résultat de la comparaison est la mesure de corrélation entre le motif et la partie de l'image. À la fin du parcours, on note l'emplacement de la partie de l'image la plus similaire au motif. Cet emplacement est le maximum de la fonction de corrélation (le "pic" de corrélation). L'emplacement du pic de corrélation est élu en tant que nouvelle position de la cible. La figure 9 illustre le principe du suivi par corrélation que nous venons de décrire.

4.2. RÉALISATION

Mesure

La recherche du motif dans une nouvelle image nécessite de "corréler" deux images : le motif et la partie de l'image à traiter. La corrélation de deux images permet d'évaluer leur *similarité*. Une façon de comparer deux images est de calculer la somme des écarts entre les pixels correspondants dans les deux images. Soit M un motif rectangulaire de taille u x v, et I l'image à traiter. L'écart entre le motif et la partie de l'image située aux coordonnées (x, y) est donné par la formule SAD (Sum of Absolute Differences) :

$$\mathrm{SAD}(x, y) = \sum_{u, v} |M(u, v) - I(x + u, y + v)| \qquad (18)$$

La valeur absolue utilisée dans la formule implique que la fonction SAD est discontinue. C'est pourquoi la somme des différences des carrés (SSD pour "Sum of Squared Difference") est souvent préférée à la SAD :

$$\mathrm{SSD}(x, y) = \sum_{u, v} (M(u, v) - I(x + u, y + v)^2) \qquad (19)$$

Deux images sont d'autant plus similaires que les valeurs de SAD et de SSD sont faibles. Dans le cas de deux images identiques, SAD = SSD = 0.

SAD et SSD sont sensibles aux variations globales de luminosité : par exemple, lorsque la luminosité ambiante diminue, l'ensemble des valeurs de pixels de l'image diminue. Cette variation provoque une différence entre le motif et l'image. Il en résulte une augmentation des valeurs de SAD et SSD. Ce phénomène est gênant car il n'est pas possible de distinguer si de fortes valeurs de SSD ou SAD sont dues à une variation de luminosité ambiante ou à une réelle différence entre le motif et l'image.

La formule de corrélation normalisée (NCC pour "Normalized Cross-Correlation") prend en compte la luminosité générale du motif et de l'image pour le calcul de similarité :

$$\mathrm{NCC}(x, y) = \frac{\sum_{u, v} M(u, v) \cdot I(x + u, y + v)}{\sqrt{\sum_{u, v} M^2(u, v) \cdot \sum_{u, v} I^2(x + y, y + v)}} \qquad (20)$$

Le dénominateur de cette équation a pour rôle de normaliser la valeur de corrélation par l'énergie du motif et de l'image. La valeur de NCC est comprise entre 0 et 1. On vérifie aisément que cette valeur est maximale (égale à 1) lorsque le motif et l'image sont identiques à un coefficient de luminosité globale près, c'est-à-dire lorsque :

$$\forall u, \forall v, \qquad M(u, v) = \alpha \cdot I(x + u, y + v) \qquad (21)$$

La figure 9 page 96 illustre le résultat du calcul de NCC entre le motif et toutes les zones de l'image de même taille que le motif.

Martin et Crowley ([Martin 95]) comparent différents calculs de similarité et notent que SSD est plus stable que NCC en présence de bruit. Toutefois, l'insensibilité de NCC aux variations de luminosité globale nous fait préférer cette mesure à SSD pour des applications destinées à être utilisées en environnement lumineux non contrôlé.

Estimation de la position

Dans chaque image, l'estimation de position de la cible est définie par la position du pic de corrélation (voir la figure 9 page 96). Notons que seules sont extraites les coordonnées de la cible dans l'image. En particulier, le suivi par corrélation ne permet par de calculer l'orientation de la cible. Nous verrons au paragraphe "Discussion" page 101 que le principe du suivi par corrélation peut être étendu pour extraire une information de position plus complète.

Validation et prédiction

La validation de l'estimation de position du suivi par corrélation est implicitement calculée par la mesure de corrélation. Celle-ci représente la similarité entre le motif et l'image. Lorsque cette valeur est faible, c'est

que le motif et l'image sont dissemblables. Dans ce cas, il est probable que le suivi a perdu la cible. La détection de l'échec du suivi se fait par contrôle de la valeur de similarité au cours du temps : lorsque cette valeur chute, l'échec du suivi est détecté.

À la différence des suivis par différence d'images et par modèle de couleur, le suivi par corrélation est d'une complexité de calcul élevée. Il est donc nécessaire de prévoir une phase de prédiction afin de limiter l'espace de recherche de la cible. Etudions la taille optimale de la zone de recherche.

Taille de la zone de recherche. La complexité de calcul du processus de recherche peut être grandement réduite si l'on considère l'effet de la fréquence de fonctionnement sur l'espace de recherche : lorsque la fréquence est élevée, le déplacement de la cible entre deux recherches est réduit (la cible n'a pas le temps de beaucoup se déplacer). La recherche de l'entité peut donc être limitée à une zone réduite centrée autour de la dernière position connue de l'entité. Jusqu'à quel point peut-on réduire la taille de la zone de recherche ? Nous présentons ici notre raisonnement sur la définition de la taille optimale ([Crowley 95]).

Nous supposons que le motif et la zone de recherche sont carrés, le raisonnement pouvant facilement se généraliser à d'autres formes. La taille de la zone de recherche optimale est celle qui maximise la vitesse de déplacement de la cible tolérée par le système. Le suivi échoue lorsque la cible effectue un déplacement entre deux images tel que sa position dans la nouvelle image est en dehors de la zone de recherche. La figure 10 illustre le cas où la cible est à la limite de la zone de recherche.

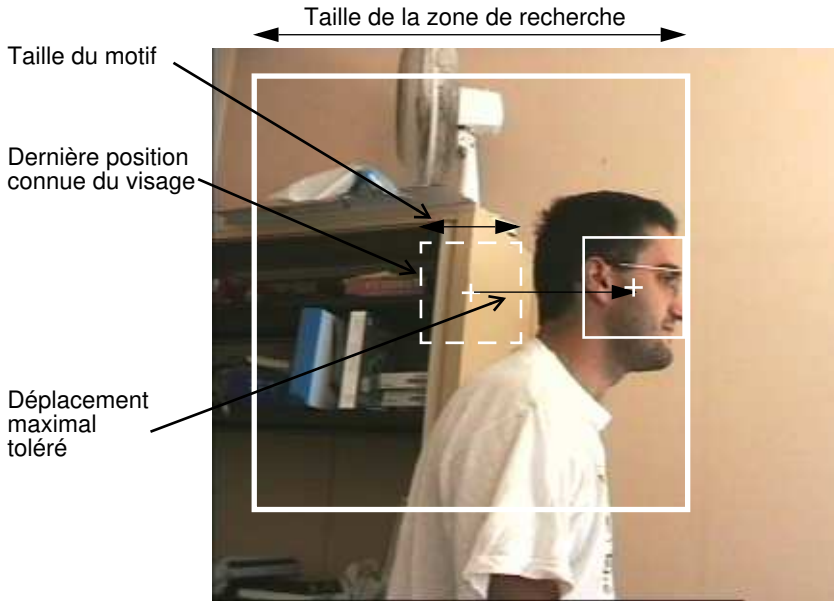


Figure 10
Déplacement maximal de l'entité en fonction de la taille de la zone de recherche
Si le visage s'était déplacé davantage vers la droite, il serait sorti de la zone de recherche, et n'aurait pu être localisé par le suivi.

Soit m la taille du motif et t la taille de la zone de recherche. Alors le déplacement maximal de l'entité entre deux images est $(t-m)/2$. Soit $F(t)$ la fréquence de fonctionnement du suivi en fonction de la taille de la zone de recherche. Nous supposons que F est inversement proportionnelle au nombre de calculs de valeur de similarité. Ce nombre correspond au nombre de positions différentes du motif dans la zone de recherche, soit $(t-m+1)^2$. La fréquence de fonctionnement est donc :

$$F(t) = k \cdot \frac{1}{(t-m+1)^2} \qquad (22)$$

Lorsque $t = m$, la recherche se résume à un seul calcul de similarité et $F(t) = k$. k est la fréquence de fonctionnement du système pour le calcul d'une valeur de similarité. Autrement dit, k est l'inverse du temps nécessaire au système pour calculer une valeur de similarité entre le motif et l'image.

La vitesse de déplacement maximale V_m a la forme suivante :

$$V_m = \frac{t-m}{2} \cdot F(t) = k \cdot \frac{t-m}{2(t-m+1)^2} \qquad (23)$$

L'équation 23 est représentée par la courbe nommée "modèle" sur la figure 11. Nous effectuons une vérification expérimentale du modèle en mesurant les temps de calcul de la recherche du motif pour une taille de zone de recherche variable. Ces mesures sont transformées sous la forme de vitesse maximale de l'entité et sont représentées sur la figure 11 sous le nom "mesures". La superposition des deux courbes témoigne de la validité du modèle.

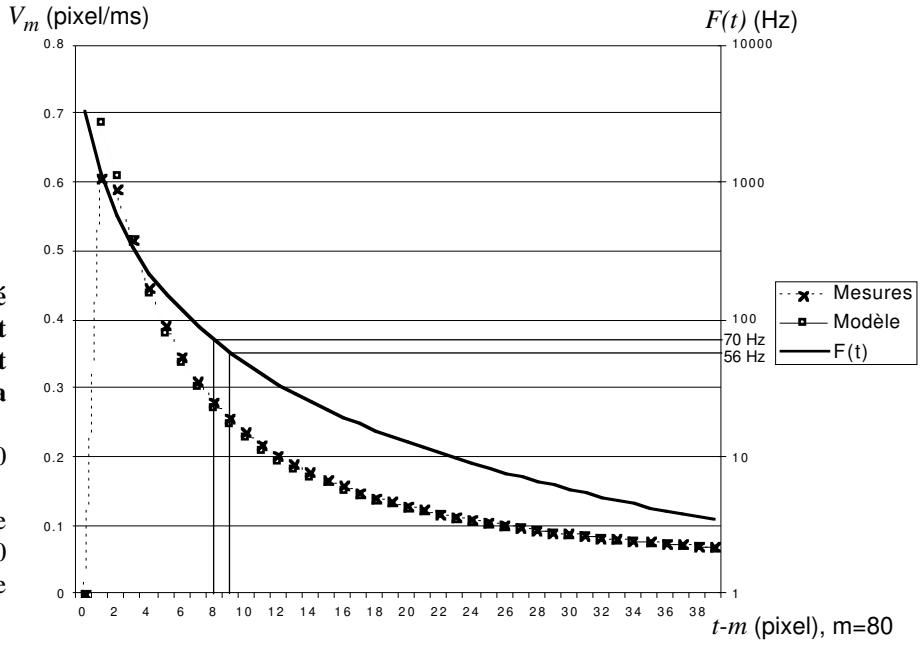


Figure 11
Vitesse maximale de l'entité (modèle et mesure) et fréquence de fonctionnement en fonction de la taille (t) de la zone de recherche
Le motif a une taille m = 80 pixels.
La taille de zone de recherche idéale pour un flux vidéo à 60 images / secondes est comprise en 88 et 89 pixels.

La forme de la courbe indique que la taille optimale de zone de recherche se situe à deux pixels de plus que la taille du motif. Nous interprétons ce résultat comme suit : l'augmentation de la taille de la zone de recherche permet de trouver le motif "plus loin" par rapport à la dernière position connue, mais a pour effet de réduire la fréquence de fonctionnement de façon quadratique. Il est donc préférable de minimiser la taille de la zone de recherche afin de maximiser la fréquence de fonctionnement. En d'autres termes, la recherche devrait s'effectuer dans une région d'un pixel de part et d'autre de la dernière position connue du motif. Toutefois, les contraintes matérielles limitent la fréquence d'échantillonnage du flux vidéo à 60 Hz (une discussion sur la fréquence d'échantillonnage peu être consultée en annexe B page 189). Or, il est inutile d'assurer une fréquence de fonctionnement du suivi supérieure à la fréquence d'échantillonnage : cela reviendrait à localiser l'entité plusieurs fois dans la même image. Le choix de la taille de la zone de recherche doit donc être effectué en fonction de la limite physique que représente la fréquence d'échantillonnage. Dans le cas présenté dans la figure 11, la taille de la zone de recherche optimale est de 88 pixels : elle correspond à la première fréquence de fonctionnement supérieure à 60 Hz.

Notre analyse sur la taille optimale de zone de recherche est reproduite par Vincze ([Vincze 96]) qui l'étend à la prise en compte de la vitesse de déplacement de la cible. Le raisonnement de Vincze est similaire au raisonnement sur l'accélération maximale rapportée au paragraphe "Prédiction" page 81. Nous motivons également, au paragraphe "Prédiction", notre choix de ne pas prendre en compte la vitesse de déplacement dans le cas du suivi des membres (main, tête) de l'utilisateur.

4.3. PERFORMANCES

L'expérimentation rapportée au paragraphe précédent permet d'estimer la valeur de la constante k de l'équation 22 à 5500 pour une taille de motif $m = 80$ pixels sur un processeur PowerPC 750 à 400 MHz. Notre système est donc capable de calculer 5500 valeurs de similarité par seconde pour un motif de taille 80 x 80 pixels. Nous ajustons la taille de la zone de recherche de façon à assurer une fréquence de fonctionnement et une latence qui répondent au requis de l'interaction fortement couplée. Le détail en est donné au chapitre V page 133 pour le suivi du doigt du **tableau magique,** et au chapitre VI page 153 pour l'extraction des déplacements du visage de la **fenêtre perceptuelle**.

À la différence des suivis par différence d'images et modèle de couleur, le suivi par corrélation considère un ensemble de pixels spatialement liés (les pixels du motif), non pas chaque pixels indépendamment les uns des autres. Le bruit de caméra affectant les pixels de façon indépendante, il est fortement improbable que l'oscillation de la valeur des pixels affecte l'ensemble des pixels du motif de façon cohérente. Ceci est d'autant plus

improbable que le nombre de pixels du motif est grand. La conséquence est que si certains pixels du motif sont perturbés par le bruit de caméra, ces perturbations sont compensées par la contribution des autres pixels du motif. Il résulte que l'ensemble des pixels constituants le motif n'est pas affecté par le bruit de caméra. On constate expérimentalement que l'information de position générée par le suivi par corrélation est statiquement stable pour des motifs dont la taille est au moins de l'ordre de 10 x 10 pixels. Lorsque l'entité suivie ne bouge pas dans le champ de la caméra, la position extraite reste strictement constante.

La résolution du suivi par corrélation, exprimée dans le repère image, est de un pixel. Etant restreint au traitement d'une petite partie de l'image (la zone de recherche), le suivi par corrélation peut être mis en œuvre sur un flux vidéo dont les images sont de grande taille sans affecter sa fréquence de fonctionnement. Par contre, la possibilité d'utiliser des images de grande taille améliore d'autant la résolution de l'information extraite dans le repère de la scène.

4.4. DISCUSSION

Les deux caractéristiques de performances du suivi par corrélation, stabilité statique et haute résolution, en font une technique de choix pour les applications nécessitant un suivi fin et stable.

En théorie, cette technique n'est applicable qu'aux translations effectuées dans un plan parallèle à l'image. En pratique, elle est capable de suivre des entités effectuant des rotations de faible amplitude et de faibles variations de distance avec la caméra. On constate que les utilisateurs, qui ne sont pas des entités contraintes, exécutent fréquemment des mouvements qui font échouer le suivi. Il est alors nécessaire de réinitialiser le suivi sur un nouveau motif. Des solutions ont été proposées dans la littérature pour étendre les capacités du suivi par corrélation à d'autres formes de déplacement de la cible. Nous rapportons le résultat de certains de ces travaux, puis nous expliquons pourquoi ces extensions ne sont pas adaptées à notre domaine applicatif.

Black et Yacoob ([Black 97]) modélisent l'apparence du visage par un plan vu en projection perspective. Ce modèle permet de suivre une entité qui effectue des mouvements plus généraux qu'une simple translation, comme par exemple des variations de distance par rapport au plan de la caméra, et certaines rotations d'amplitude limitée en dehors du plan de la caméra. Cette modélisation nécessite l'extraction de huit paramètres décrivant un plan en projection perspective. La complexité du processus d'extraction de ces paramètres est largement supérieure à l'extraction des deux paramètres d'une simple translation. La réalisation de Black et Yacoob nécessite plusieurs minutes de calcul par image, ce qui la rend inutilisable en interaction fortement couplée.

Hager et Belhumeur ([Hager 98]) proposent un algorithme capable d'extraire efficacement les paramètres de plusieurs modèles de déplacement du motif, et notamment le modèle affine qui permet de prendre en compte les rotations et changement d'échelle du motif dans l'image. L'implémentation de cet algorithme est disponible sur internet ([XVision]). Nous l'avons testée et avons constaté que le système peut suivre une entité dont l'apparence subit des rotations et changements d'échelle dans l'image. La fréquence de fonctionnement du système est de 25 Hz pour une taille de motif de 100 x 100 pixels sur un processeur MIPS R10000 à 150 Mhz. Notre expérience avec ce programme fait toutefois apparaître deux limitations :

- Les mouvements de la cible entre deux images doivent être limités : l'extraction efficace des paramètres du modèle est possible grâce à une linéarisation du problème par un développement limité dont la validité est restreinte à de faibles variations des paramètres (de translation, rotation, etc.). Cette limitation étant liée au principe fondamental de l'algorithme, il semble difficile de la contourner.
- L'extraction des paramètres est fragile. L'extraction des paramètres (six paramètres dans le cas du modèle affine) ne fournit qu'une approximation du mouvement en raison de perturbations liées aux phénomènes tels que le bruit de caméra et les formes de mouvement non pris en compte par le modèle. L'approximation faite sur chaque paramètre intervient comme nouvelle perturbation dans l'extraction des autres paramètres. Lorsque le nombre de paramètres augmente, la sensibilité globale du système aux perturbations croît également, ce qui le rend plus fragile. En pratique, le système est capable de suivre l'entité de façon plus robuste si seuls les deux paramètres de translation sont extraits. Ce constat est également l'une des conclusions du travail de Shi et Tomasi ([Shi 94]) qui préconisent, concernant le suivi, de ne considérer que les paramètres de translation par souci de robustesse.

Les applications visées pour notre système de suivi nécessitent essentiellement les paramètres de translation de l'entité suivie. C'est pourquoi nous nous restreignons à l'extraction de ces paramètres dans le flux visuel. Notre système de suivi doit pouvoir supporter des mouvements rapides de l'utilisateur. Pour cette raison, nous préférons effectuer une recherche explicite du motif dans l'image, telle qu'illustrée sur la figure 9 page 96, plutôt que d'utiliser la technique de linéarisation de Hager et Belhumeur.

Les trois techniques de suivi d'objet que nous avons étudiées, suivi par différence d'images, par modèle de couleur et par corrélation, ont en commun une complexité de calcul relativement faible qui autorise un traitement en temps réel du flux vidéo. Par contre, ces techniques sont toutes limitées par :

- l'instabilité de l'information extraite,

- le manque de précision,
- l'application restreinte à certaines formes de mouvements de la cible.

Nous étudions ci-dessous la possibilité de combiner différentes techniques de suivi afin d'aboutir à un suivi de meilleure qualité que celui offert par chaque technique mise en œuvre indépendamment.

5. Coopération de techniques

Jusqu'ici, la coopération de techniques a suscité peu d'intérêt parmi les chercheurs en vision par ordinateur. La majorité de la recherche porte sur le développement de suivis "mono-techniques" et sur leur amélioration. Nous relevons cependant quelques exceptions : les récents développements en coopération de techniques ([Coutaz 96], [Graf 96], [Toyama 96]) démontrent l'efficience de cette approche au regard de la robustesse et de l'autonomie.

La coopération de techniques repose sur une architecture qui explicite les relations que ces techniques entretiennent pour atteindre un objectif donné. Nous présentons ici deux modèles d'architecture pour la coopération de techniques de suivi en vision par ordinateur. En dernière partie, nous décrivons notre suivi de visage par coopération de techniques.

5.1. ARCHITECTURES POUR LA COOPÉRATION DE TECHNIQUES DE SUIVI

Nous avons retenu deux modèles d'architecture représentatifs complémentaires :

- le modèle de Toyama et Hager ([Toyama 96]) qui proposent une architecture en couches liées par la notion de focus d'attention incrémental (FAI),
- le modèle SERP [Crowley 94a] articulé autour d'un processus superviseur.

Tous deux visent à améliorer la robustesse du système de suivi. Toyama et Hager définissent la robustesse d'un système comme sa capacité à maintenir une dégradation *progressive* de ses performances lorsque les conditions expérimentales se dégradent, et sa capacité à retrouver ses performances optimales de façon autonome lorsque les conditions redeviennent satisfaisantes.

Selon cette définition, les suivis "mono-techniques" ne sont en règle générale pas robustes. Par exemple, notre technique de suivi par corrélation est capable de suivre une entité en translation dans le plan de projection de la caméra. Lorsque cette entité effectue une rotation hors de

ce plan, ce mouvement, qui constitue une dégradation des conditions expérimentales, provoque un échec : la dégradation de performance du suivi n'est pas progressive. Elle est totale. De plus, lorsque l'entité effectue de nouveau des translations dans le plan de projection de la caméra, le suivi par corrélation ne peut se ré-activer puisqu'il n'intègre pas de mécanisme de sélection autonome d'un nouveau motif. Le suivi par corrélation, employé seul, n'est donc pas robuste.

Focus d'attention incrémental
[Toyama 96]

Toyama et Hager proposent une architecture, nommée "Focus d'attention incrémental" (FAI)[1], destinée à assurer la dégradation progressive des performances et le retour aux performances optimales. Comme le montre la figure 12, cette architecture est structurée en couches. Le suivi est une *recherche* dans un espace de *configurations*. Une configuration peut être une description de l'état de l'entité suivie (sa position dans l'espace, sa forme, son orientation) ou de son apparence. Ainsi, le concept de configuration permet de généraliser l'utilisation de l'architecture FAI aussi bien aux approches orientées modèle (page 67) qu'aux approches par apparence (page 69).

Une couche de l'architecture a pour rôle de réduire le focus d'attention du système : elle reçoit de la couche inférieure un ensemble de configurations de la cible et fournit à la couche supérieure un ensemble moins large de configurations. La couche de plus bas niveau ne fait aucune hypothèse sur l'état courant de la cible. Au sommet de la pyramide, le

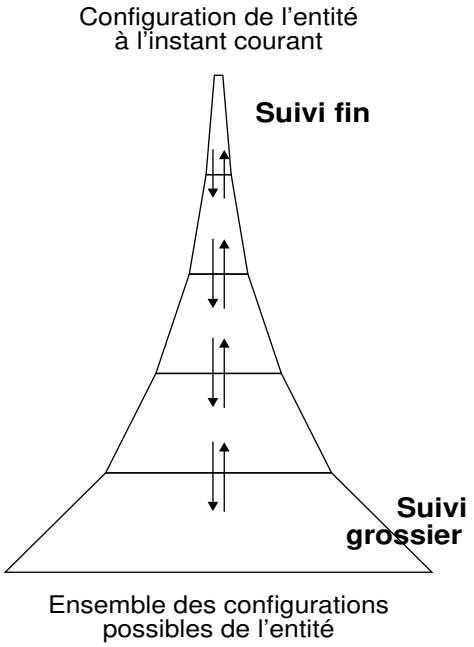


Figure 12
Architecture pour le Focus d'Attention Incrémental (FAI, d'après [Toyama 96])
Chaque couche représente une technique de suivi distincte. Le rôle de chaque couche est de réduire l'ensemble des configurations hypothétiques de la cible avant de le transmettre à la couche de niveau supérieur. Le suivi de plus haut niveau effectue la sélection de la configuration correspondant à l'état courant de la cible.

1. "Incremental Focus of Attention (IFA)"

suivi de plus grande précision est chargé d'isoler l'unique configuration sensée correspondre à la cible.

Chaque couche de la pyramide, autrement dit chaque technique de suivi, doit être capable d'estimer son état, c'est-à-dire d'estimer si la configuration de la cible fait partie de l'ensemble de configurations transmises à la couche supérieure. Si une couche estime que ce n'est pas le cas, c'est que le suivi a échoué et le contrôle est rendu à la couche de niveau inférieur afin d'élargir la recherche.

SERP
[Crowley 94a]

L'architecture SERP (Synchronous Ensemble of Reactive Visual Processes), illustrée sur la figure 13, vient de la robotique appliquée à la surveillance de phénomènes ([Crowley 94b]). Un système de suivi selon SERP est articulé autour d'un superviseur qui coordonne l'exécution d'un ensemble de processus visuels. Plusieurs de ces processus peuvent être actifs en même temps. Le choix des processus à activer dépend de la tâche en cours, des ressources computationnelles et de l'état actuel du système.

Un processus visuel exécute un algorithme de suivi (par corrélation, par différence d'images, etc.). Il réalise une transformation de l'image transmise par le processus d'acquisition en un vecteur d'observation. Ce vecteur inclut la position de la cible. Il est estampillé et décoré d'un coefficient de confiance qui traduit le degré de certitude que le processus visuel a de son observation et d'une matrice de covariance qui traduit la précision estimée par le processus de son observation.

S'appuyant sur les vecteurs d'observation fournis par les processus visuels, le processus de fusion gère la position de la cible selon les étapes "prédiction, validation, mise à jour" : l'information de position fournie en entrée au processus de fusion par un processus visuel et la position prédite sont comparées au moyen de la distance de Mahalanobis. La distance de Mahalanobis est la différence entre les valeurs de position reçue et prédite normalisées par la somme de leur covariance. Si la différence franchit un seuil donné, la position reçue en entrée est rejetée. Dans le cas contraire, elle est combinée à la position prédite comme estimation de la position courante de la cible.

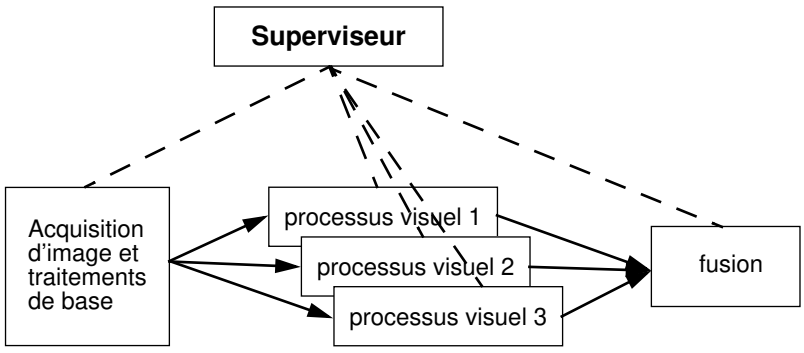


Figure 13
Architecture SERP (d'après [Crowley 94a])
Un superviseur contrôle l'ordonnancement des processus visuels. Les lignes en pointillé expriment les ordres d'activation du superviseur. Les flèches traduisent le flux principal des données

Analyse comparative de FAI et SERP

Alors que le modèle FAI répartit le contrôle d'activation entre les techniques de suivi, SERP centralise ce pouvoir de décision au sein d'un superviseur. Tandis qu'à un instant donné, une technique FAI n'a que deux interlocuteurs possibles (les deux couches adjacentes), le superviseur SERP dispose de toutes les techniques de suivi disponibles. Un contrôle distribué FAI est moins sujet aux pannes qu'un contrôle centralisé SERP. Mais une architecture en couche FAI est moins souple qu'une architecture distribuée SERP. En vérité, ces deux modèles sont complémentaires : un superviseur SERP peut régir l'enchaînement des processus visuels selon les critères de suivi fin / grossier de FAI. C'est ce que nous avons appliqué pour le suivi de visage présenté maintenant. Il est intéressant de constater que nous utilisons le principe de FAI alors que ces travaux ont progressé indépendamment à la même époque.

	Suivi par différence d'images	**Suivi par modèle de couleur**	**Suivi par corrélation**
Initialisation	*Différence entre images successives* : pas d'initialisation. *Différence avec image de référence* : mémorisation de l'image de référence. Ré-initialisation nécessaire lors de mouvement de la caméra et des changements de luminosité globale.	Construction du modèle de couleur par désignation d'un ensemble de pixels exemple. Ré-initialisation nécessaire lors de changement de conditions d'éclairage (lumière naturelle, lampe incandescente ou néon).	Désignation de la partie de l'entité à suivre. Ré-initialisation fréquemment nécessaire si la cible effectue des mouvements en dehors du plan parallèle à la caméra.
Information générée (granularité)	Contour de l'entité (gros grain)	Contour de l'entité (gros grain)	Position fine d'une petite partie de l'entité (grain fin)
Stabilité statique	Instable	Instable	Stable
Fréquence de fonctionnement	52 Hz sur image de taille 384 x 288	15 Hz sur image de taille 384 x 288 et 42 Hz sur image de taille 192 x 144	70 Hz pour un motif de taille 80 x 80 pixels et une zone de recherche de taille 88 x 88. Indépendant de la taille des images.
Contraintes	Un seul objet en mouvement dans la scène. Caméra fixe. Conditions lumineuses contrôlées.	l'entité suivie doit avoir une teinte caractéristique.	Mouvements restreints aux translations dans le plan de projection de la caméra (faibles écarts tolérés).

Table 1 : Résumé des caractéristiques des trois techniques de suivi étudiées.

5.2. NOTRE SUIVI DE VISAGE PAR COOPÉRATION DE TECHNIQUES

[Coutaz 96]
[Bérard 97]

Cette étude concerne le suivi du visage pour la réalisation d'une fenêtre virtuelle (voir le paragraphe "Dispositifs de type "Fenêtre Virtuelle"" page 20). Ce système a été notamment utilisé dans le mediaspace CoMedi ([Coutaz 99]).

La comparaison des qualités et des limitations des trois techniques de suivi présentées dans ce chapitre nous amène à les faire coopérer. L'analyse comparative présentée de manière synthétique dans la table 1 permet d'identifier des stratégies de coopération. Nous en présentons une ci-dessous.

Stratégie de coopération

Principe. On constate que le suivi par corrélation est seul, parmi les trois techniques disponibles, à satisfaire le requis de stabilité statique. Notre stratégie accordera donc une exécution prioritaire au suivi par corrélation. Cette technique nécessite une ré-initialisation du motif dès que le visage effectue un déplacement en dehors du plan parallèle au plan de l'image. Le motif, dans la fenêtre virtuelle, est une partie du visage. Il s'agit donc de localiser le visage afin d'y réinitialiser le motif sans intervention de l'utilisateur.

Le modèle de couleur, fondé sur une propriété discriminante, la couleur de peau, permet de localiser le contour du visage dans l'image. Ayant le visage, il est possible de réinitialiser le motif de la corrélation. Mais le modèle de couleur nécessite à son tour d'identifier un ensemble exemple de pixels de couleur de peau.

La technique de différence d'images permet de détecter les objets en mouvement dans la scène. Une façon d'initialiser le modèle de couleur serait de choisir les pixels correspondant aux entités en mouvement. Cette solution présente deux inconvénients : elle n'assure pas que l'entité en mouvement soit un visage, et si l'entité est effectivement un visage, le mouvement est causé par le déplacement de la tête et non par le seul visage. Les pixels de mouvement désignent aussi bien les cheveux que le visage. Nous proposons de détecter une forme de mouvement caractéristique du visage : le clignement des paupières.

Détection du clignement des paupières. Un clignement de paupière est détectable sur une image de différence entre images successives (voir page 84). Le mouvement rapide et simultané des deux paupières provoque l'apparition de deux zones de mouvement de tailles réduites et similaires, disposées approximativement sur une même horizontale. Le phénomène est illustré sur la figure 14. Cette technique nécessite toutefois que la scène observée soit globalement statique. Si, pendant le clignement, l'utilisateur est en déplacement, ou si d'autres entités de la scène se déplacent, les zones de mouvement dans l'image de différence sont trop nombreuses et le mouvement dû au clignement ne peut être isolé.

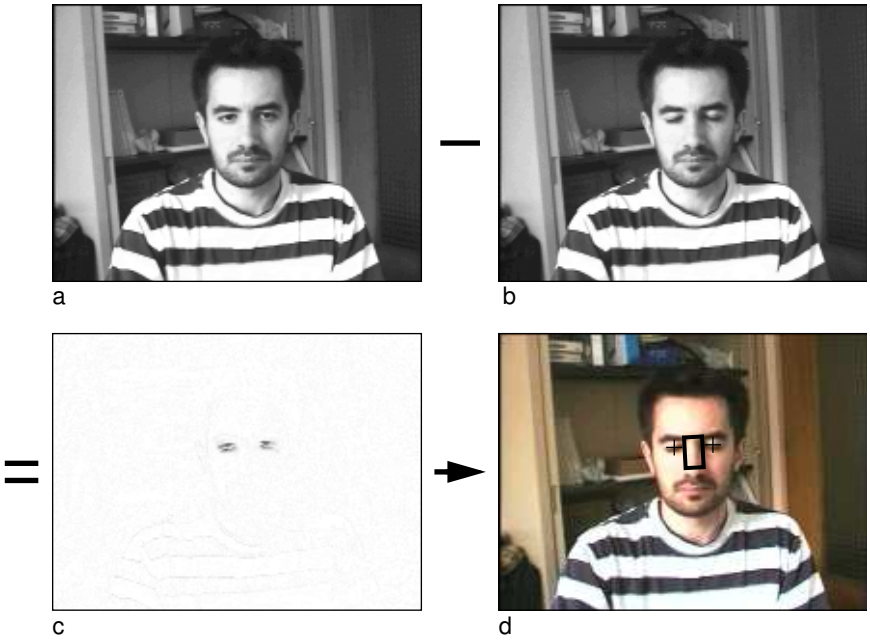


Figure 14
Détection des yeux par différence d'images pour l'initialisation du modèle de couleur de peau
Lorsque l'utilisateur cligne des yeux, la différence entre les images (a) et (b) fait apparaître la position des yeux dans l'image de différence (c). Un rectangle de pixels, situé entre les deux yeux, est sélectionné afin de constituer le modèle de couleur de peau (rectangle noir sur l'image (d)).

Lorsque la technique réussit à identifier un clignement, elle fournit deux informations : elle témoigne de la présence d'un visage dans le champ de la caméra, et elle fournit une estimation de la position des yeux. Cette dernière information sert de référence à l'extraction d'un rectangle, situé entre les deux yeux, utilisé comme exemple pour le calcul du modèle de couleur de peau (voir page 89). La validité de la position du visage extraite est calculée en appliquant la technique de calcul de validité présentée en annexe A page 183. Le vecteur d'observation comprend la hauteur et la largeur des deux boîtes englobantes des paupières, et l'écart vertical et horizontal entre ces deux boîtes.

En résumé, notre stratégie est de tirer avantage des qualités des trois techniques :

- La différence entre images successives ne nécessite aucune initialisation. Utilisée pour la détection du clignement des yeux, elle permet d'identifier et de localiser un visage. Par contre, les clignements des yeux n'étant pas fréquents, et leur fréquence étant variable, cette technique ne peut être utilisée que pour un besoin peu fréquent de la position du visage. Nous l'utilisons pour identifier un ensemble de pixels de couleur de peau afin de construire le modèle de couleur.
- Le suivi par modèle de couleur fournit en permanence la position du visage dans l'image. Par contre, cette information est peu précise et instable. Nous utilisons cette information lorsque le suivi par corrélation n'est pas fonctionnel (lors de mouvements du visage en dehors du plan parallèle au plan de l'image) et pour réinitialiser le motif du suivi par corrélation.

- Le suivi par corrélation fournit l'information la plus précise et la plus stable des trois techniques. C'est la technique utilisée la plupart du temps.

Relation avec les architectures SERP et FAI

Notre approche s'insère naturellement dans une approche FAI si l'on considère que chaque technique de suivi définit un espace de configurations de granularité donnée. La détection de mouvement constitue la couche de plus bas niveau : fonctionnant en toute circonstance, elle prend l'ensemble des configurations possibles comme données d'entrée. Le suivi par modèle de couleur constitue une couche intermédiaire : l'ensemble de configurations produites par cette couche correspond au contour de visage dans l'image. Au sommet de la pyramide, le suivi par corrélation identifie la configuration correspondant à la position précise d'une zone du visage.

Pour assurer le passage du contrôle entre les différentes techniques, nous définissons un *facteur de confiance* (FC). Chaque technique de suivi renvoie un facteur de confiance, c'est-à-dire une valeur entre 0 et 1 représentant la probabilité que l'information de position calculée correspond effectivement à la position du visage. Le facteur de confiance est naturellement associé à la mesure de validité (voir page 80). Le calcul

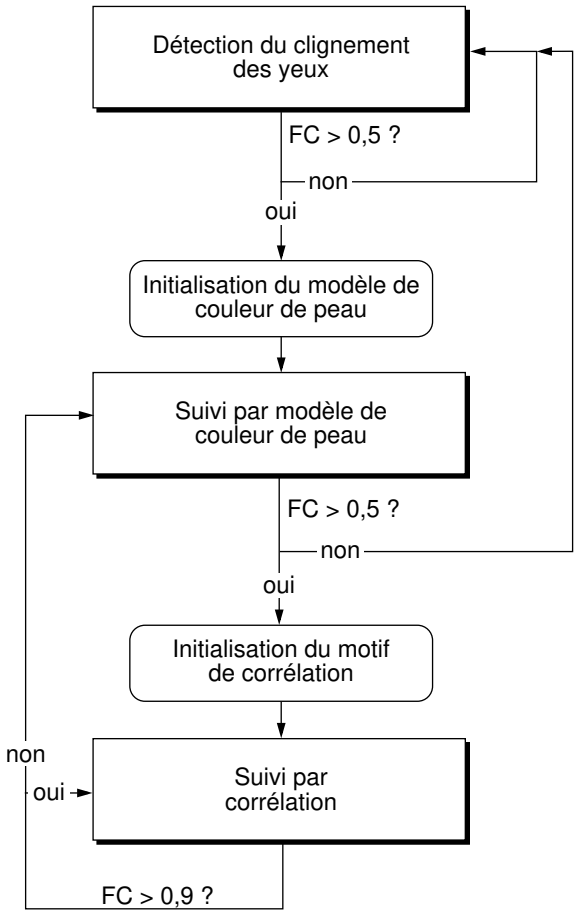
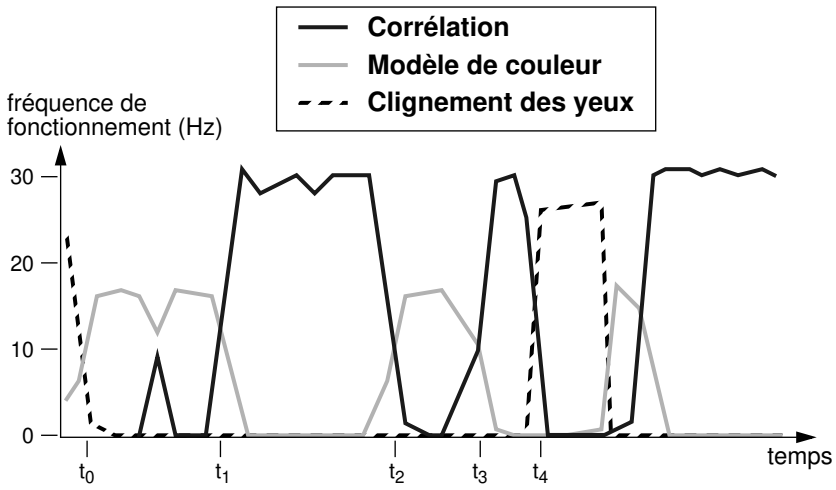


Figure 15
Automate de contrôle de la coopération des techniques de suivi
FC représente le facteur de confiance estimé par chaque technique. Les valeurs seuils sont déterminées empiriquement.

Figure 16
Fréquence de fonctionnement des différentes techniques de suivi coopérantes
La détection du clignement des yeux fonctionne jusqu'à t_0 où un clignement est détecté. Ceci permet d'initialiser le modèle de couleur qui, à son tour initialise le motif de corrélation, mais la corrélation ne devient fonctionnelle qu'à partir de t_1. Entre t_2 et t_3, le visage effectue une rotation provoquant l'échec de la corrélation. En t_4, une lampe au néon est allumée, nécessitant la reconstruction du modèle de couleur.



de la mesure de la validité est présenté à la page 107 pour la détection du clignement des paupières, à la page 93 pour le suivi par modèle de couleur et à la page 97 pour le suivi par corrélation. Le facteur de confiance permet de faire migrer le contrôle entre les différentes couches selon l'automate de la figure 15. Cet automate correspond au fonctionnement du superviseur de SERP.

La figure 16 montre la fréquence de fonctionnement relative de chaque technique de suivi relevée par une campagne de tests du système. Ce schéma illustre le partage du contrôle entre les différentes techniques en fonction des perturbations de la scène.

Discussion

La coopération de différentes techniques de suivi est une voie prometteuse pour la résolution de deux problèmes importants des systèmes de vision par ordinateur : l'autonomie et la robustesse.

Autonomie. L'autonomie est un problème complexe souvent éludée dans la littérature de vision par ordinateur : les théories et techniques sont présentées d'un point de vue général et non pas comme faisant partie d'un système. L'expérimentation nécessite l'intervention d'un opérateur pour l'initialisation.

L'initialisation manuelle est parfois acceptable voire souhaitable, pour certaines applications interactives. Par exemple, le suivi pour la désignation au doigt de notre prototype de **tableau magique** est initialisé par une action explicite de l'utilisateur (voir page 134). Par contre, dans de nombreuses applications interactives, il convient que l'initialisation du système de vision soit "transparente", c'est-à-dire que l'utilisateur n'en ait pas conscience. C'est en général le cas des systèmes immersifs présentés au chapitre I page 11, car l'objectif de ce type de système est de se substituer au monde physique d'une façon transparente. La nécessité d'actions explicites de la part de l'utilisateur a de forte chance de "casser"

la sensation d'immersion. Dans le monde physique, il n'est pas nécessaire d'initialiser les phénomènes pour qu'ils se produisent.

Robustesse. Le problème de robustesse, par contre, est au centre des préoccupations de nombreux travaux. Cependant, ces travaux sont restés majoritairement "mono-technique". Par exemple, les récents résultats de Isard et Blake ([Isard 98]) montrent qu'il est possible de réaliser un suivi d'entité robuste pour un arrière-plan encombré. Isard et Blake réalisent un suivi robuste au moyen de techniques statistiques dynamiques associant une probabilité à chaque configuration possible de l'entité. Toutefois, leur technique implique une complexité de calcul non négligeable, et la technique est robuste vis-à-vis d'un seul type de perturbation : le camouflage. Le camouflage désigne les situations où l'apparence de l'entité suivie est proche de l'apparence de l'arrière plan de la scène. Ce type de scène est particulièrement difficile à traiter car les techniques de suivi risquent à tout moment de confondre la cible et l'arrière-plan. Le camouflage n'est pas la seule cause de perturbation. Le changement brutal de la luminosité en est une autre. L'amélioration de chaque technique individuelle participe évidemment à l'objectif de robustesse. La coopération est une voie prometteuse complémentaire qu'il convient de ne pas négliger.

La coopération offre la possibilité de faire intervenir des techniques peu complexes en calcul, assurant une haute fréquence de fonctionnement du système global. Elle permet d'exécuter la technique la plus adaptée à chaque type de perturbation. Elle permet aussi d'exploiter un facteur de robustesse : la redondance entre les vecteurs d'observation. Dans SERP, c'est au processus de fusion qu'il revient de tirer parti de la redondance.

La qualité de la coopération repose sur la complémentarité des techniques mises en jeu, c'est-à-dire sur l'ensemble de leurs capacités au regard des besoins. Idéalement, la couverture devrait être totale. En pratique, on assiste à des limitations. Par exemple, notre système coopératif n'est pas parfaitement autonome :

- Les clignements de paupières sont peu fréquents, et ne sont pas toujours détectés (notamment lorsque l'utilisateur bouge en même temps). Or, lorsqu'un utilisateur entre dans le champ de la caméra, il s'attend à interagir avec le système dans la seconde qui suit. En pratique, il est préférable d'informer l'utilisateur que le système s'active grâce au clignement des paupières. Cela lui permet de déclencher l'activation du système plus rapidement par des clignements *explicites*. Ce compromis est un pas en arrière au regard de l'autonomie et de la transparence du système.
- Les techniques de différence d'images et de suivi par modèle de couleur nécessitent le réglage de paramètres : la détermination des seuils pour binariser l'image de différence (voir page 84) et l'image de

probabilité (voir page 92). Ces seuils sont déterminés lors d'une phase de calibrage initiale (voir la section "Seuillage" de l'annexe A). L'existence d'une phase de calibrage initial place le système à la merci d'un changement de condition expérimentale postérieur à la phase de calibrage. Les techniques permettant de détecter l'invalidation du calibrage et la nécessité de calibrer de nouveau font encore défaut à notre système.

6. *Résumé du chapitre*

Dans ce chapitre, nous avons décrit notre réflexion et nos travaux sur la réalisation de systèmes de vision par ordinateur dirigée par les requis de l'interaction fortement couplée.

Nous avons présenté le principe général de fonctionnement du suivi d'entité, service fondamental à l'interaction fortement couplée. Nous avons ensuite passé en revue des techniques de suivi fondées sur la vision par apparence que nous avons implémentées et évaluées : suivi par différence d'images, suivi par modèle de couleur, suivi par corrélation. Ces trois techniques partagent une faible complexité de calcul. Prises individuellement, elles constituent de bonnes candidates pour les requis de latence de l'interaction fortement couplée, mais toutes ne couvrent pas le critère de stabilité statique. Ces forces et faiblesses respectives conduisent à considérer leur coopération.

La coopération repose sur une architecture qui identifie les composants de la coopération et définit leurs relations dynamiques. Deux modèles d'architecture nous ont semblé pertinents par leur complémentarité pour mettre en œuvre la coopération de techniques de suivi sous forme d'un système global, autonome et robuste : SERP et FUI. SERP définit un cadre opérationnel général en termes de processus visuels contrôlés par un superviseur unique. FUI définit une politique de passage du contrôle entre les processus visuels par niveau de granularité (du plus large au plus étroit). Nous avons illustré le principe de la coopération de techniques de suivi avec la mise en œuvre d'une fenêtre virtuelle.

Dans la suite de ce mémoire, nous nous attachons à valider l'analyse conduite jusqu'ici. La validation s'appuie sur le développement de deux prototypes pour l'expérimentation de deux formes d'interaction fortement couplée au moyen d'un système de vision par ordinateur : le tableau magique et la fenêtre perceptuelle. Le premier met en jeu un suivi de doigt, le second, un suivi de visage.

Chapitre V *Le tableau magique*

Le **tableau magique,** héritier direct du **Bureau Digital** ([Wellner 91], [Wellner 93b], “Le bureau digital” page 34), désigne notre prototype d’interface digitale sur tableau blanc. Comme le **Bureau Digital**, il s’inscrit dans le mouvement de la réalité augmentée et partage avec lui les caractéristiques suivantes :

- Il s’appuie sur les outils physiques du monde réel non informatique. Dans le cas du **tableau magique**, il s’agit d’un tableau blanc conventionnel, de feutres de couleur à encre effaçable et d’une brosse effaceur.
- Le retour d’information du système est affiché sur la surface de travail, ici le tableau, par vidéo projection.
- La désignation se fait au doigt nu.

Le **tableau magique** a deux raisons d’être :

- Il nous sert de support expérimental à la mise en œuvre et à l’évaluation de techniques de vision par ordinateur. En particulier, nous étudions pour la désignation au doigt, une interaction fortement couplée fondée sur la vision par ordinateur.
- Il permet d’évaluer l’apport de la réalité augmentée pour un type d’activité particulière : la réflexion de groupe ou *brainstorming*.

Nous définissons cette activité dans la première section afin de motiver l’usage d’un tableau blanc et son amplification par des capacités de traitement de l’information. Nous donnons ensuite le détail de la réalisation du système dans la deuxième section. À la troisième section nous rapportons les premières expériences d’utilisation. Ce chapitre s’achève par une conclusion résumant les enseignements de ce prototype, tant du point de vue de l’implémentation que de l’apport en interaction homme-machine.

1. Motivations

Le **tableau magique** se justifie comme support à l'activité de *brainstorming* sur tableau blanc. Cette activité se traduit par la production et l'organisation d'idées par des individus œuvrant à un projet commun. Le protocole social adopté est le plus souvent de type informel : absence de locuteur principal, interruptions fréquentes. Le flux des idées est soutenu par l'ensemble des participants et une idée qui n'est pas exprimée sur-le-champ peut devenir obsolète. Le discours oral s'accompagne d'une production écrite de textes, de schémas, de formules, etc. Ces inscriptions sont réalisées en relation étroite avec le rythme des idées.

Le tableau blanc sur lequel on écrit avec des feutres de couleur à encre effaçable, est un artefact largement répandu dans les bureaux et les écoles. Nous en étudions les caractéristiques qui font de cet objet un outil adapté à l'activité de réflexion. Nous identifions ensuite ses lacunes. Avantages et lacunes définissent les requis pour la conception de notre **tableau magique** : conserver les "bonnes" propriétés du tableau blanc physique et l'amplifier par des services électroniques capables de contourner ses insuffisances intrinsèques. Au fil de l'analyse, nous comparons l'état de l'art en matière de tableaux augmentés au regard des requis retenus.

1.1. ADÉQUATION DU TABLEAU BLANC

Un tableau blanc conventionnel présente trois propriétés favorables à l'activité de réflexion collective : disponibilité immédiate, facilité et rapidité d'utilisation, surface partagée servant de mémoire collective.

Disponibilité immédiate

Une réunion de réflexion peut être planifiée ou s'engager de façon informelle et fortuite. Lorsque l'expression des idées nécessite un support écrit, les participants se dirigent naturellement vers le tableau blanc le plus proche. Disponible en permanence et sans délai de mise en route, le tableau n'interfère pas avec la tâche de réflexion. Il répond aux situations opportunistes de besoin immédiat et fortuit sous réserve, évidemment, que les instruments d'écriture (feutres et effaceurs) soient opérationnels.

Certains tableaux augmentés ne remplissent pas cette condition de disponibilité immédiate. Citons le **LiveBoard** ([Elrod 92]), le **SoftBoard** ([Microfield Graph. 99]), le **SmartBoard** ([Smarttech 99]) et le **Mimio** ([Virtual-Ink 99]). Ces systèmes enregistrent la trajectoire des outils de dessin que manipulent les utilisateurs. Mais les inscriptions produites avant la mise en marche sont perdues. L'utilisation de ces systèmes implique que le logiciel de gestion du tableau soit actif au préalable.

La disponibilité immédiate peut être "simulée" en laissant fonctionner le système en permanence. Cette solution n'est cependant pas suffisante :

- Les dispositifs de sortie (écrans ou projecteurs vidéo) ne doivent pas rester activés en permanence sous peine de détérioration. En pratique,

ils se mettent en veille au bout d'un temps donné d'inactivité nécessitant de facto un délai de réactivation.
- Le maintien d'une station de travail dédiée à la gestion du tableau limite la diffusion du tableau blanc augmenté à un nombre restreint de salles de réunions dédiées.

D'autres systèmes tels que le **ZombieBoard** ([Saund 96]) et le **BrightBoard** ([Stafford-Fraser 96a], [Stafford-Fraser 96b]) autorisent une activation *a posteriori* des services électroniques. Ces systèmes sont fondés sur un tableau blanc conventionnel qui peut être utilisé comme tel : de facto, ils satisfont la propriété de disponibilité immédiate. L'interaction en entrée se fait par capture et analyse de l'image du tableau. La capture du contenu du tableau a lieu *après* y avoir porté des inscriptions. Lorsque les services offerts par le **ZombieBoard** ou le **BrightBoard** sont jugés utiles, ceux-ci peuvent être activés alors que le travail de brainstorming est déjà avancé et que des inscriptions ont déjà été portées sur le tableau.

Considérant que la caractéristique de disponibilité immédiate est essentielle, nous retenons pour le **tableau magique** l'approche "capture a posteriori" d'une surface d'écriture conventionnelle.

Facilité et rapidité d'utilisation

L'utilisation d'un tableau blanc est facile et rapide. Cette caractéristique garantit l'accès à tout membre de l'équipe de réflexion quel que soit son profil de compétences. Les feutres et l'encre ont des propriétés dont les participants peuvent tirer parti. En particulier, ils autorisent un bon contrôle de la forme et de l'épaisseur du trait, l'échange prompt entre feutres de différentes couleurs tenus dans une main, et des corrections rapides en effaçant l'encre au doigt. La rapidité des opérations est, à notre avis, un atout essentiel du tableau blanc : alors qu'ils agissent, les participants peuvent suivre le rythme du discours. La complémentarité entre les modalités vocales et écrites peut alors avoir lieu sans discontinuité.

Concernant le tableau électronique **Tivoli** ([Pedersen 93]), Pedersen et ses collaborateurs notent la difficulté des choix de conception entre deux objectifs contradictoires : maintenir la facilité d'utilisation du tableau conventionnel, et offrir l'accès à un grand nombre de services électroniques. Ils reconnaissent que leur système a notablement dévié vers un accroissement significatif du nombre des services au détriment de la conservation d'un usage intuitif du tableau. La figure 1 illustre la dominance des éléments de contrôle. Pedersen et ses collaborateurs envisagent de corriger cet excès. **Flatland**, un tableau augmenté plus récent ([Mynatt 99a]) développé dans le même laboratoire que **Tivoli**, élimine menus et boutons et opte pour un nombre réduit de services : toutes les opérations s'effectuent au moyen de gestes simples via un stylo.

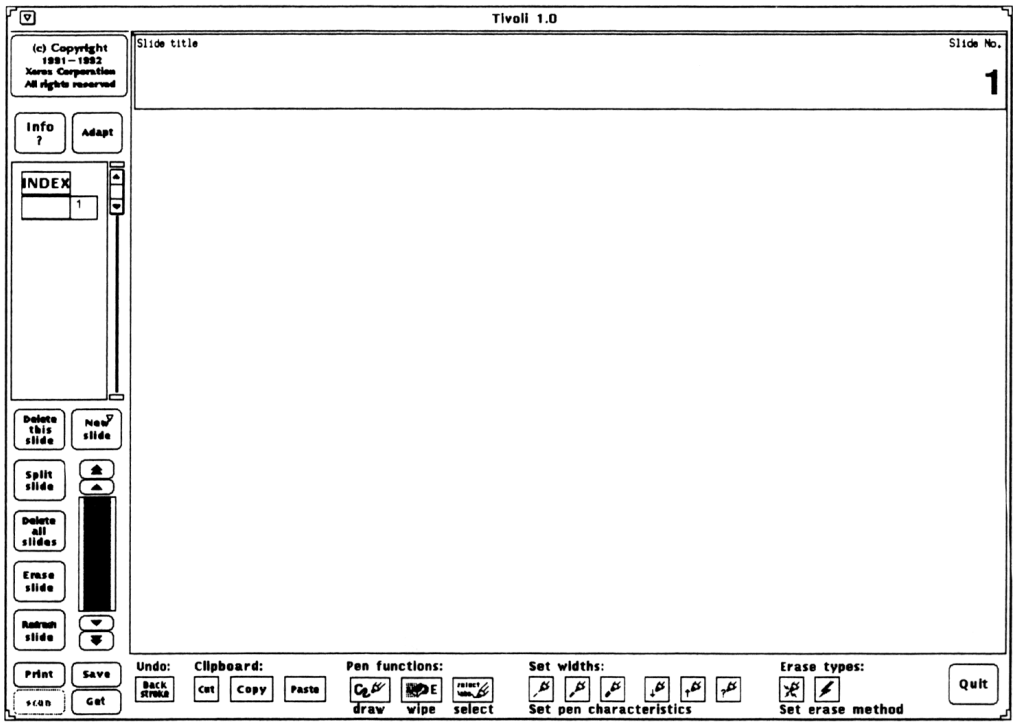


Figure 1
L'interface de Tivoli (extrait de [Pedersen 93])
L'accès à un grand nombre de fonctionnalités se fait au détriment de la simplicité d'utilisation.

Dans une étude ethnographique sur l'utilisation journalière de tableaux blancs conventionnels, Mynatt retient l'importance de conserver la simplicité d'utilisation inhérente au tableau. Ses choix de conception s'orientent vers :

> "*(...) la conservation des affordances des outils existants, même si cette contrainte nécessite de limiter les fonctions ou la complexité de l'outil augmenté.*"[1] ([Mynatt 99b]).

La conception de notre **tableau magique** adhère à ce principe de services minimalistes et de conservation des instruments familiers. Les inscriptions se feront au moyen des feutres conventionnels sur un tableau blanc lui aussi conventionnel.

Surface collective de grande taille

Les tableaux blancs offrent une surface de travail dont la taille est supérieure à celle des supports de dessin usuels : feuilles de dessin, ou grandes feuilles de papier attachées en bloc sur un présentoir. La grande taille des tableaux favorise le dialogue à plusieurs. Pedersen et ses collaborateurs nomment le tableau blanc "un tableau pour la conversation"[2] ([Pedersen 93]). Leurs résultats tirés d'expériences empiriques sur le travail collaboratif médiatisées par des surfaces de dessin partagées, démontrent que, du point de vue de la collaboration, les actions physiques

1. "(...) to retain the natural affordances of the existing tool, even if this constraint requires limiting the features or complexity of the augmented tool."
2. "conversation board"

des participants à proximité de la surface sont aussi importantes que les inscriptions. Ce résultat milite pour une surface de travail assez grande afin que plusieurs personnes puissent évoluer en sa proximité. La surface sert aussi de référentiel commun partageable à tout instant (aux occultations temporaires près, produites par les participants devant le tableau).

Le **LiveBoard**, le **SoftBoard** et le **SmartBoard** ont des tailles fixes et limitées en raison du coût de fabrication. Inversement, le **ZombieBoard** et le **BrightBoard** ont été expérimentés sur des tableaux blancs couvrant un mur entier. La conception de notre **tableau magique** relève de cette seconde approche afin que la surface utile de travail ne soit pas limitée a priori en dessous d'un requis situationnel.

S'il présente des propriétés adaptées aux activités de réflexion, le tableau blanc présente aussi des lacunes. Ces insuffisances motivent l'amplification du tableau par l'ajout de fonctions électroniques. Nous identifions les principales d'entre elles.

1.2. INSUFFISANCES

Les insuffisances du tableau blanc conventionnel sont de nature fonctionnelle : manque de support à la réorganisation spatiale des inscriptions, absence de service d'archivage et de diffusion des résultats de la discussion, collaboration distante synchrone impossible.

Réorganisation spatiale des inscriptions

L'écriture sur tableau blanc étant facile et rapide, elle est adaptée à la concrétisation d'idées. Cependant, les idées sont produites sans ordre. La phase de génération est souvent suivie d'une phase de réorganisation. L'organisation des idées est reflétée par les relations spatiales des inscriptions qui les concrétisent. Or le tableau blanc ne facilite pas cette tâche de réorganisation. En pratique, la réorganisation nécessite la recopie des inscriptions sur une partie vierge du tableau, puis l'effacement des inscriptions originales. Il s'agit d'un processus lourd, sujet à erreur (suite aux recopies) et peu adapté au changement fréquent entre phases de génération et phases de réorganisation. Il est souhaitable d'augmenter le tableau afin de fournir les moyens de réorganiser les inscriptions de façon efficace.

Dans le monde électronique, la réorganisation spatiale des objets est aisée car la copie et l'effacement des données électroniques ne présente pas les problèmes des données physiques. De fait, tous les logiciels offrent différentes méthode de réorganisation spatiale : déplacement (des icônes de fichier sur le bureau, des paragraphes dans un traitement de texte, des objets ou parties d'image dans un logiciel de dessin), iconification, gestion d'un espace de travail supérieur à l'espace physique (concept de bureaux virtuels).

Les **LiveBoard**, **SoftBoard**, **SmartBoard** et **Mimio** numérisent à la volée les inscriptions portées au tableau. La version numérique des inscriptions est immédiatement disponible pour être réorganisée à volonté. Dans le cas du **LiveBoard**, il n'existe pas d'inscription physique : on utilise des crayons optiques qui produisent immédiatement une inscription électronique. Cette approche, qui remplace l'outil usuel (le feutre à encre effaçable), ne satisfait pas le principe de conservation des outils naturels (voir page 116). L'usage de crayon optique limite, par exemple, les possibilités de contrôle de la forme du trait et l'effacement des inscriptions au moyen du doigt.

Dans le cas des **SoftBoard**, **SmartBoard** et **Mimio**, les participants produisent des inscriptions physiques avec des feutres conventionnels dont la trajectoire est enregistrée numériquement (respectivement, de façon optique, tactile ou par ultrasons). Cette approche a l'avantage de conserver l'usage d'un feutre normal pour le dessin, mais nécessite que l'utilisateur efface l'encre physique lorsque ce sont les propriétés électroniques des inscriptions qui prévalent. Notons que l'effacement de l'encre physique n'est nécessaire qu'au premier déplacement. L'inscription existe ensuite uniquement au format électronique et peut ainsi être manipulée.

La réorganisation spatiale nécessite que l'utilisateur puisse désigner des emplacements afin d'indiquer au système les informations à déplacer et leur destination. Dans le contexte du **tableau magique**, la désignation au doigt semble adaptée : en situation d'explication, on utilise l'index pour souligner une information à l'adresse de l'auditoire. Nous ferons en sorte que les participants puissent également se servir de leurs index pour désigner un emplacement au système. Par extension, la désignation au doigt servira également à la désignation des éléments de contrôle permettant d'exécuter les commandes du **tableau magique**.

Archivage et diffusion

Avec un tableau blanc conventionnel, le contenu doit être noté, sous peine d'être perdu, avant que le tableau ne soit effacé (en fin de réunion ou pour changer de thème de réflexion). Cette prise de note est fastidieuse et souvent infidèle, notamment lorsque les inscriptions comportent des schémas. La conservation du contenu du tableau permet à chaque participant de disposer d'un exemplaire et, aux absents intéressés, de prendre connaissance des résultats tangibles de la discussion. Ce contenu peut aussi servir de point de départ à une réunion ultérieure.

Tous les systèmes de tableaux blancs électroniques évoqués précédemment offrent des fonctions d'archivage et de diffusion du contenu. Toutefois, la "qualité visuelle" du service varie en fonction des techniques adoptées. Les systèmes qui numérisent à la volée la trajectoire de l'outil de dessin sont sujets aux limitations de la résolution spatiale et

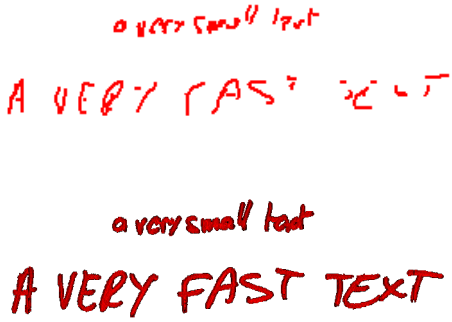


Figure 2
Capture de la trajectoire de l'outil de dessin : problèmes de résolution spatiale et temporelle
L'image du haut est un exemple de capture avec un **SmartBoard**, l'image du bas est un exemple de capture des mêmes inscriptions avec le **tableau magique**.
La première ligne de texte est "a very small text", les caractères sont de taille très réduite. La seconde ligne de texte est "A VERY FAST TEXT" les caractères ont été écrits très rapidement.

temporelle de la numérisation. La figure 2 (en haut) illustre ces deux problèmes :

- La limite de résolution spatiale est le déplacement minimal de l'outil que le tableau est capable de détecter (voir page 55). Cette limite implique que les trajectoires de petites tailles (telles que l'écriture de texte de petite taille) sont mal reproduites.
- La limite de résolution temporelle concerne la fréquence d'échantillonnage des positions de la trajectoire. En raison de cette limite, les dessins exécutés par des mouvements rapides apparaissent hachés, c'est-à-dire constitués d'une succession de segments de droite.

Les systèmes qui numérisent à la volée la trajectoire de l'outil sont également limités par la nature même de l'information capturée : seules la position et la couleur de l'outil de dessin sont enregistrées. L'épaisseur et la forme du trait ne sont pas pris en compte, de même que les modifications apportées au doigt.

Le **ZombieBoard** et le **BrightBoard** capturent l'apparence du tableau a posteriori par l'intermédiaire d'une caméra vidéo. Ils ne sont donc pas sujets aux limitations de résolution temporelle. De plus, la possibilité d'effectuer des plans très resserrés sur le tableau (correspondant à un fort facteur de zoom de la caméra) offre une grande souplesse sur le choix de la résolution spatiale. Enfin, ces systèmes capturent l'apparence des inscriptions, et en particulier l'épaisseur et la forme des traits. En bref, la capture visuelle du contenu du tableau permet d'accéder à une information plus riche que la simple trajectoire des outils et autorise une restitution plus fidèle. La figure 2 illustre la différence de qualité de capture de textes entre les deux approches : capture de la trajectoire ou capture de l'apparence.

Le **tableau magique** doit permettre de numériser le contenu du tableau avec assez de qualité et de précision pour que le contenu puisse être imprimé sur une imprimante de taille standard et distribué à tous les participants. Une fois numérisée, l'image sera stockée, par exemple pour la ré-afficher lors d'une nouvelle séance de brainstorming sur le même

sujet. L'image pourra également être diffusée par voie électronique (par courrier ou sur un site internet).

Collaboration synchrone à distance

L'usage du réseau informatique permet d'envisager le déroulement de réunions entre personnes délocalisées. La collaboration sur des activités de réflexion entre personnes distantes représente un défi car il est difficile de reproduire à distance les indices locaux (gestes, direction du regard des participants) qui permettent aux participants de se synchroniser. Ce problème touche au domaine de la communication médiatisée synchrone et dépasse le cadre de notre domaine d'étude. Cependant, disposant d'une représentation numérique du contenu du tableau et d'un flux vidéo de la scène où les participants évoluent, nous envisageons l'expérimentation de la transmission de ces informations par le réseau.

En résumé, nos motivations pour la conception du **tableau magique** sont les suivantes :

- Nous reconnaissons l'adéquation du tableau blanc conventionnel comme support de l'activité de réflexion. Nous faisons l'hypothèse que cette adéquation tient, pour l'essentiel, aux trois propriétés suivantes citées par ordre de priorité : disponibilité immédiate, facilité d'utilisation, et surface de dessin de taille adaptable à l'activité. Nous prendrons garde à conserver ces caractéristiques.
- Nous identifions trois lacunes du tableau blanc conventionnel classées par ordre d'importance : la difficulté de réorganiser spatialement les inscriptions, l'absence d'archivage et de diffusion du contenu du tableau, et l'impossibilité de conduire des réunions entre personnes situées en des lieux distants. Nous proposons de développer un système destiné à combler ces insuffisances.

2. *Le système*

Dans cette section, nous décrivons les appareils et techniques impliqués dans la réalisation du **tableau magique**. Concernant l'appareillage, le système inclut un projecteur vidéo et une caméra. Chacun d'eux définit un repère image, nécessitant une technique de transformation entre repères. Nous la présentons ici de même que son utilisation dans deux sous-systèmes : le sous-système destiné à capturer une image de bonne qualité des inscriptions du tableau, et le sous système de suivi permettant l'interaction au doigt. Ces deux sous-systèmes sont ensuite détaillés.

2.1. APPAREILLAGE

La figure 3 montre l'appareillage du **tableau magique**. L'équipement de base est un tableau blanc conventionnel sur lequel on écrit avec les feutres usuels à encre effaçable. En entrée du système, une caméra vidéo mobile SONY EVI D31 capture un flux vidéo PAL. L'orientation horizontale et verticale de l'objectif de la caméra[1], de même que son facteur de zoom sont contrôlables par l'intermédiaire d'une interface série RS232. Le flux vidéo est fourni en entrée d'une station de travail Apple Macintosh 8600 équipée d'un processeur PowerPC 604 à 350 Mhz ou d'une station de travail SGI O2 équipée d'un processeur MIPS R10000 à 150 Mhz. Ces deux stations de travail sont équipées d'origine de cartes d'acquisition vidéo. Les retours d'information du système sont assurés par un projecteur vidéo Liesgang DDV 820 à facteur de zoom variable et de définition 800 x 600 pixels.

Grâce à une fonction de calibrage autonome, le **tableau magique** est adaptable à différentes configurations spatiales autorisant des dispositions relatives approximatives entre le projecteur vidéo, la caméra et le tableau (voir plus loin, "Transformation entre repères" page 122). Pour nos expérimentations, la caméra mobile est installée de façon à ce que son objectif soit situé en dessous de l'objectif du projecteur vidéo. Avec cette disposition, la caméra ne peut "voir" le reflet de la lampe du projecteur vidéo sur le tableau. Le projecteur vidéo est installé à une distance d'environ 1,5 m. du tableau. À cette distance, la surface du tableau couverte par l'image projetée est de dimensions 1,2 m. x 0,9 m. La

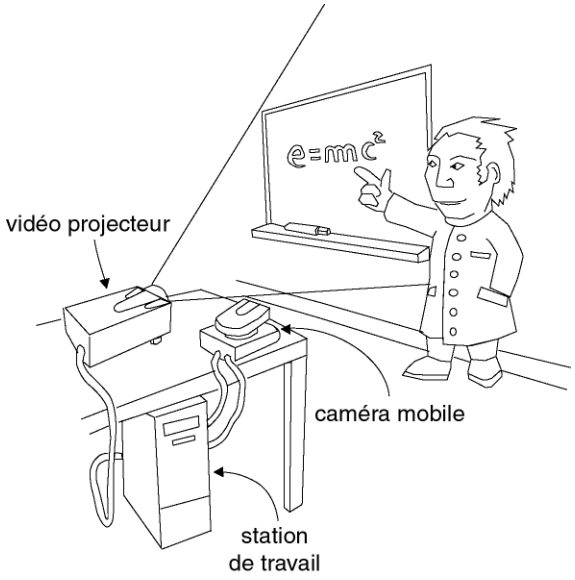


Figure 3
L'appareillage du tableau magique
Une caméra mobile capte un flux vidéo. Ce flux est traité par une station de travail qui interprète les gestes au doigt et qui capte le contenu informationnel du tableau. Le retour d'information du système est assuré par un projecteur vidéo superposant l'information numérique sur les inscriptions du tableau.

1. L'amplitude d'orientation horizontale est de 200 deg. à vitesse variable dont le maximum est 80 deg. / s. L'amplitude verticale est de 50 deg. à vitesse variable dont le maximum est 50 deg. / s.

définition du projecteur vidéo étant 800 x 600 pixels, une telle disposition assure une résolution en sortie de 0,15 cm. Il est possible d'utiliser une surface de projection largement supérieure, par exemple pour couvrir un mur entier, à condition d'utiliser un projecteur assez lumineux et de bonne définition pour assurer une résolution de sortie satisfaisante. L'utilisation d'une caméra possédant un fort facteur de zoom, telle que la SONY EVI D31, autorise la couverture d'un mur entier sans difficulté.

Nous expérimentons également une autre disposition dans laquelle le projecteur vidéo et la caméra sont fixés au plafond sur des supports métalliques. Cette disposition est satisfaisante à condition que les supports soient assez rigides pour absorber les oscillations dues aux mouvements de la caméra. Cette condition est nécessaire à la capture des inscriptions du tableau.

L'usage d'un projecteur vidéo et d'une caméra implique l'existence de deux repères image distincts : le repère de l'image projetée (ou *repère projeté*) et le repère de l'image capturée (ou *repère capturé*). Nous détaillons maintenant une technique permettant de calculer la transformation d'un repère à l'autre.

2.2. TRANSFORMATION ENTRE REPÈRES

Le montage technique du **tableau magique** et l'interaction envisagée avec cet outil nécessitent de transformer les coordonnées exprimées dans le repère capturé en coordonnées exprimées dans le repère projeté, et réciproquement. Par exemple, la position du doigt, qui est extraite des images traitées par la vision par ordinateur, est exprimée dans le repère capturé. En sortie, l'affichage d'un curseur à l'emplacement du doigt, nécessite de connaître la position du doigt dans le repère projeté. Il convient donc de transformer les coordonnées du repère capturé vers le repère projeté. Inversement, il est nécessaire de connaître la position dans le repère capturé d'un élément graphique de contrôle (par exemple, un bouton) dont on connaît les coordonnées dans le repère projeté. Nous en verrons l'application pour la réalisation d'une "zone sensible" page 134.

L'objectif des projecteurs vidéo est en général conçu pour que l'image projetée corresponde à un simple changement d'échelle de l'image générée dans la mémoire de l'ordinateur. Les pixels projetés apparaissent alors sous forme de *carrés* sur le tableau, et le repère projeté est orthonormé. Cette assertion est valide si le projecteur est installé parallèlement au plan de projection. Si le projecteur est incliné par rapport à ce plan (verticalement ou horizontalement), alors l'image projetée est déformée. Il convient donc de tenir compte de cette réalité.

Projection perspective

L'image capturée par la caméra correspond à une projection perspective de la scène. Comme le montrent les figures 4b et 5, un rectangle projeté sur le tableau apparaît sous la forme d'un polygone dans l'image capturée.

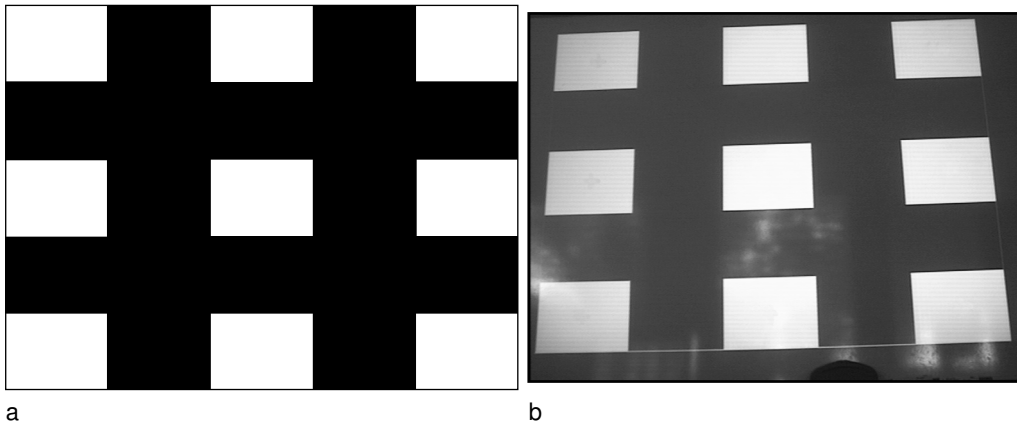


Figure 4
Image de mire projetée (a) et capturée (b).
L'image capturée par la caméra (b) correspond à une projection perspective de l'image projetée (a).

Dans le cadre du **tableau magique**, notre espace d'intérêt dans le monde physique est un tableau, qui est un plan. Nous utilisons le modèle de la projection perspective planaire présentée par Zisserman [Zisserman 97]. Cette projection est représentée par une matrice H de dimension 3 x 3 :

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = H \cdot \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \cdot \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \qquad (1)$$

où x' et y' sont les coordonnées d'un point dans le repère capturé, et x et y les coordonnées du point correspondant dans le repère projeté. Le calcul de la transformation inverse s'effectue grâce à la matrice inverse de H :

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H^{-1} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \qquad (2)$$

La transformation des coordonnées entre les deux repères nécessite donc de connaître les neufs paramètre a_{ij} de la matrice H. Zisserman propose une méthode de calcul de ces paramètres fondée sur la *mise en correspondance* d'un minimum de quatre points. Une mise en correspondance est un couple de points correspondant à la position d'un même objet dans les deux repères. Zisserman décrit une technique permettant d'utiliser un nombre de mises en correspondance supérieur à quatre afin de déterminer la matrice H avec plus de précision.

Mise en correspondance

Nous tirons avantage du projecteur vidéo pour calculer en une fois un ensemble de neuf mises en correspondance. Notre technique est la suivante :

- Nous projetons une mire de neuf rectangles dont nous connaissons les coordonnées dans le repère projeté. Cette mire est représentée sur la figure 4a.

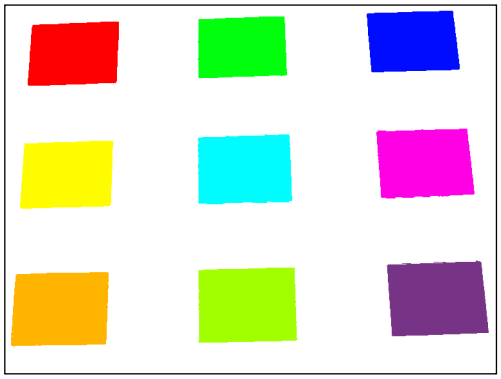


Figure 5
Analyse en composantes connexes de l'image de différence de la mire
Les neuf composantes connexes de plus grande surface sont extraite de l'image de différence seuillée.

- Une image de la mire capturée par la caméra est mémorisée. Un exemple de capture est représenté sur la figure 4b.
- La mire est effacée, une image noire est projetée à sa place. Une nouvelle image est capturée par la caméra et mémorisée.
- Une image de différence est calculée entre les deux images mémorisées (voir page 84). Les rectangles étant présents sur l'une des images et absents de l'autre image, ils apparaissent clairement sur l'image de différence. Cette image est seuillée, puis traitée par une analyse en composantes connexes (ces deux techniques sont détaillées en annexe A). Un exemple d'image résultante est donné sur la figure 5.
- Le barycentre des pixels de chaque composante connexe est calculé. Chaque barycentre, associé au centre du rectangle correspondant dans l'image projetée, constitue une mise en correspondance.

Cette technique permet d'extraire neuf mises en correspondance et de calculer la matrice de projection perspective de la caméra selon la technique de Zisserman ([Zisserman 97]). Grâce à cette matrice (et à son inverse), nous pouvons connaître les coordonnées de tout point de l'image capturée dans l'image projetée, et inversement.

Le processus décrit ici est exécuté durant une phase de calibrage à l'initialisation du **tableau magique**. Nous déterminons ainsi la matrice de projection de la caméra lorsque son point de vue est global (l'image de la caméra contient le tableau entier). Cette matrice de projection reste valide tant que la caméra ne change pas de point de vue (orientation et facteur de zoom) ou lorsqu'elle est ramenée en position initiale après avoir changé de point de vue.

Nous exécutons également ce processus lorsque la caméra est orientée vers une surface réduite du tableau afin de capturer les inscriptions en haute résolution. Cette technique fait l'objet du paragraphe suivant.

2.3. CAPTURE DES INSCRIPTIONS

Lorsque la caméra est orientée de façon à inclure le tableau entier, l'image capturée n'est pas de qualité suffisante pour les services requis de réorganisation, d'archivage et de diffusion. La figure 6 met en évidence deux

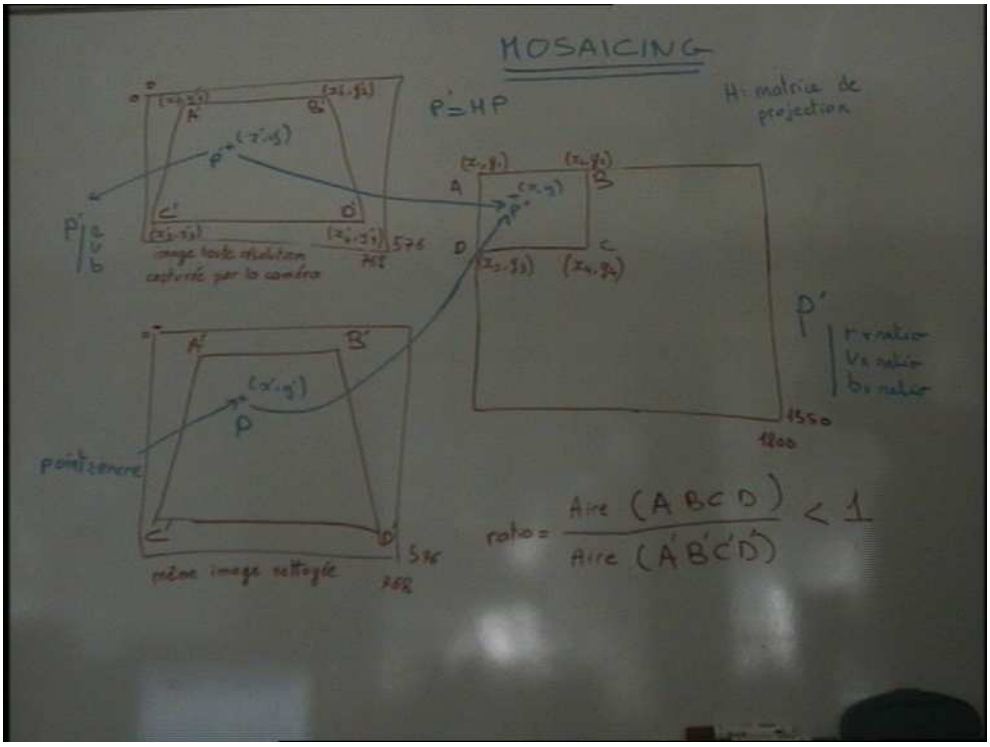


Figure 6
Image "brute" capturée par la caméra
Cette image n'a pas la qualité requise pour être imprimée, archivée ou simplement projetée de nouveau sur le tableau. Elle n'est pas de résolution suffisante (les petits caractères sont illisibles). Les variations de luminosité du fond rendent difficile la distinction entre l'encre et le fond.

problèmes qu'il convient de résoudre : insuffisance du contraste entre les pixels qui représentent l'encre et ceux qui représentent le fond du tableau, et faiblesse de résolution.

Le problème du contraste est traité en différenciant les pixels d'encre et les pixels de fond par une technique de *seuillage adaptatif*. La haute résolution est obtenue en capturant à haute résolution plusieurs images d'une petite partie du tableau, puis en les assemblant par une technique de *mosaïque*. Nous présentons maintenant ces deux techniques.

Seuillage adaptatif

La technique du seuillage adaptatif a pour rôle d'étiqueter chaque pixel de l'image avec l'une des deux valeurs : "encre" ou "fond". A la restitution d'une image étiquetée, les pixels de fond sont supprimés. Dans le cas

Figure 7
Seuillage adaptatif
Les pixels de l'image brute (en haut) dont la luminosité est inférieure à la luminosité du fond (le seuil) sont supprimés de l'image traitée (en bas).
Le seuil est ajusté en tout emplacement du tableau pour prendre en compte les variations d'illumination. On observe ici une surexposition de l'angle en bas à droite de l'image brute.

d'une impression sur papier, la valeur des pixels de fond est mise à "blanc" (le pixel n'est pas imprimé). Dans le cas d'une projection sur le tableau, la valeur des pixels de fond est mise à "noir" (le pixel n'est pas projeté). La figure 7 illustre l'effet du traitement par seuillage.

La technique s'appuie sur le constat suivant : sur un tableau blanc, les pixels du fond, parce qu'ils sont blancs, sont plus clairs que les pixels d'encre. Par conséquent, il suffit de mesurer l'intensité lumineuse des pixels de fond les plus sombres, et de choisir cette valeur comme seuil. Tout pixel dont l'intensité lumineuse est inférieure au seuil est étiqueté "encre", les autres sont étiquetés "fond".

En pratique, du fait des variations d'illumination du tableau, le seuil n'est pas valide pour toute la surface du tableau. Il arrive que les pixels d'encre appartenant à une partie fortement éclairée du tableau aient une intensité lumineuse supérieure à celle des pixels du fond d'une partie sombre du tableau. La figure 7 rend compte de ce phénomène : les pixels à l'extrémité de la lettre "c" sont plus clairs que le fond au voisinage de la lettre "M". En appliquant l'algorithme esquissé ci-dessus, des pixels de la lettre "c" seraient éliminés à tort et des pixels de fond seraient considérés comme correspondant à de l'encre.

Une solution à ce problème consiste à adapter la valeur du seuil aux différentes zones de l'image. Wellner réalise une étude de plusieurs techniques de seuillage adaptatif et propose son propre algorithme dans le contexte applicatif du **Bureau Digital** ([Wellner 93c]). L'algorithme est de faible complexité, autorisant ainsi une implémentation performante. Le seuil est calculé pour chaque pixel de l'image à partir de la moyenne de l'intensité des pixels voisins. Les pixels de fond étant supposés largement majoritaires par rapport aux pixels d'encre, la moyenne reflète l'intensité des pixels de fond. L'algorithme de Wellner est utilisé par Stafford-Fraser dans la réalisation de **BrightBoard** ([Stafford-Fraser 96a]). Nous l'utilisons également pour le **tableau magique**, mais nous la complétons par un étiquetage des couleurs pour les pixels d'encre. Notre implémentation est détaillée dans [Annedouche 99] et esquissée dans [Thevenin 99].

Ayant présenté une solution au problème du contraste des inscriptions capturées, nous détaillons maintenant la technique utilisée pour obtenir une capture en haute résolution.

Mosaïque

La représentation numérique des images capturées par une caméra vidéo est un ensemble de pixels dont la cardinalité, appelée *définition* de l'image (voir page 55), est fixe. Dans notre cas, nous utilisons une caméra qui produit un flux vidéo PAL échantillonné en 768 x 576 pixels. Comme l'illustre la figure 6, cette définition n'est pas suffisante pour assurer une

résolution satisfaisante de capture d'un tableau entier, surtout si le tableau contient des inscriptions de petite taille.

Notre approche s'appuie sur la possibilité de contrôler l'orientation et le facteur de zoom de la caméra. Elle consiste à capturer le tableau en plusieurs images, chaque image ne couvrant qu'une partie du tableau. Les *images élémentaires* sont ensuite assemblées pour constituer une seule *image globale* du tableau entier. Les images élémentaires ont la même définition que l'image représentée sur la figure 6, mais elles couvrent une plus petite surface du tableau. La densité de pixels sur le tableau est donc accrue, ce qui revient à augmenter la résolution de la capture.

Assemblage des images élémentaires. La difficulté de cette technique tient à l'assemblage des images élémentaires. Il s'agit de faire correspondre parfaitement les frontières de ces images sous peine de voir apparaître des "cassures". Le problème de la mosaïque d'images a fait l'objet de nombreux travaux ([Mann 94], [Szeliski 96], [Saund 96]). L'approche fondamentale consiste à calculer les paramètres de la transformation entre le repère de chaque image élémentaire et un repère de référence choisi. Ensuite, chaque image est projetée dans le repère de référence. Les transformations correspondent à des projections perspectives planes détaillées à la page 122.

Par exemple, nous souhaitons assembler deux images élémentaires I_1 et I_2 en choisissant le repère de I_2 pour référence. Nous calculons la matrice de projection perspective H qui projette I_1 dans le repère de I_2. Chaque

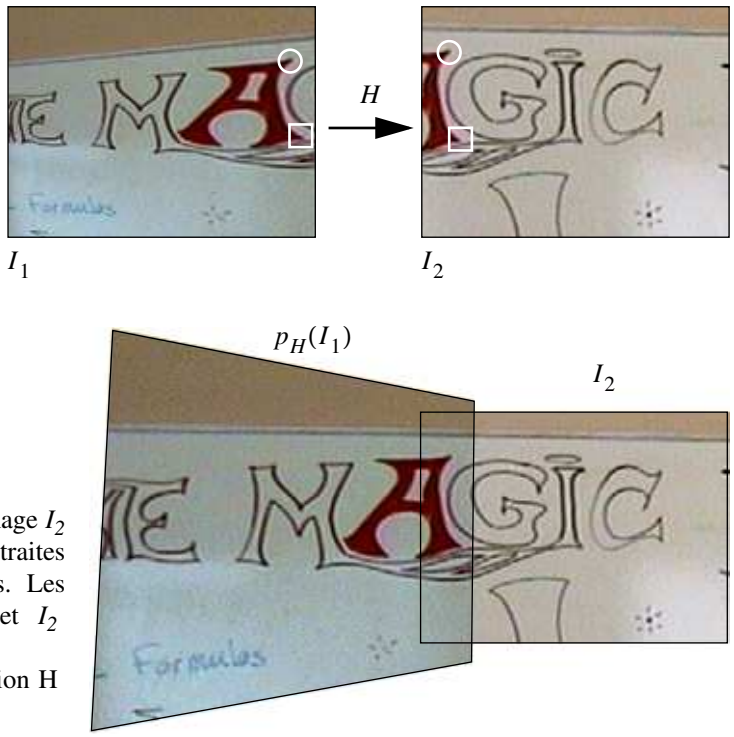


Figure 8
Principe de l'assemblage de la mosaïque
La projection perspective H de l'image I_1 vers l'image I_2 est calculée à partir de mises en correspondance extraites de la zone de chevauchement des deux images. Les cercles et carrés blancs dans les images I_1 et I_2 représentent deux mises en correspondance.
La projection $p_H(I_1)$ de l'image I_1 selon la projection H coïncide avec l'image I_2.

pixel de I_1 est ensuite projeté selon H. L'image $p_H(I_1)$ résultante coïncide avec I_2, comme l'illustre la figure 8.

Le calcul de la matrice de projection H nécessite un minimum de quatre mises en correspondances entre les deux images (voir page 123). L'approche classique ([Szeliski 96], [Saund 96]) consiste à extraire ces mises en correspondance dans la zone de chevauchement des images adjacentes. Par des calculs de similarité (voir page 96), il est possible d'identifier un même indice (par exemple, une petite partie d'une lettre) dans les deux images. Le couple des positions de cet indice dans les deux images constitue une mise en correspondance. Cette approche a ses limites :

- Les images doivent se chevaucher de façon significative afin de permettre l'extraction d'indices communs aux deux images. Ces chevauchements ont pour conséquence d'accroître le nombre d'images élémentaires nécessaires à la couverture du tableau entier.
- L'extraction d'indices dans la zone de chevauchement est un problème complexe, et peut se révéler sans solution dans le cas où la zone de chevauchement ne contient pas d'inscriptions (notamment si le tableau est blanc à cet endroit).
- L'estimation des paramètres de la matrice de projection est peu précise car les mises en correspondance sont concentrées sur une faible surface des deux images.

Notre approche. La présence du projecteur vidéo permet de mettre en œuvre une approche simple, rapide et précise pour le calcul de la matrice de projection de chaque image élémentaire. Plutôt que de calculer les projections entre les repères d'images adjacentes, nous choisissons l'image projetée comme repère de référence, et calculons les projections de chaque image élémentaire par rapport au repère projeté. La technique détaillée au paragraphe "Transformation entre repères" page 122 concerne le calcul de la projection entre une seule image globale du tableau et l'image projetée. Elle est également applicable aux images élémentaires.

Une mire de neuf rectangles est projetée sur la surface couverte par une image élémentaire. L'extraction de la position des neuf rectangles permet de calculer la matrice de projection entre le repère de cette image et le repère projeté. L'image est alors projetée selon la matrice. Cette opération est répétée pour toutes les images élémentaires afin de constituer la mosaïque. Cette approche n'est pas sensible aux problèmes de la mise en correspondance dans les zones de chevauchement :

- Le chevauchement entre les images est minimal : son seul but est de s'assurer que toute la surface du tableau est couverte (il ne doit pas y avoir de "trou" dans la mosaïque).

- Le calcul des mises en correspondance ne nécessite pas d'extraction d'indices sur les inscriptions du tableau : les rectangles de la mire projetée jouent le rôle de ces indices et sont aisément localisés par la technique de différence d'images.
- La mire est projetée sur toute la surface des images élémentaires, ce qui améliore la précision de l'estimation de la matrice de projection.

Cette approche nécessite d'avoir une connaissance approximative de la zone couverte par chaque image élémentaire afin de déterminer à quel endroit du tableau projeter les mires. Cette connaissance est construite à l'initialisation du système pendant la phase de calibrage. Les paramètres d'orientation et de zoom de la caméra sont enregistrés alors que la caméra effectue un balayage systématique de gauche à droite et de haut en bas de la surface du tableau. Le détail de cette technique est présenté dans [Annedouche 99].

Résultat

Comme le montre la figure 9, la technique de seuillage adaptatif associée à la mosaïque permet de capturer une image de qualité du contenu du tableau. Elle offre de plus une grande flexibilité de choix de la limite de résolution. La limite de résolution peut être augmentée jusqu'à la limite du facteur de zoom de la caméra. Nous avons expérimenté la création d'une mosaïque de 36 images (le tableau est alors découpé en 6 x 6 images) aboutissant à une image de 3600 x 2700 pixels pour une surface couverte de 1,2 x 0,9 m. La résolution atteinte est d'environ 0,03 cm. soit environ 30 pixels par centimètre. Un exemple de capture réalisée à cette résolution est accessible sur Internet ([Bérard 99c]). Une mosaïque de 9 images (3 x 3) représente un bon compromis entre temps de capture et

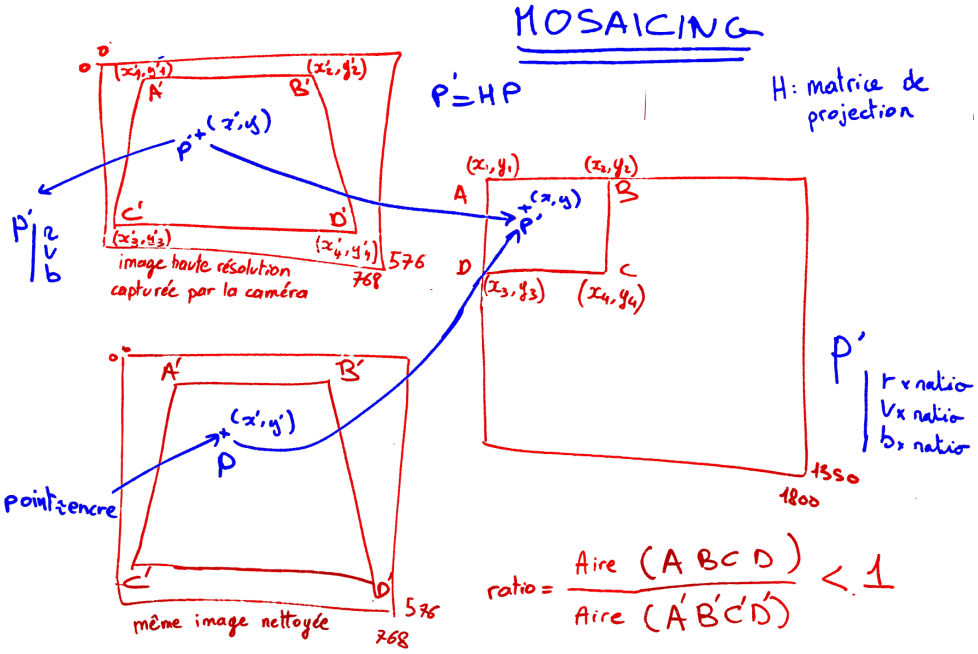


Figure 9
Exemple de capture du tableau magique
Cette image, de taille 1800 x 1350 pixels est le résultat de l'assemblage d'une mosaïque de 9 images. Contraste et résolution son largement améliorés par rapport à la figure 6 page 125.

résolution : la capture d'une image de 1800 x 1350 pixels, telle que celle de la figure 9, est exécutée en 35 secondes sur plate-forme Apple (voir page 121). La résolution est d'environ 0,07 mm., soit 15 pixels par centimètre. Le temps de capture regroupe les phases de déplacement de la caméra, d'extraction des mises en correspondances, de calcul des matrices de projection, et de projection de chaque image dans le repère projeté.

En résumé, le **tableau magique** inclut un service de capture à résolution variable utile à la mise en œuvre des fonctions d'archivage et de diffusion identifiées au paragraphe "Insuffisances".

Ayant présenté les techniques de capture des inscriptions du **tableau magique**, nous mettons maintenant l'accent sur la partie fortement couplée de l'interaction. Nous détaillons les techniques utilisées pour assurer un suivi du doigt, puis nous présentons l'interaction offerte aux utilisateurs.

2.4. SUIVI DU DOIGT

Le service de suivi de doigt a pour but de permettre à l'utilisateur d'indiquer au système des emplacements du tableau. La position du doigt est extraite du flux vidéo lorsque la caméra est dans sa position initiale. La position initiale correspond à une orientation et un facteur de zoom tels que la caméra capture l'ensemble de la surface du tableau. Le suivi du doigt est inactif pendant les phases de capture du contenu du tableau en haute résolution, phases durant lesquelles l'orientation et le facteur de zoom de la caméra sont modifiés. À l'achèvement d'une capture de contenu, la caméra est ramenée à sa position initiale. La position initiale et la matrice de projection correspondante sont mémorisées à l'initialisation du système de façon à conserver la mise en correspondance des repères "capturé" et "projeté" après chaque capture de contenu de tableau.

La matrice de projection est utilisée pour transformer la position du doigt dans le repère de l'image projetée. La connaissance de cette position permet d'afficher un pointeur à la position du doigt. Le rôle du pointeur est double :

- l'apparition du pointeur confirme à l'utilisateur que le système est en train de suivre ses gestes,
- le pointeur, de petite taille, permet de désigner les emplacements du tableau avec finesse. Le doigt nu est trop large pour permettre une désignation fine.

Le pointeur est utilisé pour désigner des inscriptions et la destination du déplacement. Dans les deux cas, la désignation met en jeu une interaction fortement couplée. Le système de suivi doit donc satisfaire les requis définis au chapitre II pour l'interaction fortement couplée. Nous présentons successivement ses performances en terme de latence, résolution et stabilité statique.

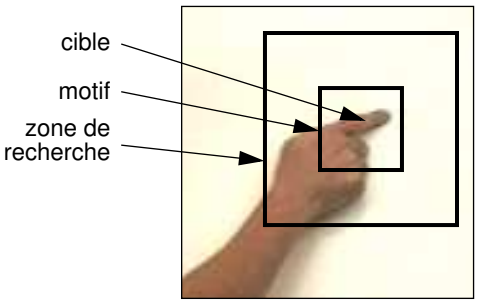


Figure 10
Rapport des tailles de cible, motif, et zone de recherche
La cible du suivi par corrélation (l'index et l'extrémité de la main) est représentée par un motif de taille 32 x 32 pixels. Le motif est corrélé dans une zone de recherche de 75 x 75 pixels.

Notre système de suivi est fondé sur le suivi par corrélation (introduit au chapitre IV page 95) que nous retenons pour ses qualités de rapidité, de précision et de stabilité statique.

Latence

D'après l'expression du requis de latence du chapitre II page 53, le système de suivi doit identifier la position du doigt en moins de 50 ms. après la capture de l'image. Ce seuil suppose que le temps d'affichage du retour d'information est négligeable. Dans notre cas, le retour d'information est soit un curseur, soit une boîte élastique lorsque l'utilisateur effectue une sélection rectangulaire. Ces retours se traduisent par la modification d'un nombre réduit de pixels dans l'image projetée. Ils s'effectuent donc en un temps négligeable.

La taille de la zone de recherche du suivi par corrélation est adaptée de façon à rendre maximale la vitesse tolérée de déplacement du doigt (voir page 98). Pour une taille de motif de 32 x 32 pixels, la zone de recherche est ajustée à 75 x 75 pixels. La figure 10 illustre le rapport de taille entre cible, motif et zone de recherche. La fréquence de fonctionnement du suivi est de 25 Hz, ce qui correspond à une latence de 40 ms. Le requis de latence est donc satisfait.

Résolution du suivi

Nous proposons à la page 55 du chapitre II une estimation de la résolution requise pour les interfaces de type **Bureau Digital**, auquel le **tableau magique** appartient. Cette estimation, de l'ordre de 0,1 cm., est fondée sur l'hypothèse que les tâches à exécuter sur ce type d'interface présentent les mêmes requis en terme de résolution que les tâches exécutées sur les interfaces graphiques classiques.

La capture de la position du doigt dans le **tableau magique** ne permet pas d'atteindre l'objectif attendu. Pour éviter le problème de l'entrelacement du flux vidéo PAL (voir l'annexe B page 189), le flux vidéo traité par le suivi a une définition de 384 x 268 pixels, soit le quart de la définition du flux PAL. Le suivi du doigt est exécuté alors que la caméra est orientée de façon à contenir le tableau entier dans son champ de vue. La caméra est ajustée de façon à couvrir à peu près la même surface que la surface couverte par le projecteur vidéo, soit environ 1,2 x 0,9 m. (il n'est pas

possible de faire correspondre exactement les surfaces projetées et capturées du fait de la déformation perspective). La résolution de capture est d'environ 0,3 cm., soit une résolution trois fois moins fine que la résolution requise. De plus, la résolution de capture se dégrade lorsque le projecteur vidéo et la caméra sont éloignés du tableau de façon à couvrir une plus grande surface de travail.

Nous verrons à la section "Évaluation" que dans notre contexte d'utilisation, la non satisfaction du requis de résolution n'est pas gênante. Par contre, il faudra apporter une solution à ce problème si le **tableau magique** devait être le lieu de tâches de désignation fines comme le dessin assisté par ordinateur. Une solution à ce problème serait, à l'instar de la capture des inscriptions, d'orienter la caméra en permanence sur un plan rapproché du doigt de l'utilisateur. Cette solution présente cependant une difficulté majeure. La caméra étant mobile, il en est de même pour le repère image dans lequel est extraite la position du doigt. Il s'agira de calculer en temps réel la transformation entre le repère capturé et le repère projeté. Or la contrainte de temps réel rend très difficile l'application de notre technique de mise en correspondance de repères.

Stabilité statique

Le suivi par corrélation assure la stabilité statique de l'information extraite (voir page 100) sous réserve que l'entité à suivre ne change pas d'apparence par rapport au motif qui sert de référentiel. Or, dans le contexte d'utilisation du **tableau magique**, l'apparence du doigt en déplacement varie. La raison essentielle est la suivante : lors d'un mouvement non contraint, l'index adopte l'orientation de l'avant-bras qui varie en fonction de la hauteur de l'inscription désignée sur le tableau. Comme l'illustre la figure 11, la variation d'orientation entraîne la modification de l'apparence de l'index.

En pratique, la position du pointeur reste stable tant que le doigt, contrairement à l'exemple de la figure 11, ne subit pas de rotation de grande amplitude. Si l'apparence varie, le pic de corrélation oscille entre les positions les plus semblables par rapport au motif. Cette oscillation, nous

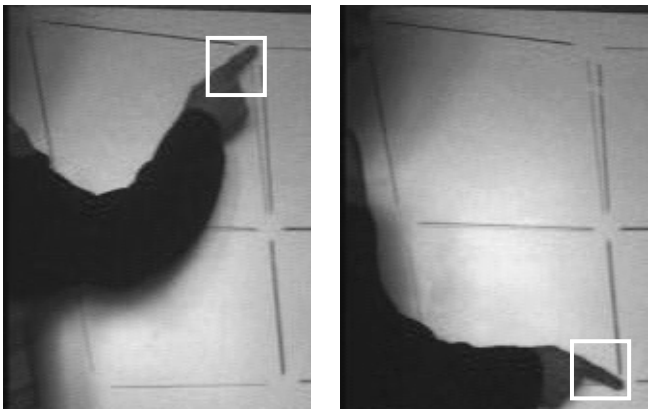


Figure 11
Variation d'apparence de l'index en fonction de l'orientation du bras
Le rectangle blanc sur les deux images symbolise la position idéale du motif du suivi par corrélation. L'apparence du doigt a varié entre les deux images.

le verrons à la section “Évaluation”, est gênante. La gêne est d’autant plus grande que l’interaction s’appuie sur la détection de *pauses* dans la trajectoire du doigt. Or une pause est détectée dans la trajectoire lorsque le curseur est statiquement stable. En l’absence de stabilité statique, la détection de pause ne peut avoir lieu.

Le problème d’instabilité statique peut être limité en appliquant notre approche introduite au paragraphe “Ajout contrôlé de contraintes” page 72. Nous imposons l’installation du tableau blanc à une hauteur telle que l’avant bras reste globalement orienté vers le haut pour la majorité des utilisateurs (à l’exception des plus grands). Cette contrainte limite le champ d’application du **tableau magique** puisqu’elle interdit l’utilisation d’un mur entier comme surface de travail. À défaut d’une technique de suivi plus générale que celle mise en œuvre ici, cette solution permet de réduire très sensiblement le problème de la stabilité statique, sans cependant anéantir l’intérêt du **tableau magique**.

Au-delà des requis généraux imposés à tout dispositif d’interaction fortement couplée, le **tableau magique** doit également satisfaire des requis qui lui sont propres : la vitesse de déplacement maximale tolérée du doigt et la possibilité d’initialiser le suivi de façon simple et rapide.

Vitesse maximale tolérée

Nous savons que le suivi par corrélation échoue si l’objet suivi a une vitesse telle que son déplacement entre deux images est supérieur à la zone de recherche (voir page 98). Nous estimons de manière empirique la vitesse maximale de déplacement du doigt pour les tâches de désignation au tableau.

Il est demandé à plusieurs personnes de désigner “rapidement” une marque à une extrémité du tableau, alors que leur doigt est initialement situé à l’autre extrémité. Le mouvement est enregistré sous forme de films numériques à la fréquence de 25 images par secondes pour des images de définition 384 x 288 pixels. La position du doigt est ensuite mesurée image par image “à la main”, c’est-à-dire de manière non automatisée. Cette expérience nous permet d’estimer que les mouvements les plus rapides sont de l’ordre de 600 pixel/s. Cette vitesse, exprimée dans le repère capturé, est dépendante de la distance de la caméra au tableau. Nous convertissons cette vitesse dans le repère du monde physique en la multipliant par la résolution de capture. Celle-ci est évaluée à $0,328$ cm. en divisant la longueur d’un objet dessiné au tableau par sa longueur en pixels dans l’image capturée. La vitesse des mouvements les plus rapides est donc de l’ordre de $600 \cdot 0,328 \approx 200$ cm/s. .

La vitesse de déplacement maximale tolérée par le suivi par corrélation est donnée par l’équation 23 page 99. Dans notre cas, la taille du motif est 32 x 32 pixels, la taille de la zone de recherche est 75 x 75 pixels, et la

fréquence de fonctionnement est 25 Hz. La vitesse maximale de déplacement du doigt est calculée ainsi :

$$V_m = \frac{t-m}{2} \cdot F(t) = \frac{75-32}{2} \cdot 25 = 537,5 \text{ pixel/s.} \quad (3)$$

La résolution de capture du **tableau magique** a été estimée à 0,3 cm. (voir page 131). Exprimée dans le repère du monde physique, la vitesse maximale tolérée par notre système de suivi est de $537,5 \cdot 0,3 = 161$ cm/s. Cette vitesse est inférieure de 18 % à la vitesse requise pour le suivi des mouvements les plus rapides. Nous verrons à la section "Évaluation" que cette limitation n'est pas réellement gênante car les utilisateurs effectuent rarement des mouvements aussi rapides. Toutefois, il apparaît que le système de suivi pourrait clairement bénéficier de la possibilité de traiter les "champs" du flux vidéo plutôt que les images entières (voir l'annexe B page 189). La fréquence des champs dans un flux PAL est de 50 Hz. Notre système de suivi fonctionne à 50 Hz. pour une taille de zone de recherche de 60 x 60 pixels. À cette fréquence, la vitesse maximale tolérée est :

$$V_m = \frac{60-32}{2} \cdot 50 = 700 \text{ pixel/s.} \quad (4)$$

soit $700 \cdot 0,3 = 210$ cm/s., c'est-à-dire une vitesse de l'ordre de la vitesse maximale estimée expérimentalement.

Initialisation

Le suivi par corrélation nécessite une phase d'initialisation durant laquelle l'apparence de la *cible* est mémorisée dans le *motif* (voir page 95). Nous mettons en œuvre une coopération des techniques de différence d'images et de corrélation inspirée de l'architecture "Focus d'Attention Incrémental" (voir page 103). Le système est constitué de deux couches : celle de bas niveau, fondée sur la technique de différence d'images, a pour rôle de détecter l'apparition d'un doigt et d'en mémoriser l'apparence qu'elle transmet à la couche de niveau supérieur. Cette dernière met en œuvre un suivi par corrélation et assure le suivi rapide et stable du doigt.

Détection du doigt par technique de différence d'images. Un petit rectangle, nommé "zone sensible", est projeté sur le tableau. La zone de l'image capturée correspondant à ce rectangle est calculée grâce à la matrice H^{-1} définie par l'équation 2 page 123. Une différence entre images successives (voir page 84) est appliquée sur cette zone. Nous mesurons l'énergie de l'image de différence résultante, c'est-à-dire la somme des carrés des valeurs de pixels. L'énergie de l'image de différence augmente de façon significative lorsque l'apparence de la zone sensible change au cours du temps. Nous faisons la supposition que seul le doigt de l'utilisateur peut être présenté sur la zone sensible et provoquer l'augmentation d'énergie. Ainsi, lorsque l'énergie dépasse un certain

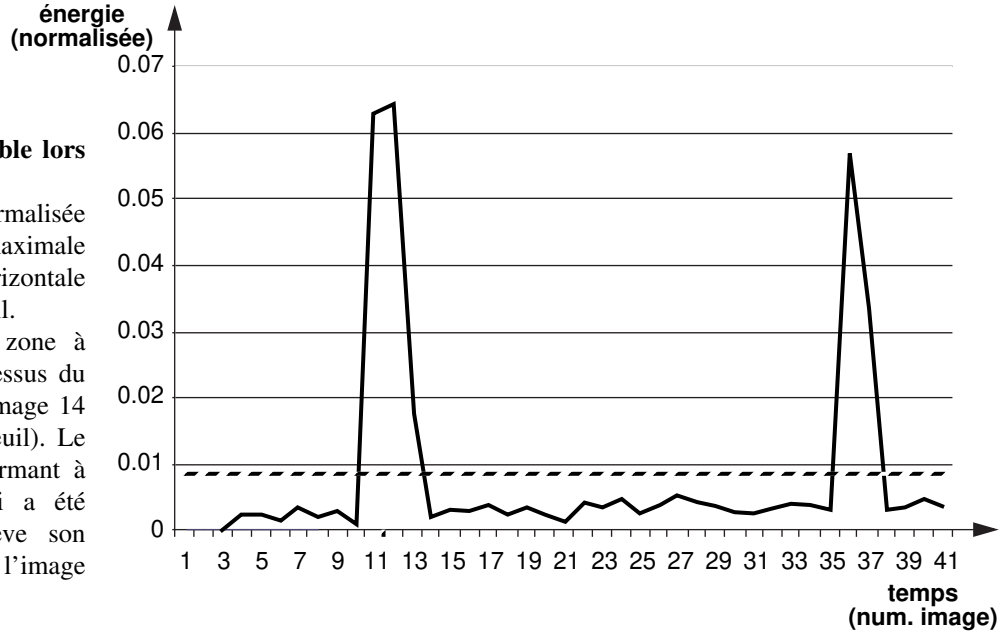


Figure 12
Énergie de la zone sensible lors d'une activation du suivi
La valeur d'énergie est normalisée par l'énergie théorique maximale de la zone. La ligne horizontale hachurée représente le seuil.
Le doigt entre dans la zone à l'image 11 (énergie au-dessus du seuil). Il se stabilise à l'image 14 (énergie en-dessous du seuil). Le curseur est affiché, confirmant à l'utilisateur que le suivi a été activé. L'utilisateur enlève son doigt de la zone à partir de l'image 36 jusqu'à l'image 38.

seuil, nous mémorisons l'apparence de la zone sensible afin de la transmettre au suivi par corrélation en tant que motif à suivre.

Le choix du seuil d'énergie est important : le seuil ne doit pas être trop haut sous peine de ne pas déclencher le suivi du doigt lorsque l'utilisateur le présente sur la zone sensible. Il ne doit pas être trop bas sous peine de déclencher le suivi intempestivement, à cause du bruit de caméra, alors que le doigt n'a pas été présenté. Nous déterminons le seuil durant une phase de calibrage à l'initialisation du système. La technique employée est similaire à celle de Stafford-Fraser détaillée en annexe A page 181. La seule différence est que la moyenne et la covariance sont calculées sur l'énergie de l'image de différence, non pas individuellement sur chaque pixel. La figure 12 détaille l'analyse de l'énergie de l'image de différence comme déclencheur du suivi par corrélation.

Passage du contrôle entre couches. Lorsque l'entrée d'un doigt dans la zone sensible est détecté, nous imposons un délai de 0,5 seconde avant l'affichage du curseur qui indique à l'utilisateur que le suivi du doigt est activé. Ainsi, l'utilisateur est incité à laisser son doigt pendant plus de 0,5 s. dans la zone sensible, avant de le retirer. Ceci nous permet de distinguer les apparitions du doigt et les *fausses alarmes*. Une fausse alarme a lieu par exemple lorsqu'un utilisateur passe devant la zone sensible : à ce moment apparaît un pic dans la courbe d'énergie de la zone sensible. Ce pic est initialement considéré comme l'entrée du doigt dans la zone sensible. Par contre, ce pic est considéré comme une fausse alarme si un deuxième pic apparaît avant une demi-seconde, ou si le deuxième pic n'apparaît pas dans les deux secondes.

Lorsque l'apparition d'un doigt est détectée, le motif de la corrélation est mémorisé et le contrôle est passé à la couche de niveau supérieur (le suivi par corrélation). Cette couche rend le contrôle à la couche inférieure lorsque le suivi par corrélation a échoué. L'échec est détecté par une faible valeur de corrélation (voir à ce propos le paragraphe "Validation et prédiction" page 97).

2.5. INTERACTION

Notre objectif de support à l'activité de réflexion privilégie, en accord avec l'analyse de Mynatt, la simplicité d'utilisation au détriment du nombre de fonctionnalités. En l'état, le **tableau magique** permet de déplacer une partie des inscriptions du tableau, d'en faire des copies, de sauver le contenu entier du tableau dans un fichier ou de l'imprimer. Il est également possible de transmettre l'image du tableau à une station de travail distante. Si celle-ci est connectée à un projecteur vidéo, elle peut reproduire le contenu du tableau local sur un tableau distant.

La fonction de suivi du doigt sert de technique d'interaction pour la réorganisation spatiale d'inscriptions. Une tâche de réorganisation comprend la sélection d'inscriptions et la désignation de leur nouvel emplacement sur le tableau. L'utilisateur doit aussi être en mesure de pratiquer des "coller" et des "couper" sur les inscriptions électroniques. Nous détaillons maintenant les mécanismes de l'interaction permettant de réaliser ces tâches.

Gestion de la sélection

La sélection et le déplacement des inscriptions sont contrôlées par la désignation au doigt. La désignation est activée lorsque l'utilisateur introduit son doigt dans la zone sensible présentée ci-dessus. Cette action explicite a pour effet d'activer le suivi. Le système affiche alors un pointeur asservi aux déplacements du doigt.

Définition de la sélection. La sélection d'une zone rectangulaire du tableau s'effectue de façon similaire à la sélection pratiquée en interface graphique classique : le pointeur est amené à l'un des angles du rectangle de sélection, puis à l'angle opposé. Durant le déplacement du pointeur, un rectangle élastique est projeté entre le premier angle et la position courante du doigt. La définition des angles du rectangle de sélection nécessite l'équivalent du "clic" de la souris.

Détection des pauses. À l'instar du **Bureau Digital**, le **tableau magique** est confronté à l'absence d'équivalent du bouton de la souris (voir à ce propos le paragraphe "Instants d'intérêt" page 35). Nous proposons une solution fondée sur la notion de pause qui a l'avantage de ne pas requérir de dispositif physique supplémentaire. Les *pauses*, assimilées à des clics souris, sont détectées par une analyse spatio-temporelle du flux vidéo. Il y a pause si le pointeur reste stable pendant un temps t_p. La détection de pause présente le risque de ralentir l'interaction si la pause est trop

longue. Or la rapidité d'interaction est l'un de nos objectifs. Nos premières expériences indiquent qu'une faible valeur de t_p de 0,5 s. est convenable. Notre solution présente également le risque de détecter des pauses non intentionnelles (fausses alarmes). En pratique les fausses alarmes sont rares. Nous en verrons l'explication à la section "Évaluation".

Un retour graphique informe l'utilisateur de la détection d'une pause par un changement de forme du pointeur. Initialement, le pointeur a la forme d'une flèche. A la première pause qui définit le premier angle de sélection, le pointeur prend la forme d'une croix. À la seconde pause, alors que le rectangle de sélection est défini, le pointeur reprend la forme d'une flèche. La détection de pause pourrait aussi être rendue plus perceptible par l'émission d'un son.

Déplacement de la sélection. Lorsqu'un rectangle de sélection est défini, son contenu peut être déplacé de la façon suivante : le pointeur est amené à l'intérieur de la surface du rectangle de sélection. Une pause est marquée, le pointeur est changé en croix pour informer l'utilisateur que le déplacement de la sélection est initié. Le rectangle et son contenu sont asservis aux mouvements du doigt. L'utilisateur marque une seconde pause pour "déposer" la sélection à l'endroit désiré.

Le déplacement de la sélection nécessite donc le déplacement des inscriptions contenues dans le rectangle de sélection. Ceci n'est pas un problème pour les inscriptions électroniques. Par contre, si le rectangle de sélection englobe des inscriptions physiques, il est nécessaire de les transformer sous forme électronique avant de pouvoir les déplacer. Cette transformation est l'objet de la technique de capture en haute résolution présentée précédemment. Cependant, cette technique nécessite de nombreuses étapes (orientation de la caméra sur la zone d'intérêt, calcul de la matrice de projection associée, constitution de la mosaïque si la sélection est de grande surface). La durée d'exécution de ces étapes, de l'ordre de la dizaine de secondes, ralentit sensiblement l'interaction et risque de nuire à la qualité du système en tant que support au brainstorming.

Nous adoptons la stratégie suivante : lorsque le déplacement est initié, une capture de la zone couverte par la sélection est effectuée à *basse résolution*. La capture s'effectue sans changer l'orientation de la caméra. Les opérations de déplacement de la caméra, de calcul de la nouvelle matrice de projection et de création de la mosaïque sont supprimées. Le temps de capture est imperceptible et une copie électronique de l'encre physique peut être immédiatement déplacée. À cette résolution, la qualité de capture est suffisante pour permettre la réorganisation spatiale, mais elle est très bruitée et de trop faible résolution pour permettre la lecture de petites inscriptions. À la demande de l'utilisateur, par exemple lorsqu'une

phase de réorganisation spatiale est terminée, le système effectue une capture en haute résolution des inscriptions physiques et met à jour toutes les sélections dont la capture s'est faite à basse résolution.

La demande de mise à jour, de même que la copie ou l'effacement des inscriptions électroniques, s'effectue par l'activation de commandes immédiates.

Commandes immédiates

Les commandes immédiates sont des commandes qui ne nécessitent pas la désignation d'un emplacement du tableau. Elles sont activées par des zones sensibles. Les zones sensibles fonctionnent de la même façon que la zone sensible servant à l'initialisation du suivi. Un ensemble d'icônes est projeté sur l'un des côtés du tableau. Le système exécute une détection de mouvement par différence d'images sur la surface de chaque icône. Lorsque l'utilisateur présente son doigt sur une icône, le mouvement est détecté et la commande associée à l'icône est exécutée. En l'état, le **tableau magique** propose les commandes immédiates suivantes :

- "Sauve" : sauve le contenu actuel du tableau dans un fichier dont le nom est fixé par défaut (si un fichier de ce nom existe déjà, il est écrasé). Une capture du tableau en haute résolution est effectuée avant la sauvegarde.
- "Imprime" : imprime le contenu actuel du tableau. Une capture du tableau en haute résolution est effectuée avant l'impression.
- "Mise à jour des sélections" : effectue une capture en haute résolution des inscriptions déplacées à basse résolution (voir le paragraphe précédent).
- "Copier" : dépose sur le tableau une copie des inscriptions de la sélection courante. Lorsque la sélection sera déplacée, une copie des inscriptions restera en place.
- "Effacer" : efface le contenu (électronique) de la sélection courante.

Le **tableau magique** est maintenant décrit de façon complète. Nous rapportons ci-dessous les premières expériences d'utilisation du prototype.

3. Évaluation

Le **tableau magique** a été évalué de façon informelle au cours d'une dizaine de sessions d'utilisation. Ces premières expériences d'utilisation ont permis de mettre à jour les principales forces et faiblesses du prototype. Nous rapportons en premier lieu nos observations sur

l'interaction avec le prototype d'un point de vue général, puis nous mettons l'accent sur la partie fortement couplée de l'interaction.

3.1. REMARQUES GÉNÉRALES

Certains choix de conception du **tableau magique** présentent le risque de compromettre la qualité de l'interaction. Nous rapportons nos observations concernant l'usage d'un projecteur vidéo, la détection de pause dans les trajectoires, et le nombre et la nature des fonctions offertes.

Usage d'un projecteur vidéo

L'utilisation d'un projecteur vidéo pour assurer le retour d'information du système entraîne un problème de gestion des *ombres*. Les participants, situés entre le tableau et le projecteur, bloquent le flux lumineux en provenance du projecteur. Ils sont ainsi responsables d'une ombre qui occulte une partie des retours d'information.

Les utilisateurs marquent une gêne face à ce phénomène. Ils se voient contraints de se placer en un endroit où la gêne causée par l'ombre est minimale. Mais cet endroit n'est pas nécessairement le plus adapté au regard d'autres contraintes, par exemple, le champ de vision des interlocuteurs. Toutefois, l'ombre n'est pas un problème spécifique à notre système. La fréquence du phénomène dans la vie courante, fait que nous savons gérer les obstructions de lumière par notre propre corps. Il s'ensuit que ce problème est plus facile à gérer dans le contexte du **tableau magique**.

Le problème de l'ombre disparaît dans le cas d'une rétroprojection (le projecteur vidéo est alors installé *derrière* un tableau translucide) ou si le tableau est remplacé par un écran (cas du **LiveBoard** [Elrod 92]). Nous rejetons ces options pour deux raisons :

- La nature du tableau blanc est changée alors que nous souhaitons le conserver tel quel.
- Le retour d'information est occulté par un post-it ou tout objet qui serait fixé sur la surface du tableau.

Nous pensons réduire de façon sensible le problème d'ombre du **tableau magique** en installant le projecteur vidéo au plafond.

Détection de pauses dans la trajectoire du doigt

L'absence d'équivalent du bouton de souris nous a dirigés vers la détection de pause dans la trajectoire du doigt. Cette solution présente a priori deux risques : ralentir le rythme de l'interaction, et augmenter le risques d'erreur due à la détection de pauses involontaires.

Après diverses tentatives, nous choisissons une pause de 0,5 s. Une étude utilisateur permettrait de cerner les effets de la durée de pause avec plus de précision. En pratique, une pause aussi courte ne semble pas ralentir l'interaction. Cette observation est cohérente avec les résultats de Maury et ses collaborateurs ([Maury 99]) qui ont montré que dans certaines

conditions, la sélection d'options de menu par l'attente d'un délai précis est plus performante que la sélection par désignation et clic à la souris.

Par contre, le choix d'une faible durée de pause accroît le risque de fausse alarme. En pratique, les fausses alarmes sont rares sur le **tableau magique**. Le suivi du doigt n'est actif qu'à la demande de l'utilisateur. Ce choix de conception permet de réduire le risque de détecter une pause alors que l'utilisateur n'a pas conscience que le système suit son doigt. Lorsque l'utilisateur active le suivi de doigt, c'est pour effectuer une tâche de sélection ou de déplacement. Il enchaîne déplacement rapide, pause, etc. jusqu'à ce que la tâche soit réalisée. Il sort ensuite sa main du champ de la caméra provoquant ainsi l'échec et l'arrêt du suivi. Durant les phases de déplacement, le doigt bouge assez rapidement pour que le risque de fausse alarme soit faible.

Nous observons cependant des fausses alarmes en fin d'exécution des tâches de sélection. Après avoir exécuté la pause qui définit le deuxième angle du rectangle de sélection, certains utilisateurs inexpérimentés ne semblent pas savoir que faire du pointeur qui reste "accroché" à leur doigt. Leur main reste parfois statique, ce qui provoque une fausse alarme. Si le pointeur est à ce moment précis en dehors de la sélection, la fausse alarme est assimilée à un clic en dehors de la sélection et a pour conséquence d'annuler la sélection (disparition du rectangle). L'utilisateur doit de nouveau définir la sélection. Si le pointeur est à l'intérieur du rectangle de sélection à ce moment là, une opération de déplacement est initiée. Elle peut être stoppée simplement en effectuant une nouvelle pause. Dans les deux cas, les conséquences d'une fausse alarme ne sont pas destructives (comme ce serait le cas d'un effacement d'inscriptions).

Fonctions

Nous avons volontairement réduit le nombre de fonctions du **tableau magique** afin de conserver une grande simplicité et rapidité interactionnelles.

Malgré cela, l'interaction nécessaire au déplacement d'inscriptions semble trop longue. Elle nécessite la présentation du doigt sur la zone sensible, la désignation des deux angles du rectangle de sélection, un "clic" à l'intérieur de la sélection, le déplacement, et enfin un "clic" pour fixer le nouvel emplacement. La réorganisation implique la définition et le déplacement de nombreux groupes d'inscriptions. Il convient donc de réduire la trajectoire d'interaction en supprimant par exemple la phase de définition de la sélection. Si le système était capable de grouper automatiquement les inscriptions, l'utilisateur pourrait déplacer un groupe par un seul geste, sans avoir à sélectionner le groupe. Nous retrouvons cette idée dans **Flatland** [Mynatt 99a].

Dans le même ordre d'idée, il serait souhaitable d'éliminer la présentation du doigt dans la zone sensible. Un suivi permanent de la position du doigt est envisageable à condition de différencier les gestes à destination des participants de ceux destinés au système. Le système pourrait identifier une forme spécifique de la main (telle que la main ouverte) comme marque de geste à son intention.

Enfin, nous avons remarqué que les utilisateurs ont tendance à ne jamais avoir assez de place sur le tableau. Il semble que la gestion d'un *espace virtuel* soit une fonction à ajouter en priorité. Les utilisateurs auraient la possibilité de faire défiler l'ensemble du contenu du tableau dans n'importe quelle direction. Le défilement provoque la disparition de certaines informations par l'un des bords du tableau, et l'apparition d'espace vierge par le bord opposé. L'information disparue pourrait être ramenée sur le tableau par un défilement dans la direction opposée.

3.2. INTERACTION FORTEMENT COUPLÉE

Nous rapportons maintenant nos observations sur le couplage interactionnel dans le **tableau magique**. L'interaction est fortement couplée lorsqu'un utilisateur effectue une tâche de désignation, soit pour définir l'un des angles du rectangle de sélection, soit pour désigner la sélection qui doit être déplacée, soit pour désigner la destination du déplacement.

Latence

Notre système assure une latence inférieure au seuil de 50 ms. estimé au chapitre II. Le délai entre mouvement du doigt et mouvement du pointeur est effectivement imperceptible.

Résolution du suivi

Nous avons estimé la résolution du suivi à 0,3 cm. (voir page 131). A priori, une telle résolution est insuffisante au vu de l'analyse des requis du chapitre II (voir page 55). Sur le **tableau magique**, la résolution d'affichage est de 0,15 cm pour une image projetée de 800 x 600 pixels sur une surface de 1,2 x 0,9 m. La résolution du suivi est donc deux fois moindre que celle de l'affichage. La conséquence est qu'un pixel sur deux, projeté au tableau, ne peut pas être désigné au doigt. Cette limitation serait gênante si les tâches de désignation sur le **tableau magique** nécessitaient une telle finesse. Or ce n'est pas le cas : le doigt n'est pas employé à *créer* des inscriptions (texte, dessins, formes géométriques), mais sert à leur réorganisation spatiale. En règle générale, la résolution de suivi est suffisante pour ce type de tâches. Le problème de résolution n'est sensible que dans le cas où les inscriptions doivent être placées avec précision (tel que sur un alignement).

Instabilité statique

Le suivi du doigt du **tableau magique** est, dans certaines conditions, statiquement instable (voir page 132). L'instabilité statique est un problème extrêmement néfaste pour l'interaction du fait de l'importance

des pauses dans notre système. Lorsque le suivi est instable, il est presque impossible de marquer une pause.

Le problème peut être résolu en élargissant la tolérance de la détection de pauses : une pause est détectée si le pointeur n'est pas sorti, au cours de la dernière demi-seconde, d'un carré virtuel de taille déterminée. Par contre, cette solution augmente la probabilité de fausse alarme (voir page 139) et diminue la résolution effective du pointage. La résolution du pointage est alors déterminée par la taille du carré, et non pas par la résolution du suivi de doigt.

Enfin, l'usage du **tableau magique** confirme que l'oscillation des retours d'information due à l'instabilité statique est visuellement dérangeante.

Vitesse maximale du doigt

L'étude rapportée à la page 133 indique que la vitesse maximale du doigt tolérée par le suivi est inférieure à la vitesse du doigt lors de certains mouvements rapides. Ce problème est peu gênant dans le contexte du **tableau magique**. Il est rare que les mouvements soient assez rapides pour faire échouer le suivi. Nous observons fréquemment des situations où l'utilisateur pense avoir désactivé le suivi (en plaçant sa main hors du champ de vue de la caméra) alors que ce n'est pas le cas. Il effectue alors un mouvement rapide et non contraint pour activer une commande immédiate, et constate que le pointeur est resté "accroché" à son doigt. Ce phénomène illustre le fait que le système est souvent capable de suivre le doigt lors de mouvements non contraints des utilisateurs, et que seuls des mouvements extrêmement rapides peuvent le faire échouer.

Nous concluons maintenant ce chapitre par une revue des enseignements du **tableau magique**.

4. Résumé du chapitre

La conception du **tableau magique** descend en droite ligne du prototype de **Bureau Digital** de Wellner ([Wellner 93b]). Alors que Wellner et ses collaborateurs exposent de façon très convaincante leurs motivations sur le concept de réalité augmentée, dont est issu le **Bureau Digital** ([Wellner 93a]), la technologie dont ils disposent ne leur permet pas d'implémenter un prototype réellement utilisable. Stafford-Fraser ([Stafford-Fraser 96a]) et Saund ([Saund 96]) s'attachent à développer les techniques permettant le passage des inscriptions du monde physique vers le monde électronique, et réalisent deux systèmes utilisés couramment. Mais ces systèmes ne mettent pas en œuvre l'interaction fortement couplée du **Bureau Digital**. À notre connaissance, le **tableau magique** est la première

création d'un prototype offrant une interaction de type **Bureau Digital** et destiné à être utilisé couramment.

Il est encore trop tôt pour juger de l'adoption du **tableau magique** par les utilisateurs. Par contre, nos premières expériences d'utilisation tendent à confirmer à la fois le bien fondé de l'approche de réalité augmentée et le bien fondé des requis non fonctionnels pour l'interaction fortement couplée énoncés au chapitre II.

Apport de l'approche de réalité augmentée. L'approche de la réalité augmentée, qui vise à limiter les modifications apportées aux outils usuels, semble être appropriée à l'informatisation d'un tableau blanc. Il est frappant de constater la faible pénétration des tableaux électroniques dans les usages. Par contre, le tableau blanc classique est quotidiennement employé, malgré sa lacune de services essentiels tels que la sauvegarde des inscriptions. Le coût des dispositifs ne peut justifier à lui seul ce phénomène. Nous l'avons constaté au laboratoire Xerox PARC où certaines salles de réunions sont équipées à la fois d'un **LiveBoard** et d'un **ZombieBoard**. Le **ZombieBoard** est systématiquement préféré au LiveBoard lors de réunions informelles, alors que la gamme de ses services électroniques est bien plus réduite que celle du **LiveBoard**.

L'approche de réalité augmentée se justifie lorsque la technologie est impuissante à reproduire certaines caractéristiques essentielles des outils courants, telles que l'ergonomie, la disponibilité, la taille. Un compromis apparaît entre la difficulté de reproduire les outils courants (approche classique), et la difficulté d'enrichir les outils courants de services électroniques (approche de réalité augmentée). Dans le cas du tableau blanc, utilisé comme support à la réflexion collective, il apparaît que le nombre de services électroniques est moins prioritaire que la conservation des outils usuels.

Requis pour l'interaction fortement couplée. Le **tableau magique** illustre, par l'expérimentation, le bien fondé de notre analyse théorique sur les requis non fonctionnels de l'interaction fortement couplée, énoncés au chapitre II. Dans certaines situations, les requis de latence, stabilité et résolution sont satisfaits par notre système de suivi du doigt. Les opérations de sélection et de déplacement s'effectuent aisément. Cependant, il existe des situations où les requis de résolution et de stabilité statique ne sont pas satisfaits. La résolution est insuffisante pour permettre le positionnement avec précision. La stabilité statique n'est pas assurée lorsque la main de l'utilisateur effectue une rotation significative. Dans ces deux cas, la non-satisfaction des requis entraîne une forte dégradation de l'interaction. Il devient difficile, voire impossible, de réaliser la tâche souhaitée. Il convient d'envisager de nouvelles solutions pour améliorer les performances du suivi de doigt sur ces deux aspects.

L'interface du prototype actuel de **tableau magique**, fortement influencée par le modèle des interfaces graphiques classiques, est une première ébauche. Les techniques développées pour le **tableau magique** réalisent un ensemble de services de base qui devraient permettre le développement d'une nouvelle interface mieux adaptée à la spécificité du dispositif. Cette interface devrait permettre de tirer tout le potentiel du **tableau magique**.

Au chapitre suivant, nous restons dans le cadre d'une interface graphique classique dans laquelle nous introduisons une interaction fortement couplée fondée sur la vision par ordinateur. Nous pouvons ainsi évaluer quantitativement le bénéfice de notre approche par comparaison avec une interface classique.

Chapitre VI La fenêtre perceptuelle

La **fenêtre perceptuelle** ([Bérard 99a], [Black 98a]) est un prototype qui met en œuvre un suivi des mouvements du visage au moyen de la vision par ordinateur. Les mouvements du visage sont capturés de façon non intrusive pour contrôler une interface graphique standard. Notre système est "perceptuel" puisqu'il est capable de réagir aux actions de l'utilisateur sans nécessité de contact physique. La **fenêtre perceptuelle** utilise les mouvements du visage pour les tâches de navigation au sein des fenêtres graphiques. La souris, ainsi libérée de la tâche de navigation, peut être utilisée à d'autres fins.

L'utilisation du suivi du visage en vision par ordinateur comme modalité d'interaction n'est pas une idée nouvelle. Cependant, la plupart des publications mettent l'accent sur les solutions techniques et négligent d'en mesurer le réel bénéfice du point de vue de l'interaction homme-machine ([Azarbayejani 93], [Oliver 97], [Yang 98a], [Toyama 98]). Le prototype de **fenêtre virtuelle** présenté au chapitre I (page 20), qui repose sur un suivi du visage au moyen de la vision par ordinateur, ne répond pas aux requis de latence et de stabilité nécessaires à l'interaction fortement couplée. L'objectif de notre prototype de fenêtre virtuelle est double :

- démontrer la faisabilité d'une interaction fortement couplée fondée sur la vision par ordinateur et satisfaisant les requis énoncés au chapitre II,
- démontrer que cette modalité d'interaction apporte un réel bénéfice du point de vue de l'interaction homme-machine.

La première section de ce chapitre s'attache à motiver l'usage des mouvements du visage comme flux d'entrée spatial en complémentarité avec le flux de la souris. Ce flux additionnel est employé au contrôle de la navigation au sein de fenêtres graphiques. Nous détaillons, en section 2, la réalisation et le fonctionnement du prototype de **fenêtre perceptuelle**.

Deux expérimentations valident notre technique. Nous détaillons ces expériences et en rapportons les résultats en section 3 et 4 respectivement.

1. Motivations

Un flux d'entrée spatial provient d'un dispositif d'entrée capable d'indiquer en permanence sa position. La souris et les capteurs magnétiques fournissent un flux spatial alors que ce n'est pas le cas pour le clavier et le bouton de la souris. Comme nous allons le voir, plusieurs expérimentations montrent qu'un flux d'entrée spatial en complément du flux de la souris permet à l'utilisateur de paralléliser ses actions et de minimiser ses mouvements. Mais ces avantages sont conditionnés par une conception de l'IHM qui respecte les capacités cognitives humaine. Nous rapportons ci-dessous ces travaux qui justifient la conception de notre **fenêtre perceptuelle**. Nous présentons ensuite un avantage spécifique au suivi du visage : la suppression d'intermédiaires dans l'interaction.

1.1. PARALLÉLISATION DES ACTIONS ET MINIMISATION DES MOUVEMENTS

Buxton et Myers ([Buxton 86]) montrent de manière expérimentale l'amélioration des performances lorsque l'interaction se fait à deux mains au lieu d'une seule. Leur scénario expérimental comprend une tâche de désignation et une seconde tâche composée de navigation et de désignation.

- Tâche de désignation : les sujets doivent ajuster la taille et la position d'un carré de façon à le superposer exactement à un carré "cible" (se reporter à la figure 1, en haut).
- Tâche composée de navigation et de désignation : un numéro de ligne d'un texte long et un mot sont présentés au sujet. Les sujets doivent faire défiler le texte pour faire apparaître dans la fenêtre le numéro de ligne où se trouve le mot demandé, puis doivent sélectionner le mot. Cette tâche est illustrée par la figure 1, en bas.

Dans la première expérience, tous les sujets mettent en œuvre une interaction à deux mains pour réaliser la tâche : l'une des mains contrôle la position du carré alors que l'autre contrôle la taille. Le résultat de l'expérience montre que les sujets adoptent naturellement une stratégie de parallélisation des opérations : les déplacements et les changements de taille sont effectués simultanément pendant 40,9 % du temps d'exécution de la tâche.

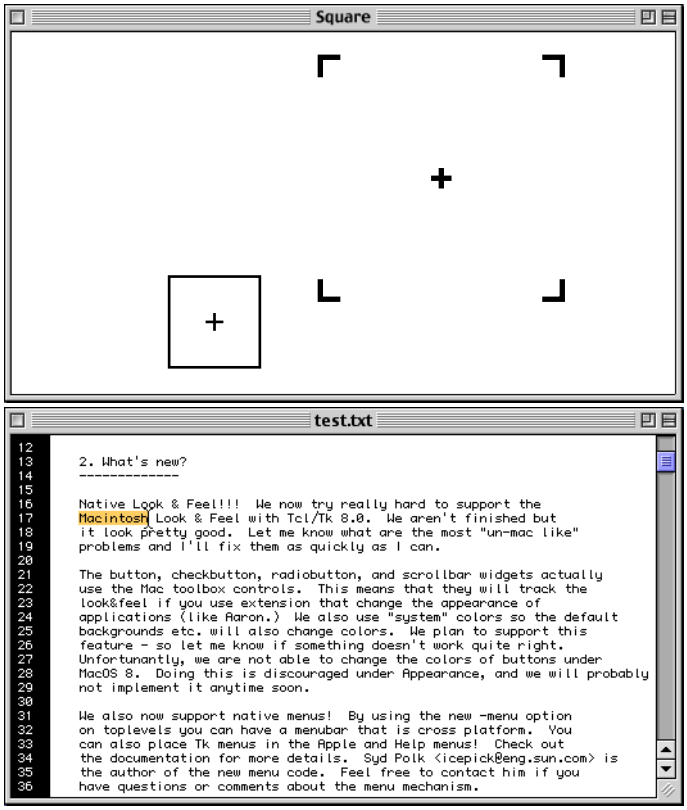


Figure 1
Tâches des expériences de Buxton et Myers ([Buxton 86])
En haut : le petit carré est à superposer sur le carré cible (dont seul le centre et les angles sont affichés).
En bas : un numéro de ligne et un mot sont présentés au sujet. Il doit faire défiler la fenêtre pour faire apparaître la ligne, puis sélectionner le mot.

Dans la deuxième expérience, les sujets sont répartis selon deux conditions expérimentales. Dans la première, les sujets utilisent uniquement la souris. Les différentes opérations nécessaires (navigation avec la barre de défilement et désignation avec le pointeur de la souris) sont multiplexées dans le temps (voir page 26). Dans la deuxième condition, les sujets répartissent les opérations sur chaque main, multiplexant ainsi les opérations dans l'espace.

Le résultat de cette expérience montre que les sujets qui mettent en œuvre une interaction à deux mains sont plus efficaces que ceux qui n'utilisent qu'une seule main. L'amélioration des performance est de 15 % pour des sujets considérés comme "experts", et de 25 % pour des sujets "novices". De plus, l'interaction à deux mains semble plus facile à apprendre que l'interaction à une main multiplexée temporellement. Pour la condition à une seule main, les experts ont en moyenne des performances de 85 % supérieures à celles des novices. Par contre, les différences de performances ne sont pas significatives entre experts interagissant à une seule main et novices interagissant à deux mains. Les auteurs concluent que l'interaction à deux mains est une interaction facile à apprendre et qu'elle permet d'améliorer les performances utilisateur dans la réalisation de tâches composées.

D'après Buxton et Myers, deux phénomènes expliquent l'amélioration des performances : la parallélisation des opérations et la minimisation des déplacements. En condition multiplexée temporellement, le pointeur de la souris effectue des déplacements inutiles du point de vue de la tâche : à chaque changement d'opération (désignation ou navigation), le pointeur doit se déplacer entre l'espace de la barre de défilement et l'espace du contenu de la fenêtre. Ce déplacement est d'autant plus coûteux en temps que le mot à désigner est à gauche de la fenêtre (la barre de défilement est à l'extrême droite, comme illustré sur la figure 1). En interaction à deux mains, les déplacements d'une main sont employés entièrement à la désignation alors que ceux de l'autre main servent uniquement à la navigation. Ainsi, il n'y a pas de déplacements inutiles.

En pratique, l'amélioration due à la parallélisation des opérations et à la minimisation des déplacements peut devenir négligeable face à la charge cognitive induite par l'introduction d'un deuxième flux d'entrée. C'est pourquoi, il est essentiel de concevoir une interaction en adéquation avec les capacités cognitives de l'utilisateur.

1.2. ASPECT COGNITIF

Une étude plus récente ([Kabbash 94]) montre que deux flux d'entrée peuvent être moins performants qu'un seul flux si leur utilisation est mal conçue. Cette étude compare les performances de sujets employant quatre formes d'interaction pour effectuer une même tâche.

La tâche est une coloration d'objet comportant deux types d'opérations : le choix de la couleur et l'application de la couleur à des objets. Une des formes d'interaction pour effectuer cette tâche est appelée le *statu quo* : les fonctions de choix de couleur et d'application de couleur aux objets sont multiplexées temporellement à la souris. Les trois autres formes d'interaction sont des interactions à deux mains. Les résultats de l'expérience montrent que les interactions à deux mains sont en général plus performantes que l'interaction à une main. En cela, cette étude confirme les résultats de Buxton et Myers présentés précédemment ([Buxton 86]).

Toutefois, l'une des interactions à deux mains se révèle être *moins* performante que l'interaction à une seule main bien qu'elle implique *beaucoup moins de déplacements* que le statu quo. Cette interaction met en œuvre deux pointeurs de souris dont les rôles sont rigoureusement symétriques : chaque pointeur est contrôlé par une souris différente et peut être utilisé aussi bien pour le choix de la couleur que pour l'application de la couleur aux objets. Les auteurs argumentent en faveur d'une parallélisation des actions qui associe une action simple, de granularité élevée, à une action plus complexe, de granularité plus fine. Cette affirmation est appuyée par la théorie de Guiard ([Guiard 87]).

Guiard propose un modèle qui caractérise l'usage complémentaire des deux mains dans des tâches dont les opérations sont complémentaires. Il décrit les rôles respectifs de la "main dominante" (la main droite pour un droitier) et de la "main non dominante" (la main gauche pour un droitier). Les relations entre les deux mains sont caractérisées selon trois points :

1. La main non dominante définit le référentiel de la main dominante. Par exemple, pour enfoncer un clou, la main non dominante tient le clou pendant que la main dominante le frappe.
2. L'ordre des mouvements consiste à utiliser la main non dominante en premier, la main dominante ensuite. Par exemple, la main non dominante positionne et maintient une feuille de papier avant que la main dominante ne commence à écrire.
3. La granularité des mouvements de la main non dominante est supérieure à celle de la main dominante. Par exemple, un peintre utilise sa main non dominante pour amener la palette de peinture à portée de main, ou pour l'éloigner, alors que sa main dominante est employée à exécuter les mouvements précis du dessin.

Zhai ([Zhai 97]) réalise une expérience empirique destinée à étudier les performances utilisateur pour une tâche comprenant la navigation dans une fenêtre de texte et la sélection d'hyperliens. Il met en avant que cette combinaison de tâches est un excellent candidat pour le modèle de Guiard : les tâches de navigation sont allouées à la main non dominante alors que la sélection des hyperliens s'effectue avec la main dominante.

Zhai montre que les dispositifs les plus efficaces pour cette tâche mettent en œuvre la combinaison d'un *joystick*[1] dédié à la navigation et d'une souris pour la désignation. Cette combinaison est réalisée de deux façons. Dans le premier cas, les deux mains de l'utilisateur sont effectivement employées à la tâche, alors que dans le deuxième cas le joystick est installé sur la souris. Dans ce dernier cas, la main dominante contrôle la désignation à la souris et un des doigts de la main dominante joue le rôle de la main non dominante, contrôlant la navigation avec le joystick.

Zhai remarque toutefois que la combinaison joystick - souris contrôlée par la même main a tendance à surcharger les fonctions associées à la main dominante. Il rapporte que les utilisateurs ont des difficultés à utiliser ce dispositif pour des tâches plus complexes qu'une simple désignation, par exemple une tâche impliquant un "glisser-déposer"[2] tout en naviguant dans une fenêtre graphique.

1. Manche à balai. Zhai utilise un TrackPoint ([Zhai 97]), c'est-à-dire un manche à balai de taille réduite manipulable avec un doigt.
2. Un glisser-déposer est une opération durant laquelle l'utilisateur déplace un objet graphique en cliquant sur l'objet et en maintenant le bouton de la souris enfoncé pendant le déplacement du pointeur.

Les travaux que nous venons de décrire ont grandement influencé la conception de la **fenêtre perceptuelle**. Notre système met en œuvre deux flux d'entrée pour réaliser des tâches combinant navigation et désignation. La navigation dans une fenêtre graphique est contrôlée par les mouvements du visage alors que la souris est utilisée pour la désignation. Nous faisons l'hypothèse que le visage peut jouer le rôle de la main non dominante puisque le visage satisfait les trois caractéristiques de la théorie de Guiard :

1. Le visage définit le référentiel pour la souris : il détermine la zone du document visible dans la fenêtre graphique. Cette zone représente l'espace de travail de la souris.
2. Les mouvements du visage ont lieu avant les mouvements de la souris : l'objet à désigner doit être visible dans la fenêtre avant que la souris ne puisse le pointer.
3. Les mouvements exécutés par le visage sont d'un niveau de granularité supérieur à ceux de la souris : la navigation se termine quand l'objet à désigner est visible dans la fenêtre. L'objet à désigner est en règle générale de taille largement inférieure à celle de la fenêtre. Il n'est donc pas nécessaire de positionner la fenêtre très précisément sur l'objet dans le document. Par contre, la désignation de l'objet à la souris nécessite plus de précision, notamment si l'objet est de petite taille.

1.3. SUPPRESSION DES INTERMÉDIAIRES DE L'INTERACTION

La **fenêtre perceptuelle** présente un atout que n'ont pas les systèmes à deux flux d'entrée étudiés jusqu'ici : la suppression d'intermédiaires entre l'utilisateur et les concepts du système. Cette capacité d'offrir une interaction plus directe est illustrée par la figure 2. La **fenêtre perceptuelle** met en œuvre un multiplexage spatial des fonctions (voir page 26) : le visage est entièrement dédié à la fonction de navigation. Il n'est donc pas nécessaire de passer par une phase d'association du visage à sa fonction logique. De plus, à l'instar du **Bureau Digital** (voir page 33), la **fenêtre perceptuelle** utilise un membre de l'utilisateur, la tête, en tant que dispositif d'entrée. Il n'est donc pas nécessaire de passer par une phase d'acquisition du dispositif physique.

L'approche de la **fenêtre perceptuelle** est difficilement généralisable : les utilisateurs ont un nombre limité de membres, et le contrôle de plusieurs membres de façon indépendante peut nécessiter un lourd apprentisssage. Cependant, nous pensons qu'il est raisonnable d'associer au visage, de façon durable, la fonction qui permet de contrôler la partie visible d'un document dans une fenêtre. Ce choix s'appuie sur une situation similaire de la vie courante : pendant la conduite d'un véhicule, le visage est en permanence orienté de façon à définir le point de vue alors que les autres membres agissent sur le véhicule.

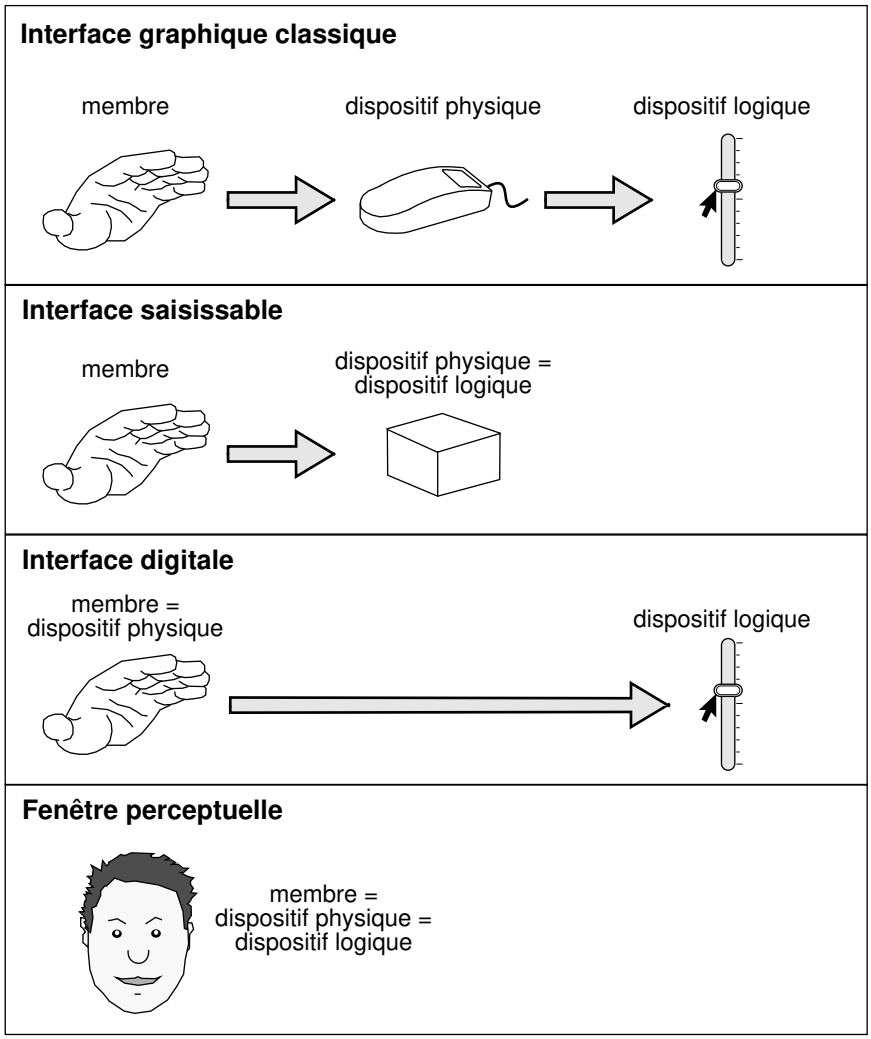


Figure 2
Illustration de l'aspect direct de l'interaction avec la fenêtre perceptuelle
Les flèches épaisses symbolisent les étapes intermédiaires de l'interaction.
Les interfaces graphiques classiques nécessitent deux étapes intermédiaires: l'acquisition du dispositif physique (souris) et l'acquisition du dispositif logique (widget) avec le pointeur de la souris.
Les interfaces saisissables suppriment l'étape d'acquisition du dispositif logique en associant aux dispositifs physiques une fonction logique durable.
L'interface digitale supprime l'étape d'acquisition du dispositif physique par l'utilisation d'un membre de l'utilisateur en tant que dispositif physique.
La **fenêtre perceptuelle** permet la suppression des deux étapes intermédiaires.

En pratique, l'utilisateur doit pouvoir activer et désactiver la navigation par le visage. Il est en effet amené à effectuer des mouvements de la tête pour regarder le clavier, le téléphone, une autre personne, etc. Dans ces cas, le système ne doit pas interpréter les mouvements du visage comme des commandes de navigation. Nous verrons au paragraphe "Activation" page 157 la solution retenue pour la **fenêtre perceptuelle**.

Ayant motivé la conception de la **fenêtre perceptuelle**, nous en présentons maintenant la mise en œuvre.

2. *Le système*

La description de notre prototype de **fenêtre perceptuelle** suit les niveaux d'abstraction de mise en œuvre : le dispositif matériel, la réalisation

Figure 3
Le dispositif de la fenêtre perceptuelle
Une caméra vidéo est installée en face du visage de l'utilisateur au-dessus du moniteur graphique. L'utilisateur s'installe normalement face à sa station de travail.

technique du suivi du visage avec ses performances et ses limitations, puis les modalités d'interaction.

2.1. DISPOSITIF

Le dispositif repose entièrement sur du matériel standard, ce qui est une originalité si on le compare à la plupart des nouveaux dispositifs d'entrée. Comme le montre la figure 3, une caméra vidéo est placée sur le moniteur graphique en face du visage de l'utilisateur. Elle fournit à une carte de numérisation un flux vidéo au format PAL. La station de travail est un Apple PowerMacintosh 8600 équipée d'un processeur PowerPC 604 cadencé à 350 MHz. et d'une carte d'acquisition vidéo d'origine.

2.2. SUIVI DU VISAGE

Nous précisons ici la technique de suivi adoptée que nous évaluons ensuite au regard des requis de l'interaction fortement couplée : latence, résolution spatiale et stabilité statique. Si les performances répondent aux attentes, le système présente néanmoins des limitations qu'il convient d'énoncer.

Technique de suivi

Les mouvements du visage sont suivis en utilisant la technique de corrélation présentée au chapitre IV page 95. La définition de l'image traitée correspond au quart du signal numérisé par la carte d'acquisition, c'est-à-dire 384 x 288 pixels. Cette définition d'image permet d'éviter le problème d'entrelacement énoncé en annexe B page 189. Un motif de taille 32 x 32 pixels est suivi dans une zone de recherche de 60 x 60

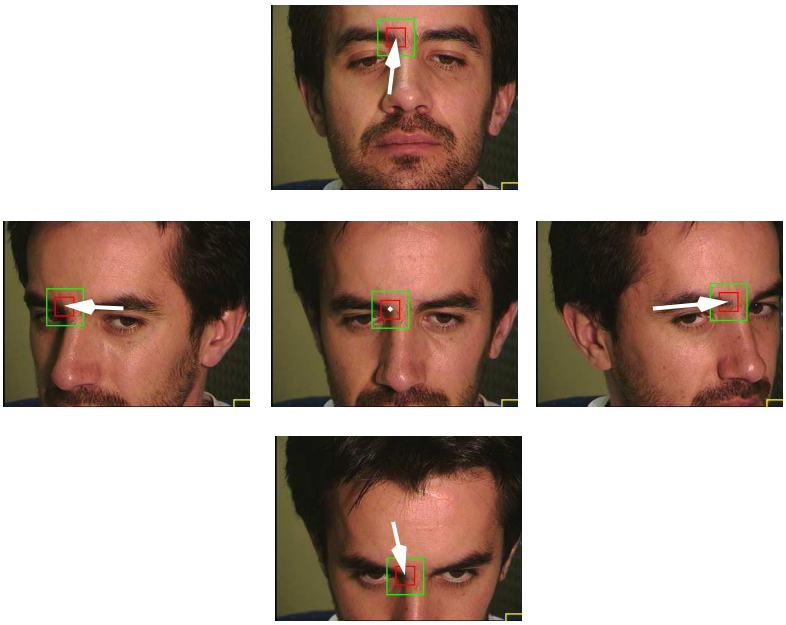


Figure 4
Translation du motif du suivi
L'utilisateur effectue des rotations de la tête qui se traduisent pas des translations du motif dans l'image. La flèche blanche indique la translation apparente du motif à partir de la position d'origine (image du centre).

pixels. Ces valeurs de paramètres sont choisies pour optimiser la vitesse maximale tolérée de la cible, selon le raisonnement présenté au chapitre IV page 98.

La cible du suivi (voir page 76) est choisie sur le visage de l'utilisateur par mémorisation d'un motif (voir page 95). Notre implémentation du suivi par corrélation ne tolère, en théorie, que les mouvements de translation de la cible dans le plan de l'image. En pratique, un choix judicieux de la cible (nous y reviendrons au paragraphe "Limitations" page 155) autorise des rotations de la cible hors du plan de l'image. Dans ce cas, les déplacements mesurés par le suivi correspondent aux translations du motif dans l'image, non pas aux translations du visage dans l'espace. Lorsque la tête accomplit une rotation, le motif du suivi effectue une translation dans l'image. Ce phénomène est illustré à la figure 4. La tolérance du suivi face aux rotations selon l'axe du cou et l'axe passant par les oreilles offre davantage de confort que des translations du visage strictement horizontales et verticales.

Latence et fréquence de fonctionnement

Pour les conditions matérielles et les paramètres définis ci-dessus, la fréquence de fonctionnement du système de suivi est de 58 Hz., soit une latence de 17 ms. Si l'on ajoute le temps nécessaire à la génération du retour d'information, notre système a une latence maximale de 65 ms., ce qui se traduit par une fréquence de fonctionnement supérieure à 15 Hz.

Les performances obtenues sont souvent supérieures à 15 Hz. Leur variabilité tient au retour d'information : on observe des chutes de 58 Hz. à 15 Hz. lorsqu'une grande surface d'écran doit être mise à jour. Le système est donc proche de satisfaire le requis des 50 ms. de latence maximum énoncé au chapitre II page 53.

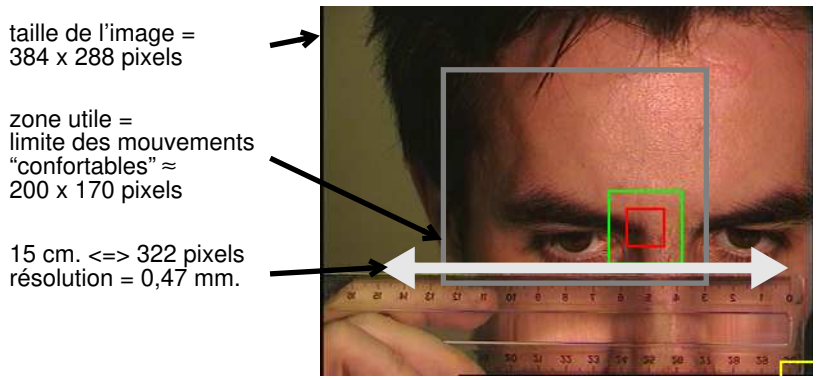
En pratique, l'interaction est fluide et la latence est difficilement perceptible. De plus, le suivi du visage est responsable d'uniquement 26 % de la latence totale du système, les 74 % restants revenant à la génération du retour d'information. Ce retour d'information est réalisé à l'aide de la boîte à outils Tk (voir annexe B page 187) dont les performances sont faibles au regard de celles des boîtes à outils graphiques *natives* des systèmes (par exemple "QuickDraw" sur Macintosh [Apple 94] ou "X-Window" sur UNIX [Nye 88]). L'utilisation d'une boîte à outils graphiques performante permettrait de réduire la contribution du retour d'information à la latence de notre système et de réduire ainsi la latence globale à une valeur inférieure à 50 ms.

Résolution spatiale

Le facteur de zoom de la caméra est augmenté manuellement afin d'améliorer la résolution spatiale. Le facteur de zoom est réglé de façon à ce que l'image du visage de l'utilisateur s'approche des bords de l'image lorsqu'il effectue les mouvements de plus grande amplitude, dans la limite des mouvements considérés comme "confortables". Un compromis est nécessaire entre augmenter la résolution selon notre technique et prendre le risque que le visage sorte du champ de vision de la caméra. En pratique, nous choisissons un facteur de zoom tel que la plupart des mouvements de l'utilisateur restent dans un rectangle dont la taille est d'environ 200 x 170 pixels, à l'intérieur d'une image de taille 384 x 288 pixels. Ceci revient à définir une *zone utile* de l'image ayant environ 50 % de la taille horizontale de l'image et 60 % de sa taille verticale. La figure 5 illustre le rapport entre une image et sa zone utile.

Le suivi par corrélation renvoie les valeurs de position de la cible avec une précision d'un pixel. La résolution du système correspond donc à la taille de ce que représente un pixel dans le monde physique. On obtient une bonne approximation de la résolution en mesurant la taille en pixels d'une règle présentée à la même distance de la caméra que le visage (voir figure 5). On constate ainsi que la résolution de notre dispositif de suivi est de l'ordre de 0,5 mm.

Figure 5
Résolution du système de suivi
Le facteur de zoom de la caméra est ajusté afin que la cible du suivi reste dans un rectangle inclus dans l'image, et ce dans la limite des mouvements confortables de l'utilisateur. Ce rectangle est appelé "zone utile".
On mesure la résolution correspondant à ce facteur de zoom en calculant la taille en pixels d'une règle présentée à la caméra et placée à la même profondeur que le visage.



Une résolution de 0,5 mm. signifie que les mouvements du visage d'amplitude inférieure à 0,5 mm. dans le plan de l'image ne sont pas détectés. Cette résolution est environ cinq fois moindre que celle d'une souris (voir page 55). Notre expérience de la **fenêtre perceptuelle** tend à montrer que les utilisateurs sont capables de contrôler très finement les déplacements de leur visage. Il semble qu'une meilleure résolution que celle réalisée par notre système aurait pu être bénéfique à l'interaction. Néanmoins, les déplacements du visage sont employés à la réalisation de tâches de navigation dont le niveau de granularité des mouvements est supérieur aux tâches de désignation. Les tâches de navigation nécessitent donc moins de finesse. En pratique, une résolution de l'ordre de 0,5 mm. s'avère suffisante pour que le système soit utilisable.

Stabilité statique

Le suivi par corrélation fournit des informations statiquement stables (voir page 57). La stabilité du suivi est vérifiée par l'expérience suivante : lorsque le visage est immobile, les données de position envoyées par le système de suivi restent strictement constantes.

Dans la **fenêtre perceptuelle**, les déplacements du visage ont pour effet de faire défiler le contenu d'une fenêtre graphique. Lorsque la surface du document est largement supérieure à la surface de la fenêtre, un déplacement du visage correspondant à un pixel peut provoquer un défilement du contenu de plusieurs pixels (c'est le cas pour l'interaction présentée au paragraphe "Contrôle de la position" page 161). Si le suivi du visage utilisé n'avait pas été stable, les oscillations de position auraient provoqué une instabilité perceptible, donc gênante, du contenu de la fenêtre. En conséquence, il est indispensable pour ce type d'interaction, de disposer d'un suivi stable. Le critère de stabilité motive à lui seul l'utilisation d'un suivi en vision par ordinateur par rapport aux suivis offerts par les systèmes magnétiques que nous savons statiquement instables.

Limitations

Notre système de suivi présente trois limitations :

1. Le suivi échoue lorsque le mouvement de la cible est trop rapide. Ce phénomène se produit lorsque la vitesse de rotation du visage est telle que le déplacement de la cible entre deux images est plus grand que la zone de recherche. Dans le contexte de notre étude, les vitesses de rotation observées sont restées inférieures à la limite de vitesse acceptée par le système.
2. Le suivi échoue si le visage effectue une rotation de trop grande amplitude ou de fortes variations de distance par rapport à la caméra. De tels mouvements induisent une modification de l'apparence de la cible dans l'image provoquant à son tour la perte de la cible (voir à ce sujet le paragraphe "Discussion" page 101 concernant les limites du suivi par corrélation). Ce phénomène est aggravé si la cible est choisie

sur une surface du visage de profondeur variable comme le nez ou le cadre des lunettes. Lors de l'étude présentée ici, la cible est indiquée manuellement en prenant soin de choisir une surface plane du visage.

3 Le suivi échoue si l'apparence dans l'image du voisinage de la cible est similaire à l'apparence de la cible. Cela peut se produire lorsque la cible est choisie sur une large surface n'ayant pas de texture, comme par exemple le front.

En pratique, la première limitation s'avère peu gênante. Seuls des mouvements du visage assez violents peuvent provoquer l'échec du suivi. En règle générale, ce type de mouvements n'est pas destiné au contrôle du système. Il a lieu par exemple lorsque l'utilisateur est surpris par la sonnerie du téléphone ou par l'entrée d'une personne dans son bureau. Le suivi ne perd pas le visage pendant les périodes d'interaction entre l'utilisateur et le système, mais il le perd entre ces périodes d'interaction. Cette situation milite en faveur d'une ré-initialisation automatique du suivi.

Le prototype présenté ici nécessite une initialisation manuelle du suivi. Une interface graphique permet de désigner, à l'aide de la souris, la région de l'image que le système mémorise en tant que motif. Nous choisissons une région du visage située entre les deux extrémités intérieures des sourcils car l'*apparence* de cette région est peu sensible aux mouvements de rotation et d'éloignement par rapport à l'écran (limitation 2) et parce qu'elle est différente de toutes les régions avoisinantes (limitation 3).

Initialiser le système manuellement serait une limitation rédhibitoire s'il devait être utilisé dans la vie courante. Comme nous l'avons vu ci-dessus, l'initialisation est fréquemment nécessaire entre les phases d'interaction. Il sera donc indispensable d'intégrer les techniques d'initialisation automatique par coopérations de techniques, présentées au chapitre IV page 103, avant de diffuser ce système. Toutefois, durant notre expérimentation utilisateur, le suivi est en règle générale initialisé une seule fois en début d'expérience. Une ré-initialisation est nécessaire uniquement si le sujet oriente son visage dans une direction très éloignée de celle du moniteur, ce qui n'arrive pas durant l'expérience. Pour cette raison, nous avons préféré porter notre effort sur la réalisation d'un suivi satisfaisant les requis de l'interaction fortement couplée plutôt que sur sa capacité d'autonomie.

2.3. INTERACTION

Le système de suivi génère les coordonnées en deux dimensions (x, y) de la cible dans le repère de l'image. La cible étant positionnée sur le visage de l'utilisateur, le système est renseigné au cours du temps sur les mouvements de son visage.

On alloue au visage la tâche de navigation au sein de l'espace des documents. Les mouvements du visage sont utilisés pour définir la zone

du document visible à l'intérieur d'une fenêtre graphique. Dans une interface graphique classique, cette tâche est assignée aux barres de défilement de la fenêtre.

Activation

L'interaction est contrôlée par une touche d'activation. Nous utilisons pour cela la touche "tabulation" ou la touche "espace" du clavier. Les touches de type "méta" ("shift", "control", "alt") seraient également de bonnes candidates pour jouer le rôle de touche d'activation.

La touche d'activation est nécessaire pour discerner les mouvements destinés à contrôler le système des autres mouvements. En effet, l'utilisateur est naturellement amené à orienter son visage dans différentes directions sans pour autant vouloir contrôler la navigation. Le problème du discernement des mouvements utiles s'apparente au problème des instants d'intérêts introduit au chapitre I page 35. Ce problème est résolu dans le prototype du **Bureau Digital** qui détecte les pics de volume audio lorsque l'utilisateur tape sur le bureau. La **fenêtre perceptuelle** pourrait également utiliser le canal audio, par exemple en mettant en œuvre une reconnaissance vocale. L'utilisateur activerait la navigation en appelant la commande "navigue" à la voix, et arrêterait l'opération au moyen d'une autre commande. Outre les problèmes liés à la réalisation d'une telle interaction multimodale, nous envisageons deux limitations du point de vue de l'utilisateur :

1. une telle activation serait une source de bruit fréquent, ce qui peut être gênant dans un contexte de bureau partagé;
2. l'appel d'une commande vocale nécessite plus d'effort et de temps que le simple appui sur une touche si la main est positionnée près de la touche. Or, on observe qu'une session de travail typique implique de nombreuses navigations, c'est-à-dire de nombreux couples activation/ désactivation de la fonction de navigation. Il est donc essentiel de réduire cet effort au minimum.

Le prototype du **tableau magique** résoud le problème de l'instant d'intérêt par une analyse spatio-temporelle de la trajectoire du doigt afin de reconnaître les *pauses*. Lorsque le doigt se stabilise pendant un certain temps, le système l'interprète comme l'équivalent d'un clic sur le bouton de la souris (voir page 139). Cette approche aurait pu être adoptée pour la **fenêtre perceptuelle** en définissant une trajectoire précise du visage comme activateur de la navigation (par exemple un cercle) et en considérant la pause comme la commande d'arrêt de la navigation. Ici aussi, le temps et l'effort nécessaires à l'exécution des trajectoires d'activation et d'arrêt ont semblé trop coûteux par rapport à l'appui sur une touche.

La **fenêtre perceptuelle** utilise donc une touche d'activation. L'utilisation d'une touche d'activation compromet certains des objectifs de la **fenêtre**

perceptuelle. Elle implique l'intervention de la main non dominante dans la tâche de navigation. Il aurait été préférable que la main non dominante reste disponible pour d'autres tâches. En pratique, la touche d'activation est enfoncée uniquement durant la navigation, la main non dominante est donc libre en dehors des phases de navigation. Durant la navigation, le contenu du document est en mouvement, ce qui réduit fortement la possibilité d'effectuer d'autres opérations en parallèle et pour lesquelles la main non dominante aurait pu intervenir.

L'aspect direct de l'interaction, présenté en page 150, est également compromis par l'utilisation d'une touche d'activation puisqu'il est nécessaire d'acquérir cette touche (c'est-à-dire d'amener un doigt dessus) avant de l'actionner. En pratique, on observe que beaucoup d'utilisateurs se servent de leur main dominante pour manipuler la souris, et que leur main non dominante reste à proximité du clavier afin d'activer les commandes ou saisir quelques caractères. C'est pourquoi nous considérons que l'acquisition d'une touche d'activation, notamment une touche de type "méta" facilement accessible sur le clavier, ne nécessite que peu de déplacements, de temps et d'effort de la part des utilisateurs. Nous pensons donc que l'aspect direct de l'interaction n'est que partiellement compromis par l'utilisation d'une touche d'activation.

La touche d'activation a deux rôles : premièrement, lorsque l'utilisateur appuie sur la touche, l'origine de la translation est enregistrée à la position courante du visage. Deuxièmement, l'appui sur la touche fait basculer le système dans le mode navigation : le contrôle du défilement de la fenêtre a lieu uniquement lorsque l'utilisateur maintient la touche enfoncée. Le contrôle s'arrête dès que l'utilisateur relâche la touche. Cet engagement dans un mode par un engagement actif de la part de l'utilisateur réduit les chances d'oublis du mode courant ([Sellen 92]).

Ce mécanisme d'activation et d'arrêt de la navigation est utilisé dans deux types d'interaction : l'une concerne le contrôle de la vitesse, l'autre le contrôle de la position.

Contrôle de la vitesse

Un ensemble de séquences vidéo numériques disponibles sur Internet illustre l'aspect dynamique du système ([Bérard 99b]) : lorsque la tête est inclinée vers le haut au-delà d'une *zone neutre*, le contenu de la fenêtre commence à défiler vers le bas. Plus la tête est inclinée, plus le défilement est rapide. Le défilement s'arrête dès que la tête est ramenée dans la zone neutre. Incliner la tête vers le bas provoque un comportement symétrique. Les rotations de la tête vers la gauche et vers la droite entraînent un défilement du contenu vers la droite et vers la gauche, respectivement. La vitesse du défilement est gouvernée par une relation exponentielle avec le mouvement de la tête plutôt que par une relation linéaire. Pour l'utilisateur, cette fonction se traduit par la possibilité de pratiquer à la

fois de fins ajustements et des défilements rapides selon le degré d'inclinaison du visage. Détaillons de manière plus formelle le calcul de la vitesse.

Dans ce qui suit, le symbole z peut représenter soit une abscisse (x) ou une ordonnée (y) selon que l'on considère les défilements horizontaux ou verticaux. Soit Δ_z la valeur de translation apparente du visage dans l'image (Δ_z est exprimée en pixels). Δ_z est calculée à tout instant par :

$$\Delta_z = z_t - z_0 \qquad (1)$$

où z_t représente la position courante du visage dans l'image, et z_0 représente la position du visage dans l'image au moment où l'utilisateur appuie sur la touche d'activation. La vitesse de défilement v_z s'exprime par l'équation :

$$v_z = f(\Delta_z) = \begin{cases} |\Delta_z| \le N, & 0 \\ \Delta_z > N, & -g \cdot (\Delta_z - N)^{\frac{1}{a}} \\ \Delta_z < -N, & g \cdot (-\Delta_z - N)^{\frac{1}{a}} \end{cases} \qquad (2)$$

où N est le rayon de la zone neutre, g représente un facteur d'échelle (la vitesse de défilement est d'autant plus rapide que g est grand), et a représente un facteur d'accélération avec $0 < a \le 1$. f est appelée *fonction de transfert* :

- lorsque a est proche de 0, la fonction de transfert est extrêmement courbée,
- lorsque $a = 1$, la fonction de transfert est une droite.

Dans le cas d'une diffusion du système, le choix des valeurs de facteur d'échelle g et d'accélération a devrait être accessible à l'utilisateur sous forme de "préférences".

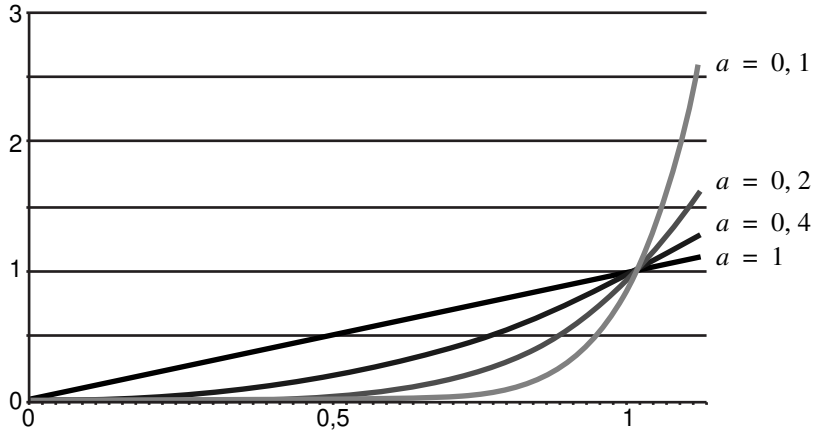


Figure 6
Courbure de la fonction de transfert en fonction du paramètre d'accélération a

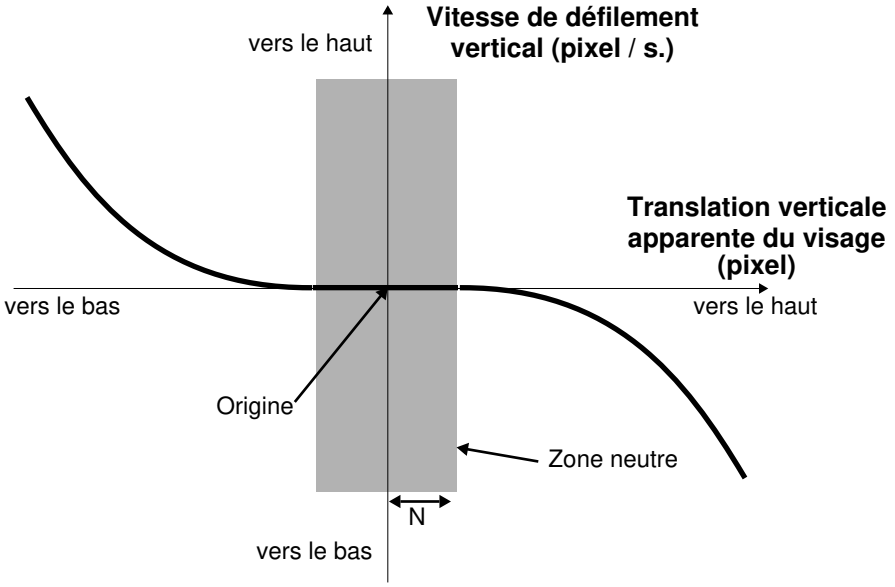


Figure 7
Forme de la fonction de transfert verticale
La courbe noire représente la vitesse de défilement verticale en fonction de la translation verticale apparente du visage.
N est le rayon de la zone neutre.
La fonction de transfert horizontale a la même forme.

La figure 6 illustre la courbure de la fonction de transfert en fonction du paramètre d'accélération. Dans nos expérimentations, nous utilisons les valeurs suivantes de g et a :

$$g = 1 \qquad a = 0,5 \tag{3}$$

La figure 7 schématise la fonction de transfert de la vitesse de défilement verticale en fonction de la translation verticale apparente du visage. La fonction de transfert horizontale est de même forme. Un défilement diagonal peut être exécuté en inclinant le visage selon les deux axes simultanément.

Finalement, à chaque cycle du système, après calcul de la translation apparente du visage, la translation du document par rapport à la fenêtre est définie par :

$$T_z = v_z \cdot \Delta_t \tag{4}$$

où Δ_t est le temps écoulé depuis le dernier calcul de T_z. Cette prise en compte du temps écoulé permet de conserver une vitesse de défilement constante lorsque la fréquence de fonctionnement du système varie (du fait des variations de la charge du système par exemple).

Prenant en compte qu'il est possible d'arrêter la navigation de deux façons différentes (en ramenant le visage dans la zone neutre, ou en relâchant la touche d'activation), nous observons deux comportements utilisateur distincts :

1 Les utilisateurs novices trouvent parfois difficile de ramener le visage à l'intérieur de la zone neutre. Le système ne générant aucun retour d'effort, la tête est libre de toute contrainte. Seul l'arrêt du défilement informe l'utilisateur qu'il a ramené son visage dans la zone neutre. Ces

utilisateurs préfèrent alors appuyer et relâcher la touche d'activation. Le rayon de la zone neutre est réglé à 0 pour ce type d'utilisateur afin que le défilement débute dès que la touche est enfoncée et que le visage s'éloigne de sa position initiale.

2 Certains utilisateurs, avec un peu d'entraînement, exécutent une navigation plus continue en maintenant la touche d'activation enfoncée. Ils utilisent la zone neutre pour arrêter le défilement quand cela est nécessaire. Pour ce type d'utilisateur, la valeur du rayon de la zone neutre est ajustée à 2 % de la hauteur de l'image traitée. La touche d'activation est utilisée par ces utilisateurs uniquement comme un contrôle marche / arrêt de plus haut niveau, au début et à la fin d'une tâche globale.

Nous avons mené une étude comparative des performances de sujets utilisant une interface graphique classique et utilisant la **fenêtre perceptuelle** en interaction de contrôle de la vitesse. Le protocole expérimental et les résultats sont présentés à la section suivante.

Contrôle de la position

La deuxième interaction réalisée par la **fenêtre perceptuelle** est de type contrôle de la position absolue. Cette interaction est similaire à celle de la souris dans la mesure où les translations effectuées par les utilisateurs sont directement reflétées par les translations du document dans la fenêtre. Le seul traitement appliqué aux données en sortie du système de suivi est une amplification, c'est-à-dire la multiplication de la translation par une constante appelée *gain*.

Si l'on reprend les notations du paragraphe précédent, la position de la fenêtre dans l'espace du document $W_{z,t}$ au moment t est donnée par :

$$W_{z,t} = W_{z,0} + g \cdot \Delta_z \qquad (5)$$

où $W_{z,0}$ est la position de la fenêtre dans l'espace du document au moment où l'utilisateur a enfoncé la touche d'activation, g est le gain, et Δ_z est défini par l'équation 1 page 159.

L'amplification est nécessaire du fait de la taille relativement réduite de la zone utile (voir page 154) par rapport à l'espace du document. Si aucune amplification n'était appliquée, de nombreux *repositionnements* seraient nécessaires pour des défilements de large amplitude. Un repositionnement est exécuté lorsque le visage est incliné à son maximum mais qu'une translation d'amplitude supérieure est nécessaire. La touche d'activation est alors relâchée afin que l'utilisateur puisse ramener son visage dans une position neutre, c'est-à-dire non inclinée, sans provoquer de défilement. La touche d'activation est enfoncée de nouveau pour continuer le défilement dans la direction initiale. Le principe du repositionnement est similaire à celui de la souris lorsque celle-ci a atteint la limite du tapis de souris : la souris doit être soulevée afin d'être déplacée à l'opposé du tapis

sans envoyer de commandes de déplacement au système. Au cours des expériences rapportées dans les sections 3 et 4 de ce chapitre, les sujets n'exécutent pratiquement jamais de repositionnement car ils peuvent, la plupart du temps, faire défiler d'un bout à l'autre le document utilisé pendant l'expérience par un seul mouvement du visage.

L'accroissement du gain permet de limiter le nombre de repositionnements nécessaires, mais il diminue la résolution en sortie du système. Lorsque le gain augmente, le déplacement minimal de la fenêtre dans l'espace du document augmente également. Le choix du gain devrait être considéré comme une préférence utilisateur. Durant nos expérimentations, la taille de la fenêtre est ajustée à 400 x 500 pixels et le document a une taille de 1600 x 2000 pixels. La valeur de gain utilisée est de 30, ce qui signifie qu'une translation apparente du visage de 1 pixel provoque une translation de la fenêtre de 30 pixels, soit 6 % de la hauteur de la fenêtre, soit encore 1,5 % de la hauteur du document. Des séquences vidéo numériques du système en fonctionnement, disponibles sur internet ([Bérard 99b]), donnent une idée concrète de l'interaction produite par le système.

Nous avons estimé le caractère utilisable de la **fenêtre perceptuelle** en mesurant les performances d'un ensemble de personnes utilisant le système. Les performances ont été mesurées durant l'exécution de deux types de tâches :

- une tâche *exploratoire* demandant aux sujets de se déplacer dans le document en suivant un chemin qui les dirige vers leur destination. Cette tâche est réalisée en utilisant l'interaction de contrôle de vitesse.
- une tâche de *glisser-déposer* demandant aux sujets de déplacer un objet de sa position d'origine vers une destination, cette destination étant immédiatement rendue disponibles aux sujets. Cette tâche est réalisée en utilisant l'interaction de contrôle de la position.

Nous les présentons l'une et l'autre dans les deux sections suivantes.

3. Performances utilisateur pour une tâche exploratoire

Les tâches exploratoires sont fréquentes en édition de documents, notamment pour les documents qui ne tiennent pas sur une page écran comme les dessins à haute résolution ou les feuilles de calcul de grande taille. Avec les interfaces graphiques usuelles, l'exploration se traduit par nombreuses tâches articulatoires. Ce constat motive l'utilisation de la

fenêtre perceptuelle, le montage d'un protocole expérimental et l'analyse des résultats de l'observation. Ces quatre points (motivation, protocole expérimental, résultats et discussion) sont ici présentés successivement.

3.1. MOTIVATIONS

Les interfaces graphiques classiques proposent les barres de défilement pour naviguer dans l'espace d'un document. Lorsque le centre d'intérêt de la tâche se porte hors de la partie visible de la fenêtre, l'utilisateur se voit contraint d'interrompre sa tâche principale afin de basculer sur une tâche de navigation. La navigation nécessite de nombreuses tâches articulatoires, notamment lorsque le défilement nécessaire est diagonal. Dans ce cas, le pointeur de la souris doit être déplacé de la zone d'édition vers une barre de défilement, l'ascenseur de la barre de défilement doit être positionné à l'endroit désiré, le pointeur doit ensuite être déplacé vers la deuxième barre de défilement afin de déplacer le deuxième ascenseur. L'échange entre les barres de défilement peut ainsi se répéter jusqu'à l'atteinte de l'état souhaité. Toutes ces opérations, qui sont des désignations de type Fitts (voir page 46), impliquent, pour l'utilisateur, un surcroît d'efforts et de temps.

Le service de contrôle de la vitesse de la **fenêtre perceptuelle** permet de réduire la charge induite par les tâches de navigation. Nous voyons plusieurs raisons à cela :

- Réduction du nombre de mouvements. Les mouvements de contrôle de la navigation sont alloués au visage au lieu de la main et cette allocation est permanente : les actions de changement de contexte enre la zone d'édition et les barres de défilement, sont alors éliminées.
- Potentiel de parallélisation. En théorie, les opérations de désignation et de navigation sont entrelacées séquentiellement dans le temps : il n'est guère possible de désigner un objet lorsque le contenu du document est en train de défiler. En pratique, ces deux opérations peuvent se chevaucher partiellement : l'utilisateur peut anticiper l'opération de navigation suivante avant même d'avoir terminé l'opération de désignation avec un clic souris : il commence à incliner son visage dans la direction de la prochaine étape de navigation. De même, il peut commencer à approcher le pointeur de la souris vers la prochaine cible à désigner dès que celle-ci apparaît dans la fenêtre et avant que l'opération de navigation ne soit terminée (par exemple lorsque la cible est centrée dans la fenêtre).

Nous concevons un protocole expérimental afin de valider ces hypothèses.

3.2. PROTOCOLE EXPÉRIMENTAL

La tâche expérimentale est conçue de façon à simuler l'édition d'un document dont la taille est largement supérieure, dans les deux dimensions, à celle de la fenêtre d'édition. La tâche nécessite des

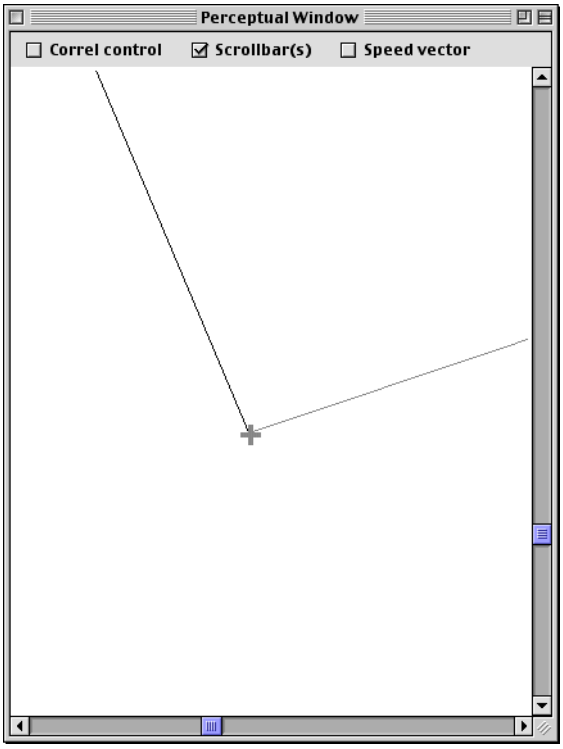


Figure 8
Tâche exploratoire au moyen de la fenêtre perceptuelle avec contrôle de la vitesse de défilement
Le fenêtre contient la prochaine cible à désigner, ainsi que deux lignes représentant la direction vers la cible précédente, et la direction qu'il faudra suivre pour atteindre la prochaine cible.
Lorsque le sujet clique sur la croix, celle-ci disparaît, ainsi que la première ligne.
Lorsque la navigation se fait grâce aux mouvements du visage, les ascenseurs des barres de défilement sont utilisés uniquement pour renseigner le sujet sur sa position dans le document.

opérations de navigation ayant la forme d'une exploration suivant un chemin : l'utilisateur connaît le chemin à suivre pour atteindre la destination en fonction du contenu de la fenêtre. Il peut anticiper. Ce type de navigation a lieu, par exemple, en suivant les lignes et les colonnes d'une feuille de calcul, ou en se repérant grâce à la forme et à la disposition des objets dans un grand dessin technique, ou encore grâce à une représentation mentale d'une grande image lorsque seulement une petite partie de l'image est visible dans la fenêtre.

La tâche expérimentale est exécutée sur un document accessible au travers d'une fenêtre d'une interface graphique classique. La taille du document est de 2400 x 3000 pixels. La taille de la fenêtre est de 400 x 500 pixels. Ainsi, seule 2,8 % de la surface du document est visible à un instant donné.

On présente aux sujets une succession de cibles représentées par de petites croix de taille 12 pixels, dessinées avec des lignes de 4 pixels de large. Le fait de cliquer sur une cible provoque sa disparition et l'apparition de la cible suivante à un autre endroit du document. La position des cibles est distribuée aléatoirement sur le document en utilisant une loi équiprobable dans les deux dimensions. Les sujets savent dans quelle direction faire défiler le document en suivant une ligne tracée entre la position d'une cible et celle de la prochaine à atteindre. Une autre ligne est également dessinée entre cette cible et la cible suivante. Cette deuxième ligne permet aux sujets d'anticiper la direction de la prochaine navigation avant même

d'avoir atteint la cible. À tout moment, les sujets connaissent leur emplacement dans le document en observant la position des ascenseurs des barres de défilement. Une copie d'écran est présentée à la figure 8.

Les sujets ont pour consigne de localiser puis de cliquer sur 50 cibles selon deux conditions expérimentales. Dans la condition appelée "statu quo", les sujets se servent de l'ascenseur des barres de défilement pour naviguer dans le document, tandis que dans la condition appelée "visage", les sujets effectuent des mouvements du visage pour naviguer dans le document selon l'interaction de contrôle de la vitesse. La séquence des emplacements des cibles est identique dans les deux conditions. Huit sujets ont volontairement participé à l'expérience. Afin de minimiser les effets d'apprentissage de la tâche, la moitié des sujets a commencé par la condition "statu-quo", tandis que l'autre moitié commençait par la condition "visage". Tous les sujets exécutent la tâche deux fois, en alternant les conditions expérimentales. L'ensemble des sujets sont experts dans le maniement des barres de défilement : ils les utilisent fréquemment dans leur travail. Excepté un, aucun des sujets n'a eu l'expérience d'une interaction fortement couplée fondée sur la vision par ordinateur, ni d'expérience sur l'interaction à deux flux d'entrée. C'est pourquoi ils sont d'abord entraînés par une séquence d'essai comprenant 30 cibles. Deux durées sont enregistrées pour chaque sujet : les durées d'accomplissement de la tâche pour les deux conditions expérimentales. Le chronométrage est déclenché au moment où le sujet clique sur la première cible et se termine au moment où il atteint la dernière cible. Lorsqu'un sujet clique à côté d'une cible, la cible ne disparaît pas et le sujet doit ré-essayer. En pratique, les sujets effectuent très peu d'erreurs de ce type. La durée moyenne de l'expérience est d'environ une demi-heure par sujet.

3.3. RÉSULTATS

Tous les sujets ayant accompli la tâche dans les deux conditions, et ne faisant aucune hypothèse a priori sur la condition la plus performante, nous traitons les données par une analyse statistique de type t-test bi-latéral avec échantillons appariés. Cette analyse révèle une différence très significative entre les moyennes des temps d'accomplissement de la tâche

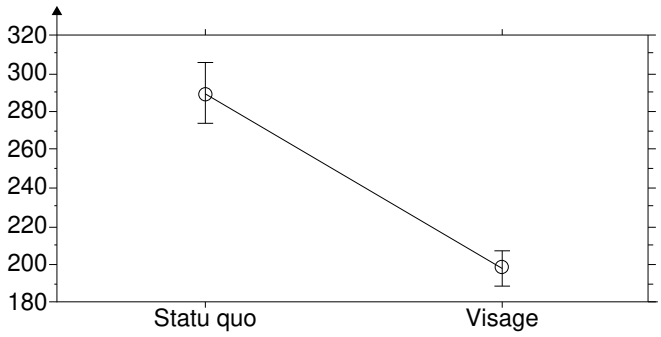


Figure 9
Moyennes des temps d'accomplissement de la tâche et leurs écarts-type en fonction des conditions expérimentales

selon les conditions expérimentales ($t(7) = 7,04$; $p = 0,000204$). Comme l'illustre la figure 9, la moyenne des temps réalisés dans la condition "visage" est de 198 s. (avec un écart-type de 26,3) et celle des temps réalisés dans la condition "statu quo" est de 290 s. (avec un écart-type de 45,6). L'amélioration de performance est de 32 %.

3.4. DISCUSSION

Le résultat le plus frappant de cette expérience est la généralité et l'amplitude de l'amélioration de performance. Les taux d'amélioration de performance varient entre 15 et 42 %, 75 % des sujets ayant amélioré leurs performances de plus de 30 %. Il est clair que la navigation au visage est mieux adaptée à la tâche étudiée que la navigation aux barres de défilement. Si un tel résultat se transpose à des tâches réelles, une amélioration de 30 % représente un avantage significatif.

L'autre fait marquant est la facilité avec laquelle les sujets ont pu utiliser une toute nouvelle modalité d'interaction. En moins d'une minute, les sujets se sont montrés capables d'utiliser avec aisance la navigation au visage. Un entretien avec les sujets fut systématiquement pratiqué après chaque session expérimentale. Ces entretiens révèlent que tous les sujets préfèrent la navigation au visage. Un des sujets émet le commentaire que le contrôle du défilement au visage est "très naturel : il suffit d'orienter le visage dans la direction de ce que l'on veut voir pour que cela apparaisse dans la fenêtre. On est alors naturellement enclin à ramener le visage vers le centre de la fenêtre ce qui a pour effet d'arrêter le défilement". Ce témoignage semble indiquer que la **fenêtre perceptuelle** a le potentiel d'améliorer la transparence de l'interaction en permettant d'interagir directement avec l'information.

La **fenêtre perceptuelle** avec contrôle de la vitesse est adaptée à la navigation exploratoire : contrairement aux barres de défilement, elle permet de spécifier finement et directement, à la fois la vitesse et l'orientation du défilement. Cette forme de contrôle est adaptée au survol de documents comme les pages internet : une fois la vitesse réglée, le contenu défile sans mouvement supplémentaire mais peut être ajusté dynamiquement si le besoin s'en fait sentir. Toutefois, lorsque l'utilisateur connaît d'emblée la localisation de la destination, la **fenêtre perceptuelle** avec contrôle de la position est plus adaptée. Nous étudions cette situation dans la section qui suit.

4. Performances utilisateur pour une tâche de glisser-déposer

Dans cette expérience, nous reproduisons les étapes nécessaires au déplacement d'un objet dans un document. Dans les interfaces graphiques à "manipulaiton directe", le déplacement se traduit par une tâche "glisser-déposer". Nous étudions ici sa mise en œuvre au moyen de la **fenêtre perceptuelle** avec contrôle de la position. Comme précédemment, nous organisons notre présentation comme suit : motivations, protocole expérimental, résultats et discussion.

4.1. MOTIVATIONS

Dans une interface graphique usuelle à manipulation directe, le déplacement d'un objet au sein d'une fenêtre a lieu selon les deux conditions distinctes :

1. soit l'objet et sa destination sont visibles,
2. soit l'objet est visible mais la destination ne l'est pas.

Dans le premier cas, l'utilisateur exécute une opération de *glisser-déposer* : l'objet est "attrapé" avec la souris et, tandis que le bouton de la souris est maintenu enfoncé, l'objet (ou son fantôme) est asservi au pointeur de la souris jusqu'au relâchement du bouton sur le lieu de destination.

Dans le deuxième cas, l'utilisateur doit naviguer dans l'espace de travail tout en "maintenant" l'objet avec la souris. Le bouton de la souris étant assigné à la tâche d'accroche d'objet, les barres de défilement ou tout autre moyen nécessitant le bouton de la souris ne peuvent être utilisés. La plupart des applications proposent l'alternative suivante : lorsque le pointeur, qui maintient l'objet, traverse l'une des bordures de la fenêtre, le contenu de la fenêtre commence à défiler en direction de la bordure opposée.

Cette méthode de navigation par contact des bordures ne permet pas d'ajuster la vitesse de défilement. Par conséquent, le défilement peut être jugé :

- soit trop rapide : la destination du glisser-déposer apparaît et disparaît dans la fenêtre avant que l'utilisateur n'ait eu la chance de stopper le défilement,
- soit trop lent : l'utilisateur doit attendre trop longtemps avant d'atteindre la destination.

En pratique, l'opération de glisser-déposer est souvent abandonnée au profit d'une suite d'opérations plus classique : désigner l'objet - couper - naviguer - désigner la destination - coller.

Avec la **fenêtre perceptuelle**, le visage servant à la navigatrion, la souris, libérée de cette tâche, peut servir à l'accroche de l'objet jusqu'à sa

destination. Une opération de glisser-déposer nécessitant moins d'actions qu'un copier-coller, nous pouvons anticiper que la **fenêtre perceptuelle** permet d'accroître les performances. Le protocole expérimental suivant vise à tester notre hypothèse.

4.2. PROTOCOLE EXPÉRIMENTAL

L'objet à déplacer, un carré noir de 20 x 20 pixels, est présenté aux sujets. Ce carré doit être déplacé dans une succession de 50 destinations représentées par des cadres noirs de 32 x 32 pixels. À un instant donné, seule une destination est présente dans le document. Les positions des destinations sont choisies aléatoirement dans l'espace du document à l'aide d'une loi équiprobable dans les deux dimensions. La taille du document est de 1600 x 2000 pixels. Lorsque le carré noir a été déposé dans un cadre de destination, le cadre disparaît et le carré noir est déplacé à un endroit choisi aléatoirement *dans la fenêtre* (il est donc toujours visible). Au même moment, le cadre de destination suivant est créé à un nouvel emplacement dans le document. Cet emplacement peut être dans la fenêtre et dans ce cas la destination est immédiatement visible, mais le plus souvent le cadre est situé en dehors de la zone visible du document, ce qui nécessite une étape de navigation. La figure 10 montre la fenêtre après l'étape de navigation. Le cadre de destination est visible dans la fenêtre. L'utilisateur n'a plus qu'à déplacer le carré noir dans le cadre et relâcher le bouton de la souris pour terminer l'opération de glisser-déposer.

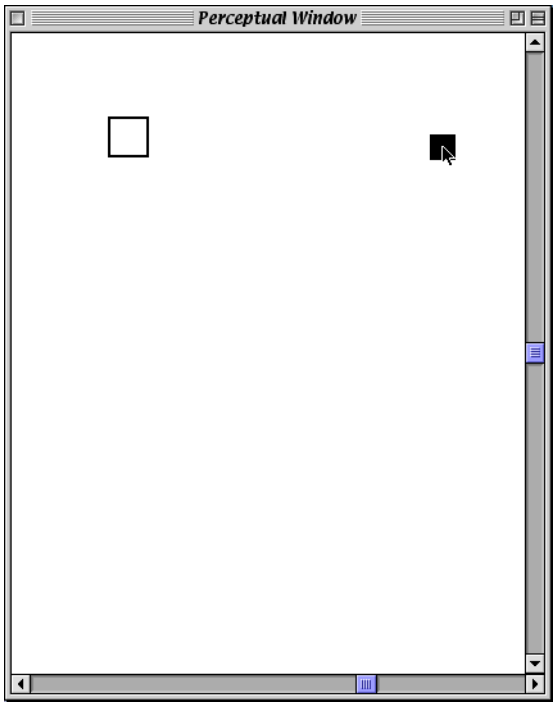


Figure 10
Fin de navigation de la tâche glisser-déposer
La destination (cadre noir) est visible dans la fenêtre. Pour terminer l'opération, l'utilisateur doit déplacer l'objet (carré noir maintenu par le pointeur de la souris) sur sa destination, et relâcher le bouton de la souris.

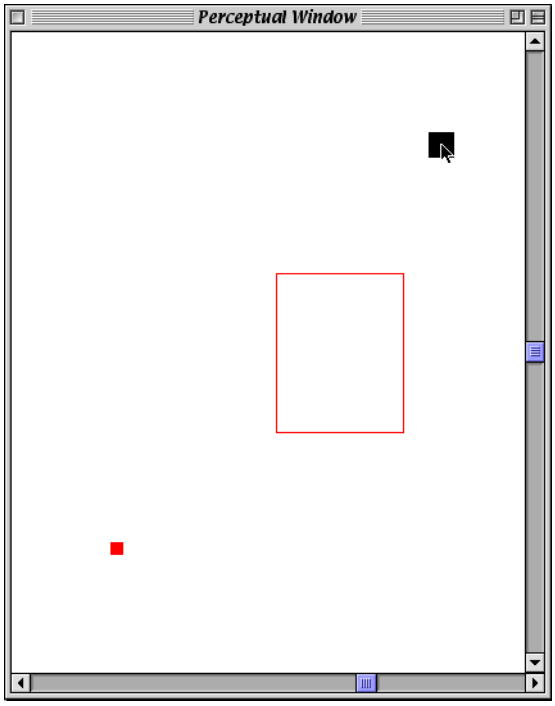


Figure 11
Vue radar active
Tant que la vue radar est active, un cadre et un petit carré rouges sont affichés dans la fenêtre. Le cadre représente la position de la fenêtre dans l'espace du document et le petit carré dénote la destination de l'opération de glisser-déposer. Pour que la destination soit visible dans la fenêtre, le cadre rouge doit être déplacé de façon à contenir le carré rouge.

À tout moment, seulement 6,2 % du document est visible au-travers d'une fenêtre de 400 x 500 pixels. Les sujets définissent la zone du document visible dans la fenêtre grâce à une *vue radar*. La vue radar représente une vue globale du document à échelle réduite. Sur cette représentation, la position courante de la fenêtre dans le document est symbolisée par un cadre rouge et la position de la destination du glisser-déposer par un petit carré rouge. Lorsqu'elle est activée, la vue radar est affichée à l'intérieur de la fenêtre, recouvrant la partie du document visible à ce moment là. La figure 11 montre la fenêtre lorsque la vue radar est activée.

La vue radar est activée lorsque la barre d'espacement est enfoncée, et reste active tant que la barre d'espacement n'est pas relâchée. Ainsi, la barre d'espacement joue le rôle de touche d'activation à la fois pour le contrôle de la navigation et pour la vue radar.

Deux conditions expérimentales sont testées dans cette expérience. Dans la première condition, appelée "statu quo", il est demandé aux sujets d'utiliser le clavier et la souris pour effectuer la tâche. Les actions nécessaires sont les suivantes :

1. Désignation de l'objet par clic souris (l'objet est alors affiché en surbrillance).
2. Appel de la commande "couper" par appui sur la combinaison de touches <commande>-X.
3. Activation de la vue radar en appuyant et maintenant appuyée la barre d'espacement. Le cadre et le petit carré rouges qui dénotent respectivement la position de la fenêtre dans l'espace global du document et la destination sont affichés.

4 Désignation par clic souris de la destination (le carré rouge). Cette action a pour effet de positionner la fenêtre à l'emplacement de la destination dans l'espace du document.
5 Désactivation de la vue radar en relâchant la barre d'espacement. La destination (cadre noir) est alors affichée dans la fenêtre.
6 Désignation de la destination en cliquant dessus. La destination est mise en surbrillance.
7 Appel de la commande "coller" par appui sur la combinaison de touches <commande>-V.

Dans la seconde condition expérimentale, appelée "visage", les sujets exécutent une opération de glisser-déposer par la séquence d'actions suivante :

1 Désignation de l'objet. L'objet est affiché en surbrillance dès que le pointeur de la souris le visite.
2 Enfoncement et maintien du bouton de la souris pour "attraper" l'objet. L'objet reste "attaché" au pointeur de la souris et suit tous ses déplacements.
3 Activation de la vue radar et de la navigation au visage en appuyant et en maintenant enfoncée la barre d'espacement. Comme dans la condition "statu quo". le cadre et le petit carré rouges sont affichés.
4 Rotation du visage pour amener le grand cadre rouge à contenir le petit rectangle rouge. Cette action a pour effet de déplacer la fenêtre à l'emplacement de la destination.
5 Désactivation de la vue radar et de la navigation en relâchant la barre d'espacement.
6 Glissement de l'objet dans le cadre de la destination. Le cadre noir qui dénote la destination est affiché en surbrillance dès que le pointeur de la souris le visite.
7 Dépôt de l'objet en relâchant le bouton de la souris.

Neuf sujets volontaires participent à l'expérience. Ils exécutent la tâche deux fois, une pour chaque condition expérimentale. Quatre sujets commencent par la condition "visage", les cinq autres commencent par la condition "statu quo". Les sujets reçoivent un entraînement minimal en exécutant une série d'entraînements sur 50 cibles avant chaque condition.

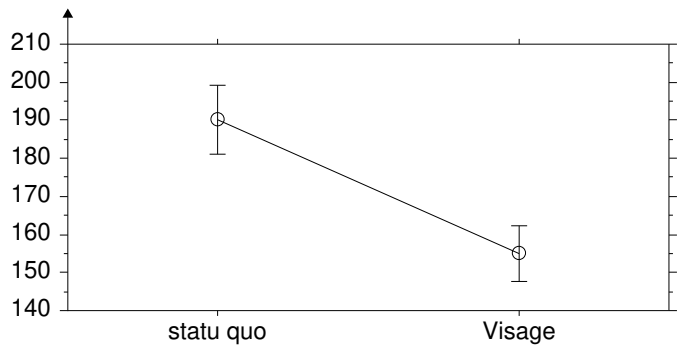


Figure 12
Moyennes des temps d'accomplissement de la tâche et leurs écarts-type en fonction des conditions expérimentales

4.3. RÉSULTATS

Tous les sujets ayant accompli la tâche dans les deux conditions, et ne faisant aucune hypothèse a priori sur la condition la plus performante, nous traitons les données par une analyse statistique de type t-test bi-latéral avec échantillons appariés. Cette analyse révèle une différence très significative entre les moyennes des temps d'accomplissement de la tâche selon les conditions expérimentales ($t(8) = 7.24$; $p = 0.00008$). Comme l'illustre la figure 12, la moyenne des temps réalisés dans la condition "visage" est de 155 s. (avec un écart-type de 22,2) et celle des temps réalisés dans la condition "statu quo" est de 190 s. (avec un écart-type de 27,2). L'amélioration de performance est de 18 %.

Un entretien pratiqué avec chaque sujet après l'expérience a révélé que 5 sujets préfèrent la condition "visage", tandis que les 4 autres préfèrent la condition "statu quo".

4.4. DISCUSSION

Comme dans l'expérience précédente, nous constatons la généralité et l'amplitude de l'amélioration de performance, mais aussi la facilité avec laquelle les sujets ont maîtrisé le système. Tous les sujets sont plus rapides dans la condition "visage" bien qu'ils n'aient jamais été exposés à cette modalité d'interaction.

Les entretiens post-expérimentaux révèlent un fort contraste entre les sujets qui ont apprécié le contrôle au visage et ceux qui ne l'ont pas aimé. Nous pensons que cette dépréciation tient à une mauvaise conception du retour d'information ayant provoqué la confusion des rôles des deux effecteurs main et visage pour les sujets du second groupe. Nous analysons ce point plus avant.

Confusion des rôles

Les sujets "réfractaires" se plaignent de ne pouvoir déterminer avec précision la position de la fenêtre dans l'espace du document et donc de perdre du temps. À la question "dans quelle condition pensez-vous avoir été le plus rapide?" les mêmes sujets répondent qu'ils pensent avoir été plus rapides dans la condition "statu quo". Les mesures expérimentales montrent que ce n'est pas le cas.

Pour ces sujets, positionner la fenêtre au bon endroit dans l'espace du document, revient à placer le carré rouge qui symbolise la destination *exactement au centre* du cadre rouge de la vue radar. Or, dans le protocole expérimental, il suffit que le carré rouge soit à l'intérieur du cadre rouge pour placer la fenêtre sur la zone voulue du document. En d'autres termes, le visage est sensé être utilisé pour le contrôle à gros grain tandis que la main dominante doit prendre le relais pour le contrôle fin des mouvements.

On constate donc chez ces sujets, une confusion des rôles qui tient sans doute à une mauvaise représentation graphique de la vue radar. Il semble

que certains sujets associent ce cadre à un curseur et le contrôlent comme tel, c'est-à-dire en essayant de le positionner avec précision. Ils utilisent donc le visage pour un contrôle fin du mouvement, capacité qui, nous l'avons vu au paragraphe "Contrôle de la position" page 161, n'est pas le propre de la **fenêtre perceptuelle**.

Pour éviter ce phénomène, il pourrait être intéressant de supprimer la vue radar en faisant la même hypothèse que dans la première expérience, c'est-à-dire en supposant que les utilisateurs connaissent à tout moment la direction de la navigation en fonction du contenu courant de la fenêtre.

Si le problème de la confusion des rôles peut être résolu, la **fenêtre perceptuelle** a le potentiel d'élargir le champ d'application des opérations de type glisser-déposer.

Généralisation des opérations glisser-déposer

En termes d'actions physiques, le glisser-déposer est plus performant que le couper-coller sous réserve que la destination soit d'emblée visible. Cette contrainte de visibilité est un facteur limitant pour l'usage du glisser-déposer dans les interfaces graphiques actuelles. La **fenêtre perceptuelle** permet d'étendre le champ d'application du glisser-déposer aux cas où la destination est située dans la même fenêtre que l'objet, même si cette destination n'est pas visible initialement. La **fenêtre perceptuelle** permet, grâce à la navigation en deux dimensions (2D), de s'affranchir partiellement de la contrainte de visibilité.

Dans de nombreuses situations, l'objet du glisser-déposer doit être déplacé entre deux emplacements situés dans des espaces de travail (donc des fenêtres) distincts. La navigation requise est alors 2,5D . La "demi" dimension de plus fait référence à l'empilement de fenêtres qui simule un effet de profondeur. Or, le problème de la navigation en 2,5D pour le glisser-déposer est en passe d'être résolu.

Apple Computer propose, depuis la diffusion du système MacOS 8.0, une navigation en 2,5D basée sur l'analyse de la trajectoire spatio-temporelle du pointeur de la souris ([Langer 98]). Une pause du pointeur sur un dossier provoque l'ouverture de la fenêtre du dossier et son affichage au premier plan, ce qui permet une navigation en "profondeur" dans la hiérarchie de fichiers. Cette technique pourrait facilement se généraliser à toutes les fenêtres présentes à l'écran. Le principe de la détection de pause dans les trajectoires est similaire à celui que nous avons mis en œuvre dans notre prototype de **tableau magique** (voir page 139).

Finalement, nous pensons que la combinaison de la **fenêtre perceptuelle** pour la navigation 2D et la technique d'Apple pour la navigation en profondeur permettrait de généraliser l'usage des opérations de glisser-déposer, ce qui peut représenter un gain appréciable pour les interfaces graphiques.

5. Résumé du chapitre

Le travail rapporté dans ce chapitre a pour but de valider expérimentalement les travaux de cette thèse. Avec la **fenêtre perceptuelle**, nous avons pu :

1. Montrer la faisabilité d'une interaction fortement couplée fondée sur la vision par ordinateur.
2. Montrer qu'une telle interaction présente de réels bénéfices pour l'interaction homme-machine.

A notre connaissance, il s'agit de la première étude qui allie mise en œuvre technique et validation expérimentale centrée sur l'utilisateur.

5.1. FAISABILITÉ

La **fenêtre perceptuelle** est fondée sur un suivi du visage en vision par ordinateur. Le prototype réalisé a une fréquence de fonctionnement supérieure à 15 Hz. correspondant à une latence inférieure à 65 ms. La **fenêtre perceptuelle** est donc proche du requis de latence de 50 ms énoncé au chapitre II. Le système de suivi est responsable de 26 % de la latence globale du système. Celle-ci peut être réduite en optimisant la génération du retour d'information. Le système de suivi, fondé sur la technique de corrélation, fournit des mesures statiquement stables de résolution de l'ordre de 0,5 mm.

Les performances de notre système satisfont les requis de l'interaction fortement couplée. Les premières expériences d'utilisation de la **fenêtre perceptuelle** indiquent que le système est réellement utilisable. Toutefois, le système ne remplit pas la condition d'autonomie indispensable à sa diffusion : l'initialisation du suivi du visage se fait manuellement. Les techniques présentées à la section "Coopération de techniques" du chapitre IV, qui visent l'autonomie, doivent être intégrées au système.

5.2. AVANTAGES

Deux études utilisateur ont été réalisées. Elles ont montré que la **fenêtre perceptuelle** permet de réaliser certaines tâches combinant navigation et désignation de façon plus efficace que les systèmes classiques.

La **fenêtre perceptuelle** s'inscrit dans le mouvement des systèmes à deux flux d'entrée. Son originalité tient à la mise en œuvre d'un nouveau type de flux d'entrée : le suivi des mouvements du visage par un système de vision par ordinateur. Si leur avantage, en terme de performance, est connu depuis longtemps, l'utilisation des systèmes à deux flux d'entrée reste marginal. Le gain de performance ne semble pas être un facteur suffisant pour faire accepter cette nouvelle forme d'interaction. Nous pensons que la **fenêtre perceptuelle** possède de ce point de vue des atouts originaux :

- Simplicité et coût minimal de mise en œuvre. La **fenêtre perceptuelle** met en œuvre uniquement du matériel standard. La caméra vidéo est un périphérique en passe de se démocratiser : de nombreuses stations de travail sont équipées en standard de dispositifs permettant l'entrée d'un flux vidéo (carte d'acquisition vidéo, entrée "DigitalVideo" IEEE 1394), et parfois même d'une caméra (stations de travail SGI Indy et O2). Dans ce contexte, la mise en œuvre de la navigation au visage peut se réduire au téléchargement et à l'installation d'un composant logiciel.
- Equivalence fonctionnelle. La **fenêtre perceptuelle** ne se substitue pas aux moyens de navigation standard. Elle constitue une alternative. La fenêtre graphique du système est équipée de barres de défilement parfaitement fonctionnelles. L'utilisateur peut choisir, entre les deux formes de navigation, celle qui correspond à son besoin actuel. L'équivalence fonctionnelle (au sens des propriétés CARE [Coutaz 95]) facilite la transition d'une forme d'interaction à une autre : comme pour les raccourcis clavier des menus, l'utilisateur n'est pas d'emblée confronté à une étape d'apprentissage obligatoire.

Conclusion

Motivés par l'absence de liens opérationnels entre les domaines de l'interaction homme-machine et de la vision par ordinateur, nous avons étudié dans cette thèse la conception et la mise en œuvre de techniques de vision par ordinateur au service de l'interaction fortement couplée.

Dans ce chapitre de conclusion, nous résumons notre contribution. Nous en soulignons ensuite les points forts et les faiblesses justifiant ainsi les prolongements de ce travail.

1. Résumé de la contribution

La contribution centrale de notre recherche relève à la fois de la vision par ordinateur et de l'interaction homme-machine :

- Nous avons démontré la faisabilité technique de dispositifs d'interaction fortement couplée fondés sur la vision par ordinateur.
- Nous avons montré que ces dispositifs offraient aux utilisateurs des gains en performances comparativement à l'usage unique de la souris, usage de référence en interaction homme-machine s'il en est.
- Nous avons mis en pratique nos idées avec la conception et la mise en œuvre de deux prototypes : le **tableau magique** avec une interaction au doigt, et la **fenêtre perceptuelle** qui permet de contrôler le défilement de fenêtre au moyen du visage.

2. Originalité et points forts

L'originalité et les points forts de notre contribution tiennent fondamentalement à notre approche résolument motivée par le rapprochement entre deux disciplines aux préoccupations disjointes : l'interaction homme-machine et la vision par ordinateur.

2.1. APPROCHE SCIENTIFIQUE

Notre approche s'appuie sur le constat suivant : les nouveaux paradigmes d'interaction se rapprochent de la *physicalité* et visent la *suppression des intermédiaires artificiels* dans les interfaces à manipulation directe. La vision par ordinateur peut, à l'avenir, tenir un rôle déterminant dans cette tendance à condition que les techniques interactives fondées sur elle répondent aux besoins humains. Dans ce contexte, nous considérons le couplage homme-machine au niveau des actions physiques, point de départ à toute interaction. Notre problème s'exprime donc en ces termes en deux étapes : quels sont les requis humains dans une situation de couplage étroit avec une machine ? Comment ces requis se traduisent-ils en vision par ordinateur ?

Pour répondre à cette question, nous avons défini le concept d'*interaction fortement couplée* que nous avons ensuite modélisé sous forme d'un système en boucle fermée. Ce modèle fait appel à deux théories issues de deux domaines distincts : la *théorie du contrôle* en Automatique et le *modèle du processeur humain* en IHM. La théorie du contrôle nous a permis d'introduire des métriques dont la sémantique est précise : la latence, la stabilité statique, la résolution. En nous fondant sur les résultats quantitatifs du modèle du processeur humain, nous avons identifié de manière analytique le seuil de latence et confirmé notre analyse par d'autres résultats issus de psychologie expérimentale.

Ainsi, nous avons joué sur l'articulation de plusieurs théories et pratiques pour définir un solide socle de requis centrés humains et utilisables à la mise en œuvre de dispositifs techniques. Parmi ces requis, la *latence tient lieu de priorité*. Son seuil est au voisinage de 50 ms. D'autres requis concernent le contexte d'usage : le déploiement d'une solution technique ne peut avoir lieu que si elle est fonctionne en dehors du laboratoire où elle a été conçue.

Cette approche descendante centrée sur l'homme nécessite cependant de faire des compromis lorsque l'état de l'art en matière de techniques ne peut répondre aux critères. En situation d'impasse, nous autorisons l'ajout de *contraintes* sur le contexte d'utilisation sous réserve que ces contraintes permettent à la technique d'atteindre les requis quantifiés (et notamment la latence) sans remetttre en cause la raison d'être du système interactif. L'ajout de ces contraintes, que nous avons justifiées, nous a conduit à faire des propositions originales en vision par ordinateur.

Vision par ordinateur : approche et techniques

Nous avons réalisé un ensemble de techniques de suivi en vision par ordinateur en visant en permanence la contrainte draconienne du temps de calcul maximal des algorithmes. Nous avons choisi la vision par apparence, moins chère en calculs que la vision orientée modèle, mais nous l'avons amendée de principes qui nous sont propres : nous préconisons l'*extraction d'information minimale* qui permettent de satisfaire les requis en autorisant l'*ajout contrôlé de contraintes sur l'environnement* s'il simplifie les problèmes de vision et s'il se révèle sans préjudice rédhibitoire pour l'utilisateur.

Suivant cette approche, nous avons réalisé différentes techniques de suivi que nous avons optimisées pour la satisfaction des requis. Concernant les techniques de suivi d'entité en vision par ordinateur, nos contributions couvrent :

- l'utilisation d'un modèle de couleur gaussien pour le suivi par modèle de couleur,
- la formalisation du calcul de la taille de la zone de recherche qui permet d'optimiser la vitesse maximale de la cible tolérée par le suivi par corrélation,
- la coopération de techniques de suivi visant l'autonomie et la robustesse du processus global de suivi,
- la détection du clignement des paupières pour l'initialisation automatique du modèle de couleur.

Validation expérimentale

Les techniques de vision par ordinateur ont été mises à profit dans la réalisation de deux prototypes de systèmes interactifs fondés sur une interaction fortement couplée.

Le **tableau magique** intègre une désignation au doigt pour la sélection et le déplacement d'inscriptions électroniques projetées sur un tableau blanc. Ce prototype incarne, à notre connaissance, la première réalisation utilisable d'un système de réalité augmentée intégrant les services imaginés pour le **bureau digital** ([Wellner 93b]). Les premières expériences d'utilisation tendent à confirmer le bien fondé de l'approche de réalité augmentée pour la tâche envisagée (la production d'idées). Notre prototype confirme le rôle déterminant de la vision par ordinateur dans la réalisation de ce type de système. Il confirme également l'apport de la coopération de techniques pour le suivi d'entité (dans le cas du **tableau magique** : détection de mouvement et suivi par corrélation).

Le prototype de **fenêtre perceptuelle** utilise une interaction fortement couplée asservie aux mouvements du visage de l'utilisateur pour contrôler la navigation dans une fenêtre d'interface graphique classique. Notre système réalise une interaction plus directe entre l'homme et la machine par la suppression d'intermédiaires : l'acquisition du dispositif physique à la main, et l'acquisition de l'élément de contrôle graphique avec le

pointeur. Deux expériences utilisateur montrent un gain quantitatif de performance comparé à une interaction classique à la souris. Outre le gain de performance, les utilisateurs de la **fenêtre perceptuelle** font preuve d'un temps d'apprentissage extrêmement réduit.

3. Limites et Perspectives

Tout au long de notre exposé, nous avons évoqué les limites de nos techniques ou de nos prototypes. Ces limites ouvrent autant de perspectives de recherche à court et moyen termes.

3.1. LIMITES ET PERSPECTIVES À COURT TERME

Nous résumons les limites en quatre points : longueur de la trajectoire d'interaction, liberté du mouvement, autonomie, boîte à outils de vision par ordinateur. Les trois premiers sujets sont centrés sur l'utilisateur, le dernier (boîte à outils) concerne le développeur.

Trajectoire d'interaction

L'expérience montre que la trajectoire d'interaction nécessaire à la sélection et au déplacement d'inscriptions sur le **tableau magique** est trop longue (voir page 140). Nous pensons intégrer l'approche de **Flatland** pour le groupement automatique des inscriptions et la reconnaissance de gestes simples tels que la main ouverte ou fermée, pour exprimer de manière directe la commande de déplacement dont la fréquence est élevée.

Liberté du mouvement

Le suivi par corrélation est limité par le modèle de mouvement : seules les translations dans un plan parallèle au plan de l'image sont tolérées (voir page 101). Cette limitation est gênante en particulier pour le suivi du doigt du **tableau magique** (page 132). Nous envisageons deux extensions de la technique de suivi par corrélation que nous présenterons plus loin avec les perspectives. Concernant le **tableau magique**, nous pensons utiliser un ensemble de motifs représentant la cible sous différentes orientations.

Autonomie

L'initialisation du suivi de visage de la **fenêtre perceptuelle** est manuelle (voir page 155). Nous souhaitons intégrer les techniques de coopération de suivis afin de rendre le prototype autonome. Cette étape est un préliminaire indispensable à l'expérimentation à grande échelle (c'est-à-dire en usage courant).

Notre coopération de techniques de suivi du visage manque, elle aussi, d'autonomie (voir page 111). Cette lacune vient de la phase de calibrage nécessaire à la détermination des différents seuils utilisés pour le bon fonctionnement des techniques de suivi. Or le recalibrage est nécessaire

dès que les conditions d'usage varient sensiblement par rapport aux conditions initiales. Il conviendrait d'étudier la détection de ces changements afin de relancer le calibrage de manière automatique.

Boîte à outils

Les techniques et composants logiciels développés au cours de cette thèse ouvrent la voie à de nouvelles applications en interaction homme-machine. Nous souhaitons rendre nos développements accessibles sous la forme d'une bibliothèque de services. Nous prendrons soin de concevoir une interface de programmation qui offre des services pertinents du point de vue de la réalisation de systèmes interactifs, masquant notamment les détails d'implémentation liés aux techniques de vision par ordinateur et aux matériels d'acquisition du flux vidéo.

Ce travail ouvre également de nombreuses perspectives de recherche à plus long terme. Nous présentons ici celles qui nous semblent les plus prometteuses.

3.2. PERSPECTIVES À MOYEN TERME

Il conviendrait de pousser plus avant notre réflexion sur nos fondements théoriques, liant l'automatique, l'IHM et la vision par ordinateur. Du côté vision, il faut obtenir davantage de robustesse et d'autonomie (sans perdre de vue la latence !). Du côté applicatif, les pistes sont nombreuses. Nous en citerons deux.

IFC : retour sur la théorie du contrôle

Concernant la modélisation de l'interaction fortement couplée, nous avons justifié la valeur maximale de latence pour les dispositifs impliqués dans une interaction fortement couplée. Nous pensons que cette modélisation peut être affinée en fouillant la théorie du contrôle. Par l'application de cette théorie, il semble possible de dériver une estimation quantitative plus fine de la latence et des conditions d'oscillation.

Côté vision : robustesse et autonomie

Concernant le suivi d'objet en vision par ordinateur, nous souhaitons enrichir l'information extraite par le suivi par corrélation et compléter sa qualité de stabilité par la qualité de robustesse. Nous pensons concevoir un suivi fondé sur les estimations de plusieurs suivis par corrélation dont les cibles seraient réparties sur toute la surface de l'entité à suivre. En considérant les relations spatiales des cibles élémentaires, nous pensons pouvoir extraire les paramètres de mouvement affine de l'entité à suivre. De plus, nous pensons réaliser un suivi plus robuste grâce à la redondance induite par la mise en œuvre de plusieurs suivis simultanément.

Côté applicatif

Parmi les nombreuses applications interactives dont la réalisation est rendue possible avec la vision par ordinateur temps réel, deux applications particulières retiennent notre attention :

1 Nous avons rapporté au chapitre I les avantages de la parallaxe de mouvement pour la perception en trois dimensions. La vision par

ordinateur a le potentiel d'offrir un couplage d'excellente qualité entre le point d'observation et l'image générée, permettant ainsi la banalisation sur les stations de travail de la perception de l'effet de profondeur. La transparence du couplage serait assurée par la coopération de techniques de suivi pour l'acquisition automatique du visage, un délai imperceptible de la réponse du système aux mouvements du visage serait assuré par l'échantillonnage du flux vidéo à la fréquence des champs (60 Hz), enfin, la stabilité statique de la position d'observation serait assurée par un suivi fondé sur le suivi par corrélation.

2 D'autre part, les travaux de Black et Yacoob ([Black 97]) méritent attention. Ces travaux ont montré qu'il était possible d'extraire avec précision certaines caractéristiques du visage (dans leur cas, l'expression) grâce à une extraction préliminaire des paramètres de mouvement planaire du visage. Par plusieurs suivis par corrélation simultanés, nous espérons extraire les paramètres de mouvement planaire du visage en temps réel (le prototype de Black et Yacoob nécessite plusieurs minutes de calcul par image). Grâce à cela, nous espérons analyser avec précision la direction du regard pour la détection du point de focalisation de l'utilisateur sur l'écran. Nous espérons également extraire les paramètres de mouvement de la bouche pour améliorer les techniques de reconnaissance vocale en environnement bruité.

En somme, avec la vision par ordinateur au service de l'IHM, "le futur ne manque pas d'avenir" (Philippe Meyer, cité dans [Coutaz 98]).

Annexe A

Techniques de vision par ordinateur à usage général

Nous présentons trois techniques de vision par ordinateur à usage général : le seuillage, l'analyse en composantes connexes, et une technique générale de calcul de validité.

1. Seuillage

1.1. PRINCIPE

Le principe du seuillage est de considérer que seules les valeurs supérieures à un certain seuil sont représentatives de la propriété mesurée dans l'image. Les valeurs inférieures au seuil ne sont pas représentatives ou sont dues au bruit.

On dit que S est l'image *seuillée* de I et on calcule S par :

$$\forall x, \forall y, S(x, y) = \begin{cases} I(x, y) \geq s, & 1 \\ I(x, y) < s, & 0 \end{cases} \qquad (1)$$

où s est la valeur du seuil. L'image S est binaire. Les pixels représentatifs de la propriété mesurée ont la valeur 1, les autres ont la valeur 0. Le problème du seuillage consiste à choisir la valeur de s. Si la valeur de s est trop basse, de nombreux pixels de l'image S sont considérés comme représentatifs alors qu'ils ont pour cause le bruit de caméra. Si la valeur de s est trop haute, certains pixels représentatifs ne sont pas classés comme tel.

1.2. CHOIX DU SEUIL

Stafford-Fraser ([Stafford-Fraser 96a]) propose une phase de calibrage initiale où l'on calcule la moyenne μ et l'écart-type σ de la variation de la mesure au cours du temps. Ces statistiques sont calculées sur une scène qui ne présente pas la propriété étudiée (en particulier, la cible n'est pas présente dans le champ de la caméra). Ces statistiques représentent donc le bruit de caméra. Une fois cette phase de calibrage terminée, on détermine un seuil par la formule suivante :

$$s = \mu + \alpha \cdot \sigma \qquad (2)$$

où μ et σ sont la moyenne et l'écart-type estimés pendant la phase de calibrage, et α un coefficient multiplicatif permettant de choisir le seuil en fonction de la probabilité que la valeur mesurée soit due au bruit. La valeur de α est choisie de telle façon que cette probabilité soit faible. Faisant l'hypothèse que le bruit de caméra suit une loi normale centrée sur μ et d'écart-type σ, la probabilité qu'une valeur supérieure au seuil s soit due au bruit est donnée par le complément à 1 de la loi Normale :

$$p = 1 - \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot x^2} \, dx \qquad (3)$$

Pour la valeur $\alpha = 2$ la probabilité est $p = 0,02$. Pour une telle valeur de α, on estime que seulement 2 % des pixels de S sont dus au bruit de caméra.

2. Analyse en composantes connexes

2.1. DÉFINITION

On dit qu'un pixel A est connexe à un pixel B si le pixel A est l'un des huit pixels du voisinage de B (la propriété est réflexive : B est alors connexe au pixel A). Une composante connexe d'une image binaire est un ensemble de pixels de valeur 1 et dont tous les membres sont connectés entre eux par au moins une chaîne de pixels connexes. La figure 1 illustre deux composantes connexes dans une image. L'analyse en composante connexe consiste à extraire les informations de toutes les composantes

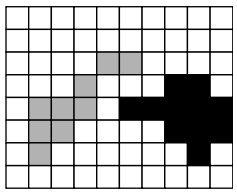


Figure 1
Analyse en composantes connexes
Les carrés blancs représentent les pixels à 0. Les pixels noirs et les pixels gris représentent les deux composantes connexes présentes dans l'image.

connexes de l'image. Nous calculons les informations suivantes pour chaque composante :

- surface (nombre de pixels),
- boîte englobante (abscisses et ordonnées maximales et minimales des pixels de la composante),
- statistiques de la distribution spatiale des pixels de la composante.

2.2. IMPLÉMENTATION EFFICACE

Nous réalisons une implémentation efficace de l'analyse en composante connexe grâce à un algorithme ne nécessitant qu'un seul parcours des pixels de l'image. Les pixels sont parcourus de haut en bas et de gauche à droite. Le traitement réalisé sur chaque pixel est le suivant :

1 * Si la valeur du pixel courant est nulle, on passe au pixel suivant.
  * Si la valeur du pixel courant n'est pas nulle, c'est que le pixel fait partie d'une composante. On parcourt le "passé" de ce pixel, c'est-à-dire l'ensemble de ses pixels connexes déjà parcourus par l'algorithme.

2 * Si tous les pixels du passé ont la valeur nulle, on crée une composante et on affecte le pixel courant à cette composante.
  * Si au moins un des pixels du passé n'a pas la valeur nulle mais que tous les pixels du passé qui n'ont pas la valeur nulle sont affectés à une même composante, alors on affecte le pixel courant à cette composante.
  * Si plusieurs pixels du passé n'ont pas la valeur nulle, et qu'ils sont affectés à des composantes différentes, alors on affecte le pixel courant à l'une de ces composantes, et on enregistre *l'équivalence des composantes* des pixels du passé : le pixel courant connecte ces différentes composantes. Elles ne forment donc qu'une seule composante connexe.

À chaque ajout d'un pixel dans une composante, on met à jour les informations de la composante. En fin de parcours, les informations des composantes équivalentes sont fusionnées. La fusion ne pose aucun problème concernant la surface (addition des surfaces) et la boîte englobante (calcul de minima et maxima). Concernant les statistiques de la distribution spatiale des pixels, nous utilisons la décomposition de la covariance en somme des carrés détaillée sur l'équation 15 page 92.

3. Calcul de validité

La technique présentée ici a pour objectif de donner une valeur de validité à une estimation de position calculée par une technique de suivi. Cette technique est applicable à toute technique de suivi fournissant la position

de la cible sous forme d'un vecteur de paramètres. La validité de cette technique est dépendante de l'hypothèse que la distribution des paramètres, pour un ensemble de vecteurs correspondant à des positions valides de la cible, suit une loi Normale.

Soit p un vecteur sur l'ensemble des paramètres décrivant la cible. Par exemple, dans le cas des techniques de suivi calculant une boîte englobante (voir le chapitre IV page 79), le vecteur p a deux composantes :

$$p = \begin{bmatrix} l \\ h \end{bmatrix} \tag{4}$$

où l et h sont la largeur et la hauteur de la boîte englobante.

Dans le cas des techniques de suivi calculant les statistiques de la distribution spatiale des pixels représentant la cible (voir le chapitre IV page 78), nous choisissons un vecteur à trois composantes :

$$p = \begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{bmatrix} \tag{5}$$

où les σ représentent les variances et la covariance de la distribution spatiale des pixels.

Durant une phase de calibrage, la technique de suivi est exécutée de façon supervisée. Un opérateur enregistre un nuage de points p correspondant aux paramètres de la cible dans de nombreux cas de localisations valides. La moyenne μ et la matrice de covariance C de cet ensemble de points sont calculées.

Durant l'exécution du suivi, la mesure de validité d'une estimation des paramètres de la cible p est donnée par la loi Normale :

$$v = \frac{1}{(\sqrt{2\pi})^n \cdot \sqrt{det(C)}} \cdot e^{-\frac{1}{2} \cdot (p-\mu)^t \cdot C^{-1} \cdot (p-\mu)} \tag{6}$$

où n est la dimension de p.

Annexe B *Considérations d'implémentation*

Au premier paragraphe, nous notons l'absence, dans la communauté de recherche en vision par ordinateur, de bibliothèque de services fondamentaux pour la vision par ordinateur "temps réel". La vision par ordinateur temps réel est destinée à traiter le flux vidéo au fur et à mesure de son acquisition, par opposition à la vision par ordinateur "hors-ligne" dont les temps de calcul imposent un pré-stockage du flux vidéo. Nous présentons notre approche qui a pour but de capitaliser les services logiciels fondamentaux, et faciliter le partage des composants logiciels qui implémentent ces services.

Dans le paragraphes suivant, nous détaillons un service fondamental qu'il serait souhaitable de voir apparaître dans une bibliothèque de services : l'acquisition du flux vidéo.

1. Bibliothèque de services

Le domaine de recherche en interaction homme-machine a une longue tradition de conception de bibliothèques de services logiciels pour la conception des interfaces graphiques ([Nye 88], [Ousterhout 94], [Apple 96], [Roseman 96], [Chan 97], [Bederson 98]). Les services présents dans les bibliothèques résultent de l'abstraction des services fondamentaux du domaine. Dans le cas des interfaces graphiques, l'abstraction concerne les dispositifs d'entrée / sortie (clavier, souris, écran), la gestion du fenêtrage, l'affichage des éléments de contrôle courant (menu, boutons), et la gestion des évènements utilisateur. Le principe de ces bibliothèques est de définir une interface de programmation pour la manipulation des concepts

fondamentaux, et d'offrir les services logiciels qui implémentent cette interface.

1.1. APPORT

L'apport est double :

- Capitalisation de code : les services sont génériques et l'interface de programmation des services est indépendante de toute application spécifique. La généricité assure que les services de la bibliothèque sont ré-utilisables dans de nombreux projets. Le concepteur d'application peut puiser dans la bibliothèque pour la réalisation des services fondamentaux, et n'a donc pas à "ré-inventer la roue" à chaque nouveau projet.
- Partage : l'interface de programmation est conçue pour assurer l'indépendance des programmes vis-à-vis des spécificités du matériel sur lequel ils sont exécutés. Un même programme peut alors être compilé sans modification sur des plates-formes matérielles différentes. La bibliothèque X-Window ([Nye 88]) a permis l'exécution, sur toute la gamme des plate-forme UNIX, de l'interface utilisateur d'un même programme. La bibliothèque Tcl / Tk ([Ousterhout 94]) perment de réaliser un logiciel possédant une interface graphique et de l'exécuter, sans modification ni re-compilation, sur la majorité des plates-formes UNIX (Linux, BSD, IRIX, Solaris, HP UX, etc...), sur plate-forme MacOS et sur les plate-forme Windows (98 / NT).

Les bibliothèques ont largement contribué au développement des interfaces graphiques dans la majorité des applications et sur toutes les plates-formes matérielles.

L'apport de bibliothèques fédérant les développements de vision par ordinateur est reconnue par la communauté. De nombreuses propositions ont déjà vu le jour : un site regroupant les ressources de vision par ordinateur sur internet ([Huber 99]) liste plus de quarante liens sur des projets de bibliothèques de service de vision. Cependant, la majorité de ces projets sont orientés vers le traitement "hors-ligne" des images ([Zuke 97]). Quelques bibliothèques offrent les services nécessaires au traitements de vision par ordinateur en temps réel ([XVision], [Microsoft]) mais ces bibliothèques ne prennent pas en compte l'aspect multi plate-forme (XVision s'exécute sur UNIX uniquement, VisionSDK sur Windows).

1.2. CONSTAT

À l'heure actuelle, l'apparition d'une bibliothèque fédérant les services logiciels pour la vision par ordinateur temps réel fait cruellement défaut. Si certains standards se sont imposés concernant le format de stockage sur disque des images, il n'en est pas de même pour la représentation et la manipulation en mémoire du flux vidéo. En pratique, les développements

sont souvent réalisés sur les services de bas niveau du système d'exploitation de la plate-forme d'accueil. Les logiciels sont dépendants de la plate-forme et ne peuvent être exécutés sur d'autres plates-formes. En conséquence, il est difficile pour une équipe de recherche d'exécuter les logiciels des autres équipes à des fins de comparaison par exemple.

Pire, la communauté ne s'étant pas encore fixé sur un standard concernant les concepts fondamentaux que sont l'image et le flux vidéo, chaque équipe définit ses propres abstractions. En conséquence, le partage de composants logiciels est trés difficile. Il est souvent plus facile de recoder une technique dont ont connait le principe plutôt que de tenter de porter le code source dans son propre environnement.

1.3. NOTRE APPROCHE

Nous avons orientés nos développements logiciels vers la conception d'une bibliothèque de services fondamentaux pour la vision par ordinateur "temps réel". Nos objectifs concernent la capitalisation de code et le partage. Concernant le partage, nous souhaitons permettre l'exécution des logiciels sur différents environnements matériels, mais aussi permettre la mise en commun de composants logiciels provenant de sources différentes dans la réalisation d'un logiciel.

Notre bibliothèque se nomme "**TclVision**" car elle est fondée sur un interpréteur **Tcl**[1] ([Ousterhout 94]). **Tcl** est un langage de script conçu dès le départ avec l'objectif de faciliter l'intégration de composants logiciels de provenances diverses. Il est utilisé dans cette optique par de nombreux travaux de recherche ([Roseman 96], [Stafford-Fraser 96a], [Zuke 97], [Bederson 98], [MacKay 98]).

Concernant nos objectifs, les avantages de **Tcl** sont les suivants :

- Le code source de **Tcl** est en libre accès.
- L'extension de **Tcl** par de nouveaux services est simple.
- **Tcl** est multi plate-forme. Il est maintenu sur la majorité des plates-formes UNIX, sur MacOS et sur Windows (98, NT).
- **Tcl** est distribué avec **Tk**. **Tk** est une extension de **Tcl** offrant les services fondamentaux pour la conception des interfaces graphiques.

Nous étendons **Tcl** par l'ajout de composants qui implémentent les services fondamentaux de la vision par ordinateur temps réel. Ces composants sont réalisés dans un langage compilé (C++) afin d'assurer des performances élevées. Le recours à l'optimisation en assembleur permettrait sans doute d'améliorer sensiblement les performances, mais limiterait fortement les possibilités de partage.

1. "Tool Command Language"

Nous illustrons maintenant l'apport de la conception d'une bibliothèque pour un service fondamental de la vision par ordinateur temps réel : l'acquisition vidéo.

2. *Acquisition du flux vidéo*

L'acquisition du flux vidéo est la fonction chargée d'acheminer le flux vidéo numérique en mémoire. En règle générale, la source du flux vidéo est une caméra fournissant un flux analogique au format PAL, ou NTSC. Une carte d'acquisition vidéo, installée sur l'ordinateur, est chargée de la conversion de ce signal analogique en un flux numérique.

2.1. INDÉPENDANCE VIS-À-VIS DU MATÉRIEL

Il y a quelques années (5 ans environ), l'acquisition du flux vidéo passait par l'achat d'une carte dédiée. La carte était livrée avec une bibliothèque permettant d'accéder aux services de la carte par l'intermédiaire d'une interface de programmation spécifique. En conséquence, le système était dépendant de la carte et ne pouvait s'exécuter que sur la plate-forme ayant servi au développement.

Depuis sont apparues plusieurs bibliothèques de services pour l'acquisition du flux vidéo. Ce sont les bibliothèques QuickTime sur plate-forme MacOS ([Apple 93b]), VideoLibrary sur plate-forme SGI IRIX et Video For Windows sur plate-forme Windows. Ces bibliothèques ont toutes en commun de proposer une interface de programmation standard à laquelle doivent se conformer les constructeurs de carte d'acquisition vidéo. Ainsi, le code source gérant l'acquisition vidéo n'est plus dépendant de la carte d'acquisition mais uniquement de la librairie à laquelle il fait appel. En pratique, de nombreux problèmes subsistent :

- Les différentes bibliothèques sont totalement incompatibles entre elles. La portabilité inter plate-forme n'est donc pas assurée.
  Apple a récemment porté la bibliothèque QuickTime 4 sur plate-forme Windows, offrant une interface de programmation unifiée avec la plate-forme MacOS. Cependant, à notre connaissance, aucun constructeur ne s'est encore conformé à l'interface QuickTime sur plate-forme Windows.
- Si l'indépendance matérielle est assurée sur plate-forme MacOS et SGI IRIX (grâce aux bibliothèques QuickTime et VideoLibrary), il n'en est pas de même sur la plate-forme Windows. Bien que de nombreux constructeurs se soient conformés à l'interface de programmation Video For Windows, certains constructeurs, tels Matrox, ont préféré continuer à distribuer leur propre interface de programmation. Il existe

donc une partition entre les logiciels supportant les différentes interfaces. La bibliothèque VisionSDK ([Microsoft]) résoud partiellement ce problème en définissant une interface de programmation au-dessus de Video For Windows et de la bibliothèque de Matrox. Cependant les difficultés ne semblent pas devoir s'arrêter là avec l'annonce de la bibliothèque Direct X dont certains services sont destinés à remplacer ceux de Video For Windows.

Notre bibliothèque, **TclVision**, propose une interface de programmation indépendante de la plate-forme pour l'acquisition du flux vidéo. Cette interface est actuellement implémentée au dessus de QuickTime sur MacOS, de la VideoLibrary sur SGI IRIX et en cours de portage au dessus de la Matrox Imaging Library sur Windows. Grâce à cette abstraction matérielle, nos prototypes fonctionnent indifféremment sur plate-forme MacOS et SGI IRIX. C'est notament le cas des prototypes de **tableau magique** (chapitre V) et de **fenêtre perceptuelle** (chapitre VI).

2.2. FORMAT DU FLUX VIDÉO

Le flux vidéo analogique, qu'ils soit au format PAL ou NTSC, est entrelacé. Le flux vidéo est consituté d'une succession de couples d'images appelées "champ pair" et "champ impair". Dans le cas du format PAL, chaque image de type "champ" est constituée de 384 lignes. Les couples d'images sont entrelacés spatialement lors de l'affichage afin de représenter une image de 768 lignes : une ligne sur deux provient du champ impair, l'autre du champ pair. La fréquence d'affichage d'une image est donc la moitié de la fréquence des champs puisqu'il faut deux champ pour afficher une image. Dans le cas du format PAL, la fréquence des champs est 50 Hz, celle des images 25 Hz[1]. Les raisons de l'entrelacement sont historiques : lorsque les formats de flux vidéo analogique ont été conçus, les concepteurs ont vu dans l'entrelacement un moyen de diffuser une image de meilleure définition au prix d'une fréquence d'affichage moins élevée, mais jugée satisfaisante.

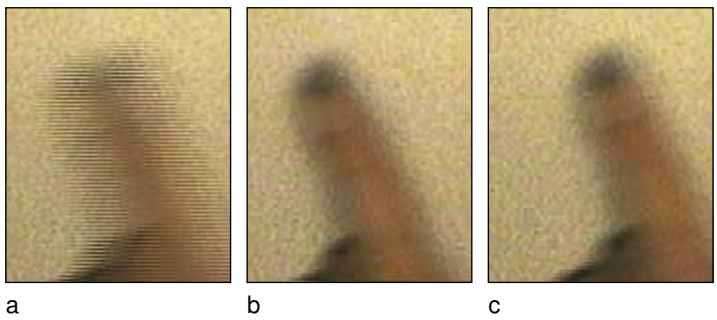


Figure 1
Entrelacement dans un flux vidéo analogique
Une image du flux vidéo (a) (un doigt en déplacement) est le résultat de l'entrelacement d'un champ impair (b) et d'un champ pair (c). Dans les images (b) et (c), les lignes des champs ont été dupliquées afin de reproduire les proportions de l'image (a).

1. Le format NTSC est constitué deux deux champs de 240 lignes à la fréquence de 60 Hz., équivalent à une fréquence d'images de 480 lignes de 30 Hz.

Du point de vue de la vision par ordinateur, l'entrelacement n'est pas un problème dans le cas d'une scène statique car la scène n'a pas varié entre les deux champs. La figure 1 illustre le problème dans le cas d'une scène dynamique (cas d'un objet en déplacement) : l'image traitée (figure 1a) représente l'apparence de la scène à deux instants différents. Le problème est en règle générale résolu en ne traitant qu'un seul champ du flux vidéo. La technique courante est de capturer une image entière du flux (les deux champs), puis de recopier un pixel sur deux dans une autre image. En conséquence, un seul des deux champs est traité et la fréquence de capture est celle des images, c'est à dire 25 ou 30 Hz. selon le format (PAL ou NTSC). De fait, les concepteurs des systèmes de vision temps réel ne cherchent pas à assurer une fréquence de fonctionnement de leur système supérieure à 25 ou 30 Hz. Pourtant, il est théoriquement possible d'atteindre une meilleure résolution de discrétisation temporelle en traitant tous les champs à 50 ou 60 Hz. Le bénéfice du traitement du flux à ces fréquences est clairement démontré au paragraphe "Taille de la zone de recherche" page 98. Nous donnons également, au paragraphe "Vitesse maximale tolérée" page 133, un exemple d'application nécessitant de traiter le flux à plus de 30 Hz.

L'implémentation de l'acquisition du flux vidéo dans notre bibliothèque **TclVision** ne permet pas, pour le moment, l'acquisition à la fréquence des champs. L'acquisition à cette fréquence ne représente pas un besoin courant en dehors du domaine applicatif de la vision par ordinateur, c'est pourquoi les bibliothèques sur lesquelles repose **TclVision** (QuickTime et VideoLibrary) n'offrent pas directement ce service.

Bibliographie

[Accot 97] Accot, J. et Zhai, S. "*Beyond Fitts' Law: Models for Trajectory-Based HCI Tasks*", ACM conference on Computer-Human Interaction (CHI), 1997.

[Ahlberg 94] Ahlberg, C. et Shneiderman, B. "*Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays*". ACM conference on Computer-Humain Interaction (CHI), pp 313-317, 1994.

[Annedouche 99] Annedouche, S. Loup, B. et Prodhomme, M. "*Le Tableau Magique*", rapport de projet de 3ème année de l'École Nationale Supérieure en Informatique et Mathématiques Appliquées de Grenoble (ENSIMAG), juin 1999. Disponible sur le site :
http://iihm.imag.fr/publs/1999/

[Apple 93a] Apple Computer Inc. "*Apple Desktop Bus Mouse II: Description and Specifications*". Apple Technical Info Library (TIL), article 11227, juin 1993. Disponible à l'adresse :
http://til.info.apple.com/techinfo.nsf/artnum/n11227

[Apple 93b] Apple Computer Inc. "*Video Digitizer Component*", dans "Quicktime Components" Collection Inside Macintosh, Addisson Wesley, 1993. Disponible sue le site :
http://developer.apple.com/techpubs/

[Apple 94] Apple Computer Inc. "*Imaging With QuickDraw*". Collection Inside Macintosh, Addisson Wesley, 1994. Disponible à l'adresse :
http://developer.apple.com/techpubs/

[Apple 96] Apple Computer Inc. "*Programmer's Guide to MacApp*". Collection Inside Macintosh, Addisson Wesley, 1996. Disponible sur le site :
http://developer.apple.com/techpubs/macos8/DevTools/MacApp/macapp.html

[Ascension 99] Ascension Technology Inc. "*Flock of Birds*" Spécification de produit. Disponible sur le site :
http://www.ascension-tech.com/

[Azarbayejani 93] Azarbayejani, A. Starner, T. Horowitz, B. and Pentland, A. "*Visually Controlled Graphics*". IEEE Pattern Analysis and Machine Intelligence (PAMI) 15 (6), Juin 1993. Également M.I.T. Media Laboratory Perceptual Computing Technical Report No. 180. Disponible à l'adresse :
http://www-white.media.mit.edu/cgi-bin/tr_pagemaker#TR180

[Azarbayejani 96] Azarbayejani, A. and Pentland, A. "*Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features*", International Conference on Pattern Recognition (ICPR), vol. IV, Vienne, Autriche, août 1996. Également M.I.T. Media Laboratory Perceptual Computing Technical Report No. 363. Disponible à l'adresse :
http://www-white.media.mit.edu/cgi-bin/tr_pagemaker#TR363

[Azuma 93] Azuma, R. T. "*Tracking Requirements for Augmented Reality*", Communication of the ACM n. 7, p. 50-51, Juillet 1993.

[Basu 96] Basu, S. Essa, I. et Pentland, A. "*Motion Regularization for Model-Based Head Tracking*", International Conference on Pattern Recognition (ICPR), p. 611-616, 1996.

[Bederson 98] Bederson, B. et Meyer, J. "*Implementing a Zooming User Interface: Experience Building Pad++*", Software: Practice and Experience, 1998. Disponible sur le site :
http://www.cs.umd.edu/hcil/pad++/papers/

[Bérard 96] Bérard, F. et Coutaz, J. "*Coopération de Techniques Sensorielles pour une Interaction Écologique*", actes des 8ème journées sur l'Interaction Homme-Machine (IHM), Grenoble, 1996. Disponible à l'adresse :
http://iihm.imag.fr/publs/1996/IHM96_Comedi.Fr.pdf

[Bérard 97] Bérard, F. Coutaz, J. et Crowley, J.L. "*Robust Computer Vision for Computer Mediated Communication*", IFIP conference on Human-Computer Interaction (INTERACT), 1997. Disponible à l'adresse :
http://iihm.imag.fr/publs/1997/INTERACT97_VirtWindow.pdf

[Bérard 99a] Bérard, F. "*The Perceptual Window: Head Motion as a New Input Stream*", IFIP conference on Human-Computer Interaction (INTERACT), p. 238-244, 1999. Disponible à l'adresse :
http://iihm.imag.fr/publs/1999/INTERACT99_PWindow.pdf

[Bérard 99b] Bérard, F. "*The Peceptual Window Movies*". Démonstration en ligne de l'équipe IIHM, 1999. Disponible à l'adresse :
http://iihm.imag.fr/demos/pwindow/

[Bérard 99c] Bérard, F. "*The MagicBoard*". Démonstration en ligne de l'équipe IIHM, 1999. Disponible à l'adresse :
http://iihm.imag.fr/demos/magicboard/

[Balakrishnan 97] Balakrishnan, R. et MacKenzie, I. S. "*Performance differences in the fingers, wrist, and forearm in computer input control*", ACM conference on Computer-Human Interaction (CHI), p. 303-310. 1997.

[Black 91] Black, M. J. et Anadan, P. "*Robust Dynamic Motion Estimation Over Time*", Proc. Computer Vision and Pattern Recognition (CVPR), p. 296-302, 1991. Disponible sur le site :
http://www.parc.xerox.com/spl/members/black/papers.html

[Black 97] Black, M. J. and Yacoob, Y. "*Recognizing facial expressions in image sequences using local parameterized models of image motion*" International Journal of Computer Vision, 25(1), pp. 23-48, 1997. Disponible sur le site :
http://www.parc.xerox.com/spl/members/black/papers.html

[Black 98a] Black, M. Bérard, F. Jepson, A. Newman, W. Saund, E. Socher, G. and Taylor, M. "*The Digital Office: Overview*", AAAI Spring Symposium on Intelligent Environments, Stanford, California, 1998.

[Black 98b] Black, M. J. et Jepson, A. D. "*Recognizing temporal trajectories using the condensation algorithm*". International conference on Automatic Face and Gesture Recognition (AFGR), 1998. Disponible sur le site :
http://www.parc.xerox.com/spl/members/black/papers.html

[Buxton 86] Buxton, W. et Myers, B. "*A Study in Two-Handed Input*". ACM conference on Computer-Human Interaction (CHI), p. 321-326, 1986.

[Cadoz 96] Cadoz, C. "*Réintroduire les sensations physiques*", La Recherche n. 285, p. 80-84, Mars 1996.

[Card 83] Card, S. Moran, T. Newell, A. "*The Psychology of Human-Computer Interaction*", Lawrence Erlbaum Associates, 1983.

[Chan 97] Chan, P. et Lee, R. "*The Abstract Window Toolkit*", dans "The Java Class Libraries, Second Edition, Volume 2", Adisson Wesley, 1997.

[Chomat 99] Chomat, O. et Crowley, J. L. "*Probabilistic Recognition of Activity using Local Appearance*", IEEE conference on Computer Vision and Pattern Recognition (CVPR), 1999.

[Cipolla 98] Cipolla, R. et Pentland, A. "*Computer Vision for Human-Machine Interaction*". Cambridge University Press, Cambridge, UK, 1998.

[Collet 99] Collet, C. "*Capture et suivi du regard par un système de vision dans de contexte de la communication homme-machine*", Thèse de Doctorat de l'École Normale Supérieure de Cachan, 15 janvier 1999.

[Coutaz 95] Coutaz, J. Nigay, L. Salber, D. Blandford, A. May, J. et Young, R. "*Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE properties*", IFIP conference on Human-Computer Interaction (INTERACT), p. 115-120, 1995.

[Coutaz 96] Coutaz, J. Bérard, F. et Crowley, J.L. "*Coordination of perceptual processes for Computer Mediated Communication*", Second International Conference on Automatic Face and Gesture Recognition (AFGR), 1996.

[Coutaz 98a] Coutaz, J. "*Interfaces Homme-Machine : le Futur ne Manque pas d'Avenir*", Conférence invitée, Proc. ERGO-IA'98, Ed. ESTIA/ILS, p. 43-55, 1998.

[Coutaz 98b] Coutaz, J. Bérard, F. Carraux, E. Crowley, J. "*Early experience with the mediaspace CoMedi*", IFIP Working Conference on Engineering for Human-Computer Interaction (EHCI), 1998. Disponible sur le site : http://iihm.imag.fr/publs/1998/

[Coutaz 99] Coutaz, J. Bérard, F. Carraux, E. Astier, W. et Crowley, J.L. "*CoMedi: Using Computer Vision to Support Awareness and Privacy in Mediaspaces*". ACM conference on Computer-Human Interaction (CHI), extended abstracts (Video demo), p.13-14, 1999. Disponible sur le site : http://iihm.imag.fr/publs/1999/

[Crowley 81] Crowley, J. L. "*A Representation for Visual Information*", Doctoral Dissertation, Carnegie Mellon University, Nov 1981.

[Crowley 94a] Crowley, J. L. et Bedrune, J. M. "*Integration and Control of Reactive Visual Processes*", European Conference on Computer Vision, (ECCV), 1994.

[Crowley 94b] Crowley, J. L. et Christensen, H. I. "*Vision as Process*", Springer Verlag, Heidelberg, 1994.

[Crowley 95] Crowley, J.L. Bérard, F. et Coutaz, J. "*Finger Tracking as an Input Device for Augmented Reality*", International Workshop on Automatic Face and Gesture Recognition (AFGR), 1995.

[Deering 92] Deering, M. "*High Resolution Virtual Reality*". ACM conference on Computer Graphics (SIGGRAPH), pages 195-202, 1992.

[Douglas 99] Douglas, S. A. Kirkpatrick, A. E. et Scott MacKenzie, S. "*Testing Pointing Device Performance and User Assessment with the IS0 9241, Part 9 Standard*". ACM conference on Computer-Human Interaction (CHI), p. 215-222, 1999.

[Dubois 99] Dubois, E. et Nigay, L. "*Classification Space for Augmented Surgery, an Augmented Reality Case Study*", IFIP conference on Human-Computer Interaction (INTERACT), pp.353-359, 1999.

[Elrod 92] Elrod, S. Bruce, R. Gold, R. Goldberg, D. Halasz, F. Janssen, W. Lee, D. MacCall, K. Pedersen, E. Pier, K. Tang, J. et Welch, B. "*LIVEBOARD: a Large Interactive Display Supporting Goup Meetings, Presentations and Remote Collaboration*", ACM conference on Computer-Human Interaction (CHI), p. 599-607, 1992.

[Essa 95] Essa, I. "*Analysis, Interpretation, and Synthesis of Facial Expressions*", PhD Thesis of the Massachusetts Institute of Technology, février 1995.

[Fitts 53] Fitts, P. M. "*The information capacity of the human motor system in controlling the amplitude of movement*", Journal of Experimental Psychology, pp. 381-391, 47, 6, 1953.

[Fitzmaurice 95] Fitzmaurice, G. Ishii, H. et Buxton, W. "*Bricks: Laying the Foundations for Graspable User Interfaces*". ACM conference on Computer-Human Interaction (CHI), 1995. Disponible sur le site : http://www.dgp.utoronto.ca/people/GeorgeFitzmaurice/

[Fitzmaurice 96] Fitzmaurice, G. "*Graspable User Interfaces*". PhD Thesis, Computer Science Deparment, University of Toronto, 1996. Disponible sur le site : http://www.dgp.utoronto.ca/people/GeorgeFitzmaurice/

[Fjeld 98] Fjeld, M.Bichsel, M. et Rauterberg, M. "*BUILD-IT: An Intuitive Design Tool Based on Direct Object Manipulation*". Dans I. Wachsmut & M. Frölich (eds.) Gesture and Sign Language in Human-Computer Interaction (GW), Lecture Notes in Artificial Intelligence, Vol. 1371, pp. 297-308. Berlin: Springer-Verlag, 1998.

[Foley 82] Foley, J. et Van Dam, A. "*Fundamentals of Interactive Computer Graphics*", Addison-Wesley, 664 p., 1982.

[Gaver 95] Gaver, W. Smets, G. and Overbeeke, K. "*A Virtual Window on a Media Space*", ACM conference on Computer-Human Interaction (CHI), p. 257-264, 1995.

[Graf 96] Graf, H. P. Cosatto, E. Gibbon, D. Kocheisen, M. et Petajan, E. "*Multi-Modal System for Locating Heads and Faces*". IEEE conference on Automatic Face and Gesture Recognition (AFGR), 1996.

[Granlund 78] Grandlund, G. H. "*In Search of a General Picture Processing Operator*", Computer Graphics and Image Porcessing, vol. 8, p. 155-173, 1978.

[Guiard 87] Guiard, Y. "*Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model*". Journal of Motor Behavior 19(4). p. 486-517, 1987.

[Hager 98] Hager, G. D. et Belhumeur, P. N. "*Efficient Region Tracking With Parametric Models of Geometry and Illumination*". IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), vol. 20, no. 10, octobre 1998.

[Harris 92] Harris, C. "*Tracking with Rigid Models*". Dans "Active Vision" p. 59-73, the MIT Press, Cambridge, Massachusetts, 1992.

[Harrison 98] Harrison, B. L. Fiskin, K. P. Gujar, A. Mochon C. et Want, R. "*Squeeze me, Hold me, Tilt Me ! An exploration of Manipulative User Interface*", ACM conference on Computer Human Interaction (CHI), p. 17-24, 1998.

[Hartson 90] Hartson, H. R. Siochi, A. C. et Hix, D. "*The UAN: a user-oriented representation for direct manipulation interface designs*", ACM Transactions on Information Systems, p. 181-203, vol. 8, n. 3, juillet 1990.

[Huber 99] Huber, D. "*The Computer Vision Home-Page*", Site internet. Disponible à l'adresse : http://www.cs.cmu.edu/~cil/txtvision.html

[Isard 98] Isard, M. et Blake, A. "*CONDENSATION: Conditional Propagation for Visual Tracking*". International Journal for Computer Vision, 29(1), p. 5-28, 1998.

[Ishii 97] Ishii, H. et Ullmer, B. "*Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms*", ACM conference on Computer-Human Interaction (CHI), 1997.

[Kabbash 94] Kabbash, P. Buxton, W. et Sellen, A. "*Two-Handed Input in a Compound Task*". ACM conference on Computer-Human Interaction (CHI), p. 417-423, 1994.

[Kalman 60] Kalman, R. E. "*A New Approach to Linear Filtering and Prediction Problems*", Transaction of the ASME, Journal of Basic Engineering, p. 35-45, 1960.

[Kang 98] Kang, S. B. "*Hands-free navigation in VR environments by tracking the head*". International Journal on Human-Computer Studies, p. 247-266, vol 48, 1998.

[Kass 87] Kass, M. Witkin, A. and Terzopoulos, D. "*Snakes: Active Contour Models*", First International Conference on Computer Vision (ICCV), p. 259-268, 1987.

[Koenderink 87] Koenderinck, J. J. et Van Doorn, A. J. "*Representation of Local Geometry in the visual system*", Biological Cybernetics, vol. 55, p. 367-375, 1987.

[Krueger 90] Krueger, M. W. "*Artificial Reality II*". Addisson Wesley Publishing, 1990.

[Lamping 94] Lamping, J. et Rao, R. "*Laying out and Visualizing Large Trees Using a Hyperbolic Space*". ACM conference on User Interface System and Toolkits (UIST), 1994.

[Langer 98] Langer, M. "*Spring-loaded Folders*", dans "Mac OS 8.5: Visual QuickStart Guide". Peachpit Press, 1998. Extrait disponible sur le site : http://beta.peachpit.com/vqs/K5814/excerpt/71.html

[Liang 91] Liang, J. Shaw, K. et Green, M. "*On Temporal-Spatial Realism in the Virtual Reality Environment*". ACM conference on User Interface System and Toolkits (UIST), pages 19-25, 1991.

[Lindeberg 96] Lindeberg, T. "*Edge detection and ridge detection with automatic scale selection*", IEEE conference on Computer Vision and Pattern Recogntion (CVPR), p. 465-470, 1996.

[Mackay 93] Mackay, W. Velay, G. Carter, K. MA, C. et Pagani, D. "*Augmenting Reality : Computational Dimensions to paper*", Communication of the ACM n. 7, p. 96-97, Juillet 1993.

[Mackay 95] Mackay, W. E. Pagani, D. S. Faber, L. Inwood, B. Launiainen, P. Brenta, L. et Pouzol, V. "*Ariel: Augmenting Paper Engineering Drawings*", ACM conference on Computer-Human Interaction (CHI), p. 421-422, 1995.

[Mackay 96] Mackay, W. E. "*Réalité augmenté : le meilleur des deux mondes*", La Recherche n. 285, p. 80-84, Mars 1996.

[MacKay 98] Mackay, W. E. et Beaudouin-Lafon, M. "*DIVA: Exploratory Vido Analysis with Multimedia Streams*". ACM conference on Computer Human Interaction (CHI), p. 416-423, 1998. Disponible sur le site : http://www-ihm.lri.fr/~mbl/DIVA/

[MacKenzie 92] MacKenzie, I. S. "*Fitts' law as a research and design tool in human-computer interaction*", Human-Computer Interaction, 7, 91-139, 1992.

[MacKenzie 93] MacKenzie, I. S. Ware, C. "*Lag as a determinant of Human Performance in Interactive Systems*". Conference on Human Factors in Computing Systems (INTERCHI), pages 488-493, ACM Press, 1993.

[Maes 94] Maes, P., Darrell, T., Blumberg, B. and Pentland, A. "*The ALIVE System: Wireless, Full-body Interaction with Autonomous Agents*", M.I.T. Media Laboratory Perceptual Computing Technical Report No. 257, 1994. Disponible à l'adresse : http://www-white.media.mit.edu/cgi-bin/tr_pagemaker#TR257

[Mann 94] Mann, S. et Picard, R. W. "*Virtual Bellows: Constructing High Quality Stills from Video*", IEEE first international conference on Image Processing, p. 363-367, vol. 1, 1996. Également disponible comme M.I.T. Media Laboratory Perceptual Computing Technical Report No. 259, 1994. Disponible à l'adresse : http://www-white.media.mit.edu/cgi-bin/tr_pagemaker#TR259

[Martin 95] Martin, J. et Crowley, J. L. "*Experimental Comparison of Correlation Techniques*", 3rd International Symposium on Intelligent Robotic Systems (SIRS), p. 287-294, 1995.

[Mauriac 99] Mauriac, L. "*Ces fenêtres qui font écran*". Les cahiers du vendredi. Libération Multimédia. Vendredi 26 mars 1999. Disponible sur le site : http://www.liberation.fr/multi/cahier/articles/sem99.13/cah990326a.html

[Maury 99] Maury, S. Athènes, S. et Chatty, S. "*Rhytmic menus: toward interaction based on rhythm*", ACM conference on Computer-Human Interaction (CHI) extended abstracts, p. 254-255, 1999.

[Maybeck 79] Maybeck, P. S. "*Stochastic models, estimation and control*", Volume I. Academic Press, New York, 1979.

[Microfield Graph. 99] Microfield Graphics Inc. "*SoftBoard: technical specifications*", site internet. Disponible sur le site : http://www.softboard.com/

[Microsoft] Microsoft Inc. "*The Vision Software Development Kit*". Disponible sur le site : http://www.research.microsoft.com/research/vision/

[Mynatt 99a] Mynatt, E. Igarashi, T. Edwards, W. K. et LaMarca, A. "*Flatland: New Dimensions in Office Whiteboards*". ACM conference on Computer-Human Interaction (CHI), p. 346-353, 1999.

[Mynatt 99b] Mynatt, E. "*The Writing on the Wall*". IFIP conference on Human-Computer Interaction (INTERACT), p. 196-204, 1999.

[Newman 92] Newman, W. et Wellner, P. "*A Desk Supporting Computer-based Interaction with Paper Documents*". ACM conference on Computer-Human Interaction (CHI), p. 587-592, 1992.

[Nye 88] Nye, A. "*Xlib programming manual*". X-Window system volume 1, O'Reilly & Associates, 1988.

[Oliver 97] Oliver, N. Pentland, A.P. and Bérard, F. "*LAFTER: Lips and Face Real Time Tracker*" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Puerto Rico, 1997.

[Ousterhout 94] Ousterhout, J. "*Tcl and the Tk Toolkit*", Addison-Wesley Professional Computing, 1994.

[Pedersen 93] Pedersen, E. R. McCall, K. Moran, T. P. et Hulas, F. G. "*Tivoli: An Electronic Whiteboard for Informal Workgroup Meetings*". Conference on Human Factors in Computing Systems (INTERCHI), p. 391-398, ACM Press, 1993.

[Polhemus 99] Polhemus Inc. "*Isotrack II*" Spécification de produit. Disponible sur le site :
http://www.polhemus.com/

[Rasmussen 86] Rasmussen, "*Information processing and Human-Machine Interaction, an approach to cognitive engineering*". Series, Vol. 12, North-Holland, 1986.

[Rauterberg 98] Rauterberg, M. Fjeld, M., Krueger, H. Bichsel, M. Leonhardt, U. et Meier, M. "*BUILT-IT: A Planning Tool for Construction and Design*". ACM conference on Computer-Human Interaction (CHI) Conference Companion, 1998.

[Robert 67] Robert, P. "*Le petit Robert : dictionaire de la langue française*", Paris, 1967.

[Roseman 96] Roseman, M. et Greenberg, S. "*Building Real Time Groupware with GroupKit, A Groupware Toolkit*", ACM Transactions on Computer Human Interaction (TOCHI), 3(1), March 1996.

[Russel 95] Russel, S. and Norvig, P. "*Artificial Intelligence. A Modern Approach*", Prentice-Hall series in Artificial Intelligence, 1995.

[Saund 96] Saund, E. "*Image Mosaicing and a Diagrammatic User Interface for an Office Whiteboard Scanner*", Site internet, Xerox Palo Alto Research Center. Disponible sur le site :
http://www.parc.xerox.com/spl/members/saund/

[Scapin 89] Scapin, D. L. et Pierret-Golbreich, C. "*MAD : Une méthode analytique de description des tâches*". Colloque sur l'ingénierie des Interfaces Homme-Machine (IHM), p. 131-148, 1989.

[Schiele 95] Schiele, B. et Waibel, A. "*Gaze Tracking Based on Face Color*". International Workshop on Automatic Face and Gesture Recognition (AFGR), 1995.

[Shneiderman 87] Shneiderman, B. "*Designing the user interface: Strategies for effective human-computer interaction*", Addison-Wesley, 1987.

[Sellen 92] Sellen, A. Kurtenbach, G. P. et Buxton, W. "*The Prevention of Mode errors Through Sensory Feedback*", Human Computer Interaction, p. 141-164, Lawrence Erlbaum, vol.7, n. 2, 1992.

[Shi 94] Shi, J. et Tomasi, C. "*Good Features to Track*". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994.

[Smarttech 99] Smart Technologies Inc. "*SmartBoard: technical specifications*", Site internet. Disponible sur le site :
http://www.smarttech.com/

[Smets 88] Smets, G. Stratmann, M. et Overbeeke, C. "*Method of causing an observer to get a three- dimensional impression from a two-dimensional representation*". US Patent 4, 757, 380.

[Smets 95] Smets, G. J. F. "*Designing for telepresence: the Delft Virtual Window System*". Dans : Hancock, P. Flach, J. Caird, J. et Vincente, K. "*Local applications of the ecological approach to human-machine systems*", pages182-207, Lawrence Erlbaum Associates, 1995.

[Socher 95] Socher, G. Merz, T. et Posch, S. "*3-D Reconstruction and Camera Calibration from Images with known Objects*". Proceedings of the 6th British Machine Vision Conference (BMVC-95), p. 167-176, D. Pycock (Ed.), 1995.

[Stafford-Fraser 96a] Stafford-Fraser, J. Q. "*Video-Augmented Environments*" Doctor of Philosophy Thesis, Gonville & Caius College, University of Cambridge, Février 1996. Disponible sur le site :
http://www.uk.research.att.com/~qsf/

[Stafford-Fraser 96b] Stafford-Fraser, Q. et Robinson, P. "*BrightBoard: A Video-Agmented Environment*". ACM conference on Computer-Human Interaction (CHI), p. 134-141, 1996.

[Swain 91] Swain, M. J. et Ballard, D. H. "*Color indexing*". International Journale on Computer Vision, p.11-32, Vol. 7, No. 1, 1991.

[Szeliski 96] Szeliski, R. "*Video mosaics for virtual environments*", IEEE Computer Graphics and Applications, p. 22-30, vol, 16, N. 2, Mars 1996.

[Thevenin 99] Thevenin, D. Bérard, F. et Coutaz, J. "*Capture d'Inscriptions pour la Réalité Augmentée*", actes de la 11ème conférence francophone sur l'Interaction Homme-Machine (IHM), 1999. Disponible

[Toyama 96] Toyama, K. et Hager, G. D. "*Incremental Focus of Attention for Robust Visual Tracking*". IEEE conference on Computer Vision and Pattern Recognition, 1996, également rapport technique de l'Université de Yale, disponible sur le site :
http://www.cs.yale.edu/~toyama/

[Toyama 98] Toyama, K. "*'Look, Ma – No Hands!' Hands-Free Cursor Control with Real-Time 3D Face Tracking*". Proceedings of Workshop On Perceptual User Interfaces (PUI), San Francico, Novembre 1998. Disponible sur le site : http://research.microsoft.com/PUIWorkshop/Proceedings/Proc-Start.htm

[Turk 91a] Turk, M. "*Interactive-Time Vision: Face Recognition as a Visual Behavior*", Doctor of Philosophy Thesis, MIT, 1991.

[Turk 91b] Turk, M et Pentland, A. "*Eigenfaces for Recognition*", Journal of Cognitive Neuroscience, p. 71-86, MIT Press, 1991.

[Ullmer 97] Ullmer, B. et Ishii, H. "*The MetaDESK: Models and Prototype for Tangible User Interfaces*", ACM conference on User Interface Systems and Toolkits (UIST), 1997.

[Underkoffler 98] Underkoffler, J. et Ishii, H. "*Illuminating Light : An Optical Design Toll wih a Luminuous-Tangible Interface*", ACM conference on Computer-Human Interaction (CHI), pp 177-178, 1998.

[Vincze 96] Vincze, M. "*Optimal Window Size for Visual Tracking*". Applications of Digital Image Processing, SPIE proceedings vol. 2847, p. 106-117, 1996.

[Virtual-Ink 99] Virtual Ink Inc. "*MIMIO: Technical Specifications*". Site internet, 1999. Disponible sur le site : http://www.virtual-ink.com/

[Voorhorst 98] Voorhorst, F. A. "*Affording Action. Implementing Perception-Action coupling for Endoscopy*". Thèse de doctorat de l'université de Delft, Hollande, 1998. Disponible à l'adresse : http://www.iha.bepr.ethz.ch/pages/forschung/MMI/projects/laparoscopy/PS/thesis.pdf

[Ware 93] Ware, C. Arthur, K. et Booth, K. S. "*Fish tank virtual reality*". Conference on Human Factors in Computing Systems (INTERCHI), p. 37-42, ACM Press, 1993.

[Ware 94] Ware, C. et Balakrishnan, R. "*Reaching for Objects in VR Displays: Lag and Frame Rate*". ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 1, No. 4, pages 331-356, Décembre 1994.

[Wellner 91] Wellner, P. "*The DigitalDesk Calculator: Tangible Manipulation on a Desk Top Display*". ACM conference on User Interface Systems and Toolkits (UIST), ACM publ, p. 27-33, 1991.

[Wellner 93a] Wellner, P. Mackay, W. E. et Gold, R. "*Back to the Real World*", Communication of the ACM n. 7, p. 24-27, Juillet 1993.

[Wellner 93b] Wellner, P. "*Interacting with paper on the DigitalDesk*". Communication of the ACM n. 7, p. 87-96, Juillet 1993.

[Wellner 93c] Wellner, P. "*Adaptive Thresholding for the DigitalDesk*". Xerox Research Center Technical Report n. EPC-1993-110. Disponible sur le site : http://www.xrce.xerox.com/

[Wren 97] Wren, C. R. Azarbayejani, A. Darrell, T. et Pentland, A. "*Pfinder: Real-Time Tracking of the Human Body*". IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol 19, no 7, pp. 780-785, Juillet 1997. Disponible à l'adresse :
http://www-white.media.mit.edu/cgi-bin/tr_pagemaker#TR353

[XVision] "*XVision*". Site WWW. Disponible à l'adresse :
http://www.cs.yale.edu/AI/VisionRobotics/XVision/

[Yang 98a] Yang, J. Stiefelhagen, R. et Waibel, A. "*Visual Tracking for Multimodal Human Com-puter Interaction*". ACM conference on Computer-Human Interaction (CHI), p. 140-147, 1998.

[Yang 98b] Yang, J. Weier, L. et Waibel, A. "*Skin-Color Modeling and Adaptation*". Third Asian Conference on Computer Vision (ACCV'98), 1998.

[Zhai 97] Zhai, S. Smith, B. A. et Selker, T. "*Improving Browsing Performances: A study of four input devices for scrolling and pointing tasks*". IFIP conference on Human-Computer Interaction (INTERACT), p. 286-293, 1997.

[Zisserman 97] Zisserman, A. "*Projective transformation between two images of a planar scene*", dans "Computer Vision Online: Vision Geometry and Mathematics", édité par Fisher, R. B. à l'Université d'Edinburgh. Disponible à l'adresse :
http://www.dai.ed.ac.uk/CVonline/

[Zuke 97] Zuke, M. et Umbaugh, S. E. "*CVIPtools: A Software Package for Computer Imaging Education*", Computer Applications in Engineering Education, p. 213-220, John Wiley & Sons, NY, vol. 5, n. 3, 1997.

Résumé Cette thèse traite de l'usage de la vision par ordinateur pour des situations d'interaction fortement couplée (IFC) entre l'Homme et la machine. Une interaction est fortement couplée sur un intervalle de temps donné lorsque les systèmes humain et artificiel sont engagés de manière continue dans l'accomplissement d'actions physiques mutuellement observables et dépendantes sur cet intervalle. Le déplacement d'un objet graphique avec la souris relève de l'IFC. Nous modélisons l'IFC sous la forme d'un système en boucle fermée constitué de deux sous-systèmes de type stimulus-réponse. Ce modèle permet d'identifier des requis applicables à la conception, à la réalisation ou à l'évaluation de dispositifs utilisables en IFC. En particulier, nous recommandons une latence inférieure à 50 ms., une résolution adaptée à la tâche utilisateur et la satisfaction de la stabilité statique. Nous considérons ensuite l'usage de la vision par ordinateur dans ce contexte.

Une revue des deux approches dominantes du domaine, vision orientée modèle et vision par apparence, nous permet de justifier notre choix de la seconde dont les techniques, de plus faible complexité de calcul, sont susceptibles de satisfaire le requis de latence. Nous présentons ensuite les techniques de vision par ordinateur que nous avons réalisées en adoptant une approche résolument dirigée par la tâche utilisateur. Les deux derniers chapitres détaillent nos expérimentations à la fois techniques et ergonomiques avec la mise en œuvre de deux prototypes : le tableau magique et la fenêtre perceptuelle. Le premier utilise un suivi du doigt en vision par ordinateur pour la désignation d'inscriptions sur un tableau blanc physique amplifié de services électroniques. La fenêtre perceptuelle, quant à elle, utilise un suivi du visage comme nouveau flux d'entrée spatiale dans une interface graphique usuelle. Ce flux est utilisé pour la navigation dans une fenêtre.

Mots-clés Interaction homme-machine, vision par ordinateur, interaction fortement couplée, suivi de doigt, suivi de visage, réalité augmentée, dispositif d'entrée, interaction à plusieurs flux d'entrée spatiale.

Abstract This thesis focuses on the use of computer vision in the context of tightly coupled interaction (TCI) between people and computers. The interaction is tightly coupled within a time interval when the human and artificial systems are continuously engaged in the accomplishment of physical actions that are mutually observable and dependent on this interval. Moving a graphical object with a mouse involves a TCI. We model the TCI as a closed-loop system composed of two stimulus - response subsystems. This model permits the identification of requirements relevent to the conception, the realization or the evaluation of devices in terms of their ability to support TCI. In particular, their ability to operate with a latency of less than 50 ms., with both a resolution and a static stability suitable for the user's task. We then consider the use of computer vision in this context.

A review of the two dominant approaches in the domain, model-based vision and appearance-based vision, justifies our choice of the latter. Its techniques are more suitable because they are less costly in terms of computational complexity and consequently more likely to satisfy the latency requirement. We present computer vision techniques that we have developed in accordance with our resolutely task-driven approach to design. The two final chapters present our technical and ergonomic investigations of two prototype systems: the magic board and the perceptual window. The former uses a computer-vision finger tracker to manipulate drawings in order to implement electronic services on an ordinary physical whiteboard. The latter uses a computer-vision face tracker as a new kind of spatial input stream for an ordinary graphical user interface. This input stream is used to navigate in a graphical window.

Key words Human-computer interaction, computer vision, tightly coupled interaction, finger tracking, face tracking, augmented reality, input devices, multiple input streams.