Rapport de magistère d'informatique 2ème année

Université Joseph Fourier

Utilisation réflexive du flux vidéo en interaction homme-machine

Christophe Lachenal

Tutrice : Laurence Nigay

20 septembre 1999

Remerciements

Je tiens à remercier toutes les personnes sans lesquelles cette année de magistère n'aurait été que le pâle reflet de celle que j'ai passée. Par qui débuter ces remerciements si ce n'est par Laurence qui a passé beaucoup de temps à relire mon pauvre français et cela même si les conditions de modems interposés avec six heures de décalage ne simplifiaient point les choses. Ensuite je tiens à remercier Frédéric qui a eu un jour l'idée sur laquelle se sont basé mes travaux de recherche et qui a été d'une aide précieuse pour l'implémentation du système. Merci aussi à François qui m'a donné quelques pointeurs toujours très intéressants et qui m'ont permis d'approfondir ma réflexion. De plus je remercie tous les membres de l'équipe Ingénierie de l'Interaction Homme-Machine qui ont contribué à une très bonne ambiance de travail. Enfin je voudrais remercier Joëlle, qui a suivi mes travaux de recherche pendant l'été et qui a contribué grandement à l'élaboration des axes de mon espace de conception. Je la remercie vivement d'avoir accepté de m'encadrer pour l'année de DEA, et j'espère que, sous sa responsabilité, je pourrais poursuivre mes recherches au sein de l'équipe dans une direction qui me tient à cœur.

Table des matières

1 Contexte du stage

1.1 Equipe IIHM du laboratoire CLIPS

Mon stage de magistère s'est effectué au sein de laboratoire CLIPS (Communication Langagière et Interaction Personne Système) à Grenoble. Ce laboratoire s'est constitué autour des thèmes concernant les interfaces homme-machine, les système interactifs, les systèmes "multimédia" et les "réalités virtuelles".

Les axes de recherche du laboratoire CLIPS sont la langue comme objet d'étude mais aussi comme mode de communication dans le dialogue homme-machine ou la traduction automatique, les systèmes d'interaction pour des usages finalisés et les systèmes multimédias fournissant les modèles et les outils de base.

Mon projet s'est déroulé sous la responsabilité de Laurence Nigay dans l'équipe IIHM (Ingénierie de l'Interaction Homme-Machine). L'équipe IIHM a pour objectif l'ergonomie cognitive par l'innovation logicielle. Ses thèmes de recherche sont :

- la conception, la mise en œuvre et l'évolution des systèmes interactifs,
- l'architecture logicielle des systèmes interactifs,
- l'interaction multimodale,
- l'interaction multi-utilisateurs.

Les deux principaux domaines applicatifs de l'équipe sont l'exploration des grands espaces d'information et la communication interpersonnelle médiatisée.

1.2 Sujet du stage

La caméra est un dispositif d'entrée (de l'utilisateur vers le système interactif) de plus en plus utilisé. Ce dispositif commence à être utilisé chez les particuliers. Devant ce phénomène récent, de nombreuses études visent à cerner les futures attentes des utilisateurs vis-à-vis des caméras.

La caméra permet d'acquérir en temps réel les informations visuelles du monde situé dans son point de mire. Ces informations sont codées images par images et envoyées au système sous la forme d'un flux vidéo numérisé via la carte d'acquisition. Grâce au codage de l'information au sein de la vidéo, les données peuvent être automatiquement traitées. En fin de traitement, l'information contenue dans la vidéo est décodée et affichée sur un dispositif de restitution permettant à l'utilisateur de visualiser la scène. La vidéo est alors un média car elle est support de l'information.

Au sein du domaine de l'IHM (Interaction Homme-Machine), la vidéo est utilisée pour avoir accès à un contenu informationnel centré sur l'utilisateur. En effet la vidéo permet d'obtenir des informations pertinentes pour l'interaction, à propos de l'utilisateur, de sa tâche et de son environnement proche. Il est important de noter qu'il n'est pas nécessaire de faire des traitements complexes sur les images provenant de la vidéo pour en obtenir un contenu interactionnel de premier ordre.

Les informations contenues dans la vidéo en provenance d'une caméra pointée sur l'utilisateur peuvent être utilisées par celui-ci à des fins personnels. L'objectif est d'améliorer l'interaction, en rendant observable des informations à propos de l'utilisateur. Ce type d'utilisation est à rapprocher d'une propriété d'ergonomie appelée Réflexivité. La réflexivité se définit comme la capacité du système à rendre observables les informations de soi-même qui sont envoyées à autrui.

Dans ce contexte, mon sujet de stage consiste à étudier le couplage possible entre la vidéo et la propriété de réflexivité au sein du domaine de l'IHM. Mon étude vise trois objectifs principaux qui sont :

- Identifier les usages de la vidéo en IHM.
- Définir un espace de conception des systèmes utilisant la vidéo.
- Concevoir et mettre en œuvre de nouvelles techniques basées sur l'utilisation de la vidéo et le potentiel de l'apport de la réflexivité à l'interaction.

Mon rapport est organisé selon ces trois objectifs. Dans une première partie, je présente les axes de conception des systèmes qui font usage de la vidéo. J'illustre ensuite mon espace de conception, formé de ces axes, avec des systèmes existants. Dans une deuxième partie du rapport, je présente un système que j'ai mis en œuvre : ce système intègre la vidéo de façon réflexive.

2 Aspects conceptuels de l'étude

2.1 Introduction

Le flux vidéo est de plus en plus utilisé dans les applications interactives. Il véhicule des informations du monde réel, qui par traitements informatiques pourront être introduites dans le monde virtuel de l'ordinateur. De ce point de vue informationnel, le flux vidéo est un média au même titre qu'un fichier contenant des données car il est un support de diffusion de l'information. Le développement de l'utilisation de ce média est dû à plusieurs raisons :

- la chute des prix des dispositifs matériels liés à la vidéo (caméra, carte d'acquisition et carte de compression) qui rend cette technologie accessible à un grand nombre d'utilisateurs,
- l'augmentation de la puissance de calcul des processeurs qui permet aujourd'hui de faire des traitements en temps réel sur les images provenant de la vidéo, sans impliquer une perte sur la résolution de celle-ci,
- l'explosion des technologies communicantes qui offrent les capacités suffisantes pour la transmission de données.

Ainsi de nombreux systèmes utilisent la vidéo avec des finalités très différentes telles que la communication interpersonnelle médiatisée, le partage de données visuelles entre plusieurs utilisateurs pour produire un artéfact commun (application aux collecticiels), ou encore la vidéo surveillance. De plus les traitements des images vidéo sont souvent complexes et variés : multi-résolution, filtres de protection de l'espace privé, détection de mouvements,,

identification d'objets. Enfin la vidéo est utilisée dans des systèmes mono ou multi-utilisateurs.

Cette diversité met en évidence le besoin d'un espace de conception permettant de situer, de comparer, d'évaluer les techniques existantes ainsi que d'en concevoir de nouvelles. La première contribution de mon étude est la définition d'un tel espace.

2.2 Espace de conception pour les systèmes utilisant la vidéo

L'espace visé doit permettre de caractériser tous les systèmes fondés sur l'usage de la vidéo. L'entrée et la sortie d'un système sont des éléments de caractérisation usuels d'un système interactif. Aussi dans mon espace, un axe est dédié aux entrées et caractérise une source vidéo tandis qu'un deuxième axe décrit le dispositif de restitution, les sorties du système.

Entre l'acquisition de la source vidéo et sa restitution selon une modalité d'interaction de sortie qui soit adaptée à la tâche en cours de l'utilisateur, des traitements informatiques sont effectués. Ces traitements sont réalisés en appliquant différents opérateurs comme la juxtaposition ou la sélection. Le troisième axe de mon espace est dédié à ces opérateurs, ces derniers établissant le lien entre les entrées (axe "Source vidéo") et les sorties (axe "Dispositif de restitution).

Il convient enfin de considérer un quatrième axe, noté "Propriétés", qui identifie les propriétés ergonomiques que le système doit vérifier. Certaines de ces propriétés sont centrées sur la perception humaine ou encore sur l'interaction homme-machine. Nous verrons que ces propriétés ont des incidences sur le choix des opérateurs par exemple.

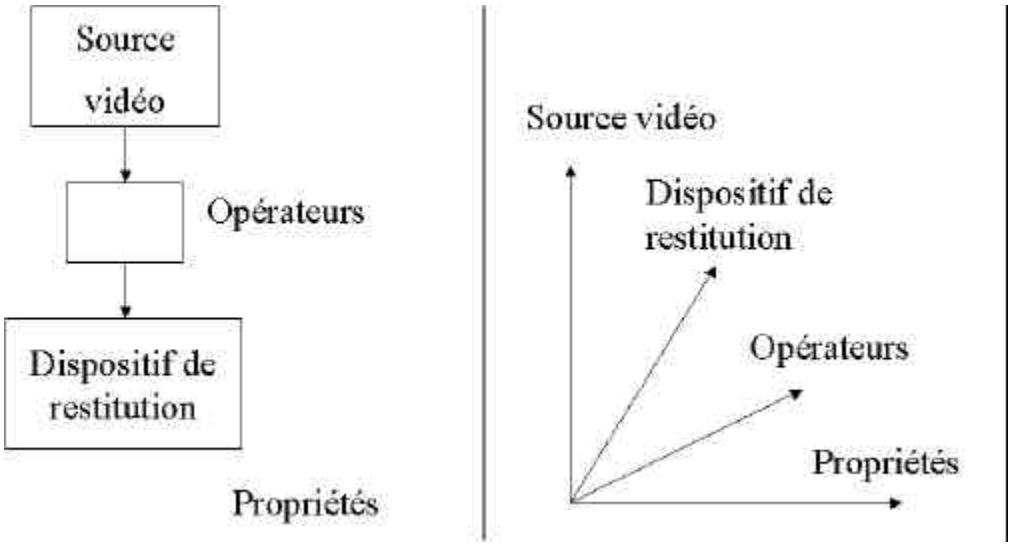


Figure 1 : Axes de l'espace de conception et leurs relations.

La Figure 1 résume les axes de mon espace et leurs relations. Dans le paragraphe suivant, je détaille chacun de ces axes.

2.2.1 Axes de l'espace de conception

Je présente les axes de l'espace en considérant la boucle d'interaction, des entrées vers les sorties : aussi je présente d'abord l'axe "Source vidéo" puis l'axe "Opérateurs" et enfin l'axe "Dispositif de restitution". Cependant il convient de détailler en premier lieu l'axe "Propriétés" qui caractérise de façon globale la boucle d'interaction.

2.2.1.1 Axe "Propriétés"

Un système doit vérifier certaines propriétés pour qu'il soit utilisable et permette à l'utilisateur de réaliser sa tâche. Dans ce paragraphe, je présente les propriétés qu'un système utilisant de la vidéo doit respecter pour que l'interaction homme-machine ou homme-homme qu'il met en œuvre soit satisfaisante. Je regroupe ces propriétés en trois catégories : conscience de groupe, continuité et équivalence.

- *Conscience de groupe* : la conscience de groupe (en anglais "awareness") désigne la connaissance qu'un individu participant à une entreprise collective, a de l'activité de ses collaborateurs et de leur environnement. Autrement dit, la conscience de groupe fournit à l'action individuelle un contexte situationnel collectif. Elle peut être subdivisée en trois propriétés que sont le contact visuel, la perception de l'activité d'autrui et la réflexivité.

  Le contact visuel est une propriété objective qui garantit en situation de dialogue entre collaborateurs un sentiment d'attention d'autrui pour celui qui parle [Brittan 94]. Si la propriété est vérifiée alors l'orateur doit voir son auditoire le regarder. Cette propriété est difficile à satisfaire. Pour que ce soit le cas, il faut que chaque personne participant à l'entreprise collective ait comme focus direct la caméra qui le filme. Cependant peu de dispositifs permettent cela ; car l'utilisateur doit aussi regarder la personne qui lui parle sur son dispositif de restitution. On admettra donc une légère erreur.

  Il est important lors d'une communication médiatisée interpersonnelle de voir ce que font ses collaborateurs. Cette propriété est la perception de l'activité d'autrui. Elle est très importante pour savoir si l'on peut parler à une personne ou non (présence, occupation…). Cependant cette propriété va à l'encontre du respect de l'espace privé d'autrui d'où la nécessité de filtrer l'information et de ne montrer dans certains cas que ce qui est essentiel.

  Enfin vient la propriété de réflexivité. En communication médiatisée, la réflexivité est la capacité du système à rendre observables les informations de nous-même que nous envoyons à autrui. En effet elle sert à savoir ce que les autres voient de nous et de notre environnement afin de ne pas être lésé par la transmission d'information non désirée. Un cas particulier de la réflexivité est l'utilisation de la métaphore du miroir. Notons que dans le cas du miroir, l'utilisateur est à la fois le focus de la vidéo et le destinataire.

- *Continuité* (au sens perception): La continuité garantit que la restitution est perçue comme un tout et non comme la composition de parties séparées. Elle est constituée de trois sous propriétés : la latence (continuité temporelle), la continuité visuelle (caractéristique spatiale) et l'homogénéité.

  La latence est le temps que met un mécanisme de stimulus réponse pour réaliser un cycle. Ce temps est constant (avec un écart type très faible). Dans cet espace, la latence est le

temps qui sépare l'acquisition d'une image source et sa restitution en sortie après traitement. La latence est donc une métrique. Une latence trop grande est dommageable pour l'interaction [Bérard 99].

La continuité visuelle est une propriété visant à garantir que l'utilisateur n'aura pas à explorer du regard plusieurs zones pour obtenir l'information suffisante à la réussite d'une tâche simple. Il est en général plus facile de définir la discontinuité (par exemple, un écran divisé en deux avec d'un coté le plan du cadastre d'une ville et de l'autre la répartition de la population) que la continuité (combinaison des deux cartes dont il est difficile de prouver la suffisance). Le principe à suivre est le non éparpillement des informations principales, le mieux étant de combiner sur la même surface toute l'information souhaitée. Nous verrons par exemple que l'opérateur de fusion semble adapté pour résoudre ce problème [Ishii 91].

L'homogénéité est une propriété visuelle qui vise à vérifier l'absence de discontinuité entre les caractéristiques de restitution de plusieurs sources vidéo.

- *Equivalence* (équipollence): La symétrie est une propriété visant à juger l' « équivalence » d'un dispositif dans le cas d'une utilisation collaborative. En effet il est important de savoir si des modalités et des dispositifs différents sur des lieux distants peuvent restituer à leur utilisateur les mêmes informations. Cette propriété est importante à vérifier afin qu'aucun utilisateur ne soit défavorisé par rapport à un autre. Je distingue 4 niveaux d'équivalence : similitude, équité, équilibre et symétrie.

  La similitude est une comparaison fondée sur l'existence de qualités communes à deux choses (dans notre cas peut être plus). Cette propriété est la plus faible des propriétés d'équivalence. La similitude peut être explicitée de manière plus formelle. L'image restituée en sortie est la représentation d'un concept. Soient deux représentations R1 et R2 (où Ri est la représentation d'un concept via une image sur le site i). La similitude s'exprime comme suit : si R1 différent de R2 alors intersection de R1 et R2 = R3 est non vide(ou intersection de interprétation(R1) et interprétation(R2) non vide : analyse sémantique). En effet, si R3 était vide il n'y aurait pas de similitude visuelle.

  L'équité consiste à mettre chacun sur un pied d'égalité. Cette propriété est subjective. La notion d'équité est rattachée à celle de justice naturelle dans l'appréciation de ce qui est dû à chacun. L'équité est une propriété plus forte que la similitude. En effet elle vérifie la similitude ; en plus elle doit satisfaire que pour toute tâche T, R3 (définit au paragraphe au dessus) soit suffisante (contienne toutes les informations nécessaires) pour que les utilisateurs réussissent la tâche T.

  L'équilibre est par définition une équivalence par compensation. Comme toutes ces propriétés sont liées à l'analyse du système en sortie, l'équilibre garantit que chaque utilisateur a sur son dispositif de restitution le même contenu informationnel même s'il est véhiculé par une modalité différente. De manière plus formelle si la représentation R1 (du site 1) est différente de la représentation R2 (du site 2) alors ces deux représentations donnent une même interprétation à chacun des deux utilisateurs (interprétation(R1) = interprétation(R2)).

La symétrie est la propriété d'équivalence la plus forte. Elle garantit que chaque restitution est en tout point identique quelle que soit sa localité. La symétrie est rattachée à la notion de WYSIWIS (ce que tu vois est ce que je vois). En formalisant cela revient à R1 = R2 et donc aussi interprétation(R1) = interprétation(R2).

Pour que ces propriétés soient vérifiées, les interfaces en entrée et en sortie doivent être construites avec des opérateurs particuliers.

2.2.1.2 Axe "Source vidéo" ou dispositif physique d'acquisition du flux vidéo

Cet axe est dédié aux entrées du système, et dans le cas de mon étude, la vidéo. Cet axe identifie cinq critères ou facteurs de conception :

TYPE : analogique / numérique.
La source vidéo peut être analogique ou numérique. Le type influence les possibilités postérieurs de traitements du flux vidéo. La puissance de calcul actuel des ordinateurs permet d'envisager l'utilisation de la vidéo sous forme numérique (contrairement à ce qu'il se faisait au début des années 90, cf. paragraphe 2.3 Illustration de l'espace).

DEFINITION : <taille, codage>.
La définition détermine la qualité instantanée de l'image vidéo. Elle est constituée d'un doublet <taille, codage>, comme par exemple<640*480 ,8 bits>. La taille de l'image détermine le nombre de pixels constituant celle-ci. Le codage est le nombre de bits utilisé pour coder la couleur d'un pixel (couleur : 24 bits ou 32 par complétude ; niveau de gris : 8 bits).

PROVENANCE : locale / distante.
La source vidéo peut être d'origine locale ou distante par rapport au lieu de la restitution. La provenance a des incidences sur la fréquence de restitution et sur la propriété de latence définie ci-dessus.

CONTROLE : humain / système.
Une source peut être contrôlée par un opérateur qui est soit un humain soit une machine. Le contrôle consiste notamment à pouvoir changer le focus de la caméra (pan, tilt et zoom). Par défaut le contrôle est effectué par le système. Le contrôle est une caractéristique qui varie suivant la finalité du système. En effet dans un système de vision par ordinateur avec suivi d'objets, le système contrôle la source vidéo automatiquement, tandis que dans un mediaspace, le contrôle de la source vidéo est laissé à l'utilisateur afin de contrôler les informations qu'il souhaite transmettre à autrui ou encore de pouvoir explorer l'environnement d'un collaborateur distant.

FREQUENCE : <10 hertz / 10 hertz < f < 25 hertz / >25 hertz.
La fréquence détermine le nombre d'images par seconde acquises par la source dans un but de transmission ou de traitement. Ce nombre peut varier dans un intervalle compris entre 1 (0 = pas d'image donc pas de source) et 60 pour le format NTSC. Il est donc difficile de qualifier ce facteur. Cependant, je propose trois catégories que sont : >25 hertz (fluidité pour l'œil humain) , entre 10 et 25 hertz (suffisant pour pouvoir bien analyser les mouvements d'une personne située dans le focus de la source) et <10 hertz (grande saccade des mouvements).

2.2.1.3 Axe "Opérateurs"

Les systèmes considérés dans mon étude utilisent une ou plusieurs sources vidéo. Chaque source subit des traitements plus ou moins complexes avant d'être affichée selon une modalité d'interaction en sortie. Un traitement peut être soit une opération sur une image source, soit une opération sur plusieurs images sources. De ce fait un opérateur est soit un opérateur unaire (par exemple l'opérateur de zoom) soit un opérateur N-aire (par exemple l'opérateur de juxtaposition). Les traitements possibles sur une image sont variés, aussi la liste des opérateurs que je propose n'est certes pas exhaustive mais contient ceux qui sont les plus utilisés dans les systèmes existants. De plus il est possible de définir de nouveaux opérateurs par composition d'opérateurs de base.

- Opérateurs unaires

Un opérateur unaire est une fonction qui prend en paramètre une seule image et qui en sortie rend une image. Dans les opérateurs unaires on trouve deux opérateurs très utilisés que sont l'opérateur de sélection et de zoom.

- *Sélection :* La sélection permet de prendre dans une image une sous partie de celle-ci vérifiant une certaine propriété P. Par exemple seront retenus tous les pixels (sous constituants d'une image) qui ont leur composante Rouge supérieure à un nombre fixé. Cette propriété de sélection est choisie en fonction du traitement visé. Cet opérateur permet de faire des opérations très complexes comme la superposition. De manière semi-formelle, l'opérateur de sélection s'exprime comme suit :
  sel(A,P)=B où A et B sont des images de même taille et B est l'image A dont certaines sous-parties sont invisibles (c.-à-d. transparentes comme dans le format GIF).

- *Zoom :* L'opérateur de zoom permet d'agrandir ou de réduire la taille d'une image. Cet opérateur est très largement utilisé pour réduire la taille de plusieurs sources vidéo afin, par exemple, de les mettre côte à côte sur l'interface de restitution (effet mosaïque des systèmes de téléconférence). Cet opérateur s'exprime comme suit :
  zoom(A,X)=B où A et B sont des images et X le facteur de conservation de la taille. Si X=100 alors A=B, si X<100 alors A est plus grande que B et si X>100 alors B est plus grande que A.

Il existe une multitude d'autres opérateurs unaires de traitement d'images, comme les filtres utilisés dans les médiaspaces, et de changement de modalités comme l'image d'un histogramme de couleur d'une image. De plus il est facile de créer de nouveaux opérateurs en utilisant l'opérateur de sélection avec une propriété P adéquate et un opérateur N-aire.

- Opérateurs N-aires

Dans le contexte de mon étude, un opérateur N-aire est une fonction qui prend en paramètre plusieurs images, qui les combine et qui en sortie rend une seule image. Ces opérateurs sont très nombreux. Comme pour les opérateurs unaires, je ne décris ici que les plus utilisés dans la littérature : la juxtaposition, la fusion et la superposition.

- *Juxtaposition :* La juxtaposition permet de mettre côte à côte plusieurs images au sein d'une image donnée. On distingue la juxtaposition verticale et la juxtaposition horizontale. La juxtaposition verticale (respectivement horizontale) permet de mettre côte à côte verticalement (respectivement horizontalement) les images filles spécifiées dans l'image mère. Cet opérateur s'exprime comme suit :
  Comp(T, mère, (Lfilles)) où T est le type de composition (ici V pour verticale et H pour horizontale), mère l'image de destination et Lfilles la liste des images à composer dans l'ordre de placement dans l'image mère. Si T est aussi issue d'une composition on peut obtenir une mosaïque d'images.

- *Fusion* : La fusion permet de combiner plusieurs images dans une image donnée. Cet opérateur est basé sur le principe de transparence de N couches d'images. Il est intéressant d'utiliser cet opérateur pour combiner sur une même zone visuelle plusieurs types d'informations sans discontinuité. Cet opérateur s'exprime comme suit :
  Comp(F, mère, (Lfilles)) où F spécifie une composition de type fusion (les autres paramètres sont analogues à ceux de la juxtaposition).

- *Superposition :* La superposition permet d'appliquer plusieurs images les unes sur les autres sans transparence et en réservant à chaque image une sous zone d'affichage dans l'image mère. Ces sous zones sont définies grâce à l'opérateur de sélection. Cet opérateur permet de réaliser l' « overlay » et d'incruster une image dans une autre comme on le fait en cinématographie en utilisant un fond bleu comme propriété de sélection.

2.2.1.4 Axe "Dispositif de restitution"

Le dispositif de restitution (interface de sortie) est un élément crucial du système. En effet, c'est grâce à ce support que l'utilisateur peut percevoir les données pertinentes pour la réalisation de la tâche en cours. L'analyse se faisant sur le dispositif qu'utilise un utilisateur du système et comme ce dernier peut avoir à sa disposition plusieurs dispositifs de restitution (comme un écran et un rétroprojecteur) plusieurs axes "Dispositif de restitution" seront nécessaires à la description d'un système donné. Cet axe contient trois facteurs.

DEFINITION
En sortie, l'image restituée à l'utilisateur est produite grâce aux opérateurs et peut donc être composite. Ce facteur est donc un N-uplet de couple <taille, codage>. La taille est un doublet <x, y> où x est le nombre de pixels composant une ligne et y le nombre de pixels composant une colonne de l'image considérée. Si la restitution est la juxtaposition de deux sources vidéo et si celles-ci ont des fréquences très différentes alors cela peut entraîner une non homogénéité.

CONTROLE
Le contrôle couvre différents aspects : contrôle de la fréquence de restitution d'une source vidéo, contrôle du type de dispositif de sortie, etc. Le contrôle peut soit être délégué au système soit à l'utilisateur.

FREQUENCE
La fréquence permet l'analyse du temps de rafraîchissement d'une zone vidéo en sortie. Comme pour le facteur fréquence de la source, la fréquence en sortie peut être de trois

types : <10 hertz , 10 hertz < f < 25 hertz , >25 hertz. Ce facteur est donc aussi liée à la propriété d'homogénéité.

SUPPORT
Le support peut être de différents types, comme un écran, un mur (vidéo projection) ou un casque, chacun ayant des avantages et des inconvénients.

2.2.2 Conclusion

Il est important de noter que mon espace permet l'analyse d'un système en considérant :

- un seul utilisateur,
- plusieurs sources vidéo,
- plusieurs dispositifs de restitution.

Au paragraphe suivant, j'illustre mon espace de conception en l'appliquant à des systèmes représentatifs de l'état de l'art des systèmes utilisant la vidéo.

2.3 Illustration de l'espace de conception

Pour chacun des deux systèmes considérés, nous caractérisons d'abord les interfaces en entrée et en sortie conçues (axes "Sources vidéo" et "Dispositifs de restitution") puis nous étudions les opérateurs nécessaires à la réalisation de l'interaction (axe "Opérateurs") et enfin nous considérons les propriétés ergonomiques offertes par le système (axe "Propriétés").

2.3.1 Système TWS (Team Work Station)

Le premier système considéré est le système TWS (Team Work Station) développé de 1989 à 1994 par Hiroshi Ishii [Ishii 94] et son équipe. TWS a été construit dans le but de permettre à plusieurs personnes (au maximum quatre) de travailler ensemble via un espace informatique partagé tout en permettant d'utiliser des sources informationnelles réelles ou virtuelles. En effet l'utilisation de l'ordinateur n'exclut pas l'intérêt de la manipulation de livres et autres matériels non informatiques qui sont des sources indispensables d'informations.

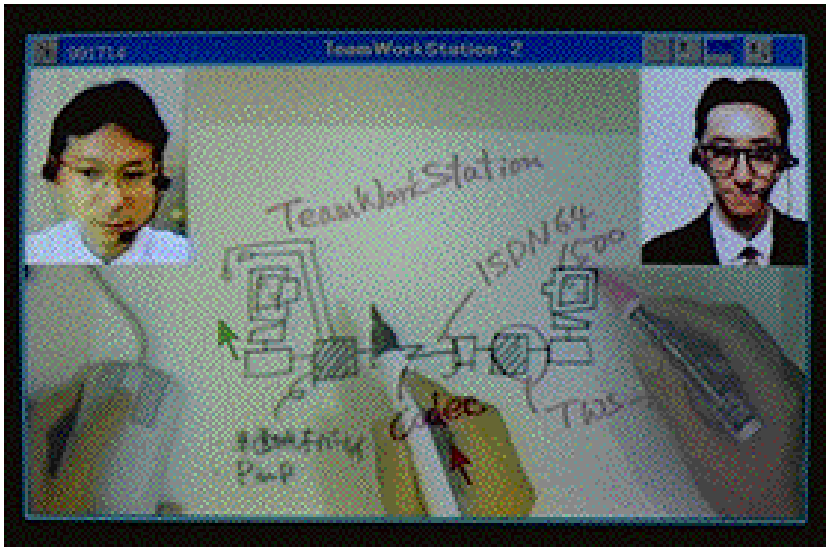


Figure 2 : Image du système TWS.

Comme le montre la Figure 2, l'interface est partagée en plusieurs parties. Le centre est dédié à l'espace partagé pour la collaboration et les images des participants sont affichées sur les côtés supérieurs. Ce système utilise deux caméras par site (par utilisateur) ; une caméra est pointée sur le visage de l'utilisateur (et est situé en face de l'utilisateur), l'autre est pointée sur la main de l'utilisateur (et est située à la verticale de la main).

Selon mon espace de conception, TWS se caractérise comme suit.

2.3.1.1 Système TWS : axe "Sources vidéo"

Dans le système TWS, quatre sources vidéo sont nécessaires. Ces sources sont toutes identiques et chacune est représentable par l'axe suivant. Cependant deux d'entre elles sont des sources distantes tandis que les deux autres sont locales.

TYPE	analogique
DEFINITION	format NTSC= <640*480,24 bits>
PROVENANCE	locale / distante
CONTROLE	humain
FREQUENCE	entre 10 et 25 hertz

2.3.1.2 Système TWS : axe "Dispositifs de restitution"

Sur chaque site, le support de restitution est un écran. Comme le montre la Figure 2, l"écran est divisé en plusieurs parties ayant des définitions différentes. Selon l'axe de la restitution, on obtient les caractéristiques suivantes.

DEFINITION	<640*480,24> (surface partagée) <130*200,24> (pour les images des visages)
CONTROLE	humain
FREQUENCE	entre 10 et 25 hertz
SUPPORT	écran

2.3.1.3 Système TWS : axe "Opérateurs"

Les opérateurs appliqués aux sources vidéo sont simples :

- réduction des images de la source vidéo contenant le visage des participants pour obtenir des images notées A et B,
- fusion des 2 sources vidéos contenant l'espace de travail pour obtenir l'image notée C,
- superposition des images A et B sur C dans les coins supérieurs.

2.3.1.4 Système TWS : axe "Propriétés"

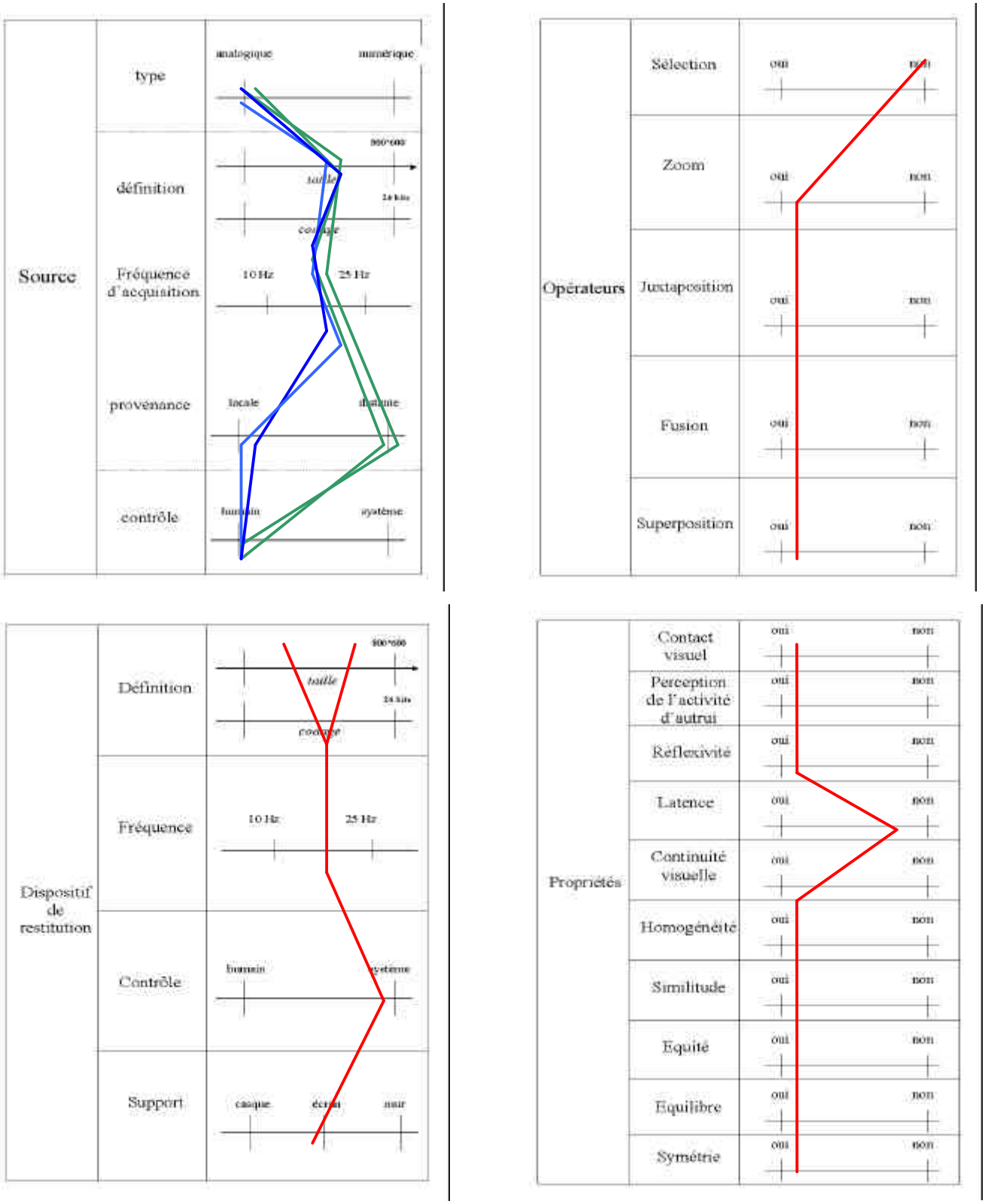
Nous considérons d'abord les propriétés liées à la conscience de groupe. La propriété de contact visuel est respectée car la caméra qui filme le visage de l'utilisateur est positionnée en face de celui-ci. Comme l'utilisateur regarde son écran, l'angle constitué par la caméra, le regard et l'écran est donc minime. La propriété de perception de l'activité d'autrui est aussi valide, chaque utilisateur percevant sur l'écran l'activité de l'autre grâce à la fusion des flux vidéo pointés sur les espaces de travail. La réflexivité est aussi respectée car les utilisateurs peuvent tous observer la même chose sur leur écran, leurs propres activités et image ainsi que celles des autres participants. .

Les propriétés en relation avec la continuité sont aussi respectées. La continuité visuelle est garantie par l'opération de fusion sur les informations liées à l'espace collectif, les autres zones d'affichage n'ayant point de liaison informationnelle directe entre elles. La latence est t faible car tous les traitements se font de manière analogique. L'homogénéité est satisfaite car les fréquences de restitution sont identiques.

Enfin nous analysons les propriétés d'équivalence. Comme je l'ai souligné précédemment, les utilisateurs peuvent tous observer la même chose sur leur écran, la symétrie est donc garantie. Par définition de la similitude, de l'équité et de l'équilibre, ces propriétés sont vérifiées car la symétrie les implique.

2.3.1.5 Système TWS : synthèse

Je propose une représentation graphique synthétisant les valeurs prises par un système donné dans mon espace de conception. Cette représentation est inspirée de celle utilisée dans [Fitzmaurice 95] pour l'application d'un espace de conception. Elle permet d'obtenir une vue synthétique des caractéristiques d'un système et de facilement comparer des systèmes entre eux. La \Figure 3 synthétise mon analyse du système TWS.



\Figure 3 : Analyse de TWS selon mon espace. Les sources vidéo contenant les images des utilisateurs sont représentées en traits pointillés, celles contenant l'espace partagé en traits pleins. Une couleur foncée est utilisée pour les deux sources vidéo locales tandis qu'une couleur claire dénote les sources vidéo distantes (ici deux sources car nous considérons deux utilisateurs).

2.3.2 Système GestureCam

Une analyse similaire à celle faite pour le système TWS, peut être conduite pour le système GestureCam [Kuzuoka 94]. Cette analyse va permettre de faire une évaluation comparative de ce système par rapport à TWS.

GestureCam est un système qui offre une collaboration spatiale entre deux utilisateurs. Ce système permet un partage d'informations visuelles entre deux utilisateurs. La Figure 4 schématise le dispositif complet de ce système. Ce système est basé sur un protocole d'action instructeur/opérateur.

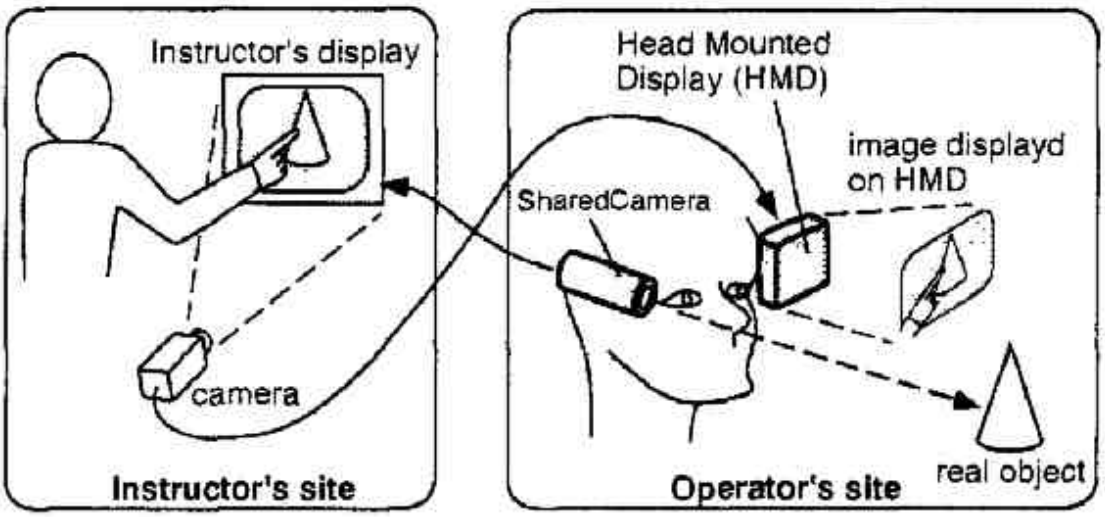


Figure 4 : Dispositif du système GestureCam (Figure issue de [Kuzuoka 94]).

L'opérateur doit agir sur des objets réel. Néanmoins ne sachant pas toujours comment mener à bien sa tâche, l'opérateur peut demander de l'aide à l'instructeur. Ce dernier voit la scène et lui montre comment faire. L'intérêt de ce système réside dans le type d'interaction mis en œuvre mais aussi dans la non symétrie au niveau des dispositifs et de l'informations observables. Ces différences impliquent que l'analyse de ce système soit conduite par rapport à chacun des utilisateurs. Dans mon espace de conception, ces différences sont localisées au niveau des axes "Propriétés" et "Dispositifs de restitution". En effet la propriété de réflexivité n'est valide que pour l'opérateur.

2.3.2.1 Système GestureCam : les axes

Pour mon analyse je considère l'utilisateur opérateur. La Figure 5 présente notre anlyse sous forme graphique. A la Figure 5, il est facile de constater que trois propriétés ne sont pas vérifiées par le système GestureCam. Il est intéressant de noter que la réflexivité est néanmoins vérifiée même si ce n'est qu'à travers un niveau d'indirection supplémentaire, dû au fait que la source est distante aussi bien du coté instructeur qu'opérateur.

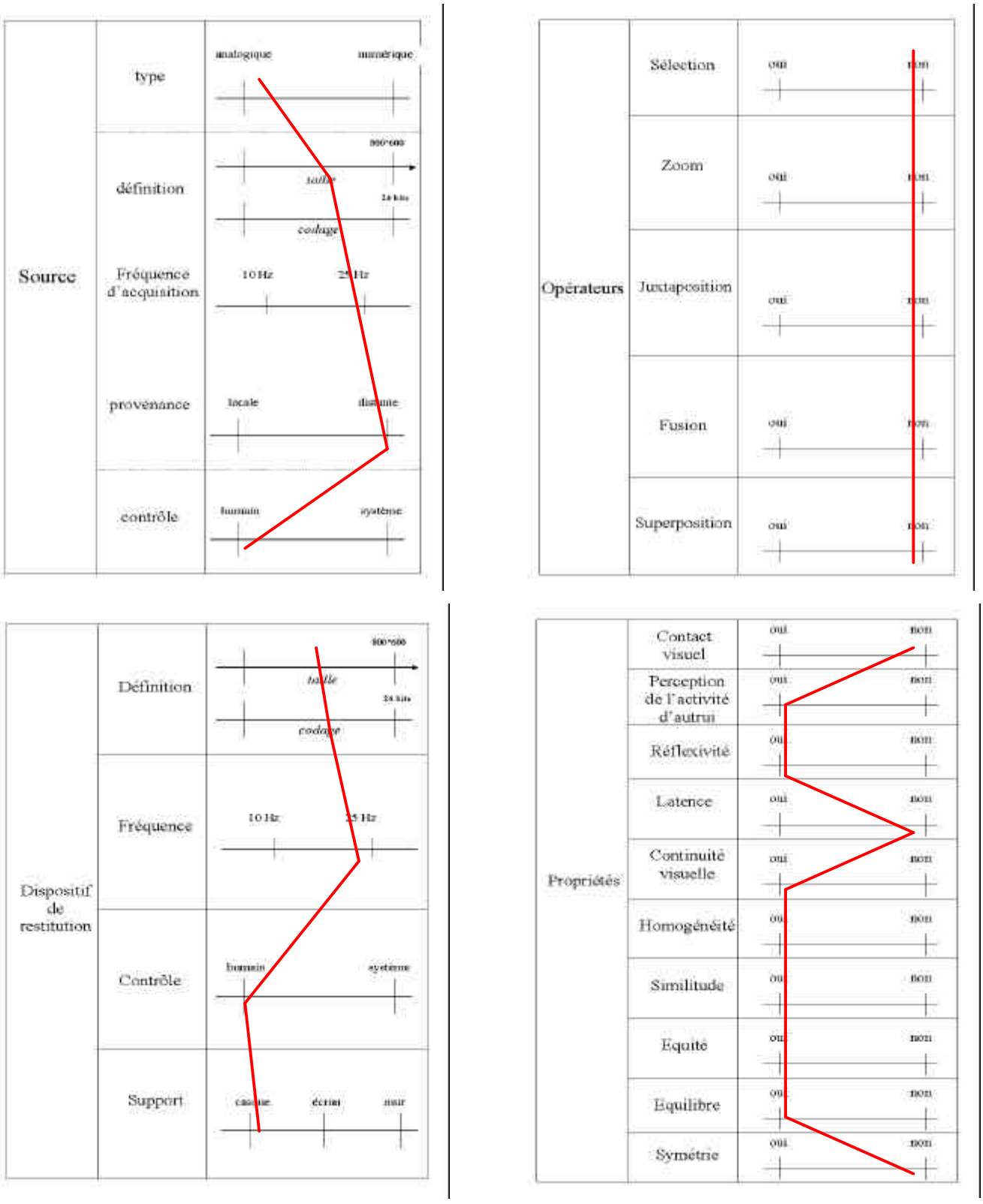


Figure 5 : Analyse de GestureCam (utilisateur = opérateur) selon mon espace.

2.3.3 Comparaison de systèmes

La superposition de la \Figure 3 et de la Figure 5 permet de constater très rapidement les différences entre les deux systèmes étudiés, notamment au niveau des traitements effectués. Par l'analyse des deux systèmes, je souligne l'étendu des possibilités.

La superposition des schémas obtenus selon mon espace de conception permet de rapidement évaluer les différences mais aussi d'identifier les zones vides définies par les caractéristiques peu appliquées ou les propriétés difficiles à vérifier. Pour ce dernier objectif,, il convient néanmoins de conduire cette analyse pour de nombreux systèmes, l'étude de deux systèmes n'étant pas suffisante.

La présentation de mon espace de conception a permis d'introduire et d'organiser les concepts pertinents pour la conception d'un système exploitant la vidéo. Au paragraphe suivant, je présente un système que j'ai conçu et développé et dont l'objectif est l'utilisation réflexive de la vidéo. La réflexivité est l'une des valeurs identifiées sur l'axe « Propriétés » de mon espace.

3 Utilisation réflexive de la vidéo : réalisation logicielle

Cette partie de mon rapport est dédiée à la conception et à la mise en œuvre d'un système basé sur l'utilisation réflexive de la vidéo. Mon approche consiste à vérifier la propriété de réflexivité afin d'aider l'utilisateur à accomplir sa tâche. Pour la réalisation effective de sa tâche, l'utilisateur peut avoir besoin de se voir ainsi que d'observer son environnement de travail. Mon système repose sur l'utilisation de la métaphore du miroir.

Je décris mon système selon deux volets complémentaires : la réalisation du système et les applications possibles.

3.1 Techniques utilisées

Dans ce paragraphe, nous décrivons les mécanismes de technique de fusion des pixels pour réaliser l'effet miroir et augmenter la qualité de l'image finale obtenue. Je décris complètement ce mécanisme dans [Vernier 99].

3.1.1 Effet miroir

Toute l'information visuelle liée à l'utilisateur est véhiculée par les images en provenance de la vidéo. Afin de restituer à l'utilisateur son image capturée par une caméra, mon approche consiste à fusionner le flux vidéo de la caméra avec l'image de l'interface. L'effet miroir réalisé est un cas particulier de réflexion. L'image résultante de la fusion permet la visualisation sur une même surface de deux types d'informations différentes dont la combinaison garantit une continuité visuelle. L'intérêt de cette approche réside dans la combinaison de deux repères visuels : le repère de l'interface et le repère ambiant (de l'utilisateur dans le monde physique) capturé par l'image vidéo de l'utilisateur. L'effet recherché est l'immersion visuelle de l'utilisateur avec l'objet de sa tâche par un effet miroir de l'interface. L'hypothèse faite est que l'utilisation d'une métaphore du monde réel (miroir) doit rendre le système facile à appréhender et à utiliser puisqu'il repose sur des connaissances acquises de l'utilisateur.

Figure 6 : Fusion de l'image de l'interface et d'une image vidéo.

L'effet miroir consiste en la fusion sur une même image (image résultante) de l'image de l'utilisateur (image B) et de celle de son interface graphique (image A), comme le montre la Figure 6. L'interface miroir permet à l'utilisateur de positionner ses mains au sein de celle-ci. Ainsi, grâce à cette métaphore, l'utilisateur peut pointer des objets situés dans son interface de manière très naturelle.

Le dispositif matériel utilisé est le suivant : une caméra placée sur l'écran en face de l'utilisateur reliée à une carte d'acquisition. Le centre d'intérêt de l'utilisateur se situe sur l'écran ; le fait de faire pointer la caméra sur l'utilisateur n'est pas anodin. En effet cela permet de minimiser l'angle constitué par la caméra, le regard et l'écran et donc de garantir la propriété de contact visuelle. La fusion entre les deux images est une opération coûteuse en ressource du processeur. Le mécanisme de fusion s'effectue pixel par pixel en appliquant la formule de la Figure 7 à chacune des composantes (Rouge, Verte, Bleue) des deux images. A la Figure 7, RVBRes désigne la valeur du pixel de l'image résultante, RVBInter, la valeur du pixel de l'image de l'interface et RVBVid, la valeur du pixel de l'image vidéo. Les coefficients de composition CoeffInter et CoeffVid définissent les proportions relatives « d'interface et de vidéo » dans l'image résultante.

$$RVBRes = \frac{CoeffInter * RVBInter + CoeffVid * RVBVid}{CoeffIhm + CoeffVid}$$

Figure 7 : Calcul de la valeur d'un pixel dans l'image résultante.

Pour que la métaphore du miroir soit parfaitement perçue par l'utilisateur, il convient de garantir une grande fluidité au niveau de l'interface de sortie ainsi qu'une latence très faible. La fusion étant appliquée au niveau du pixel, le calcul exprimé par la formule de la Figure 7 est exécuté près de 30 millions de fois par seconde pour des images sources de 640 par 480 pixels à la fréquence de 30Hz. Le calcul de la formule nécessite donc de nombreux cycles d'horloge, et il est donc nécessaire d'améliorer le coût algorithmique de ce calcul. En particulier, la multiplication est une opération très coûteuse en ressource processeur. Une optimisation simple consiste à utiliser des coefficients de la forme $m/2^n$, par exemple (1/2, 1/2), (1/4, 3/4) ou (3/8, 5/8). Ce faisant, j'exploite les capacités d'optimisation intrinsèque des processeurs : les divisions par puissance de 2 se font par instructions de décalage bit à bit. De plus, comme le montre l'expression de la Figure 8, l'utilisation d'un masque spécifique permet d'appliquer le décalage sur les trois composantes Rouge, Verte et Bleue à la fois. L'expression de la Figure 8 traduit notre version optimisée de la formule initiale de fusion avec comme coefficients (CoeffInter = 1/2, CoeffVid = 1/2).

```
RVBRes = ((RVBInter >>1) & 0x007F7F7F) + ((RVBVid>>1) & 0x007F7F7F);
```

Figure 8 : Fusion optimisée de deux pixels avec coefficient miroir égal à ½.

La valeur du masque 0x007F7F7F se justifie par le codage XBGR de l'image en mémoire. Lorsque l'un des coefficients est de la forme $m/2^n$ avec m différent de 1, je le transforme en une somme de coefficients de la forme $1/2^n$. Par exemple dans le cas où CoeffInter = ¾ et CoeffVid = 1/4, je transforme CoeffInter en deux coefficients respectivement égales à 1/2 et 1/4 (en effet 1/2 + 1/4 = 3/4). L'expression est alors celle de la Figure 9. Grâce à cette optimisation, l'effet fonctionne avec un taux de rafraîchissement proche de 20 hertz.

```
RVBRes = ((RVBInter>>1) & 0x007F7F7F) +
         ((RVBInter>>2) & 0x003F3F3F) +
          ((RVBVid>>2) & 0x003F3F3F) ;
```

Figure 9 : Fusion optimisée de deux pixels avec coefficient miroir égal à 1/4.

Selon l'application à mettre en œuvre (paragraphe 3.2 intitulé Applications), le choix du coefficient à appliquer a son importance. Cependant il ne faut pas que ce choix compromette l'interaction en ralentissant le processus et donc en perdant la fluidité nécessaire. Pour pallier à cela, j'ai implémenté 32 niveaux de pourcentage miroir optimisé. Cela offre à l'utilisateur la possibilité de choisir le pourcentage de vidéo qu'il souhaite utiliser sans remettre en cause la fluidité de l'interaction. Dans le système développé, le coefficient miroir peut être changé par l'utilisateur grâce à deux touches clavier dédiées à cet effet.

De plus, comme 32 niveaux ont été réalisés, une différence infime existe entre deux niveaux ; le fondu continu de la vidéo à l'interface est donc possible. L'image de la Figure 10 illustre ce fondu en présentant un dégradé de coefficients miroir. A la Figure 10, le dégradé s'effectue sur l'axe vertical de l'image : en haut de l'image le pourcentage de l'image de l'interface est maximal tandis qu'en bas, c'est celui de la vidéo qui est maximal. En effet, le bas de l'image dévoile que la vidéo, l'interface a complètement disparu.

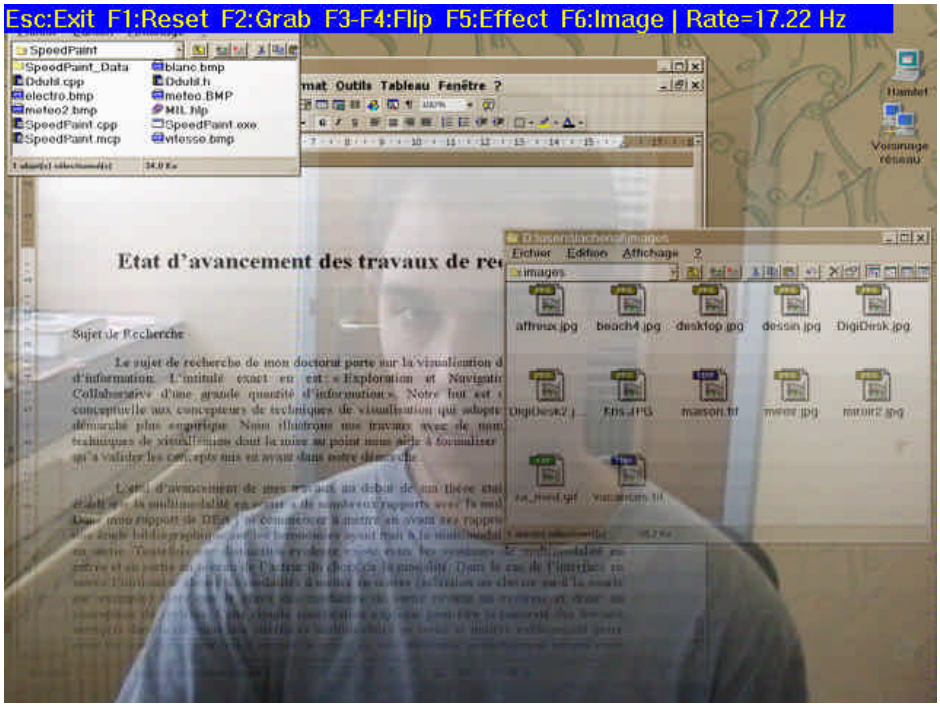


Figure 10 : Dégradé de l'effet miroir.

3.1.2 Amélioration de la qualité de l'image finale

L'image résultante, doit, comme pour un miroir réel, rester lisible. Nous avons choisi d'exploiter les notions de contraste et de luminosité par le truchement de la technique des histogrammes d'intensités lumineuses pour résoudre cette contrainte. Un histogramme d'intensités lumineuses est un graphique, représentant pour chaque valeur d'intensité lumineuse possible, le nombre de pixels de l'image qui ont cette intensité.

Figure 11 : Image avant traitement.

Figure 12 : Histogramme d'intensités lumineuses correspondant à la Figure 11.

L'histogramme de la Figure 12 caractérise l'image de la Figure 11 résultant d'une fusion. On constate que la plupart des pixels d'une telle image sont centrés sur les valeurs moyennes. Ceci se justifie ainsi : un pixel très sombre (resp. très clair) de l'image finale s'obtient par fusion de pixels sources très sombres (resp. très clairs). Ce cas se présente rarement dans la fusion de deux images indépendantes. Aussi, l'histogramme d'intensités lumineuses de l'image finale se resserre-t-il autour des valeurs moyennes. Ceci se traduit visuellement par le voile terne de l'image de la Figure 11.

Figure 13 : Image après traitement.

Figure 14 : Histogramme d'intensités lumineuses correspondant à la Figure 13.

Pour obtenir une image plus lumineuse et plus contrastée, j'utilise une technique appelée en infographie « technique du point noir et du point blanc ». Cette technique consiste à calculer une valeur minimale (resp. maximale) qui cerne au mieux les bords des pics de l'histogramme. Pour cela, je calcule une nouvelle valeur relative du blanc et du noir. Une fois ces deux valeurs calculées, l'intervalle [Min , Max] regroupe la majorité des valeurs des pixels. Nous définissons enfin la fonction affine adéquate par laquelle nous étendons l'intervalle : un pixel ayant la valeur Min (resp. Max) aura, après application de la fonction, la valeur 0 (resp. 255). La formule de calcul correspondante est présentée à la Figure 15.

P[m, n] = si P[m, n] <= Min alors 0
sinon si P[m, n] >= Max alors 255
*sinon (255 * (P[m, n]- Min) / (Max - Min))*

Figure 15 : Calcul de la nouvelle luminosité en fonction de la valeur de Min et Max.

Cette technique, qui accentue les composantes RVB d'un pixel, permet d'obtenir une image comme celle de la Figure 13. Dans cette image, les objets de l'IHM graphique se distinguent davantage. L'effet de voile terne de la Figure 11 est nettement atténué. La Figure 14 présente l'histogramme correspondant à l'image de la Figure 13. Par rapport à l'histogramme de la Figure 12, on peut constater que l'intervalle de répartition des pixels est plus étendu.

Pour implémenter cette technique, le traitement s'effectue sur chaque composante (R, V et B) et consiste à transformer une valeur sur 8 bits (de 0 à 255) en une autre valeur sur 8 bits. Mon algorithme consiste à pré-calculer un tableau de 256 entrées qui fait correspondre aux 256 valeurs de la Figure 12, les 256 valeurs de la Figure 14. Une fois ce calcul effectué, notre technique ne rajoute qu'une indirection dans l'algorithme initial. Ainsi le résultat à l'écran reste tout aussi fluide. Comme pour le coefficient miroir, l'utilisateur a la possibilité de changer les valeurs de Min et de Max. Cela lui permet de gérer la luminosité et le contraste de l'image qu'il souhaite utiliser.

3.2 Applications de l'effet miroir

Après avoir expliqué comment réaliser l'effet miroir avec une bonne qualité de l'image résultante tout en garantissant une fréquence supérieure à 30 hertz, je présente dans ce paragraphe trois applications possibles de cette technique, notée Pixels Miroirs. Les trois applications reposent donc sur la réflexivité et la métaphore du miroir

3.2.1 Geste explicatif

Une explication fait souvent usage de gestes déictiques à l'adresse d'objets, sujet du discours. En coprésence, l'orateur et l'observateur partagent un même univers physique. A distance, une interface augmentée fondée sur la métaphore du miroir, permet de pallier la discontinuité des espaces. Dans les scénarios d'utilisation que je considère, l'objet du discours est un document électronique. De plus les différents utilisateurs se trouvent dans des lieux distants, mais sont connectés entre eux par leurs systèmes reliés en réseau. Je regroupe ces scénarios d'utilisation sous le terme de « télé-explication gestuelle ».

Le scénario d'utilisation consiste en un expert dans un domaine donné qui souhaite expliquer un schéma, un graphique ou plus généralement une image à un auditoire éloigné et utilise pour cela un outil informatique de visioconférence. Dans le cas de la télé-explication, l'expert et ses gestes sont les entités pertinentes du monde réel : ils seront donc intégrés à l'interface augmentée. L'expert ou orateur se voit dans l'interface : comme devant un miroir, il peut désigner du doigt les concepts électroniques présentés dans l'interface. Il lui suffit de placer la main de façon à ce que l'image de son doigt se superpose à l'objet en question. A distance, l'observateur reçoit l'image fusionnée. L'espace électronique, objet de la discussion, et l'orateur ne font qu'un.

La Figure 16 montre un exemple de télé-explication inspirée de la présentation de la météo et réalisée avec notre système. Dans cette situation interactionnelle, le présentateur voit sur un moniteur (ou projeté en face de lui) l'objet du discours (la carte) et sa propre image inscrite dans la scène. Cette interface augmentée offre deux avantages sur le système en vigueur à la télévision :

- le présentateur ne masque pas la carte puisqu'il est transparent. En conséquence, le téléspectateur peut à tout instant inspecter la région qui l'intéresse,
- le présentateur, qui voit la carte devant lui, n'est pas contraint de se tourner sur le côté pour en consulter le contenu : son espace interactionnel (carte et téléspectateurs) est devant lui et il en fait partie intégrante.

Figure 16 : Présentation de la météo avec effet miroir.

La source du document fusionné avec l'image de l'orateur définit le type d'application. Si le document est défini par l'orateur, l'application s'apparente à une présentation à distance comme l'exemple de la météo. Au contraire, si le document provient de l'auditoire, par exemple l'image d'une photocopieuse en train d'être réparée par un non expert, l'application consiste alors en une aide à distance.

J'envisage d'autres domaines d'application de la télé-explication fondés sur notre système : le dépannage à distance et le télétutorat. Un dépanneur reçoit l'image électronique d'un objet à réparer. Sur l'interface augmentée par effet miroir, le dépanneur montre à distance comment réparer l'objet en désignant sur l'image les manipulations à effectuer. Il peut en particulier reproduire les gestes à effectuer sur l'objet réel. Le télé-enseignement constitue un autre domaine applicatif privilégié de ma technique. L'enseignant peut fournir des explications à l'étudiant distant qui rencontre des difficultés dans l'utilisation d'un logiciel. Il montre la façon de faire directement sur l'interface du logiciel. L'étudiant effectue les actions en même temps qu'il reçoit les explications. De fait, l'étudiant est en situation d'apprentissage actif.

3.2.2 Interface 3D miroir

Une autre application possible de ma technique consiste à exploiter l'effet miroir pour accentuer un effet 3D dans une interface multi-fenêtrée. Les interfaces multi-fenêtrées, que nous connaissons, reposent sur la métaphore du bureau. Les documents en cours d'utilisation sont empilés les uns sur les autres. Un document appartient à une fenêtre manipulable par l'utilisateur. Chaque fenêtre est composée d'interacteurs (en anglais « widgets »), comme les boutons et les barres de défilement. Ces interacteurs sont construits de façon à faire ressortir un effet 3D : les fenêtres les plus anciennement utilisées sont placées sous les plus récemment manipulées. La fenêtre au premier plan est celle active sur laquelle l'utilisateur effectue sa tâche. Il existe donc une troisième dimension dans les interfaces multi-fenêtrées qui organise les fenêtres. Cependant l'effet 3D est simulé et deux fenêtres non actives et non superposées ne sont pas différenciables selon l'historique de leur utilisation. Il peut en résulter une mauvaise perception de la dimension de profondeur, qui peut être dommageable pour

l'interaction. Pour pallier ce problème et donc accentuer l'effet 3D, je propose d'utiliser l'effet miroir. Je considère que chaque fenêtre est une surface ayant des propriétés de réflexion. Ainsi dans les fenêtres proches de l'utilisateur (les plus récemment utilisées), l'image réfléchie de l'utilisateur est plus grosse que dans les fenêtres situées plus loin dans l'espace. La Figure 17 illustre ce principe. On constate à la Figure 17 que l'on peut facilement différencier des fenêtres non superposées : la fenêtre en haut à gauche est nettement plus ancienne et donc plus loin que celle en haut à droite.

Figure 17 : Interface 3D augmentée par effet miroir.

Afin de réaliser cet effet de profondeur, j'utilise une fonction de zoom logicielle qui permet de grossir l'image vidéo affichée dans les fenêtres plus proches. Le zoom réalisé est très faible entre deux fenêtres proches dans l'espace (la position dans l'espace étant calculée selon leur historique d'utilisation). Cependant le fait que le texte doit rester lisible dans toutes les fenêtres, atténue cet effet de profondeur. Les premiers tests informels ont néanmoins montré que les personnes perçoivent sans trop de difficulté l'effet désiré de profondeur.

3.2.3 Interaction gestuelle de dessin

Une troisième application des Pixels Miroirs concerne un système de dessin ou de retouche d'images. La Figure 18 montre un exemple de retouche d'images au moyen d'une interface augmentée par effet miroir de la main. Dans cette application, la caméra observe, comme dans le cas du bureau digital [Wellner 93], une feuille de papier sur laquelle l'utilisateur dessine avec les instruments de son métier (pinceau, crayon, gomme, etc.). Face à lui, est présentée sur un écran, l'image de l'interface augmentée. Comme le montre la Figure 18, l'utilisateur y voit l'objet de la tâche (son dessin ou l'image électronique), sa main et les inscriptions produites sur la feuille de papier. Au départ, la feuille de papier est blanche. Lorsque l'utilisateur amène son outil (crayon, pinceau, stylo, etc.) sur la feuille, l'image de la main est affichée à l'écran et fusionnée avec le dessin ou l'image à l'écran. Cette juxtaposition permet de déposer de l'encre au bon endroit dans le dessin électronique. A la demande de l'utilisateur, le système met en œuvre un algorithme d'extraction de l'encre déposée sur la feuille. Ce faisant, l'encre physique de la feuille de papier pénètre sans rupture interactionnelle dans l'espace électronique. Chaque extraction d'encre donne lieu à la création

d'un calque électronique qui se superpose à l'image électronique. L'utilisateur termine son dessin ou sa retouche en fusionnant les calques.

L'apport majeur de cette interface augmentée est la souplesse offerte dans le choix des modalités d'entrée : pour une tâche donnée, l'utilisateur a le choix entre les instruments du métier et les outils informatiques. De plus, ce choix est décidable à tout instant : l'utilisateur peut dessiner avec son feutre favori et changer ensuite la couleur du trait en utilisant les services du logiciel.

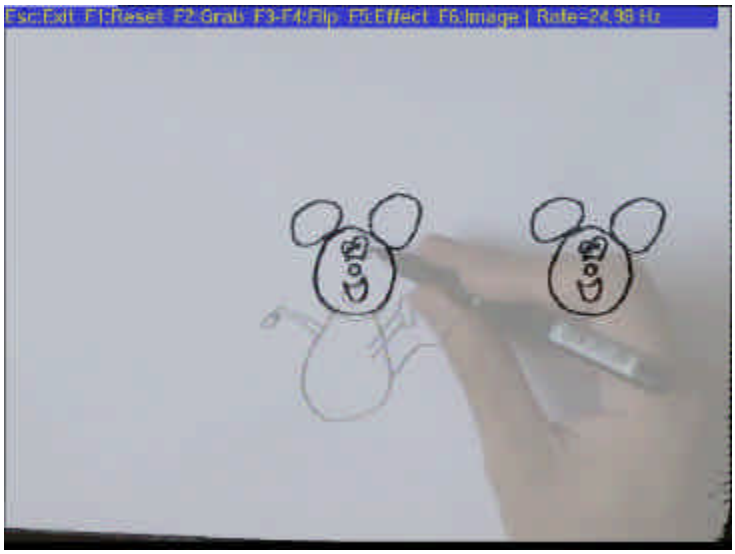


Figure 18 : Combinaison d'un dessin en cours d'édition avec la main de l'utilisateur.

Comme je l'ai indiqué, ce système met en œuvre un algorithme d'extraction de l'encre déposée sur la feuille de papier. En effet, l'image acquise par la caméra, comme celle présentée à Figure 19, est une image couleur contenant beaucoup de bruits, par exemple des ombres. Pour distinguer l'encre du fond de la feuille de papier, l'image doit être nettoyée. Le but est donc, à partir d'une image capturée par la caméra, d'obtenir une image noir et blanc où les points noirs correspondent à de l'encre. Une méthode immédiate et simple consiste à définir un seuil de valeur de la luminosité d'un pixel, au dessous duquel un pixel est déclaré comme étant un pixel d'encre (car plus sombre). Cependant cette méthode n'est pas satisfaisante car un seuil identique sur toute l'image peut induire de grosses erreurs sur son analyse. En effet, une ombre sur la feuille peut être considérée comme de l'encre et de l'encre bien éclairée comme du papier.

Figure 19 : Image d'un dessin sur papier acquise par la caméra.

Peter Wellner propose un algorithme utilisant un seuil adaptatif local pour remédier à ce problème [Wellner 93]. Le principe de base consiste à modifier, pendant le parcours de l'image, la valeur du seuil, qui est calculée par rapport à la moyenne des intensités des derniers pixels rencontrés. Quand l'intensité d'un pixel est significativement inférieure à ce seuil, le pixel est considéré comme représentant de l'encre. Cette solution n'est pas satisfaisante quand une trop grande zone d'encre est présente sur l'image car la fin de la zone d'encre n'est plus considérée comme de l'encre. En effet, lors du parcours d'une ligne sur cette zone d'encre, le seuil diminue et devient trop faible pour que les points suivants soient déclarés comme étant de l'encre. Ce problème peut être résolu en effectuant un parcours sur l'image selon chacune des directions cardinales. Ainsi, une grande zone horizontale d'encre sera détectée par les parcours verticaux et une grande zone verticale d'encre par les parcours verticaux. Cependant avec un tel algorithme, on obtient une image nettoyée avec des traits en créneau comme le montre la Figure 20, car les nuances au bord d'un trait ont été prises comme faisant partie du papier.

Figure 20: Image issue de l'acquisition d'encre selon la méthode de Wellner [Wellner 93].

Pour limiter cet effet, il est possible d'appliquer à chaque pixel une matrice de convolution adéquate. Ainsi la valeur d'un pixel sera la moyenne pondérée de sa valeur et de celle de ses voisins. Un pixel situé en partie limitrophe entre des pixels noirs d'encre et des pixels de la feuille blanche deviendra donc gris. La matrice utilisée pour faire ce lissage (« antialiasing ») est la suivante : (1,2,1)*(1,2,1). Par exemple, en appliquant cette matrice à tous les points de l'image de la Figure 20, on obtient l'image de la Figure 21. Ce lissage entraîne néanmoins une image résultante qui est floue. Le résultat n'est donc pas satisfaisant pour un système de retouche précise de dessin.

Figure 21 : Image après lissage.

L'approche, que j'ai adoptée, n'utilise donc pas de lissage sur l'image originelle. Mon algorithme est le suivant :

- Acquérir l'image à traiter.
- Créer une image 4 fois plus grande que celle acquise (agrandissement de celle-ci).
- Appliquer l'algorithme de Wellner sur l'image agrandie.
- Réduire l'image résultante au format initial.

Mon approche est issue de la méthode utilisée par les dessinateurs de bandes dessinées pour obtenir d'infimes détails. En effet, les dessins ne sont que très rarement faits à l'échelle auquel nous l'achetons, mais ils sont conçus à une taille deux ou trois fois plus grosses avant de subir une réduction. Les détails obtenus sont dus à cette phase de réduction. Ainsi ma solution consiste dans une première étape à obtenir une image vidéo deux fois plus grosses de bonne qualité afin qu'après l'avoir traitée selon l'algorithme de Wellner la phase de réduction fournisse une image de bonne qualité.

Ma méthode pour agrandir l'image doit au mois préserver les pixels qui sans agrandissement auraient été considérés comme de l'encre en appliquant l'algorithme de Wellner. De plus ma méthode doit rajouter des pixels d'encre pour lisser l'image au moment de l'agrandissement. En effet si l'agrandissement consiste uniquement à dupliquer l'image initiale pixel par pixel, l'application de l'algorithme de Wellner donnerait les mêmes résultats qua sans agrandissement car l'algorithme effectue les calculs soit sur les lignes soit sur les colonnes. Pour rajouter de l'encre lors de l'agrandissement, je propose la solution suivante :

- Comme la hauteur et la longueur de l'image initiale est doublée, un pixel dans l'image initiale (image A) donne naissance à 4 pixels dans l'image finale (image B).
- Parmi les quatre pixels de l'image B associés à un pixel de l'image A, le pixel en haut à gauche de l'image B prend la même valeur que celui provenant de l'image A. J'applique cela sur toute l'image B.
- Pour chaque pixel de l'image A, il convient alors de déterminer la valeur des trois pixels restant dans l'image B. Pour cela, le pixel en haut à gauche de chaque bloc de 4 sert de repère pour compléter les 3 autres. Tout d'abord je complète les lignes impaires en calculant la valeur de chaque pixel restant (Pixel(2*i) pour tout i compris entre 1 et longueur de l'image sur 2) suivant la formule suivante :

  Pixel(2*i) = si plus_sombre(pixel(2*i-1),pixel(2*i+1))
  alors moyenne(pixel(2*i-1), pixel(2*i+1)) ;
  sinon pixel(2*i-1).

  Ainsi l'agrandissement se fait, comme le montre la Figure 22, selon la méthode de [Stafford-Fraser 96] en rajoutant le test de la formule ci-dessus pour limiter le flou. En effet le flou est dû au fait que par le calcul la valeur du pixel le plus sombre diminue. Pour pallier à cela, j'effectue donc un test par rapport à une valeur de référence qui est le pixel en haut à gauche au sein d'un zone de quatre pixels dans laquelle une méthode directe d'agrandissement aurait dupliquer la valeur initiale. De ce fait dans l'image grossie un pixel encre aura au moins trois voisins contigus considérés comme encre.

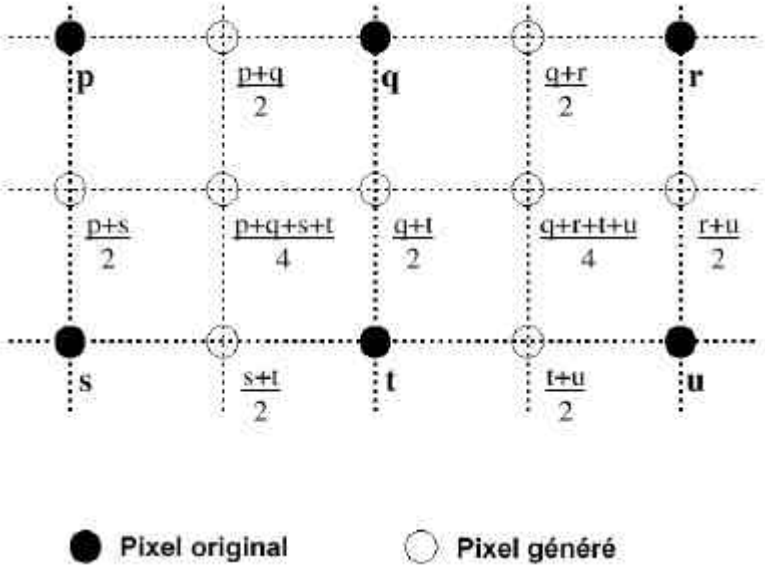


Figure 22 : Algorithme de [Stafford-Fraser 96] pour augmenter la résolution.

Les pixels des lignes impaires sont ensuite calculés de la même manière en gardant comme repère les pixels du haut. Ainsi si le pixel est très sombre dans l'image A alors ces 4 pixels correspondant dans l'image B seront eux aussi très sombres. De plus, un pixel très clair situé à côté d'un trait d'encre dans l'image A, va avoir ses voisins dans l'image B plus sombres. L'intérêt de cette méthode est qu'en réduisant la taille de l'image B après l'avoir traitée avec l'algorithme de Wellner, on obtient une image avec plus de détail et non floue.

La phase de réduction permet de « créer » des détails et de lisser l'image finale. Pour cela, la réduction s'effectue en affectant à un pixel de l'image finale, la moyenne de quatre pixels de l'image initiale. Le lissage peu être amélioré en décalant légèrement la « zone de réduction », comme le schématise la Figure 23 : au lieu de réduire selon les carrés aux bords en pointillé, la réduction se fait par rapport aux carrés aux bords pleins.

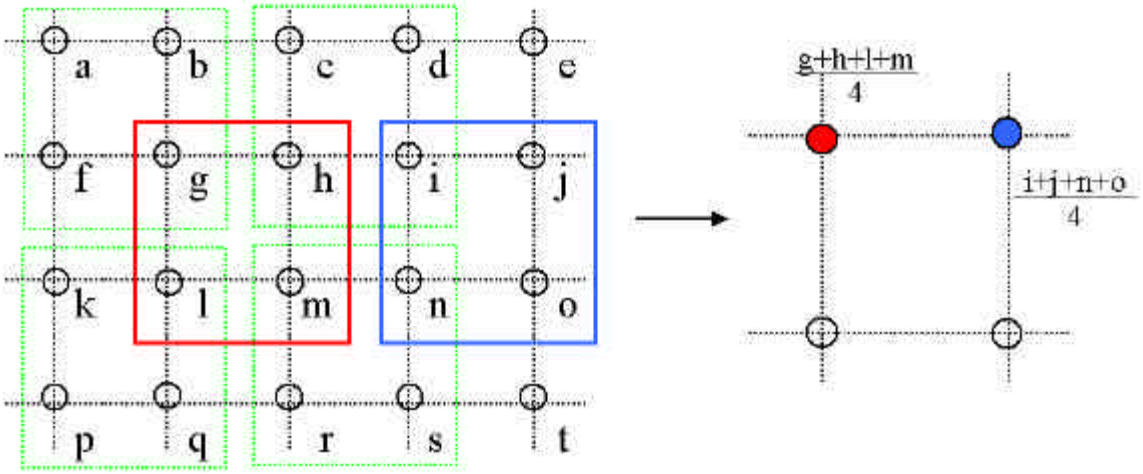


Figure 23 : Méthode de réduction utilisée pour lisser.

La Figure 24 présente l'image obtenue après application de mon algorithme. Par rapport à l'image de la Figure 21, celle obtenue est plus nette et aussi plus proche de ce que l'utilisateur a dessiné.

Figure 24 : Image obtenue en apliquant mon algorithme.

A la Figure 25, je récapitule les résultats obtenues par les trois algorithmes cités ci-dessus. De gauche à droite les images grossies du haut du sapin selon la méthode du seuillage simple puis avec la technique de lissage et enfin avec mon algorithme.

Figure 25 : Récapitulatif des trois solutions d'acquisition d'encre.

4 Conclusion

4.1 Contribution

L'espace de conception, que j'ai défini, vérifie les objectifs fixés lors de son élaboration. En effet il organise, selon des axes, les concepts pertinents à la conception d'un système exploitant la vidéo. Comme je l'ai montré, il permet aussi de caractériser un système existant mais aussi de comparer plusieurs systèmes. Son élaboration a été basée sur une étude des techniques et des systèmes existants décrits dans la littérature du domaine. Cet espace n'est pas exhaustif, mais il est un premier pas vers la définition d'un espace cohérent de conception des systèmes exploitant la vidéo.

L'étude du couplage de la vidéo et de la propriété de réflexivité a permis la conception et l'élaboration d'une technique permettant une utilisation réflexive de la vidéo. Cette technique a été implémentée de façon à être portable comme un service d'une boîte à outils. Elle offre la fusion en temps réel entre un flux vidéo numérique en haute résolution et une autre source d'informations visuelles. Pour souligner l'intérêt de cette technique, j'ai développé trois prototypes d'applications interactives : la télé-explication, l'interface 3D et la retouche d'images. J'ai de plus exposé de nombreux autres domaines d'application de ma technique de Pixels Miroirs.

4.2 Perspectives

Les perspectives à cette étude sont nombreuses.

A court terme, je souhaite complètement développer les trois applications qui démontrent l'intérêt de ma technique de Pixels Miroirs. Pour la complétude de l'application de télé-explication, il convient maintenant de développer la couche réseau qui va permettre la collaboration entre plusieurs utilisateurs. L'application de la technique des Pixels Miroirs pour traduire la profondeur dans une Interface 3D nécessite plus de développement car pour l'instant le prototype développé est statique : en effet l'utilisateur ne peut pas pour l'instant manipuler ni déplacer les fenêtres. Enfin pour complètement développer l'application de retouche d'images, l'interface utilisateur doit être conçue et développée. Pour l'instant les changements de paramètres et l'extraction d'encre se font par des touches clavier dédiées. Il convient donc de développer l'interface graphique de cette application. Lorsque les applications seront complètement finalisées, il conviendra alors de conduire des tests ergonomiques afin de valider expérimentalement les propriétés ergonomiques identifiées dans mon espace de conception.

A plus long terme, j'envisage de nombreuses extensions à l'application de retouche d'images qui a mes yeux est très intéressante aussi bien pour la création graphique que pour l'enseignement. J'envisage en particulier de concevoir une technique de colorisation de l'encre acquise, la couleur étant choisie par l'utilisateur. De plus l'application gagnerait en souplesse d'utilisation si une pile de calques (correspondant à chaque acquisition d'encre) était gérée. En effet dans l'état actuel du prototype, la fusion successive de l'encre acquise est définitive et se fait dans l'ordre d'acquisition.

Plus généralement, les perspectives à cette étude consistent en la création de nouvelles modalités d'interaction fondées sur l'image vidéo. La vidéo est largement utilisée et à des fins diverses. L'espace de conception permet d'analyser et de concevoir de nouvelles applications

basées sur ce média. La vidéo peut être à la base de nouvelles méthodes d'interaction. L'effet miroir est une de ces nouvelles méthodes d'interaction. Ces perspectives s'inscrivent dans mon étude de DEA qui fait suite à cette étude de magistère 2ème année.

Références

[Brittan 94] David Brittan ; Being There : the promise of multimedia communications. Technology review 92 / CSCW 94 p57.

[Ishii 91] Hiroshi Ishii, Naomi Miyake ; Toward An Open Shared Workspace : computer and video fusion approach of teamworkstation. Communications of ACM december 1991 , Vol 34 , No 12.

[Ishii 94] Hiroshi Ishii ; TeamWorkStation 1989-1994. http://www.media.mit.edu/~ishii/TWS.html

[Bérard 99] François Bérard ; Vision par ordinateur pour l'interaction fortement couplée. Thèse de doctorat de l'université Joseph Fourier, à paraître.

[Wellner 93] Peter. D. Wellner ; Adaptative Thresholding for the DigitalDesk, EuroPARC Technical Report EPC-93-110.

[Kuzuoka 94] Hideaki Kuzuoka, Toshio Kosuge, Masatomo Tanaka; GestureCam : A Video Communication System for Sympathetic Remote Collaboration. ACM 1994 p35.

[Fitzmaurice 95] George W.Fitzmaurice, Hiroshi Ishii, William Buxton ; Bricks : Laying the Foundations for Graspable User Interfaces.

[Stafford-Fraser 96] James Quentin Stafford-Fraser ; Video-Augmented Environments. A disertation submitted for degree of Doctor of Philosophy.

[Vernier 99] Frédéric Vernier, Christophe Lachenal, Laurence Nigay, Joëlle Coutaz ; Interface Augmentée par effet miroir. A paraître dans IHM99.