

# Audio-visual Privacy Protection for Video Conference

M. Vijay Venkatesh, Jian Zhao, Larry Profitt and Sen-ching S. Cheung

Center for Visualization and Virtual Environments

University of Kentucky, Lexington KY-40507

Email: (mvijay,cheung)@engr.uky.edu,(Jian.Zhao,drprofitt)@uky.edu,

**Abstract**—Group video-conferencing systems are routinely used in major corporations, hospitals and universities for meetings, tele-medicine and distance learning among participants from very distant locations. As the use of video-conferencing becomes widely prevalent, the privacy concern's raised by this technology becomes an important issue to be addressed. In this paper we propose a real-time privacy preserving video conferencing system which protects the visual and audio privacy of selected individuals. In our proposed system we differentiate between the general participants and private participants (PP) whose privacy needs to be protected. We further divide the private participants into two different categories and provide a varying level of privacy protection based on the requirements. Specifically, among private participants, we have Active Private Participants (APP) who interactively participate in the meeting and Passive Private Participants (PPP) who play a passive observatory role. The video and audio privacy of the APP are protected by obfuscating their visual information by simple black boxing and real-time pitch modification process respectively. For the PPP, we completely protect their privacy by continuously detecting their presence and erasing them with a real-time adaptive background replacement process.

## I. INTRODUCTION AND RELATED WORK

Recent advances in multimedia communications technology have made video-conferencing a standard tool adopted by many organizations for interactive meetings among participants from distant locations. While the highly interactive nature of this medium of communication offers an indisputable advantage, there are also valid questions raised about the breach of privacy of individuals participating in those meetings. Privacy is of obvious concern when video-conferencing systems are used in organizations like hospitals or law firms where the identities of participants often need to be concealed. In law enforcement domain, there are situations where audio privacy of individuals have to be preserved along with the visual information. In corporate meetings, employees involved with the project will actively participate in a video conference while administrative assistants and interns may also be present for support and training. While privacy of these passive participants are not of major concern, their presence serve as a distraction to the remote participants. Considering various circumstances described above, it is clear that a system that provides a layered privacy protection mechanism based on the specific requirement would be greatly beneficial.

The goal of the privacy protected video-conferencing system is to provide a balanced approach that allows interactive

behavior of general participants while protecting the privacy of the selected participants based on different requirements. We identify three important criterion that need to be addressed in designing such a system: First, we need to identify how to protect the visual and audio privacy of selected individuals. Second, the privacy protection mechanism should not restrict the normal behavior of both the general and private participants. Third, the technique should be capable of running at real-time to be useful in such an interactive medium.

A large variety of video obfuscation techniques have been proposed in the literature to protect visual privacy of individuals. These techniques range from the use of black boxes, large pixels (pixelation), face masking, scrambling, complete object removal by background replacement and dynamic object inpainting [1], [2], [3]. These techniques based on their sophistication provide increasing levels of privacy protection. While complete object removal provides a definite advantage in terms of privacy protection, its full-scale implementation requires inpainting of occluded motion and dynamic background which is a very time consuming process.

Audio privacy protection by cryptographic approaches provide secure processing capabilities but are time consuming due to computational overhead. Automatic speech transcription offer good level of protection but their performance is not reliable under noisy conditions. Signal obfuscation based approach like pitch shifting and voice conversion are being increasingly used for privacy protection. Recent studies have shown that privacy protection by pitch shifting offers efficient audio privacy protection and at the same time maintain the intelligibility of the modified data [4].

The contribution of our work is to provide an efficient mechanism for audio and visual privacy protection by pitch shifting and video obfuscation schemes respectively under real-time constraints. The rest of the paper is organized as follows: Section II provides a high level description of the proposed privacy protecting video-conferencing system. In Section III we explain the modified adaptive background subtraction process that allows us to segment and track different participants in real time. We then discuss our audio privacy protection mechanism in Section IV. Finally we provide the experimental results in Section V and conclude with the future work.

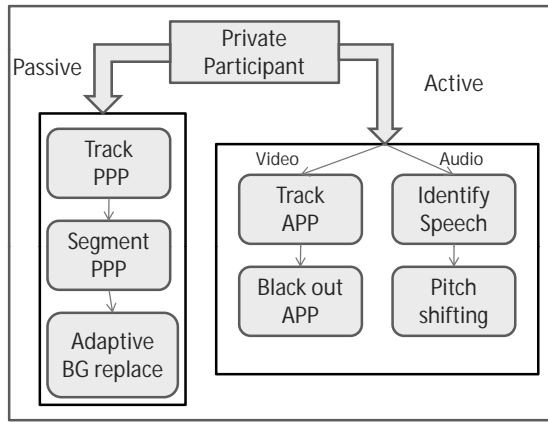


Fig. 1. Block diagram of the privacy protection scheme

## II. SYSTEM DESIGN OVERVIEW

Our system provides a layered privacy protection based on requirements by recognizing two classes of private participants, the APP and PPP. The block diagram of the protection mechanism provided in our system is shown in Figure 1.

We allow the APP to interactively participate in the conferencing setting without any constraints. Their private visual information is protected by simply blacking out the visual appearance and distorting the corresponding audio information using a pitch shifting process. As a result of the simple black out process, the visual information of the other participants will also be blocked when there is an occlusion with the APP. In case of PPP, we completely protect the video privacy by continuously detecting their presence and erasing them with a real-time adaptive background replacement process. The advantage is that the privacy protection mechanism completely erases the presence of the PPP. To accomplish this task under real time constraints, we impose some movement restrictions on the PPP. Such restrictions are not limiting due to their inherently passive observatory role.

## III. VIDEO PRIVACY PROTECTION SYSTEM

We begin by describing the essential operations that are needed to provide privacy protection for visual information.

- *Background-foreground separation*: Identify all the participants in the video as foreground regions by comparing with a background model of an empty room.
- *Object segmentation*: Segment the individuals when there is an occlusion between participants in the video.
- *Private participant identification*: Identify the private participants at all time instances.
- *Privacy Protection*: Provide a layered privacy protection based on classification between APP and PPP respectively.

As illustrated in the Section I, most of the algorithms performing these operations come with a high computational cost and therefore are not suitable for real-time video-conferencing systems. We exploit the nature of typical video-conferencing

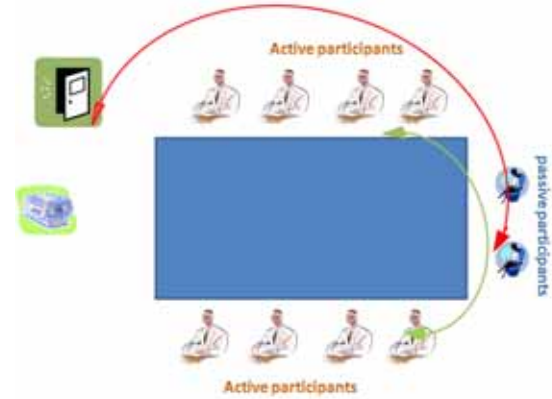


Fig. 2. Seating arrangements for PPP in video conference setting

scenarios and make two assumptions to alleviate some of the computational burden:

- 1) We restrict the seating and movement of PPP's in such a way that they never occlude other participants as shown in Figure 2. This condition alleviates the need to perform the time consuming process of dynamic foreground inpainting. We only need to locate the position of the PPP's at all time instances and replace them with the background. We do not have this restriction for the APP since we use a simple black-out.
- 2) In most circumstances, the participants stay relatively stationary and occlusions between two moving participants are relatively uncommon. Specifically, we assume that *there is no occlusion between a moving PPP and other moving participants*. This is a reasonable assumption as the participants usually appear disjoint in the video and the occlusion occur only when some participants enter or leave the conference.

In our system, we develop a fast object segmentation scheme by combining a stationary background model to detect all the participants and an adaptive background model to identify moving participants. Based on our second assumption, each of the foreground blobs detected by the adaptive background model corresponds to a single moving participant. All the blobs are then tracked from frame to frame. Using our first assumption, the genesis of the object track can provide information of the passive participant. Objects classified APP and PPP are accorded with their respective levels of privacy protection. The details of the adaptive background subtraction process along with the object segmentation process is explained below.

### A. Background modeling and object segmentation

Each pixel in the stationary background model is modeled as a 4-tuple  $(E_i, \sigma_i, a_i, b_i)$  where  $E_i$  is the expected color value,  $\sigma_i$  is the standard deviation of color value,  $a_i$  is the variance of the brightness distortion and  $b_i$  is the variance of the chromaticity distortion of the  $i^{th}$  pixel as described in [5]. An improved version of this method which can adapt

to varying conditions was introduced in [6]. However it comes with high computational cost due to the overhead required for adapting  $a_i$  and  $b_i$ .

Most of the adaptive background subtraction algorithms use a small the learning rate  $\alpha$  to update the background model. The background model is therefore updated as an interpolation between the previous background  $B^{t-1}$  and the current frame  $I^t$ . That is,

$$B^t = (1 - \alpha)B^{t-1} + \alpha I^t \quad (1)$$

This value of  $\alpha$  needs to be carefully tuned to provide the appropriate level of adaptation. Inspired by [7], we adopt a Time Out Map (TOM) to keep track of how long a pixel has been classified as a foreground object. Instead of a consensus based approach in [7] which requires buffering a large number of frames to compute the static foreground, we use the static background subtraction method described in [5] to estimate the foreground. Note that this modification will only enable us to identify the foreground object. It does not provide any further information to differentiate between an existing static foreground pixel or a newly added dynamic foreground pixel due to object motion. This greatly simplifies our model and is adequate for our purpose where occlusions occur briefly most of the time. We compute the running mean and standard deviation for pixels whose TOM value are greater than 0.

$$\mu_{i,j}^t = \frac{((TOM_{i,j}^t - 1) \cdot \mu_{i,j}^{t-1} + I_{i,j}^t)}{TOM_{i,j}^t} \quad (2)$$

$$\sigma_{i,j}^t = \sqrt{\frac{s_{i,j}^t - TOM_{i,j}^t \cdot (\mu_{i,j}^t)^2}{TOM_{i,j}^t - 1}} \quad (3)$$

where

$$s_{i,j}^t = \sum_{k=1}^{TOM_{i,j}^t} (I_{i,j}^{t-TOM_{i,j}^t+k})^2$$

The pixels that have a TOM larger than a threshold, which we set to 15, will be used to update an adaptive background model with their correspondent means and standard deviations. This threshold plays a similar role as the buffer size of median filter background subtraction. In general this choice of threshold is not sensitive and we show in our experiment that it is generally safe to set a large threshold as most individuals turn to be static most of the time during a conference. The foreground object mask thus obtained by using this adaptive background model denotes the foreground object with significant motion.

$$MF_{i,j}^t = \begin{cases} 0 & |I_{i,j}^t - \mu_{i,j}^t| < a \cdot \sigma \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

$a$  is a control parameter typically set between 2 and 3. This mask can be used to segment the static and moving foreground objects under occlusion by simple difference.

To summarize, we have two sets of background models with different color features. The static background model calculates intensity variance and color distortion as discriminative features and estimates a stable foreground mask with

all the individuals. The adaptive model uses color value as feature to estimate moving object in the scene. It uses a time out map to keep statistics of each pixel and update them respectively, which enables the incremental update of the background statistics.

### B. Object identification and Tracking

Identifying the passive participants requires human interaction. The simplest way is to mark the specific positions of the PP's in the scene at the beginning of the video conference. Foreground blobs are tracked throughout the entire sequence and their initial positions indicate whether they are PP's or other regular participants. In our experiments, blob tracking is achieved by identifying the overlapping area between blobs in successive frames. This simple approach is extremely fast and can solve our problems well since most blobs are relatively static. Also, identification of moving foreground provides us with additional information to greatly improve the segmentation and tracking. After locating all the PP blobs, a simple black boxing in case of APP or a background replacement in case of PPP is used to protect the privacy.

### IV. AUDIO PRIVACY PROTECTION BY PITCH SHIFTING

The first step in audio privacy protection is to detect the instances in which the APP is speaking. In a typical conference setting, there is an omni-directional microphone that records the voice of all the participants. In our system we require the APP to wear a standard bluetooth wireless microphone which helps us to robustly identify the instances in which the APP is speaking. Along with the voice captured by the omni-directional microphone, we record the audio signal transmitted by the wireless bluetooth microphone. The setup is shown in Figure 3. We partition the signal from the bluetooth microphone into equal duration frames of size  $T = 1024$  samples at 16Khz, and perform an amplitude thresholding process to detect the speech activity by the APP. Once we detect speech by the APP, we use the time domain pitch shifting technique to modify the speech throughout its duration, thereby protecting the privacy [8]. To reduce the rapid switching between detection during pauses, we delay the decision by 24 buffer periods, equivalent to 1.5s, before we switch back from distorting the APP signal. The pitch shifting procedure is simple enough to be implemented in real-time but can provide improved privacy protection and maintain the intelligibility of the modified data. In the first step, the input signal of length  $N_1$  undergoes time scaling by a Synchronous Over Lapped Addition (SOLA) process resulting in compression or expansion of the the signal to length  $N_2$ . This is followed by resampling stage that resizes the signal back to its original size  $N_1$ , modifying the pitch of the given signal. As recommended in [4], [9], we set the pitch shifting parameter to  $\alpha = 1.4$ .

### V. EXPERIMENTAL RESULTS

We present results of experiments in which we protect the privacy of a PPP using background replacement

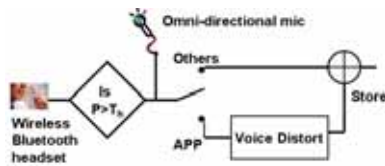


Fig. 3. Audio privacy protection setup for APP

and privacy of the APP using black box. Detailed information on unmodified and privacy protected videos, results of segmentation and tracking, along with the results of audio privacy protection involving APP are available in our project website <http://vis.uky.edu/mialab/VideoConferencing.html>.

We tested the efficiency of the detection process by capturing an audio sequence with three speakers. Out of the total duration of 104 seconds, the speech from the designated APP occupies 28.5 seconds. Our algorithm misses only 0.1 seconds of the APP's speech, resulting in a recall of 0.98. 0.2 seconds of speech from a non-APP is also distorted, resulting in a precision of 0.99.

Figure 4(a) shows an example frame in which the designated PPP is seated at the back along with four other regular participants. We can also observe that a participant is moving across the hall. Figure 4(b) depicts the result of the privacy protection after erasing the PPP and replacing it with the background.

A more challenging situation happens when there is an occlusion between other participants and the PPP as shown in the Figure 4(c). We can see that the moving person partially occludes the presence of the stationary PPP. The corresponding privacy protected image which preserves the participant while protecting the partially occluded PPP by replacing it with the estimated background model is shown in Figure 4(d).

Figure 4(e) shows the actual frame in which the person at the center is designated as the APP and we can see a regular participant walking behind. The corresponding privacy protected video by a black boxing is in Figure 4(f). The privacy protected audio of the PPP along with other details can be found in our project webpage.

## VI. CONCLUSION AND FUTURE WORK

In this paper we proposed a practical video-conferencing system that can provide layered privacy protection for selected individuals based on requirements. We develop a fast object segmentation scheme by combining two sets of background models: a stationary background model which provides stable foreground regions with all individuals and an adaptive model, which identifies the moving objects in the scene. This approach allows us to perform robust segmentation as a two pass background subtraction process which can be implemented in real-time. We also provide efficient audio privacy protection by using a time-domain pitch shifting technique. Future work would involve making this system more robust by improving the tracking process to allow for movement of multiple par-



Fig. 4. Privacy protection under both schemas: The first row shows the privacy protection of the unoccluded stationary PPP. The 2nd row shows privacy protection of the partially occluded PPP. The 3rd row shows the privacy protection of the APP by black boxing.

ticipants and to improve upon the segmentation of both audio and video sources.

## REFERENCES

- [1] E. N. Newton, L. Sweeney, and B. Main, "Preserving privacy by de-identifying face images," *IEEE transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, February 2005.
- [2] J. W. et.al, "Privacy protecting data collection in media spaces," *ACM Multimedia*, pp. 48–55, October 2004.
- [3] M. V. Venkatesh, S.-C. Cheung, and J. Zhao, "Efficient object-based video inpainting," *Pattern Recognition Letters : Special issue on Video-based Object and Event Analysis*, 2008.
- [4] J. Chaudhari, M. V. Venkatesh, and S. C. Cheung, "Experimental studies on audio privacy protection," *IEEE ICASSP, Under Submission*, 2009.
- [5] T. Horprasert, D. Harwood, and L. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. IEEE Frame-Rate Applications Workshop*, 1999.
- [6] P. A. et.al, "Adaptive parametric statistical background subtraction for video segmentation," in *VSSN '05: Proc. of the third ACM international workshop on Video surveillance & sensor networks*, New York, NY, USA, 2005, pp. 63–66.
- [7] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern Recognition*, vol. 40, no. 3, pp. 1091–1105, March 2007.
- [8] U. Zoelzer, Ed., *Dafx: Digital Audio Effects*. New York, NY, USA: John Wiley & Sons, Inc., 2002.
- [9] J. Chaudhari, S. C. Cheung, and M. V. Venkatesh, "Privacy protection for life-log video," *IEEE SAFE Workshop on signal processing applications for public security and forensics*, 2007.