Vis Comput (2009) 25: 25–37
DOI 10.1007/s00371-008-0283-3

ORIGINAL ARTICLE

Mobile phone-based mixed reality: the Snap2Play game

**Tat-Jun Chin · Yilun You · Celine Coutrix ·
Joo-Hwee Lim · Jean-Pierre Chevallet · Laurence Nigay**

Published online: 13 August 2008

Abstract The ubiquity of camera phones provides a convenient platform to develop immersive mixed-reality games. In this paper we introduce such a game which is loosely based on the popular card game "Memory", where players are asked to match a pair of identical cards among a set of overturned cards by revealing only two cards at a time. In our game, the players are asked to match a "digital card", which corresponds to a scene in a virtual world, to a "physical card", which is an image of a scene in the real world. The objective is to convey a mixed-reality sensation. Cards are matched with a scene identification engine which consists of multiple classifiers trained on previously collected images. We present our comprehensive overall game design, as well as implementation details and results. We also describe how we constructed our scene identification engine and its performance. Finally, we present an analysis of player surveys to gauge the potential market acceptance.

Keywords Mixed reality · Scene identification · Memory game · Mobile phone

1 Introduction

The increase in functionality and processing power of mobile phones in conjunction with the massive bandwidth afforded by advanced telecommunication networks allow a plethora of innovative and interesting game applications to be developed for mobile phone users, e.g., [1, 11]. In addition, the existence of high-resolution cameras on many mobile phones opens up another interesting direction for games based on image or video processing techniques.

In this paper we describe an innovative game application which exploits cameras on mobile phones. Our game consists of using the phone camera as an input interaction modality (a way for the user to interact with the game). The aim of our game, apart from the obvious role of providing entertainment, is to convey a feeling of "mixed reality" (i.e., a seamless interchange between a virtual world and the real world) to the user. Indeed our design approach is based on providing similar input modalities for interacting with the virtual world and with the real world. The game is inspired by the popular card game "Memory", where players are asked to match a pair of identical cards among a set of overturned cards by revealing only two cards at a time. In our game, the players are asked to match a "physical card", which represents the real world, to a "digital card", which is a token of the virtual world.

The game begins with the process of collecting a digital card: The player goes to a pre-determined location, and, with

T.-J. Chin (✉) · Y. You · J.-H. Lim · J.-P. Chevallet
Image Perception, Access and Language Lab, Institute
for Infocomm Research, 21 Heng Mui Keng Terrace,
Singapore 119613, Singapore
e-mail: tatjun@gmail.com

Y. You
e-mail: ylyou@i2r.a-star.edu.sg

J.-H. Lim
e-mail: joohwee@i2r.a-star.edu.sg

J.-P. Chevallet
e-mail: viscjp@i2r.a-star.edu.sg

C. Coutrix · L. Nigay
Laboratoire d'Informatique de Grenoble, Université
Joseph Fourier, B.P. 53, 38041 Grenoble cedex 9, France

C. Coutrix
e-mail: celine.coutrix@imag.fr

L. Nigay
e-mail: laurence.nigay@imag.fr

the aid of a custom software on the phone which exploits orientation sensing hardware, points the phone camera at a pre-determined direction and orientation (i.e., at a "virtual scene"). This is to simulate a photo taking experience in the virtual world, and it is emphasized that the corresponding real-world scene at which the player is inadvertently guided to aim is inconsequential for the game. Upon "snapping" the virtual scene at the right vantage point, the player receives the digital card which is actually an image of a real-world scene in a separate location.

Drawing from his familiarity of the local geography or guidance from the system, the player proceeds towards the location inferred from the digital card. Upon reaching the correct area, the player attempts to capture the scene with an image (the physical card) which resembles the digital card contents as closely as possible. The physical card is then transmitted to a service provider which employs a scene identification engine to verify the card. If verification is obtained, the player has successfully matched a pair of cards, and he can continue to collect the remaining card pairs (by first receiving directions to the next digital card). In a competitive setting, the player who collects all cards first is the winner.

We call this game "Snap2Play", and the overall success of the game is indicated when the player, apart from being entertained, is immersed during the interchange between the physical and virtual world, i.e., he is only mildly aware of the differences between capturing the digital card and the physical card. The main components to achieve this are interaction techniques and devices that provide mixed-reality feelings, and a image-based scene identification engine.

The rest of the paper is organized as follows: Sect. 2 describes in detail the rules and flow of the game. Section 3 describes a very important aspect of the game which is how we design and implement interaction techniques and devices that are capable of simulating a photo taking experience on a mobile phone for collecting the digital card. Section 4 explains the other major aspect of our game which is how our scene identification engine is constructed. Section 5 reports empirical results of the scene identification engine, and also an analysis of player surveys. Finally, the conclusion is drawn in Sect. 6.

2 How the game works

Snap2Play can be implemented as a single- or multiplayer (co-operative or competitive) game. The player interacts with the game through a mobile phone installed with the Snap2Play application. The overall flow of the game is coordinated through the mobile phone network (i.e., GPRS) by a game server (henceforth, the "game system"). Upon starting, the player is first introduced to the rules and objectives of the game, and is prompted to select his preferred game trail based on the descriptions provided by the application. The design of the game trail allows the incorporation of various vested interests, e.g., introduction to tourist spots or shopping precincts, promotion of physical fitness. The player then proceeds to hunt for the first digital card by receiving, from the game system through the phone network, a message with the rough location of the card. The aim is for him to enter the "Primary Search Area (PSA)" of the card; see Fig. 1.

The system will automatically detect when the player enters the PSA, after which a 2D digital compass which points to the direction of the digital card is activated on the phone. Based on the guidance by the compass, the player moves towards the digital card until he eventually reaches a "Reduced Search Area (RSA)". This is the pre-determined physical location, with a certain level of tolerance, where the digital card is embedded at a pre-defined orientation in space. The above steps are achieved through using GPS navigation.

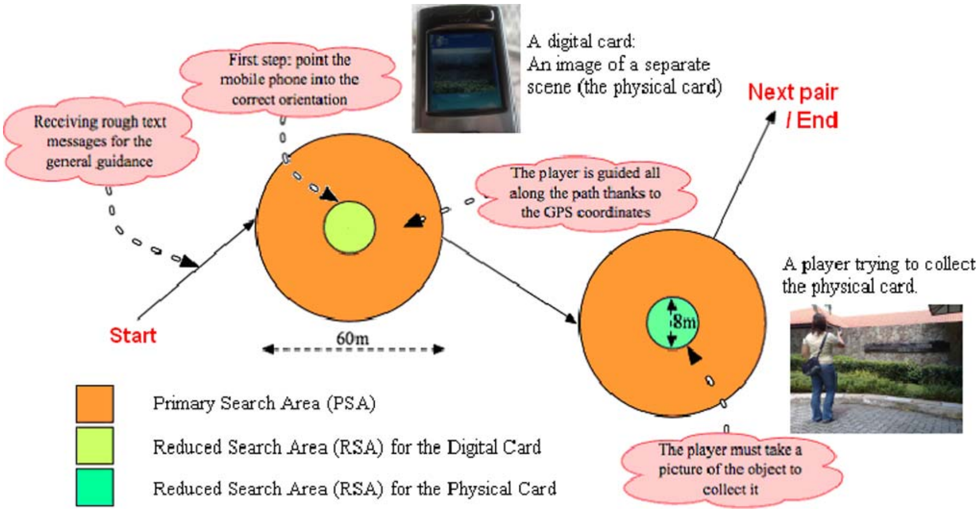

A digital card:
An image of a separate scene (the physical card)
First step: point the mobile phone into the correct orientation
Receiving rough text messages for the general guidance
Next pair / End
The player is guided all along the path thanks to the GPS coordinates
A player trying to collect the physical card.
8m
Start
60m
The player must take a picture of the object to collect it
🟧 Primary Search Area (PSA)
🟩 Reduced Search Area (RSA) for the Digital Card
🟩 Reduced Search Area (RSA) for the Physical Card

Fig. 1 The Snap2Play game scenario

Once the player enters the RSA, a text notification is given to activate the phone camera and to use it to locate and snap the digital card. The player is guided with a software that exploits attached orientation sensors on the phone to aim the camera at a pre-determined vantage point of the digital card. When the correct view-point is established, the player is notified through a superimposed image on the video feed or through tactile feedback, and the player can trigger the camera to collect the digital card. The elaborate manner in which the digital card is obtained is aimed at simulating the experience of capturing an image in a virtual world. Section 3 describes in detail how this is achieved.

The digital card is actually an image of a scene in a separate location. The player will need to find the location inferred from the digital card based on either his familiarity of the local geography or guidance from the system. When the player reaches the PSA of the physical card, the compass will be invoked to guide the player towards the corresponding RSA; see Fig. 1. The player proceeds by trying to snap an image of the scene (the physical card) to match the contents of the digital card. The physical card is sent (e.g., via MMS) to the system to be verified by a scene identification engine. Section 4 describes how the scene identification engine is constructed. If the cards are deemed matching, the player can continue to collect the remaining cards (by receiving a message with the rough location of the next digital card) if the game has not finished.

3 Collecting a digital card

The goal of this subsystem is to simulate a photo-taking experience of a virtual object, i.e., the digital card. It is a "mixed object", since the digital card is also augmented with a physical location. We aim to simulate a photo-taking experience of a digital card as being similar to taking a picture of a physical object or building. To this end we design and implement interaction techniques and devices that impart sensations of performing the intended actions in the physical world. The following describes how we achieved this.

3.1 Designing the interaction technique

For the conceptual design of such an interaction technique, we used the Mixed Interaction Model [3]. To describe mixed systems, the model focuses on mixed objects. It helps in designing and identifying task objects and tools. In this case, the task object is the augmented digital card and we modeled interaction techniques for collecting it.

According to the Mixed Interaction Model, a mixed object is defined by its physical and digital properties as well as the link between these two sets of properties. The link between the physical and the digital parts of an object is defined by "linking modalities". As shown in Fig. 1 (bottom part), a linking modality includes the two levels of an interaction modality [12], i.e., a pair (device, language). As opposed to interaction modalities used by the user to interact with mixed environments, the modalities that define the link between physical and digital properties of an object are called linking modalities. The mixed object augmented digital card in Fig. 2 (bottom part) has two types of linking modalities: Input linking modalities acquire and interpret a subset of physical properties and output linking modalities are in charge of generating physical properties based on the set of digital properties. The model helps in designing such a mixed object. For more details, refer to [3, 4].

3.2 Embedding and locating a virtual scene

Here we explain how the mixed interaction model described previously is materialized. Three parameters are required to embed a virtual scene in the physical world: a 2D coordinate (i.e., a GPS location), the direction (relative to the north) to face the scene, and the orientation (relative to the ground) to view the scene. The three different sensing instruments to achieve this, respectively, a GPS receiver, a compass, and a tri-axis accelerometer, are described in Sect. 3.3. When designing the game trail, the three parameters of all digital cards are recorded. Locating a digital card during play is a matter of re-producing its embedding parameters.

Acquiring the first parameter involves the straightforward process of reading from a GPS receiver and digital compass attached to the phone by the Snap2Play application. When locating a digital card during play, the system iteratively monitors the GPS location and heading of the player through the GPRS network and notifies the player when he is sufficiently close to a PSA or RSA (cf. Sect. 2 and Fig. 1). Once the correct GPS coordinate is obtained, the system notifies the Snap2Play application to activate the video feed and begins to monitor the compass and accelerometer reading in order to provide guidance to the pre-determined "vantage point" (orientation) from which to view a virtual scene. Feedback to the player is given visually through the video feed, whereby a variable sized image of the digital card (to signify divergence from the correct orientation) is superimposed. Feedback via video is crucial to compel the player to acquire a digital card as though he is acquiring a physical card. Once the right orientation is obtained, the player triggers the phone to collect the digital card.

3.3 The hardware

In our implementation, two standalone devices are required apart from the mobile phone. First is the Holux GPSlim 236 Bluetooth GPS receiver which contains a SiRF-Star-III chipset, allowing it to have an accuracy within 3 m; see

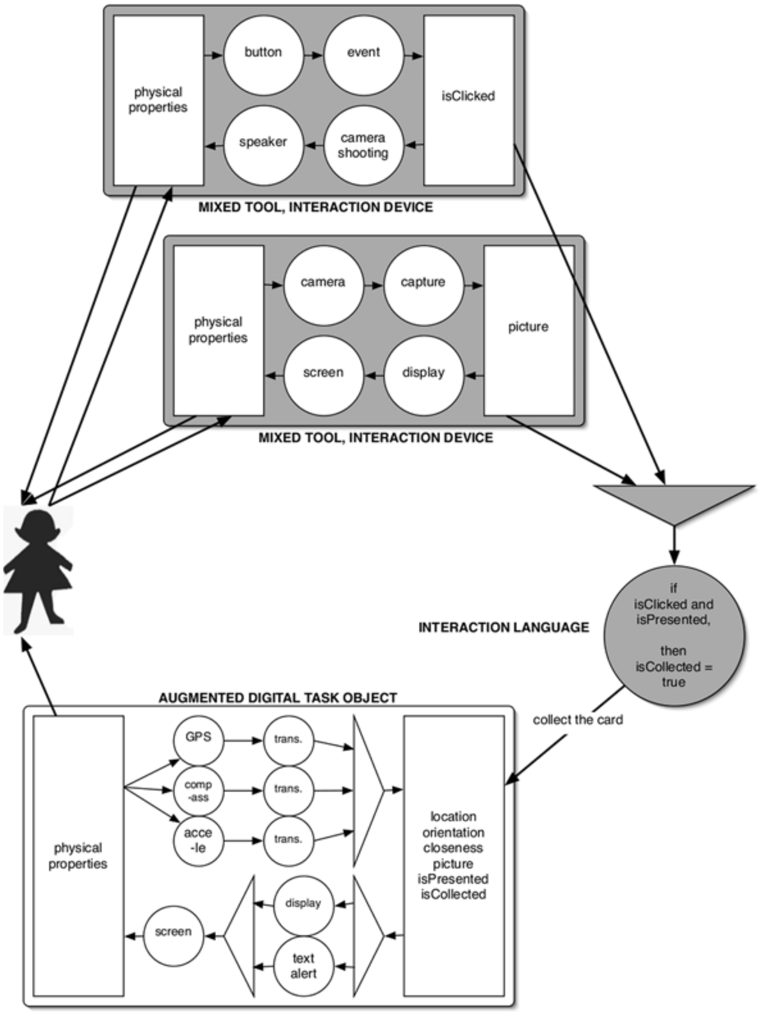
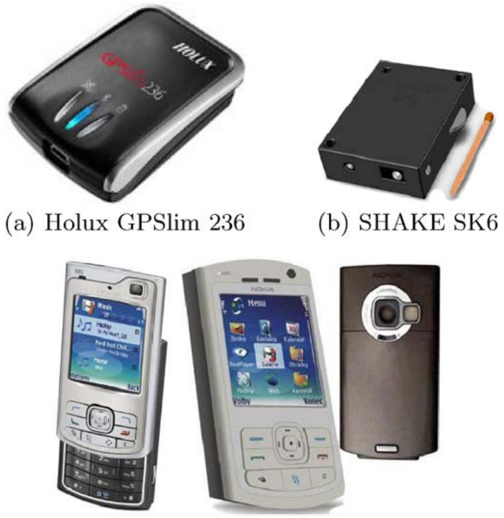


Fig. 2 The Mixed Interaction Model from which we designed the interaction technique for collecting the digital card



(a) Holux GPSlim 236

(b) SHAKE SK6

(c) Nokia N80

Fig. 3 The hardware components in our implementation

Fig. 3(a). This is digitally connected to the mobile phone via the Bluetooth protocol. Its small size allows the player to carry it conveniently in a pocket. Despite the impressive accuracy, the existence of high-rise buildings and frequently overcast sky in our game trails make GPS reading unrepeatable, hence the inclusion of the PSA and RSA to allow some degree of tolerance of positioning error in the game. Note that more advanced phone models with onboard GPS receivers will render such an extra device unnecessary.

In addition, a compass and a tri-axis accelerometer are also needed for the game. While the compass gives heading relative to the magnetic north, the tri-axis accelerometer provides orientation-sensitive readings. These functions can be conveniently provided by an instrument called the "Sensing Hardware Accessory for Kinaesthetic Expression (SHAKE)" device. The "SK6" model is used in our game. The accelerometer needs to be attached to the phone, and fortunately the small form factor of the SK6 allows this to be easily achieved; see Fig. 3(b). It is also digitally interfaced to the phone via the Bluetooth protocol. Finally, the mobile phone used in our system is the Nokia N80 model[1] which has a 3 Megapixel camera. Figure 3(c) shows the actual phone used in our system.

4 Collecting a physical card

The collection of a physical card requires the determination, without soliciting manual attention, of whether a physical card (an image) corresponds to the scene at which a player should be seeking to match the current digital card he is holding. Many previous works on object and scene category recognition (e.g., [5, 7, 8, 10]) provide excellent potential solutions. We take a *discriminative* machine learning approach for this task. Our scene identification engine consists of classifiers trained on previously collected sample images of the scenes. A discriminative approach is favored here due to its simplicity, in that less parameter selection effort is required, and effectiveness, as is evident in [8, 10]. Note that the scene identification engine resides on the coordinating server of Snap2Play and not on the mobile device.

4.1 Multiple classifiers for scene identification

Our scene identification engine consists of N classifiers H_n, where N is the number of physical cards to be collected in a particular route, and $1 \leq n \leq N$. Given a physical card $\mathbf{I}$, the following result is obtained:

$$n^* = \operatorname*{argmax}_n H_n(\mathbf{I}), \tag{1}$$

where $H_n(\mathbf{I})$ evaluates the confidence of $\mathbf{I}$ belonging to the nth scene. Provided that $H_{n^*}(\mathbf{I})$ is larger than a predetermined threshold, $\mathbf{I}$ is assigned the label n^*. Otherwise, the system is programmed to decline classification. If n^* matches the digital card that is currently pending, then $\mathbf{I}$ is successfully paired.

Before training H_n, a set of sample images of the N scenes have to be collected. Ideally, the samples should be captured in a manner that includes, as much as possible, the variations expected from images taken by the players as Fig. 4 illustrates. Secondly, the type of feature to be extracted from input images $\mathbf{I}$ on which H_n operates has to be determined. Experimental results from object and scene category classification [5, 7, 8, 10] suggest that local features which are invariant to affine transformations are very suited for the task. In particular, the SIFT [9] and SURF [2] frameworks have been proven to be effective and robust against distortions caused by rotation, scaling, affine transformations, and minor lighting changes. They comprise two common stages, namely keypoint detection and local descriptor assignment, although both stages were developed from different physical considerations. Each descriptor is a feature vector, typically of 64 or 128 dimensions, which represents a keypoint. Matching descriptors of different keypoints amounts to comparing the underlying visual patterns which gave rise to the keypoints. Figure 4 shows the SIFT and SURF keypoints of a sample image. In our current Snap2Play prototype we used the SIFT method.

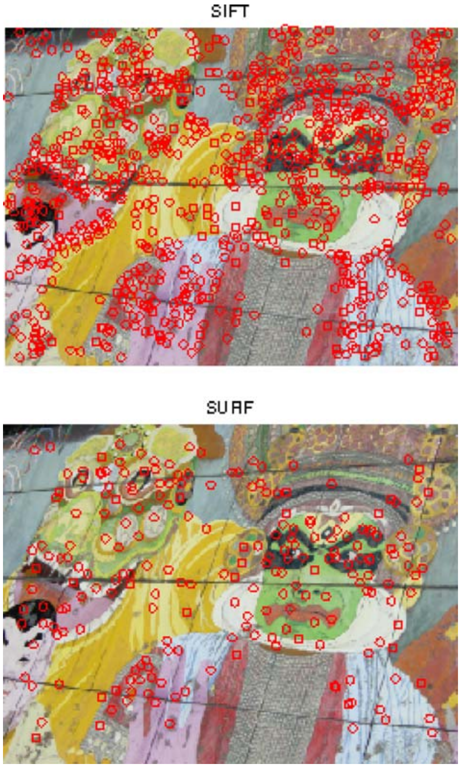


Fig. 4 SIFT and SURF keypoints on a sample image. Since these methods have different characteristics, the detected keypoints do not necessarily overlap

4.2 Acquiring and condensing sample image sequences

Frequently the physical cards correspond to large man-made structures, for example, shopping malls, monuments, or even large murals. In practical circumstances the player is likely to take and submit any physically accessible facade of the structure as an answer. Consequently data collection becomes complicated: Capturing from afar to fit a structure in an image causes important visual features (e.g., SIFT, SURF) to disappear, so naturally we prefer close range images. However, close distances allow a large number of distinct viewing positions and viewpoints, and hence, distinct facades of the structure.[2] The data collector is presented

[1] Effort is currently underway to migrate to the Nokia N95.

[2] Here, "viewing position" differs from "viewpoint". The former implies an x–y coordinate on the surface of the Earth, while the latter refers to the direction towards which to observe a structure given a viewing position.

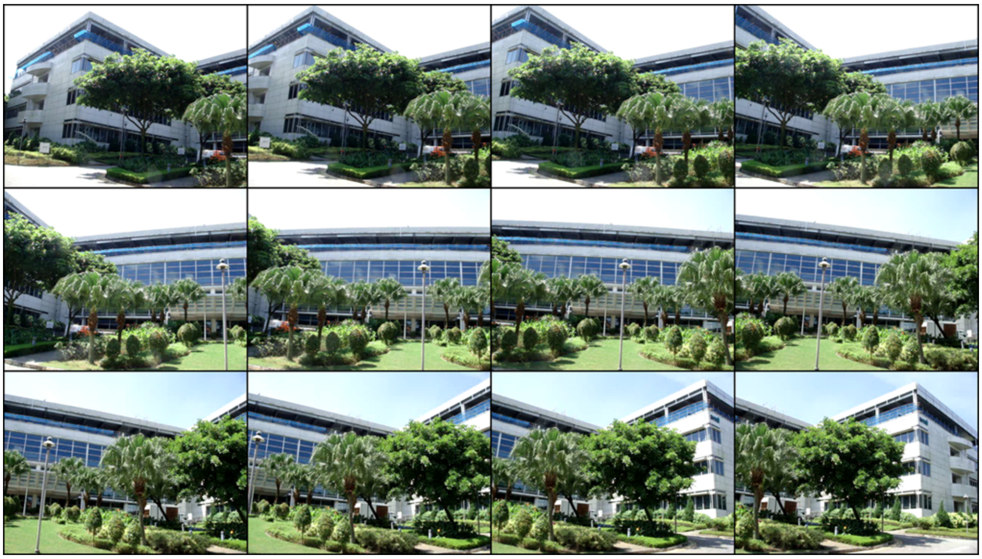


Fig. 5 Capturing the facades of a large structure by recording a video. The video frames contain a lot more information about the structure than a few images

with the dilemma of choosing among them, especially if each facade has the equal probability of being presented for validation.

To alleviate this problem we propose to capture large structures with video recordings. Instead of snapping just a few images, the collector pans smoothly to capture a structure in video, i.e., in a "pure planar motion" [6]. For a given viewing position, a set of video recordings can acquire much more information about a large structure than a few still images. Figure 5 illustrates the idea.

Collectively a massive number of keypoints are detected in the frames (images) in one video, and we must reduce the number of keypoints to consider for training image classifiers. Since our videos are recorded in a smooth panning motion, many of the keypoints in a frame will be re-occurrences from the previous frames (but in slightly differing views). We can track keypoints across the sequence to identify the overlaps.

Let $\{(\mathbf{x}_i, \mathbf{p}_i)\}$ and $\{(\mathbf{y}_j, \mathbf{q}_j)\}$ be the sets of keypoints detected in two successive frames. Symbols $\mathbf{x}_i$ and $\mathbf{y}_j$ denote the keypoint positions and $\mathbf{p}_i$ and $\mathbf{q}_j$ their descriptors. Since the images represent two views of the same scene, a homography $\mathbf{H}$ exists between corresponding points:

$$\mathbf{H}\tilde{\mathbf{y}} \times \tilde{\mathbf{x}} = 0, \tag{2}$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ indicate the homogenous coordinates of $\mathbf{y}$ and $\mathbf{x}$. Our aim is to find the best homography $\mathbf{H}^*$.

To achieve this, we first compute a pairwise similarity matrix using the Euclidean distance between $\mathbf{p}_i$ and $\mathbf{q}_j$. All possible corresponding keypoints between the two frames are identified by considering that a pair of keypoints are matching if the distances of their descriptors are below a predefined threshold. $\mathbf{H}^*$ is determined as the $\mathbf{H}$ that allows the largest number of corresponding keypoints to overlap (i.e., the distance between $\tilde{\mathbf{x}}$ and $\mathbf{H}\tilde{\mathbf{y}}$ is below a certain threshold). We perform a RANSAC procedure to estimate $\mathbf{H}^*$. For more details, refer to [6].

The process is repeated successively on each frame pair, and overlapping keypoints are accumulated into the same track, while keypoints without matches are simply discarded. The descriptors of keypoints in the same track are summarized with the mean vector of the descriptors. This is achieved in an incremental manner during keypoint tracking:

$$\boldsymbol{\mu}_{t+1} = \frac{m}{m+1}\boldsymbol{\mu}_t + \frac{1}{m+1}\mathbf{p}_{t+1}, \tag{3}$$

where $\boldsymbol{\mu}_t$ is the current mean vector of a track consisting of m overlapping keypoints, $\mathbf{p}_{t+1}$ is the descriptor of a keypoint newly assigned to the track, and $\boldsymbol{\mu}_{t+1}$ is the updated mean vector. The mean vector, essentially a descriptor itself, is regarded as the representation of a particular consistently re-occurring visual pattern.

Figure 6 illustrates the idea, and Fig. 7 shows an example result. As an indication of effectiveness, a typical structure to be registered as a physical card is fully captured by a 25-frame video. This collectively contains a total of about 30,000 SIFT keypoints. Among these, only about 4,000 are determined as unique by the method. Hence, the procedure condenses a large number of keypoints from a large structure to a much smaller and manageable number of keypoints. This will speed up the image classifier training algorithm introduced next.

4.3 Training classifiers via boosting

Many possibilities exist to construct H_n given a set of keypoint descriptors condensed from videos of structures intended as physical cards. We apply the technique of *boosting* to train H_n. A popular boosting algorithm, the AdaBoost

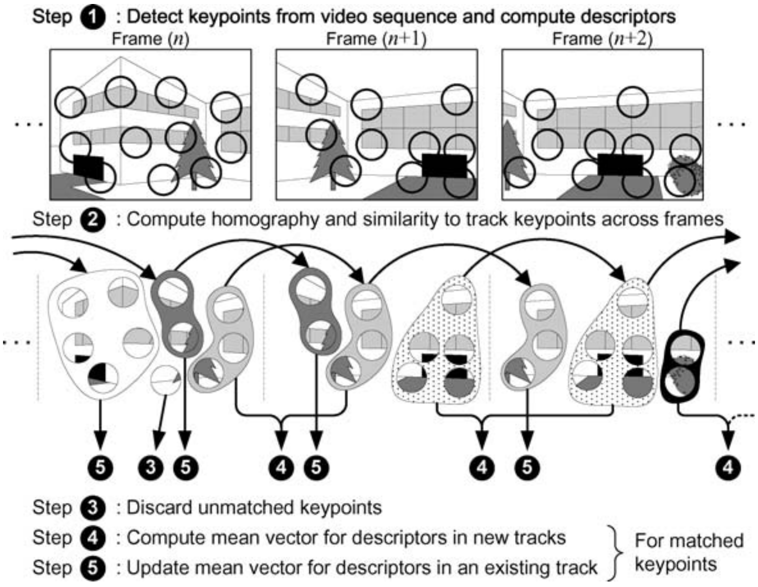

Step 1 : Detect keypoints from video sequence and compute descriptors
Frame (n)
Frame (n+1)
Frame (n+2)
Step 2 : Compute homography and similarity to track keypoints across frames
5
3
5
4
5
4
5
4
Step 3 : Discard unmatched keypoints
Step 4 : Compute mean vector for descriptors in new tracks
Step 5 : Update mean vector for descriptors in an existing track
For matched keypoints

Fig. 6 Finding keypoint overlaps using temporal continuity

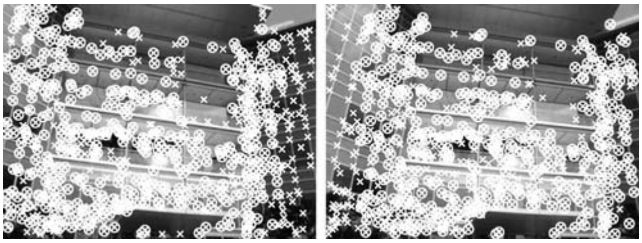


Fig. 7 Overlapping keypoints are re-occurrences of the same local feature. In this pair, crosses are detected keypoints, and those with bounding circles indicate that an overlap is found

procedure, was adapted in [10] for object-class recognition with impressive results, and we apply their method here. Basically boosting constructs the desired classifier H_n by linearly combining T *weak* classifiers:

$$H_n(\mathbf{I}) = \frac{1}{\sum_{t=1}^{T} \alpha_n^t} \sum_{t=1}^{T} \alpha_n^t h_n^t(\mathbf{I}), \tag{4}$$

where h_n^t is the tth weak classifier of H_n, and α_n^t is its corresponding weight ($\alpha_n^t \geq 0$).

A weak classifier h_n^t determines whether query image $\mathbf{I}$ belongs to scene n. Each h_n^t must be able to classify correctly at least only half of the time (hence "weak classifiers"), but when their decisions are aggregated a competent overall classifier H_n can be obtained. For our task, a weak classifier is defined as

$$h_n^t(\mathbf{I}) = \begin{cases} 1 & \text{if } \min d(\mathbf{v}_n^t, \mathbf{p}) \leq \theta_n^t \text{ for all } \mathbf{p} \text{ from } \mathbf{I}; \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\mathbf{v}_n^t$ is the defining feature of h_n^t, and $\mathbf{p}$ is one of the keypoint descriptors detected in $\mathbf{I}$. Function $d(\cdot,\cdot)$ is a predetermined distance metric (e.g., Euclidean) in the descriptor space, while constant θ_n^t is a closeness criterion for h_n^t. More intuitively, h_n^t is activated (returns '1') if there exists at least one keypoint in $\mathbf{I}$ that is sufficiently similar to $\mathbf{v}_n^t$. Finally, it can be seen that $H_n \in [0, 1]$.

Table 1 The AdaBoost algorithm to train a classifier H_n for the nth physical card

Input: Condensed training videos $(\mathbf{S}_1, l_1), \ldots, (\mathbf{S}_K, l_K)$ with labels, where $l_k = +1$ if $\mathbf{S}_k$ belongs to the nth physical card, and $l_k = -1$ otherwise. The total number of weak classifiers T to combine.

Initialization: Set weights $w_1 = \cdots = w_K = 1$.

Perform

```
for $t = 1, \ldots, T$ do
    Find best weak hypothesis $h_n^t$ with respect to
    $w_1, \ldots, w_K$ using the weak hypothesis finder
    routine in Table 2.
    Compute $\mathcal{E}_n^t = (\sum_{k=1,h_n^t(\mathbf{S}_k)\neq l_k}^{K} w_k)/(\sum_{k=1}^{K} w_k)$,
    where $h_n^t(\mathbf{S}_k)$ means subject all the condensed
    descriptors of $\mathbf{S}_k$ to weak classifier $h_n^t$.
    Compute $\beta_n^t = \sqrt{(1-\mathcal{E}_n^t)/\mathcal{E}_n^t}$ and $\alpha_n^t = \ln \beta_n^t$.
    Update $w_k \leftarrow w_k \cdot (\beta_n^t)^{-l_k \cdot h_n^t(\mathbf{S}_k)}$.
end for
```

Output: T weak classifiers h_n^t with corresponding weights α_n^t. Threshold of h_n^t is given by the weak hypothesis finder routine in Table 2.

For a particular physical card, any descriptor mean vector condensed from its video frames can be used to define a weak classifier, i.e., to be used as the $\mathbf{v}_n^t$ in (5). The goal of AdaBoost is to select a subset of all available weak classifiers for which the defining visual features, as far as possible, simultaneously exist in the facades of that physical card while at the same time do not appear in the facades of other physical cards. The algorithm which is shown in Table 1 iteratively chooses weak classifiers depending on how well they perform on sample videos with different weights, which are in turn computed based on how well previously selected weak classifiers perform on these samples. The closeness criterion for a weak classifier is provided by the weak hypothesis finder routine in Table 2.

5 Results

5.1 Scene identification performance

We first examine the performance of the scene identification method independently of the game. To achieve this we collected an image database of landmarks that can potentially be used as physical cards. The landmarks are mainly buildings with prominent facades in our campus. They were recorded in video in the manner described in Sect. 4.2 and Fig. 5. The device used is an off-the-shelf consumer digital camera. In our database, the length of the videos range from 1 s to 10 s depending on the size of the place, while the framerate is kept at 25 fps. Also depending on the physical size, 3 to 6 videos were recorded for each landmark. In general, larger buildings require not only more videos, the videos are also lengthier. Figure 8 illustrates the types of landmarks we have collected. We recorded 44 different landmarks which amount to about 21,000 image frames or 1.5 GB of data. Videos of the same landmark were assigned the same class label.

Table 2 The weak hypothesis finder routine

```
Input: Labeled condensed descriptors $(\mathbf{v}_k^f, l_k)$ and weights $w_k$, where
$1 \le k \le K$ and $1 \le f \le F_k$. $F_k$ is the total number of condensed descriptors in the $k$th video $\mathbf{S}_k$.

1. Define distance metric $d(\cdot, \cdot)$ for descriptors.
2. For all descriptors $\mathbf{v}_k^f$ and all videos $\mathbf{S}_j$, find the
   minimal distance between $\mathbf{v}_k^f$ and local features
   in $\mathbf{S}_j$:
       $d_{k,f,j} = \text{argmin}_{1 \le g \le F_j} d(\mathbf{v}_k^f, \mathbf{v}_j^g)$.
3. Sort the minimal distances as such: For all $(k, f)$,
   find a permutation $\pi_{k,f}(1), \ldots, \pi_{k,f}(K)$ such that
       $d_{k,f,\pi_{k,f}(1)} \le \cdots \le d_{k,f,\pi_{k,f}(K)}$.
4. Select the best weak hypothesis as such: For all $\mathbf{v}_k^f$,
   compute the following value
       $\text{argmax}_s \sum_{j=1}^{s} w_{\pi_{k,f}(j)} l_{\pi_{k,f}(j)}$
   and select $\mathbf{v}_k^f$ for which the above value is the
   largest.
5. Compute threshold for best weak hypothesis as
       $\theta = \frac{1}{2}(d_{k,f,\pi_{k,f}(s^*)} + d_{k,f,\pi_{k,f}(s^*+1)})$,
   where $s^*$ is the value of $s$ at the maximum in
   Step 4.

Output: Best weak classifier defined by the selected $\mathbf{v}_k^f$ and its corresponding $\theta$.
```

For each landmark, a separate testing set of *still images* (collectively 1349 images) were also captured in an *unconstrained* manner on different days, from different viewing positions and viewpoints. They are labeled according to the landmark class labels. Note that our digital camera, like most consumer models, captures video in a much lower resolution (480 × 640) than still images. The pre-processing steps we applied include resizing the images to 240 × 320 pixels and a color to greyscale conversion.

It is quite improbable that a game trail includes 44 physical cards since it would be too physically demanding. Therefore, we randomly sample the database to form subsets of 5 landmarks each. The landmarks in a subset represent the physical cards that a player would encounter in a game trail. The testing set is sampled accordingly. A total of 24 such subsets were created. The classification accuracy and training duration of the proposed method were evaluated and averaged across each subset. This is compared against applying [10] directly on the database images without employing the video summarization procedure. Figure 9 depicts the results.

From the Receiver Operation Characteristic (ROC) curves of the classifiers, compared to applying [10] directly, it is evident that the proposed method can train more accurate scene identification engines. This is due to the fact that

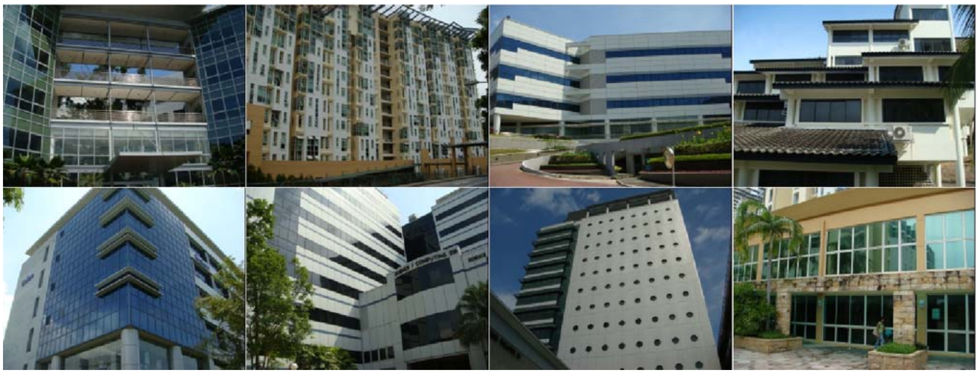
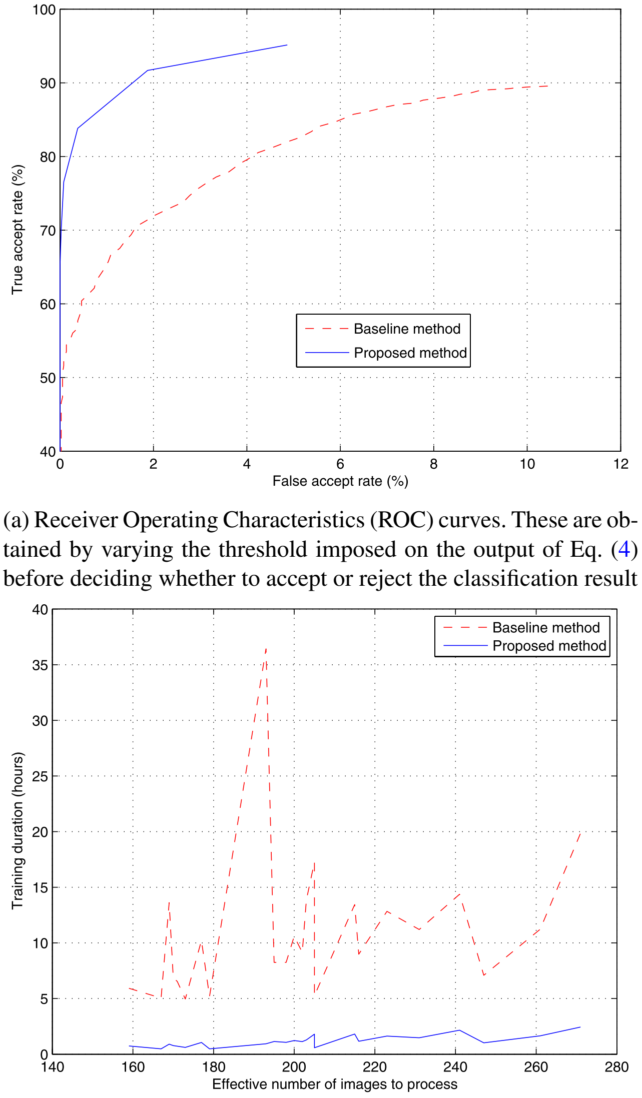


Fig. 8 Types of landmarks in our database

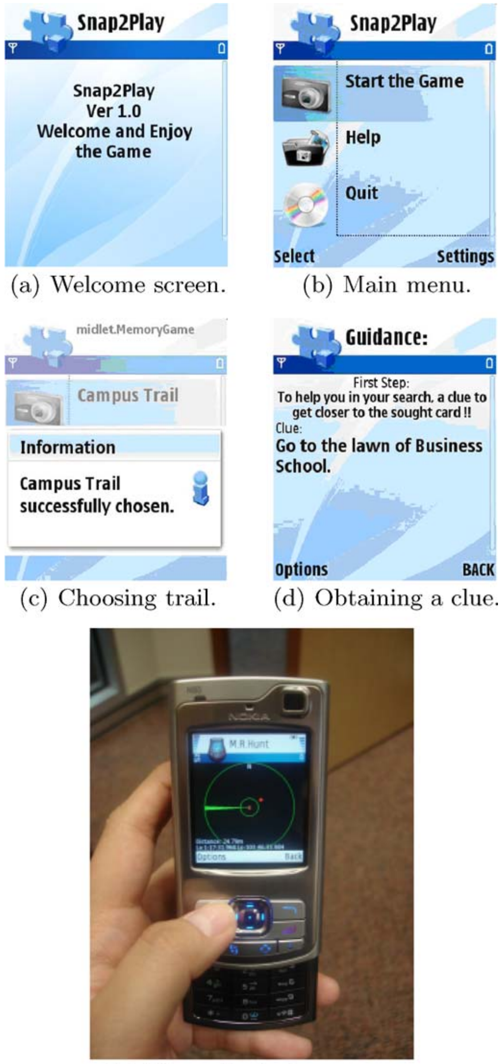


(a) Receiver Operating Characteristics (ROC) curves. These are obtained by varying the threshold imposed on the output of Eq. (4) before deciding whether to accept or reject the classification result

(b) Training durations. The X-axis corresponds to the number of training images in a subset

Fig. 9 Scene identification results and training durations. "Baseline method" indicates applying [10] directly without employing the video summarization technique

summarizing images of the training videos according to the proposed method allows us to filter out spurious keypoints, leaving only keypoints that are consistently reoccurring across frames to be used for boosting. Conceding a false acceptance rate of 2% provides a true acceptance rate of more than 90%—this is more than sufficient for Snap2Play where in the course of a game most players would query only once or twice for each physical card. This also means that we can impose a high threshold to guard against "impostor" images (images not corresponding to any of the physical cards) intentionally taken by the players to fool the system. In terms of training durations the proposed method is *at least* 10 times quicker than applying [10] directly. This is because the video summarization procedure reduces, by a big margin, the number of descriptors to consider for boosting. Thus, preparing the scene identification engine for a trail with 5 physical cards takes just about 1 hour.



Snap2Play
Ver 1.0
Welcome and Enjoy
the Game

Campus Trail
successfully chosen.

First Step:
To help you in your search, a clue to
get closer to the sought card !!
Clue:
Go to the lawn of Business
School.

(a) Welcome screen. (b) Main menu.

(c) Choosing trail. (d) Obtaining a clue.

(e) The activated compass.

Fig. 10 Several screenshots of the user interface of Snap2Play on the Nokia N80 of our system

5.2 Our system and an example trail

We implemented Snap2Play as a single-tier Java Platform Micro-Edition (J2ME) client application. The client was developed in a modular approach to allow future versions to incorporate new input modalities and sensors. Snap2Play requires an interface to capture and perform scene identification on the physical card. We built an interface with a client–server architecture which allows transmission of the image to the game server via GPRS or WiFi. The communication layer of the our interface is developed using Java Platform Standard-Edition (J2SE) and the scene identification engine

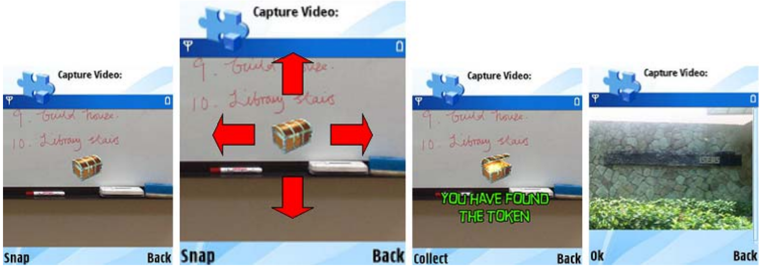


Fig. 11 When attempting to collect a digital card an animation is superimposed on the video feed which reacts according to the phone movement. The player must point the phone at the right orientation and direction so as to position the animation in the middle of the screen before the card can be collected. Upon collection a digital card which serves as a clue to the corresponding physical card is given to the user

is built in C++ for performance. Figure 10 illustrates the actual user interface on the mobile phone of our system, while Fig. 11 depicts the process of collecting a digital card.

Next we present an example of a Snap2Play trail. This trail is designated along a path in our campus, as shown on a map in Fig. 12, and it contains three pairs of cards. Representative buildings of institutes in our campus were chosen as the physical cards, as depicted in Fig. 13(a). Actual images submitted by the trial players for physical card verification are shown in Fig. 13(b). Large magnitudes of variations in lighting condition and pose as well as unintentional occlusions exist in these images. The robustness of the scene identification engine allows it to successfully recognized images taken in such conditions. However, severe pose changes and affine distortions can still result in failure of identification.

5.3 Player surveys

Thirty trial players with ages in the range of 15 to 25 were recruited to play and complete a survey at the end of the game. The trail to follow was the campus trail described in Fig. 12. The objective of the survey is to gauge the market acceptance level of Snap2Play among early adopters, and also to find areas for further improvement. To ensure that the surveys were completed with independent opinions, only one player was allowed to follow the trail at one time. The players begin after some briefing, and most took around one hour to collect all three pairs of cards. Some crucial survey results are shown in Fig. 14.

Figures 14(a) and 14(b) allow us to conclude that, from the players' point of view, the performance of the scene identification engine in terms of accuracy and speed was excellent, since a majority of them find that the number of attempts required to validate a physical card and the speed to achieve that validation is acceptable. An overwhelming majority of the players agree that the user interface of Snap2Play on our Nokia N80 was easy to use, as depicted in Fig. 14(c). According to Fig. 14(d) most players find the game entertaining, and would not mind playing it again, albeit there are some who remarked that they would only attempt it again if the trail was different. Hence, it is confirmed that Snap2Play has enormous potential, but the design of the trail and the places at which the digital and physical cards are embedded play a very integral role in ensuring the continued popularity of the game. Finally, we attempt to estimate, from the players' point of view, to what extent the mixed reality feeling is conveyed. Results in Fig. 14(e) show that the players find the process of collecting both types of cards easy, suggesting that both digital and physical cards require similar amounts of effort to complete. A more direct question in Fig. 14(f) reveals that most players agree that the experience of collecting the digital and physical card is similar, thus allowing us to conclude that the goal of mixed reality is fairly successful.

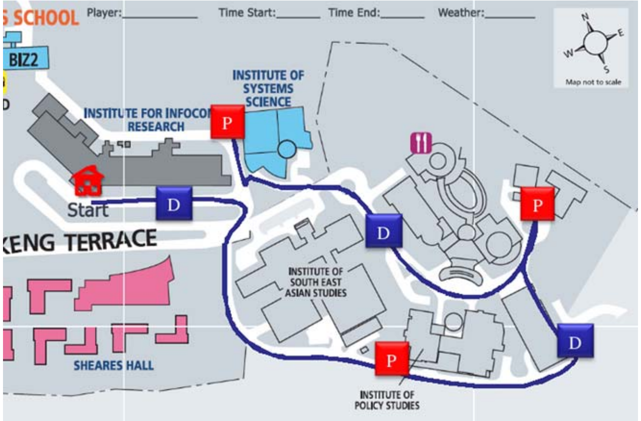
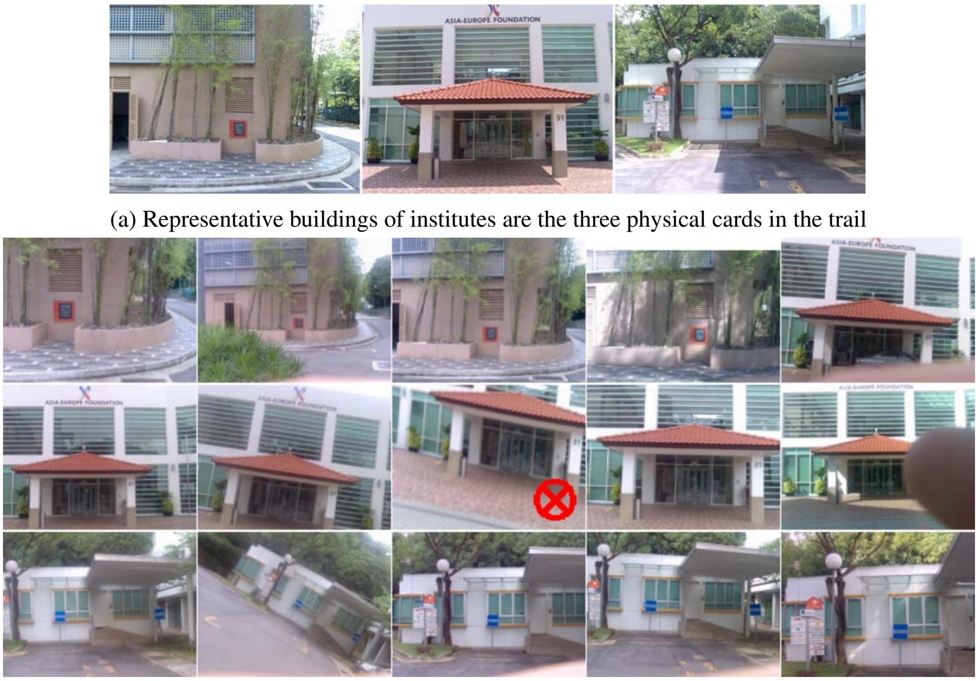

S SCHOOL
Player:
Time Start:
Time End:
Weather:
N
E
W
S
Map not to scale
BIZ2
INSTITUTE OF SYSTEMS SCIENCE
INSTITUTE FOR INFOCOMM RESEARCH
P
Start
D
ENG TERRACE
D
P
INSTITUTE OF SOUTH EAST ASIAN STUDIES
D
P
SHEARES HALL
INSTITUTE OF POLICY STUDIES

Fig. 12 A Snap2Play trail in our campus. "D" and "P", respectively, indicate digital and physical cards. This trail is approximately 1 km long

6 Conclusion and future plans

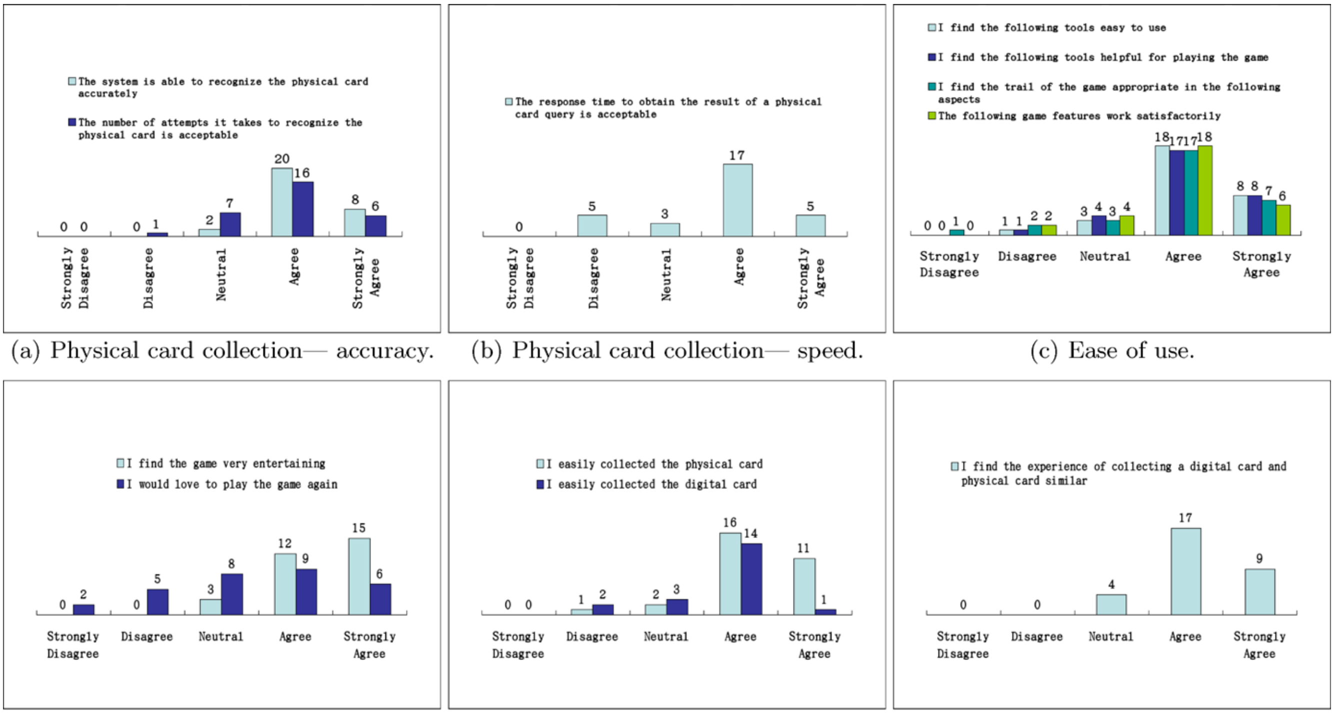
In this paper we propose a novel mixed-reality game which is implemented on mobile camera phones. The game requires the player to match a physical card in the real world to a digital card in a virtual world. This was achieved by exploiting orientation sensing hardware and a scene identification engine. Experimental results show that boosting de-



(a) Representative buildings of institutes are the three physical cards in the trail

(b) Actual images submitted by players for physical card verification. All these are correctly recognized except the one with the cross due to severe distortions

Fig. 13 Collecting physical cards of the trail



(a) Physical card collection— accuracy.

(b) Physical card collection— speed.

(c) Ease of use.

(d) Overall acceptance.

(e) Mixed reality— card collection.

(f) Mixed reality— experience.

Fig. 14 Some of the crucial survey results

scriptors to construct discriminative classifiers is an effective technique for constructing the scene identification engine, while the proposed video processing method allows us to use large landmarks for physical cards without consuming a prohibitive amount of training time. Player surveys obtained from thirty trial players disclose that our implementation of Snap2Play has mostly achieved our objectives, including being entertaining and conveying the sensation of mixed reality. The surveys also allow us to conclude that Snap2Play is a promising game that can find wide market adoption.

Our next step is to implement Snap2Play in commercial hotspots, e.g., shopping precincts, tourist spots, historical trails, where various commercial interests can be incorporated in the game, e.g., players obtaining promotional offers upon successful completion, trails that involve cards at shops of advertisers. We aim to gather data on the acceptance level of Snap2Play from the public at large, and also to invite industry partners to provide their inputs.

Acknowledgements We convey our enormous gratitude to Emelie Annweiler, Wee-Siang Tan, Kelvin Kang-Wei Goh, and Clement Yuan-Loong Tung for their tireless effort in developing the various aspects of our game. We also reserve countless thanks to Siti Raudah Abdul Aziz and Mohammad Nuremi bin Kamsani of Temasek Polytechnic for inviting, managing and instructing our trial players, and also for compiling the surveys.

References

1. Ballagas, R.A., Kratz, S.G., Borchers, J., Yu, E., Walz, S.P., Fuhr, C.O., Hovestadt, L., Tann, M.: REXplorer: a mobile, pervasive spell-casting game for tourists. In: CHI '07 Extended Abstracts on Human Factors in Computing Systems (2007)
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: European Conference on Computer Vision (2006)
3. Coutrix, C., Nigay, L.: Mixed reality: a model of mixed interaction. In: Conference on Advanced Visual Interface (2006)
4. Coutrix, C., Nigay, L.: Balancing physical and digital properties in mixed objects. In: Conference on Advanced Visual Interface (2008)
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Computer Vision and Pattern Recognition (2003)
6. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2003)
7. Li, F.F., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition (2005)
8. Lim, J.H., Chevallet, J.P., Gao, S.: Scene identification using discriminative patterns. In: International Conference on Pattern Recognition (2006)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
10. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. Pattern Anal. Mach. Intel. **28**(3), 416–431 (2006)
11. Strachan, S., Williamson, J., Murray-Smith, R.: Show me the way to Monte Carlo: density-based trajectory navigation. In: Proceedings of ACM SIG Computer-Human Interaction (CHI) (2007)
12. Vernier, F., Nigay, L.: A framework for the combination and characterization of output modalities. In: International Workshop on Design, Specification and Verification of Interactive Systems (2000)

Tat-Jun Chin received the his Ph.D. degree in 2007 from the Department of Electrical and Computer Systems Engineering (ECSE), Monash University, Victoria, Australia. He is currently working as a Research Fellow in the Institute for Infocomm Research (I^2R) of the Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include computer vision and machine learning.

Yilun You is a research officer in I^2R, A*STAR, and IPAL lab and a participant in the ICT-Asia Project: MoSAIC. She is the engineer in the Snap2Tell and Snap2Play team. She received her B.Sc. (Hons) in Computing from University of Tasmania, Australia in 2005. Her research interests include mobile development, human-factor engineering, and mixed-reality interaction.

Celine Coutrix is a Ph.D. Student at the Joseph Fourier University (UJF, Grenoble 1) and is a member of the HCI research group of the Grenoble Informatics Laboratory (LIG). She graduated with honors in 2005 at ENSIMAG, a leading French Graduate School for advanced Applied Mathematics and Computer Science studies. Her thesis focuses on design and software development of mixed interactive systems that smoothly merge physical and digital worlds. In particular, she presented the Mixed Interaction Model and developed mobile user interfaces.

Joo-Hwee Lim received his B.Sc. (Hons I) and M.Sc. (by research) degrees in Computer Science from the National University of Singapore and his Ph.D. degree in Computer Science and Engineering from the University of New South Wales. He has joined Institute for Infocomm Research and its predecessors, Singapore in Oct. 1990. He has conducted research in connectionist expert systems, neural-fuzzy systems, handwriting recognition, multi-agent systems, and content-based retrieval. He was a key researcher in two international research collaborations, namely the Real World Computing Partnership funded by METI, Japan and the Digital Image/Video Album project with CNRS, France and School of Computing, National University of Singapore. He also contributed technical solutions to a few industrial projects involving pattern-based diagnostic tools for aircraft and battleship navigation systems and knowledge-based post-processing for automatic fax/form recognition. He has published more than one hundred refereed international journal and conference papers in his research areas including content-based processing, pattern recognition, and neural networks. He is currently the Department Head of the Computer Vision and Image Understanding Department, with staff strength of over fifty research scientists and engineers, at the Institute of Infocomm Research, Singapore. He is also the co-Director of IPAL (Image Perception, Access and Language), a French-Singapore Joint Lab (UMI 2955, Jan. 2007 to Dec. 2010).

Jean-Pierre Chevallet received the B.Sc. and M.Sc. degrees in computer science from Grenoble University France (U. Joseph Fourier), the M.Sc. by Research at Grenoble Polytechnic Institute, and the Ph.D. degree in Computer Science in 1992 from Grenoble University. He is Associate Professor at the Grenoble University (U. Pierre Mends-France) since 1993. Since September 2003, he is also Director of IPAL CNRS Mixed International Unit between I2R, and NUS based in Singapore. His research interests are in Information Retrieval, including natural language processing for information indexing and retrieval, multilingual document indexing, logical model of information retrieval, structured document indexing, and multi-media indexing and retrieval. He has participated in several European projects and working groups, in major Information Retrieval competitions including TREC, AMARYLLIS, and CLEF. He is the co-founder of the French Association and Conference for Information Retrieval (ARIA and CORIA) and also reviewer in several top IR international conferences. Dr Chevallet has contributed himself to more than 60 conference and journal papers in the field of Information Retrieval.

Laurence Nigay is a Professor of Computing at the University Joseph Fourier (UJF, Grenoble 1) and at the Institut Universitaire de France (IUF) and is a co-leader of the HCI research group of the Grenoble Informatics Laboratory (LIG). She has received several scientific awards (including the CNRS Bronze medal in 2002 and the UJF gold medal in 2003 and in 2005) for excellence in her research. Her research interests focus on the design and development of user interfaces. In particular, her research studies center on new interaction techniques, Multimodal and Augmented Reality user interfaces. She is currently a coordinator of the European STREP OpenInterface (FP6-35182) on multimodal interfaces. She was vice-chair of the IFIP working group WG 2.7 User Interface Engineering from 1998 to 2004 and a visiting scientist at the University of Glasgow (2001–2002). She has published more than 140 articles in conferences, journals, and books.