RAPPORT DE RECHERCHE

L R I

PELE-MELE, A VIDEO COMMUNICATION SYSTEM SUPPORTING A VARIABLE DEGREE OF ENGAGEMENT

GUEDDANA S / ROUSSEL N

Unité Mixte de Recherche 8623
CNRS-Université Paris Sud – LRI

03/2006

Rapport de Recherche N° 1437

CNRS – Université de Paris Sud
Centre d'Orsay
LABORATOIRE DE RECHERCHE EN INFORMATIQUE
Bâtiment 490
91405 ORSAY Cedex (France)

Pêle-Mêle, a video communication system supporting a variable degree of engagement

Sofiane Gueddana
Projet In Situ, LRI & INRIA Futurs
Bâtiment 490, Université Paris-Sud
91405 Orsay Cedex, France
gueddana@lri.fr

Nicolas Roussel
Projet In Situ, LRI & INRIA Futurs
Bâtiment 490, Université Paris-Sud
91405 Orsay Cedex, France
roussel@lri.fr

ABSTRACT
Pêle-Mêle is a multiparty video communication system that supports a variable degree of engagement. Pêle-Mêle combines computer vision techniques with spatial and temporal filtering of the video streams and an original layout to support synchronous as well as asynchronous forms of communication ranging from casual awareness to focused face-to-face interactions. This note presents the system's design concept and discusses some of its implementation details.

Categories and Subject Descriptors
H.1.2 [Models & Principles]: User/Machine Systems – *Human factors*, H.4.3 [Communications Applications]: Computer conferencing, teleconferencing, and videoconferencing, H.5.2 [User Interfaces]: User-centered design, H.5.3 [Group and Organization Interfaces]: Collaborative computing.

General Terms
Design, Human Factors.

Keywords
Video-mediated communication, computer-mediated communication, multi-scale communication, coordination, communication, collaboration.

1. INTRODUCTION
Video communication systems are most often used for short, synchronous and highly-engaged face-to-face interactions. Previous work on mediaspaces has demonstrated the potential value of long-term video links for casual awareness and informal interaction [5]. Yet, few video systems manages to effectively support both general awareness and face-to-face interactions. Two notable exceptions are Community Bar [6] and MirrorSpace [8], which are systems that provide users with simple ways of choosing the level of engagement that best suits their needs from a discrete (Community Bar) or continuous (MirrorSpace) set of possibilities.

We believe that this notion of variable degree of engagement is a particularly important one for mediated communication. Before showing the user, the system allows him or her to negotiate his or her display and eventually to withdraw. Instant messaging applications, for example, make it easy for users to indicate their current "status" and to adapt the pace of the remote conversation to their local context. They also support transparent transitions between synchronous and asynchronous communication. Existing video systems lack this ability to seamlessly transition from loosely-coupled interactions to highly-coupled ones.

As part of a new research project funded by a major telephone company, we are currently designing a new generation of domestic communication systems. The first and exploratory one, Pêle-Mêle is designed for close groups use in domestic situations, with little risk for privacy (e.g. family, friends). The system functions in two main stages: Firstly, the system automatically detects interesting situations and allows users to adjust details intentionally. Secondly, it uses spatial and temporal composing techniques to display different levels of detail.

The next section presents the design concept behind Pêle-Mêle. We then describe some of its implementation details and discuss the problems encountered.

2. OVERVIEW AND CONCEPT
Pêle-Mêle[1] is a multiparty video communication system. Physically, it consists in a screen equipped with a video camera and connected to a small, unnoticeable computer. The screen displays first an overview of all the connected places and further a more detailed view of the places where someone is actually engaged in a communication. Each representation of a place potentially combines both live images and recorded ones showing previous activity. The layout of all the images is shared among Pêle-Mêle instances on a strict WYSIWIS basis to help users relate one to another and support gaze awareness .

Pêle-Mêle analyses the activity of local users and classifies it according to a three-level scale: *away*, *available* and *engaged*. The activity level observed at each place determines the nature of the video images that represent it:

> *away* – the place is represented by filtered images that illustrate its past;

[1] Pêle-Mêle is a French word for a decorative frame used to display an eclectic set of pictures

available – the place is represented by video clips that show its past and a filtered live stream that illustrates its present;

engaged – the place is represented by video clips and a live stream that simultaneously show both its past and present.

At this point, it is important to understand that these filters, combined with the screen layout, help users perceive the differences between the activity levels. The next section provides the details concerning the filters used on the clips and lives streams.

The layout uses a focus-plus-context approach. The position of a particular video stream depends on the activity level of the corresponding place. Live images from people engaged in a communication are overlaid in the middle of the screen, while images of other available people are shown on the periphery of the display (Figure 1). Smooth animated transitions between these two representations ease perception and understanding the state changes. Time is represented by the z axis (i.e depth). Thus, recorded video clips slowly shrink and drift toward the center of the screen to represent the passage of time (Figure 2).

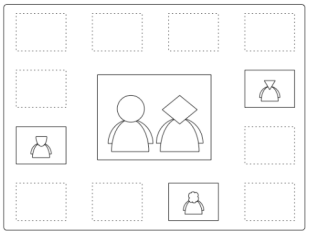


Figure 1. Focus-plus-context view of live video streams.

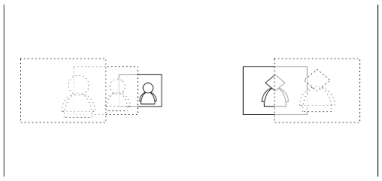
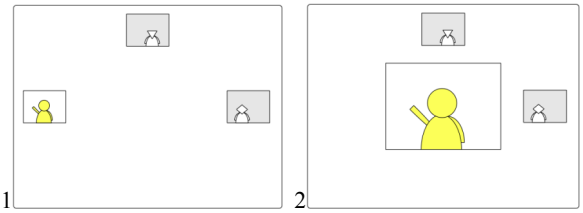


Figure 2. Perspective timeline effect used for recorded images.

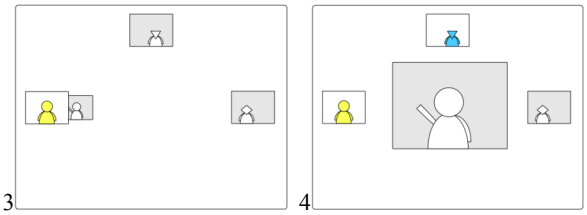
The following scenario further illustrates the differences between activity levels:

Joey, Chandler and Ross each have a Pêle-Mêle at home. Joey has some tickets for tonight's game he would like to share with his friends, but Chandler and Ross are not there. Joey waves at the Pêle-Mêle, which switches from *available* (1) to *engaged* (2). His video stream is automatically recorded while he shows the tickets to the camera.

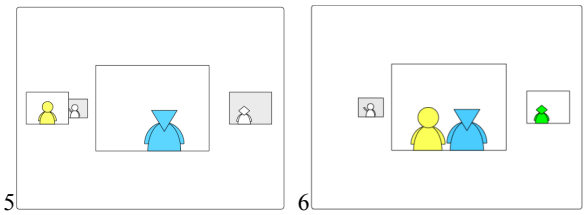


Joey goes back to his comfortable armchair and favorite tv show, which triggers a transition back to *available* (3). The clip that shows him with the tickets has been added to the display, starts to drift with time and is automatically played in the focus area from time to time.

Chandler comes home. His Pêle-Mêle switches from *away* to *available*. Chandler notices Joey's clip as it is played in the focus area (4).



Chandler now wants to talk to Joey about the tickets. He moves towards the Pêle-Mêle, which switches to *engaged* (5). Joey gets up and moves closer to his Pêle-Mêle, which also switches to *engaged*. Their video streams are now superimposed (6) and an audio connection is automatically set up. At the same time, Ross comes home, which switches his Pêle-Mêle to *available*.



3. IMPLEMENTATION DETAILS

Pêle-Mêle is implemented in C++ on an Apple Mac mini computer. It uses the Núcleo[2] toolkit for video capture, recording and transmission. Several spatial and temporal filtering techniques have been implemented to enrich or impoverish the displayed image according to the perceived activity level of the corresponding places. The OpenCV[3] toolkit is used in conjunction with Núcleo to implement basic computer vision techniques to determine these activity levels.

We use OpenCV's face detector based on simple Haar-like features to locate people and estimate the distance that separates them from the device. This technique has a few problems. First, it assumes a "standard" face size, which produces incorrect estimations for people who don't fit that standard (e.g. children). Second, optimal detection occurs for people who are facing the camera at a close distance. Nevertheless, under these particular conditions, it is pretty reliable. The current image is subtracted from the previous one or from a reference image in order to detect motion or presence. More robust techniques based on optical flow computation have also been implemented. However, the simple image difference is considerably faster and is accurate enough for the purpose of this system.

Presence, motion and distance values provided by the computer vision techniques are used to regularly update the activity level (details follow in the three subsections below). Level changes trigger slow animated transitions on the display. Though the chosen techniques are not particularly stable or robust, they proved to be adequate as false positives were more frequent than false negatives (i.e. the perceived activity level might be higher

[2] http://insitu.lri.fr/~roussel/projects/nucleo/

[3] http://www.intel.com/technology/computing/opencv/

than it really is, but is rarely lower). The system thus incorrectly promotes background communications to the foreground from time to time, which can somehow be seen as a feature rather than a bug. Preliminary testing indicates that users quickly understand this tendency and sometimes even take advantage of it.

Figure 3 shows a simplified version of the activity detection algorithm. We will now describe more precisely the operation modes corresponding to each level.

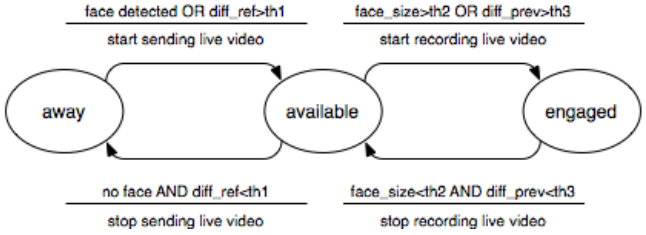


Figure 3. Simplified view of the activity detection algorithm.

3.1 Away

This level is activated when no face is detected in the local image and the difference with a regularly-updated reference frame stays below a certain threshold. These conditions are fulfilled when the camera is covered, the system is stopped or when the user is out of range. The user's space is represented on-screen by a static image, the last captured at *available* level that slowly drifts with the passage of time. The image is filtered to make it clear it is not live and mitigate privacy concerns. Videos previously recorded at the *engaged* level are displayed without filtering. They also drift along time axis and are randomly played from time to time in the focus area. At this level, all displayed pictures are small. When its visible area is reduced, user who wants to watch a picture must approach and enter camera range. In this case, the system detects him and enlarges his picture.

Shown in figure 4, the filter used on the static image is both spatial and temporal: it produces an oil painting effect that gradually removes details over time. This filter is esthetic but has a high computational cost. It reduces detail without suppressing all visible information. For example, it can prevent from identifying an unknown person, and in the same time can provide an idea of recorded activity.

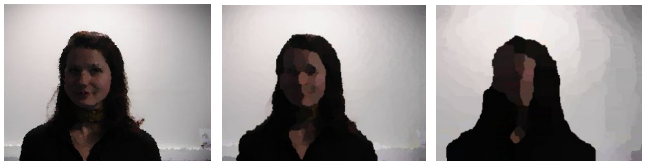


Figure 4. Evolution of filter result with time.

When changing from *away* to *available* level, there are risks for privacy if sharing its picture was immediate. Someone can be unaware of being filmed and could want to prevent his picture from being shown to others. To lower this risk, we introduce a delay in showing pictures captured in this level. This delay change allows the user to prevent publishing intimate situations. The user is warned of being captured by an audio feed-back[blur-filtration]. If someone on front of the camera doesn't want to be seen, he has 15 (or more) seconds to leave or to cover the camera. In this case, accumulated pictures are never shown.

3.2 Available

This level is activated when a face is detected at more than 2 meters or when something new is on front of the camera. For the same reasons as in the previous level, there is a delay when displaying pictures. If an event occurs in the available level that the user prefer to not show, he or she has 15 seconds to select *away* level by covering or turning away the camera. In this case, the most recently captured pictures are not published.

While changing the state from away to available level, the size of the picture grows increasing its visibility (cf. Figure 5). Hence user become more visible to other group members.

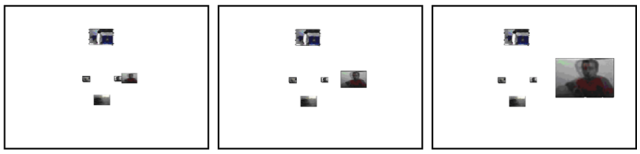


Figure 5. Picture size growth when changing from away to available level.

The *available* level is associated to a moderate level of engagement, in which communication is not the user primary task. Pictures from this level have a medium size and are positioned at the border of the screen. Pêle-Mêle shows the delayed video with time composing [3]: Latest picture is blended with pictures of previous moments. At regular interval, photos are taken and blended with the delayed video stream. These photos disappear progressively with time. With a glance at an *available* user, others can see what is happening at that time and in the recent moments (cf. Figure 6). To blend photos alpha blending was used with transparency increasing over time. This technique quickly and easily conveys photos temporal order.

Figure 6. Two examples of time composed pictures.

We were presented with two main challenges when we tried to improve the visualization permitted by this time window. First, we required a blending technique more adapted to mix several frames from the same stream, because picture rendered by apha blending have a low contrast. Rendering can be improved by using histogram-cutting technique [9]. Second, choosing the most representative moment of the activity to take photos is not trivial. Chance photos can be void of interesting content. To resolve this problem we are investigating techniques developed by artistic installations as LastClock[1] or Fantomes[2]. For example, we are developing a technique to mix a stream with the same stream sent at increasing delays (e.g. 1 sec, 10 sec, 1 min, 10 min, etc).

3.3 Engaged

This level is activated when a face less than 2 meters away from the camera or when a substantial amount motion is detected. It is assumed that if the picture contains a high level of motion or a nearby face, it is worth showing with more detail.

When a place is displayed at this level, its display is big and contains blended images from other sources that are also actively *engaged* (cf. figure 7). Blending more than three different sources is ticky and the results can be confusing. It is difficult to distinguish faces and their associated background. Partial superposition has the potential to limit this problem. To minimize this problem, Pêle-Mêle uses a more transparent alpha for local than for distant pictures. It is the only exception to the WYSIWIS base.

Figure 7. Two then three users engaged together.

While changing from the *available* to the *engaged* level, the image moves to screen center and its size grows. The delay is progressively suppressed. Recording video start in this transition.

We have developed a technique to smoothly resynchronize a delayed stream. The delayed stream is accelerated in order to catch up the present. For example, when somebody is near the device, his picture is big and the stream is live. When he moves farther away, his picture becomes smaller and the stream smoothly decelerate. The video stream become more and more delayed, and finally gets back to normal speed when maximum delay is reached. We used frame dropping to accelerate the video stream. This technique degrades the visual temporal information in two ways. Firstly, motion rhythms are deformed when accelerated; secondly, frame dropping suppresses intermediary frames containing motion-related visual information. We couldn't use lossless acceleration technique, like interpolation, due to the high computational cost not adapted to real time processing.

Preliminary user testing showed tendancy for users to associate low frame rates or delayed video stream to a software or hardware performance problem. This problem can be resolved by using a time indicator that shows explicitly the introduced delay. For example, we can use grayscale picture for delayed stream and colored picture for live stream. For low frame rates, each picture can smoothly disappear before being replaced by the next one.

4. CONCLUSION

Existing video communication systems lack the ability to seamlessly change from loosly-coupled to highly-coupled interaction and from asynchronous to synchronous communication. The Pêle-Mêle system addresses these issues using the notion of variable degree of engagement, with smooth animated transitions. Pêle-Mêle combines computer vision techniques to determine users distance and level of motion, in order to find representative pictures to display.

Pêle-Mêle screen layout and filtering helps users to perceive the differences between the activity levels through a focus-plus-context approach. Filtering and video stream delay were incorporated to mitigate privacy concerns. PêleMêle implements a new time composing technique to improve awareness over time. PêleMêle smoothly changes between asynchronous to synchronous communication. We found that users tended to misunderstand the delayed display as system malfunctions. Further work will address this concerns by expliciting the delay through various visualizations. Globally, permanent video link raise huge privacy concerns. Video representation over time also poses delicate problems.

Further work will look into enriching engaged level to support voice communication. We are developing a command establishing an audio link through Skype[4] started when users superimpose their faces. We are designing now a series of prototypes deriving from the Pêle-Mêle.

5. REFERENCES

[1] Angesleva, J. and Cooper, R. Last Clock. *IEEE Computer Graphics and Applications*, 25(1):20–23, 2005.

[2] Braun, C-A. April in Paris: Signs of Artificial Life along the Seine. *IEEE MultiMedia*, 12(3):14–18, 2005.

[3] Gutwin, C. Traces: Visualizing the Immediate Past to Support Group Interaction. In *Proceedings of Graphics Interface*, pages 43–50, May 2002.

[4] Lee, A., Schlueter, K. and Girgensohn, A. Sensing Activity in Video Images. In *CHI'97 Extended Abstracts*, pages 319-320, 1997. ACM Press.

[5] Mackay, W. Media Spaces: Environments for Informal Multimedia Interaction. In M. Beaudouin-Lafon, editor, Computer-Supported Co-operative Work, Trends in Software Series, John Wiley &Sons Ltd, 1999.

[6] McEwan, G. and Greenberg, S. Supporting social worlds with the community bar. In *Proceedings of GROUP '05*, pages 21–30, 2005. ACM Press.

[7] Neustaedter, C., Greenberg, S. and Boyle, M. Blur Filtration Fails to Preserve Privacy for Home-Based Video Conferencing. *ACM Transactions on Computer-Human Interaction* (TOCHI), to be published.

[8] Roussel, N., Evans , H. and Hansen, H. Proximity as an interface for video communication. *IEEE Multimedia*, 11(3):12-16, July-September 2004.

[9] Vernier, F., Lachenal, C., Nigay, L. and Coutaz, J. Interface augmentée par effet miroir. In *Actes d'IHM'99*, pages 158-165, Novembre 1999. Cépaduès.

[4] http://www.skype.com/