Pêle-Mêle, a Video Communication System Supporting a Variable Degree of Engagement

Sofiane Gueddana
LRI (Univ. Paris-Sud - CNRS) & INRIA Futurs *
Bâtiment 490, Université Paris-Sud
91405 Orsay Cedex, France
gueddana@lri.fr

Nicolas Roussel
LRI (Univ. Paris-Sud - CNRS) & INRIA Futurs *
Bâtiment 490, Université Paris-Sud
91405 Orsay Cedex, France
roussel@lri.fr

ABSTRACT

Pêle-Mêle is a multi-party video communication system that supports a variable degree of engagement. It 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 some of its implementation details.

Categories and Subject Descriptors

H.4.3 [**Communications Applications**]: Computer conferencing, teleconferencing, and videoconferencing

General Terms

Design, Human Factors

Keywords

Video-mediated communication, variable degree of engagement, smooth transitions

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 [4]. Yet, few video systems manage to effectively support both general awareness and face-to-face interactions. Two notable exceptions are Community Bar [5] and MirrorSpace [7], which both 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.

Instant messaging applications make it easy for users to indicate their status and adapt the pace of the conversation to their current context, supporting transparent transitions between synchronous and asynchronous communication. Existing video systems lack this ability to seamlessly transition back and forth between loosely-coupled interactions and highly-coupled ones. We believe that the notions of *variable degree of engagement* and *smooth transitions between degrees* are particularly important for mediated communication and should be taken into account by communication systems designers.

As part of a research project funded by a major telephone company, we are designing a series of image-based communication systems to explore these two notions in the context of the home environment. This work builds on experiences and results from a previous multi-disciplinary project that investigated the communication patterns and needs of distributed families [3, 7]. This project particularly pointed out the importance and difficulty of coordination between and within households, and the need for more subtle, less intrusive forms of communication than the telephone.

This note describes Pêle-Mêle, the first new system we developed. The next section provides an overview of its design concept. We then present some implementation details and conclude with directions for future work.

2. OVERVIEW AND CONCEPT

Pêle-Mêle is a video system designed for between-home close-knit group interaction (e.g. family, friends). It physically consists in a screen equipped with a video camera and connected to a small, unnoticeable computer. The display layout follows a focus-plus-context approach: the screen shows both an overview of all the connected places and a detailed view of the ones where someone is actually communicating through the device. The layout is shared among Pêle-Mêle instances on a strict WYSIWIS basis to help users relate one to another and support gaze awareness.

Pêle-Mêle constantly monitors the activity of local users and classifies it according to a three-level scale: *away*, *available* and *engaged*. The activity observed at each place determines the nature of its on-screen representation, which potentially combines live images and pre-recorded ones that are filtered, delayed or displayed as-is:

away The place is represented by video clips showing past activity and a filtered view of the last image it transmitted.

* projet in|situ| (`http://insitu.lri.fr`), Pôle Commun de Recherche en Informatique du plateau de Saclay (CNRS, Ecole Polytechnique, INRIA, Université Paris-Sud)

CSCW'06, November 4–8, 2006, Banff, Alberta, Canada.

available The place is represented by video clips showing past activity and a delayed live stream.

engaged The place is represented by video clips showing past activity and a live stream which is recorded for later use.

Live images from people engaged in using the Pêle-Mêle are overlaid in the middle of the screen, while available people are shown on the periphery (Figure 1, left). Auditory feedback and smooth animated transitions between these two representations ease perception and understanding of the state changes. Images showing past activity are also displayed on the periphery along a perspective timeline: they slowly shrink and drift toward the center of the screen over time (Figure 1, right).

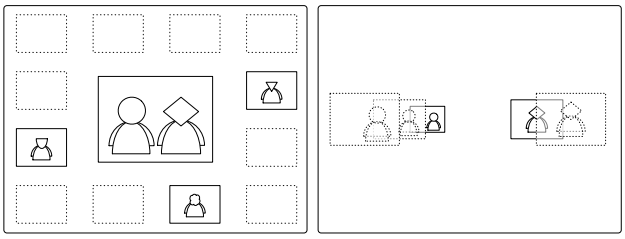
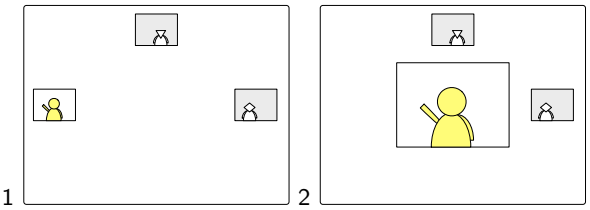


Figure 1: Focus-plus-context view of live streams and perspective timeline effect used for recorded images.

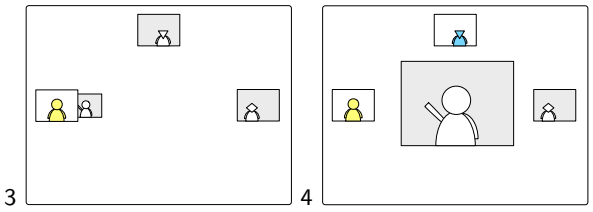
The following scenario further illustrates the concept:

> 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.

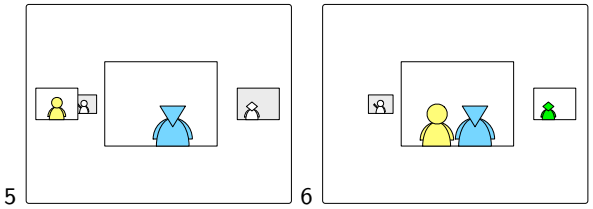


> Joey goes back to his comfortable armchair and favorite TV show, which triggers a transition back to *available* (3). The clip that shows him with the tickets has been added to the display. It slowly drifts in perspective over time and is automatically played in the focus area from time to time. Chandler 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[1] toolkit for video capture, recording and transmission as well as simple presence and motion estimation. OpenCV[2] is used for more complex computer vision techniques such as face detection or optical flow computation. Finally, Pêle-Mêle implements spatial and temporal filtering techniques similar to those proposed by Hudson & Smith [2] or Gutwin [1]. These filters are used, for example, to degrade or delay images to mitigate privacy concerns, or to compose them over time to increase the understanding of each other's activities.

Presence is detected by subtracting a reference image from the current one. Motion is estimated by comparing successive images. More robust techniques based on optical flow computation have also been implemented. However, simple image difference is considerably faster and accurate enough for our purpose. We use OpenCV's face detector to estimate the distance that separates people from the device. This assumes a "standard" face size, which produces incorrect estimations for people who don't fit that standard (e.g. children), and works best for people facing the camera at a close distance. Nevertheless, under these particular conditions, it is pretty reliable.

Presence, motion and distance estimations are used to constantly assess the local activity level (Figure 2). Transitions between levels trigger auditory feedback and slow animated transitions on the display that add some hysteresis into the system. Though the chosen computer vision techniques are not particularly stable or robust, they seem to be adequate. Informal testing indicates that users quickly understand how the system works and how they can adjust their level of engagement through simple movements.

We will now describe more precisely the operation modes corresponding to each activity level.

[1] `http://insitu.lri.fr/~roussel/projects/nucleo/`
[2] `http://www.intel.com/technology/computing/opencv/`

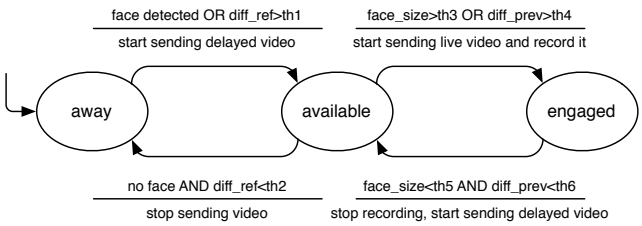


Figure 2: Simplified view of the activity detection algorithm.

3.1 Away

Away corresponds to the situation where no face is detected and the difference with a reference image stays below a certain threshold. At this level, no image is transmitted to the other instances. The place is represented by the last image transmitted at the *available* level and clips recorded at the *engaged* one. These images are displayed in a small size on the periphery of the screen and slowly drift in perspective over time.

Clips are displayed as grayscale images. They are normally represented by a single image, the first one, but get promoted to the focus area from time to time to be played at a larger scale if none of the places is at the *engaged* level. The last *available* image degrades over time to make it clear it is not live and mitigate privacy concerns. As illustrated by Figure 3, the filter produces an oil painting effect that rapidly removes details without suppressing all visible information.

Figure 3: Image degradation over time (one minute, two minutes).

The small size of the images displayed at this level invites users who want to see them to move close to the display. If they come close enough, their own Pêle-Mêle will switch to the *engaged* mode and start recording them. We anticipate that this will in turn support the asynchronous creation of common knowledge by reciprocal exchange of video clips (e.g. I saw you watching me opening the present you sent).

3.2 Available

The Pêle-Mêle switches from *away* to this level when it detects a face or an important change between the current scene and the reference image, in which case this reference is also updated. The assumption we make is that such a change is probably caused by incoming people or previously undetected ones. An auditory feedback is generated and the size of the image representing the place on the periphery is slowly increased to a medium size (Figure 4). If a face is found close enough or if significant motion is detected, the Pêle-Mêle switches further to the *engaged* level. If no face is found and the scene doesn't change anymore, it falls back to *away*.

The *available* level is the one for which privacy concerns are the greatest, as it corresponds to situations where someone might be seen by the Pêle-Mêle without being actively engaged in a communication. In order to reduce the risk of unintended privacy exposure, we introduce a delay of several seconds between the capture of the images and their display. The images, however, are immediately processed by the activity detection algorithm. This allows users to prevent the public display of a particular situation by moving out of the camera's field of view to trigger a transition back to the *away* level before the delay expires. As explained above, the last image transmitted will also be rapidly degraded when used at the *away* level to provide some information without unnecessarily exposing privacy.

In a way similar to what Hudson & Smith or Gutwin proposed [2, 1], the delayed video stream is temporally composed to provide awareness of recent activity. Selected past images are alpha-blended with the current one before it is displayed. The alpha value of each image is inversely proportional to its age, which makes it easy to perceive their temporal order (Figure 5). This technique tends to produce composited images with a low contrast, but histogram stretching techniques can be used to alleviate this problem [8]. The selection of past images is a more complex problem. Our current implementation selects a new image every two seconds and uses the last four ones. But these images, like randomly-selected ones, are often void of interesting content. Video summarization techniques could be useful, but they are usually designed for scenarised videos created from multiple sources and associated to an audio channel, while the image streams we process are taken from a unique and fixed viewpoint.

Figure 5: Examples of time composed pictures.

3.3 Engaged

During the transition between the *available* and *engaged* levels, the size of the video stream slowly increases while it moves toward the center of the screen. Auditory feedback accompanies the transition and the delay is progressively suppressed. To achieve this, the stream is accelerated by dropping some of its images in order to catch up with the present. This technique degrades the visual-temporal information in two ways: it deforms motion but also suppresses intermediary frames containing motion-related information. Lossless acceleration techniques (e.g. frame interpolation) were not used due to their high computational cost.

When the transition is finished, live images are displayed in a big size in the focus area and recorded for later use at the *away* and *available* levels. The images of all the places at the *engaged* level are actually alpha-blended together (Fig-

Figure 4: Image size growth during the transition from *away* to *available*.

ure 6). Although the combined display of local and remote participants is known to improve the co-presence feeling [6], the blending of multiple video sources can be quite confusing, e.g. making it difficult to associate faces and backgrounds. To minimize this problem, Pêle-Mêle uses a lower alpha value for local images. This is the only exception to our WYSIWIS design principle.

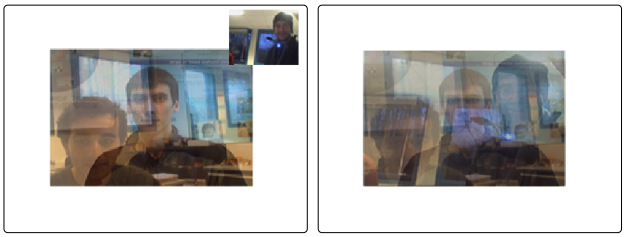


Figure 6: Two then three users engaged together.

The transition back from *engaged* to *available* triggers when a close face and a significant motion are not detected anymore. During this transition, the size of the video stream decreases while it moves back to the periphery. The stream also temporarily decelerates to introduce the few seconds delay mentioned previously.

4. SUMMARY AND FUTURE WORK

Existing video communication systems lack the ability to move from loosely-coupled to highly-coupled interactions, from casual asynchronous awareness to synchronous face-to-face communication. Pêle-Mêle is our first attempt at addressing this problem using the notions of *variable degree of engagement* and *smooth transitions between degrees*. In this note, we presented the general concept of this system and briefly described the computer vision techniques, the screen layout and the image filtering techniques it uses.

Informal testing indicates that users quickly perceive the three levels of engagement currently supported by the system. The layout, the auditory feedback and the animations help them perceive the transitions, and they quickly understand how these transitions can be triggered by simple movements. The effectiveness of the spatial and temporal filters in mitigating privacy concerns and supporting better awareness over time is more difficult to assess. Long term use and participatory (re)design workshops should help us get user feedback on these important issues and improve the system in the future.

Informal testing already showed that users sometimes misunderstand the delayed display of images at the *available* level as system malfunctions. Future work will address this concern by exploring ways of expliciting the delay, by altering the images or enriching them with abstract visual representations. Future work will also investigate different image-based representations of past activity. We are already investigating new ways of selecting past images and composing them to better support awareness of recent activity at the *available* level. Finally, future work will also seek to support additional degrees of engagement. As an example, we are thinking of using a VoIP application to enrich the current *engaged* level with an audio link when users superimpose their faces.

5. ACKNOWLEDGMENTS

This work has been supported by France Télécom R&D as part of the DISCODOM project. Thanks to Danielle Lottridge for her helpful comments on an earlier version of this note.

6. REFERENCES

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

[2] S. E. Hudson and I. Smith. Techniques for Addressing Fundamental Privacy and Disruption Tradeoffs in Awareness Support Systems. In *Proc. of CSCW'96*, pages 248–257. ACM Press, Nov. 1996.

[3] H. Hutchinson, W. Mackay, B. Westerlund, B. Bederson, A. Druin, C. Plaisant, M. Beaudouin-Lafon, S. Conversy, H. Evans, H. Hansen, N. Roussel, B. Eiderbäck, S. Lindquist, and Y. Sundblad. Technology probes: Inspiring design for and with families. In *Proc. of CHI 2003*, pages 17–24. ACM Press, Apr. 2003.

[4] W. Mackay. 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.

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

[6] O. Morikawa and T. Maesako. HyperMirror: toward pleasant-to-use video mediated communication system. In *Proc. of CSCW'98*, pages 149–158. ACM Press, 1998.

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

[8] F. Vernier, C. Lachenal, L. Nigay, and J. Coutaz. Interface augmentée par effet miroir. In *Proc. of IHM'99*, pages 158–165. Cépaduès, Nov. 1999.