Privacy Protection for Life-log Video

Jayashri Chaudhari, Sen-ching S. Cheung, M. Vijay Venkatesh
Center for Visualization and Virtual Environments, Department of Electrical and Computer Engineering
University of Kentucky, Lexington, KY 40507, USA

***Abstract*— Recent advances in wearable cameras and storage devices allow us to record the holistic human experience for an extended period of time. Such a life-log system can capture audio-visual data anywhere and at any time. It has a wide range of applications from law enforcement, journalism, medicine to personal archival. On the other hand, there is a natural apprehension towards such an intrusive system as the audio-visual information of unsuspecting subjects captured in the life-log video may be misused. Thus, along with the technical challenges, the privacy and legal issues arising in such recordings must be carefully addressed. In this paper, we describe a wearable life-log system that combines real-time audio distortion and visual blocking to protect the privacy of the subjects captured in life-log video. For audio, our system automatically isolates the subject's speech and distorts it using a pitch-shifting algorithm to conceal the identity. For video, our system uses a real-time face detection, tracking and blocking algorithm to obfuscate the faces of the subjects. Extensive experiments have been conducted on interview videos to demonstrate the ability of our system in protecting the identity of the subject while maintaining the usability of the life-log video.**

I. Introduction

It has been fifty years since the pioneering vision of Vannevar Bush to build a system that can record all aspects of human experiences [1]. Recent advancements in wearable technologies, wireless networks and storage devices have finally enabled us to build such a life-log system that can continuously record almost all human experience for days at a time [2]. While there have been notable advances in the availability of hardware components needed to build a practical system, key challenges are being actively addressed in the software technology required to manage and analyze the huge amount of unstructured multimedia data captured in the process.

Privacy is one of the most neglected issues that need to be carefully handled in a system that records everything, everywhere, and at every moment. The natural apprehension towards such an intrusive system stems from the fact that the audio-visual information of unsuspecting individuals may be misused. Legal and privacy related issues in a life-log system called "Total Recall", which captures all experiences of a person, has been previously studied [3]. There are prototype systems for privacy protection in video surveillance [4], [5] and in face recognition [6]. However, to the best of our knowledge, there has been no automatic solution that protects privacy information in life-log video recordings.

This paper presents a wearable system that implements real-time privacy protection for human subjects captured in a life-log video. Our system protects the visual identities of all subjects by first identifying and tracking all faces captured in the video, and then blocking them using solid-color boxes. In our previous work, we have demonstrated that, even in dynamic background, it is possible to completely remove the presence of an individual in a video using video inpainting [7]. However, for this life-log application, we choose to block only the face and leave the rest of the body unobstructed. This is based on the assumption that blocking the face is sufficient for privacy protection while conveying the body language of the subject is conducive in maintaining the utility of such types of video. Our system also protects the audio identity of the subject by distorting his/her voice using a time-based pitch shifting algorithm. We measure the performance of the audio distortion algorithm in concealing the speakers identity by having both human subjects and speaker-identification software to match the original and distorted voices. To ensure the usefulness of the audio recordings after distortion, we also measure the subjective intelligibility of the distorted speeches. To the best of our knowledge, this is a novel study on audio privacy protection that has not been conducted before.

The rest of the paper is organized as follows: we first describe the architecture of the proposed wearable life-log system in Section II. In Section III, we describe our privacy protection mechanism for audio-visual information. In Section IV, we analyze the experimental results of privacy protection scheme and we conclude the paper in Section V.

II. Life-log System

Our wearable life-log system shown in Figure 1 consists of shoulder mounted S.C.O.U.T. camera, omnidirectional microphone, Archos Portable Media Assistant (PMA) which can store 150 hours of video and a VAIO Micro-PC which serves as a processing unit. Various factors such as comfort-to-wear, ability to capture both audio and video with high resolution, and availability of wireless connectivity are taken into consideration in designing the system. The camera can capture NTSC resolution video at 30 FPS with a field of view of 72.5^o and light sensitivity of 0.2 Lux. The microphone is configured to sample the audio at 44 kHz. In order to obtain a stable and unobstructed view of the scene, the camera is mounted on the shoulder. Most of the existing wearable systems use head-mount cameras which result in jittery video due to unconscious head movement. A fully-charged PMA can capture up to 6 hours of data continuously. The Sony VAIO Micro PC uses 1.2 GHz Intel Core Solo Processor with 1GB memory. It weights 1.2 pounds and can last up to 5 hours on battery.

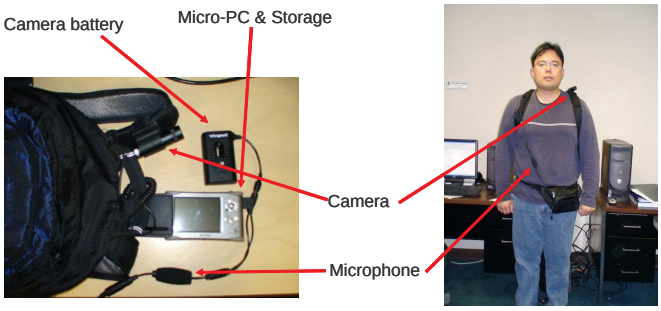

Fig. 1. Main components of the proposed wearable life-log system: a small camera mounted on the shoulder, a microphone, and the processing, storage and browsing unit in the small backpack.

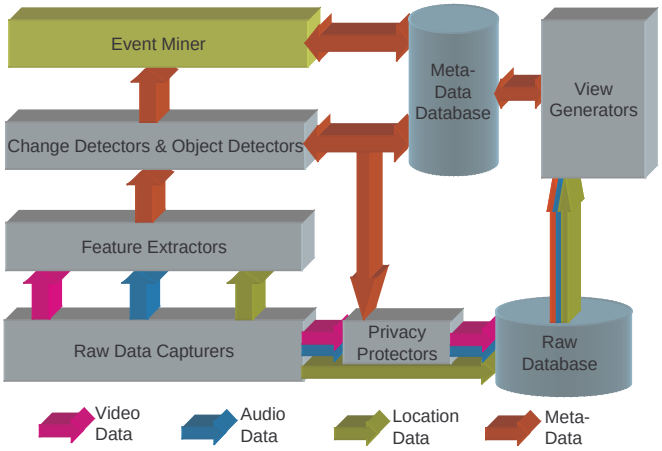

Fig. 2. *Software architecture of the Life-log System.*

Figure 2 shows the software architecture of our proposed life-log system. It consists of six main processing components and two databases. The processing components are Raw Data Capturers, Feature Extractors, Change Detectors and Object Detectors, Privacy Protectors, Event Miner and View Generators. A processing component is a high-level behavioral description of a particular kind of input/output processing. The Raw Data Capturers are responsible to interact with the hardware and obtain the raw audio, video and location information. The Feature Extractors remove noises from the audio and video and extract various types of attributes that are amenable to analysis. These features are fed to various Change Detectors and Object Detectors. A change detector builds an online statistical model of various features and reports the time instance when there is a significant change in the input data. It is useful for partitioning the video in temporal dimension into logical segments which can be much more efficiently manipulated than individual video frames. Object detectors detect and track specific visual and audio objects such as a face or the voice of an individual. These information are then consumed by the Privacy Protectors. When the privacy protection mode is on, the audio of the captured subject will be distorted and all the faces detected in the video will be blocked. Finally, the Event Miner is responsible to extract a semantic structure over a prolong period of recordings, and the View Generator is used to generate user-interface views in support of various types of browsing. It is beyond the scope of this paper to cover every aspect of the system. In the sequel, we will focus on the privacy protector and describe its algorithmic details and performance.

III. Privacy Protection Methodology

In this section, we discuss the privacy protection scheme for our life-log system. Our current design focuses primarily on interview videos in which the producer, i.e. the person wearing the life-log system, is speaking with a single subject in a relatively quiet room. Such a scenario can be found in environments where doctors interviewing patients or police officers interviewing a crime witness [8]. It is of paramount importance to protect the privacy of the subjects in these types of interviews.

There are three main objectives in implementing our privacy protection mechanism:

1) Accuracy versus usefulness: it is important to achieve closed to perfect accuracy in the privacy protection scheme because even a small inaccuracy might disclose the person's identity. On the other hand, the protection scheme should provide enough information to make the data useful for review.
2) Anonymity: Another important criterion is called k-anonymity where k is a small integer denoting the number of distinctive individuals that can be identified after privacy protection [9]. For example, a 1-anonymous privacy protection scheme distorts the data in such a way that every individual will look and sound identical. Even if an attacker knows the identity of the subject in one video, he/she will not be able to gain any knowledge about other videos. In practice, it is difficult to achieve 1-anonymity and thus the goal is to make k as small as possible.
3) Speed: It is highly desirable that the protection mechanism can work in real time. The small physical size of a life-log system limits its computational power and thus it is imperative to design highly optimized algorithms.

With these objectives in mind, we present the overall privacy protection design of the life-log system. Under the privacy protection mode of our system, the subject's face is continuously detected, tracked and blocked with a solid-color box in real time. We choose not to block out the entire body such as that in [7] as it will essentially remove all visual information in an interview sequence. Face de-identification scheme described in [6] that re-paints the face with a generic face is too complicated to be implemented in real-time. To protect the audio identity of the subject, the system identifies the subject's voice by a segmentation algorithm and then distorts it using the PitchScaleSOLA algorithm described in [10]. The distortion is performed in such a way to conceal the identity of the speaker, to maintain the intelligibility of the speech, and to make different distorted voices sounded as

much alike as possible to ensure anonymity. The details of these algorithms are described in the next two sections.

A. Voice Segmentation and Distortion

We propose a simple segmentation algorithm to detect the subject's voice by using the audio signal power. The microphone in our life-log system is closer to the producer than to the subject. As a result, the audio signal power will generally be higher when the producer is speaking. Therefore we can separate the audio signal into the voices of the producer and the subject based on thresholding the signal power. Let s_i be the audio sample at time i. We compute the power by first partitioning the audio signal into equal-duration frames of size T and compute its power P_k as follows:

$$P_k = \frac{1}{T} \sum_{i=kT+1}^{kT+T} {s_i}^2 \quad (1)$$

where k is the frame index.

The classification is based on two thresholds: a silence-threshold T_S to identify the ambient noise and an producer-threshold T_P to identify the producer's voices. The segmentation scheme is illustrated in Figure 3. If the power P_k is smaller than T_S, this frame indicates a pause so the output classification will not change. If the power P_k is between T_S and T_P, this frame is likely the voice from the subject which needs to be distorted. If it exceeds T_P, it is the producer's voice and the audio will not be distorted.

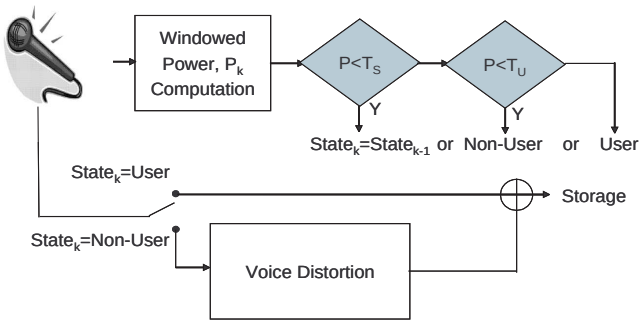

Fig. 3. Audio Segmentation

The next step is to conceal the identity of the subject by distorting the subject's speech while preserving the intelligibility of the conversation. This dual purpose is achieved by shifting the pitch of the speech signal. We use the time-domain pitch shifting method called PitchScaleSOLA as discussed in [10]. Compared with other pitch shifting algorithms based on frequency domain analysis or delay line modulation, this time-domain method is computationally less complex and thus more amenable to real-time implementations. This algorithm works by first time-stretching the audio signal followed by a re-sampling process to maintain the same length. The time-stretching algorithm expands the input signal from length N_1 to N_2. To preserve the speech structure, the input signal is divided into overlapping blocks of size N with hop size S_a, then the pitch of the overlapping blocks are shifted according to scaling factor $\alpha = N_1/N_2$. α lies between 0.25 and 2 with value 1 signifies no pitch shifting. Discrete-time lag of maximum similarity is calculated in the overlapping region. At the point of maximum similarity, the overlapping blocks are weighted by a window function and then summed together. The re-sampling process is performed with an inverse sampling ratio of N_1/N_2 so as to undo the changes in the number of samples. All the parameters, including the scaling factor α, block length N, and hop size S_a, affect the quality of the distortion and the intelligibility of the distorted speech. In section IV, we discuss experiments to study the effect of using different parameters in the audio distortion algorithm.

B. Face Detection and Blocking

Our face detection and blocking module is based on the efficient implementation of the Adaboost face classifier by Viola and Jones [11] in the OpenCV software package [12]. This implementation is very efficient – on our micro PC, it is capable of identifying most of the upright frontal faces under good lighting condition at the rate of 15 frames per second for a frame size of 352×288. Applying this classifier in a frame-by-frame basis, however, is not accurate enough for privacy protection. Whenever a person is turning his/her head or making any hand gesture that partially occludes the face, the classifier fails to detect the face. Furthermore, the performance of the classifier is adversely affected by the movement of the camera, which is inevitable as the camera is mounted on the shoulder of the producer. Such momentary relapse is usually sufficient for a viewer to identify the subject. To further improve the performance, we have added a temporal tracking component using the classifier's outputs as observations.

The tracking component is based on tracking the skin color measured by the dominant hue color in the face region identified by the classifier. If the classifier fails to provide any observation, we search for all pixels that match the skin color in an area slightly larger than the last-observed face region. The new face region is defined as the bounding box containing these pixels. If no such bounding box is found, the face is declared to be disappeared. Occasionally, there are background objects that resemble the skin color and the proximity of these objects with a face may introduce false tracks. To limit the lifetime of these false tracks, we mandate that all face tracks must be validated by a true face observation from the classifier within a certain time limit, empirically set to three seconds in our system. If there are multiple face observations in a scene, we match these observations to existing tracks based on minimizing the sum of their distances on the image plane.

The final step is to obfuscate the identified face region. We choose to color the entire region using a single color as it reveals no information about the underlying face. In order to provide a visual cue on whether the subject is speaking, we utilize the output from the audio segmentation algorithm described in Section III-A to change the blocking color from black to red when the subject starts talking. This simple step provides better visual feedback and indicates that the accom-

panied voice is being distorted. We currently do not perform selective blocking if multiple faces are present. However, this could be done quite easily by having the producer to select the particular subject of interest.

IV. Experimental Results

To evaluate the performance of our system, we have performed various experiments based on two different sets of data. First, we use our life-log system to capture three interviews with three different subjects in a quiet meeting room. The contents of all the interviews are identical and each lasts for roughly 1 minute and 30 seconds. In Section IV-A, we use this data to qualitatively demonstrate the operations of the entire system[1] and to quantitatively measure the segmentation algorithm. Second, we collect voice samples of 11 people with two audio clips of each person's voice, one of them is used to train the speaker identification software, and the other one is used for testing purpose. These data is then used to test the audio distortion algorithm and the test results can be found in Section IV-B.

A. Privacy protected video interviews

Firstly, we evaluate the performance of the segmentation algorithm discussed in section III-A. We use precision and recall measures to statistically analyze the segmentation output. The results are shown in Table I. The precision and recall metrics are computed by counting the number of transitions between producer-voice segments and subject-voice segments in the output and comparing them with those found in the ground truth. Their definitions are shown in Equations (2). The ground truth is manually measured from the videos.

$$\begin{aligned} \text{Recall} &= \frac{\text{\# correctly-identified transitions}}{\text{\# transitions in ground-truth}} \\ \text{Precision} &= \frac{\text{\# correctly-identified transitions}}{\text{\# identified transitions}} \end{aligned} \tag{2}$$

As shown in Table I, our segmentation algorithm produces

TABLE I
Segmentation Results

Meeting#	Precision	Recall
1	0.375	0.8571
2	0.583	1
3	0.353	1

good recall performance but rather poor precision values. Our algorithm generates extra transitions because it sometimes confuses pauses in a person's voice as transitions. This problem can potentially be alleviated by adaptively adjusting the size of the window in measuring the signal power.

The face detection and blocking module works well. Selected frames are shown in Figure 4. In all three sequences, the faces are blocked at all time. There are occasional false alarms that linger for a short period of time. This does not have any adverse effect in terms of privacy protection.

[1]These video clips are available for download in our group website `http://www.vis.uky.edu/mialab`.

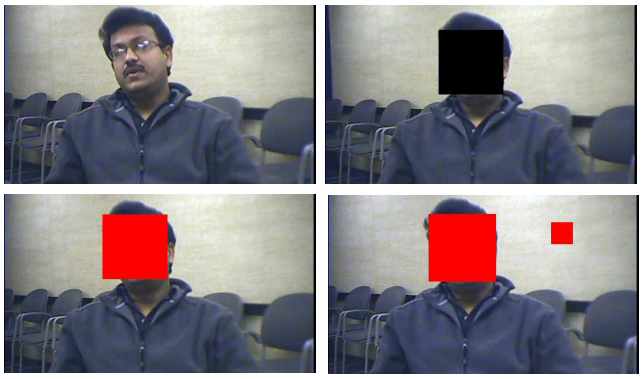

Fig. 4. Top-left: a frame from the original sequence; Top-right: face blocked when the subject is not speaking; Bottom-left: face blocked when the subject is speaking; Bottom-right: false alarms on the background wall.

B. PitchScaleSOLA voice distortion algorithm

We first analyze the performance of the audio distortion algorithm based on how well a speaker can be identified after the distortion. We use a public domain text-independent speaker recognition software [13]. We collect two voice samples from 11 test subjects. The first sample is used to train the speaker identification software, and the other one is used for testing. The results are shown in Table II. We run the speaker identification program on four sets of data: original test data without distortion and three different sets of distorted data with parameters as indicated in the table. These three sets of parameters are chosen so that they result in vastly different distorted sound. To measure their performances, we compute the error rate in identifying the correct speaker and the number of distinct speakers k found by the speaker identification software.

We find that setting α around 1.5 produces good error rate (column 4 and 6) and maintains reasonable intelligibility of conversation. The 100% error rate shows that the voice-distortion algorithm works well in hiding the identity of the speaker. On the other hand, the first two schemes give anonymity of 6 (out of 11) while the last scheme gives only 2, while $k = 1$ is the ideal case. Thus, we conclude that the parameters $N = 1024$, $S_a = 128$ and $\alpha = 1.5$ produce the best overall privacy protection results in our audio distortion algorithm.

Second, we evaluate the distortion algorithm on how intelligible the conversation is after distortion. We have attempted to produce the transcripts for both the original and distorted audio by using a speech recognition software. Nevertheless, we are unable to obtain any reasonably correct transcription on the distorted speech due to its unnatural audio characteristics. As a result, we have to resort to manual transcription. Using two different speech samples from eight test subjects, we use the best distortion algorithm to distort one sample while keeping the other unmodified. Five human testers are asked to transcribe the distorted and non-distorted audio files of the eight subjects. Word-error rate (WER) is calculated for each

TABLE II

SPEAKER RECOGNITION RESULTS

Testing (personId)	Ground Truth (PersonId)	Without Distortion	Distortion 1 $N = 2048, S_a = 256$ $\alpha = 1.5$	Distortion 2 $N = 2048, S_a = 300$ $\alpha = 1.1$	Distortion 3 $N = 1024, S_a = 128$ $\alpha = 1.5$
1	1	1	5	8	5
2	2	2	6	8	6
3	3	3	5	3	5
4	4	4	6	6	5
5	5	5	3	10	6
6	6	6	8	6	5
7	7	7	5	2	5
8	8	8	10	11	5
9	9	9	5	8	5
10	10	10	5	2	5
11	11	11	4	8	5
Error Rate		0%	100%	90.9%	100%
Measured k		11	6	6	2

transcription by each tester to measure the intelligibility of words before and after distortion. The average WER for each subject is shown in Figure 5. WER of distorted speech range between 3% to 43%, which shows that while the distortion has affected the clarity of speech, it maintains certain level of intelligibility. For certain subjects (4,5,7,8), the difference in WER between undistorted and distorted voices is small. On the other hand, the difference is large for other subjects (1,2,3). One possible reason for this large difference is that the first three subjects have strong accents as none of them are native speakers of English. The effect of accents on audio distortion is a subject that deserves further investigation.

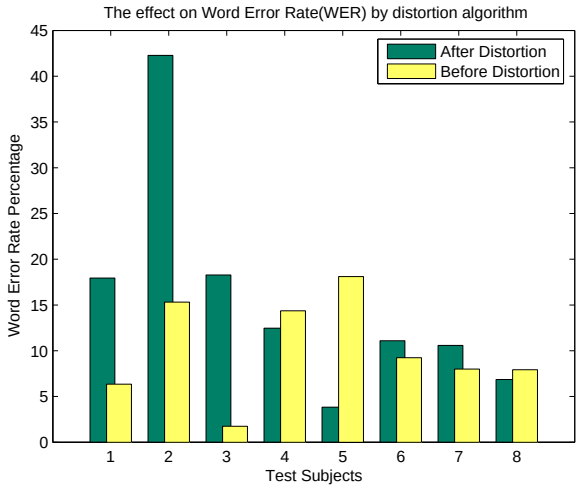

Fig. 5. The effect of Distortion on Word Error Rate

V. CONCLUSIONS

In this paper, we present a real-time implementation of voice-distortion and face blocking as a solution for privacy protection in life-log systems. Our preliminary results show that the voice-distortion algorithm works well in hiding speaker's identity and achieves a small k-anonymity of two, while maintaining the distorted speech reasonably intelligible. For interview-style video, the face detection and blocking module works well by blocking the face during the entire duration of the video. We are now conducting a larger-scale study with more test subjects and diverse testing environments.

ACKNOWLEDGMENT

The authors would like to thank the support of Department of Justice under the grant numbered 2004-IJ-CX-K055 and the constructive comments from the anonymous reviewer.

REFERENCES

[1] V. Bush, "As we may think," *The Atlantic Monthly*, vol. 176, pp. 101–108, 1945.

[2] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong, "Mylifebits: Fulfilling the memex vision," in *Proceedings of ACM Multimedia*, 2002, pp. 235–238.

[3] D. K. William Cheng, Leana Golubchik, "Total recall: Are privacy changes inevitable?" in *Proceedings of the 1st ACM workshop on Continuous Archival and Retrieval of Personal Experiences (CAPRPE '04)*. New York, New York, USA: ACM Press, 2004, pp. 86–92.

[4] J. Wickramasuriya, M. Datt, S. Mehrotra, and N. Venkatasubramanian, "Privacy protecting data collection in media spaces," in *ACM International Conference on Multimedia*, New York, NY, Oct. 2004.

[5] W. Zhang, S.-C. Cheung, and M. Chen, "Hiding privacy information in video surveillance system," in *Proceedings of the 12th IEEE International Conference on Image Processing*, Genova, Italy, Sept. 2005.

[6] 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.

[7] S.-C. Cheung, J. Zhao, and M. V. Venkatesh, "fficient object-based video inpainting," in *Proceedings of the 13th IEEE International Conference on Image Processing*, Atlanta, GA, Sept. 2006.

[8] "Cylon systems," http://www.cylonsystems.com/.

[9] L. Sweeney, "k-anonymity: a model for protecting privacy," in *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002, pp. 557–570.

[10] U. Z. et al., *DAFX - Digital Audio Effects*. John Wiley and Sons, LTD, 2002.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[12] "Open source computer vision library (opencv)," http://www.intel.com/technology/computing/opencv/.

[13] L. Rosa, "Text-independent speaker recognition based on neural networks," http://www.advancedsourcecode.com/neuralnetspeaker.asp.