Text Driven Face-Video Synthesis Using GMM and Spatial Correlation

Dereje Teferi, Maycel I. Faraj, Josef Bigun

School of Information Science, Computer, and Electrical Engineering (IDE)
Halmstad University
P.O.Box 823, SE-301 18
Halmstad, Sweden
{Dereje.Teferi,Maycel.Faraj,Josef.Bigun}@ide.hh.se

Abstract. Liveness detection is increasingly planned to be incorporated into biometric systems to reduce the risk of spoofing and impersonation. Some of the techniques used include detection of motion of the head while posing/speaking, iris size in varying illumination, fingerprint sweat, text-prompted speech, speech-to-lip motion synchronization etc. In this paper, we propose to build a biometric signal to test attack resilience of biometric systems by creating a text-driven video synthesis of faces. We synthesize new realistic looking video sequences from real image sequences representing utterance of digits. We determine the image sequences for each digit by using a GMM based speech recognizer. Then, depending on system prompt (sequence of digits) our method regenerates a video signal to test attack resilience of a biometric system that asks for random digit utterances to prevent play-back of pre-recorded data representing both audio and images. The discontinuities in the new image sequence, created at the connection of each digit, are removed by using a frame prediction algorithm that makes use of the well known block matching algorithm. Other uses of our results include web-based video communication for electronic commerce and frame interpolation for low frame rate video.

1 Introduction

People have to be uniquely identified to get access to an increasing number of services; their office, their bank account, their computer, their mail, even to enter a country etc. To this effect, person identification is done in many ways, one of which is biometrics.

Biometrics is the study of automated methods for uniquely recognizing humans based upon one or more intrinsic physiological or behavioral traits. Biometric systems are in use in many applications such as security, finance, banking etc [1], [2], [3]. In spite of their high level of accuracy, biometric systems have not been used massively in the aforementioned areas. One of the main drawbacks is that biometric data of a person (such as face, speech, etc) are not secret and cannot be replaced anytime the user wants to or whenever they are compromised by a third party for spoofing. This problem is minimal if the authentication system works with the help of a human supervisor as in border control where the presented trait can be visually checked to see if it is genuine or fake. However, this risk is high for remotely controlled biometric applications such as those that use the internet [4]. The risk of spoofing on biometric systems can be reduced by combining multiple traits into the system and incorporating liveness detection.

Recent technological advances such as those in audio-video capture and processing have enabled researchers to develop sophisticated biometric systems. As such it has also given spoofers the chance to become more vicious system attackers. Some spoofers can impersonate clients even in multimodal biometric applications. Impersonation is done, for example, by audio-video playback and application of image processing techniques without the presence of the real client.

The actual presence of the client can be assured to a certain degree by liveness detection systems. Liveness detection is an anti-spoofing mechanism to protect biometric systems [5], [6], [7]. It is performed by, for example, face part detection and optical flow of lines to determine liveness score [8], analysis of fourier spectra of the face image [9], lip-motion to speech synchronization[10], body temperature, on the spot random queries such as pronouncing of random sequences of digits, iris size scanning under varying illumination etc.

Random text-prompted liveness detection or audio-video based recognition systems use random digits as their text prompts. This is because random numbers are easier to read for the client than random texts and also easier to synthesize. Accordingly, the digit speaking image sequences of the XM2VTS database is used in this work for the experiment[11]. In our approach we develop a text prompted face-video synthesis system for testing the performance of liveness detection as well as audio-visual recognition systems.

We use GMM based speech recognizer to identify the locations of each digit spoken by the subject. The pre-recorded image sequence is then reshuffled according to system prompt (or digits entered). However, the process of shuffling creates discontinuities between digits in the new image sequence. This discontinuity between the last frame of a digit and the first frame of the next digit is compensated by estimating the frames in between. The GMM based speech recognizer is used to identify the locations of the digits in a video of new training data.

A number of motion estimation techniques are discussed in [12],[13],[14], and [15]. The method applied here for motion estimation is the well known block matching which uses the Schwartz inequality.

2 Speech Recognition

Here we present a speech recognition system using the well known Gaussian Mixture Model(GMM).

Figure 1, illustrates text-dependent speech recognition of a digit for a specific person. The speech analysis is represented by Mel-Frequency Cepstral feature

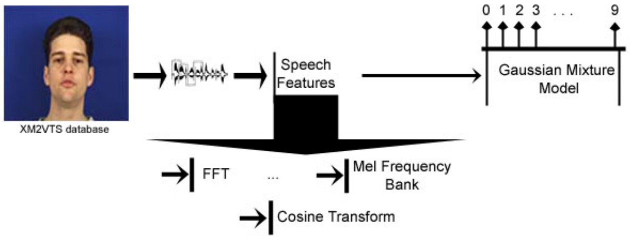

Fig. 1. Speech Recognition System

analysis where the extracted feature set is put into Gaussian Mixture Model system.

2.1 Speech features

The vocal tract structure of a person influences the speech spectrum significantly which in turn can be used to distinguish a user from other users. Spectral representations of a person's speech can be extracted with several methods, and in this paper we implemented the commonly used Mel-Frequency Cepstral Coefficients(MFCC). Our acoustic features were constructed by MFCC according to [16]. The sampled wave files in the XM2VTS database were processed by using the HTK(Hidden Markov Model Toolkit)[17],[18].

The sampled waveform is converted into a sequence of acoustic parameter blocks, where the output sampling period is set to 10 ms and the window size is set to 25 ms, [10]. From each block we extract a 39 dimensional feature vector, consisting of 12 cepstral coefficients with normalized log-energy, 13 delta coefficients and delta-delta coefficients. The delta and delta-delta coefficients correspond to the first and second order time derivatives of extracted MFCC, also known as velocity and acceleration respectively.

2.2 Gaussian Mixture Model

A Gaussian Mixture Model(GMM) can be represented as a weighted sum of multivariate Gaussian distributions [16]. Here we use the GMM to model the person specific features [16]. In this paper, we developed the GMM using the HTK toolbox[17]. Each person is described with Gaussian model parameters that are learned from a training set, that are the mean and variances of a number of Gaussian distributions, as well as their associated weights which are used to linearly combine the individual Normal components to finally yield a multi-dimensional distribution for the features. More details about the set-up can be found in [10].

3 Motion Estimation

We use motion estimation technique to predict frames to reduce the discontinuity between frames. The discontinuity occurs due to the rearrangement of sequence of frames according to the prompt of the biometric system.

Motion estimation is a common technique used in video compression. A video is a sequence of frames and there is always some redundant data between adjacent frames within a certain period of time ($t_1 - t_0$). The redundancy is small for fast paced and complex motion of objects and background in the video and high otherwise. This redundancy is exploited to compress the video. That is a reference frame (sometimes known as the independent frame) is taken from a sequence every n frame apart. Then the middle frames are predicted from these frames as well as from previously predicted frames by the codec. The codec actually uses the Motion vector and prediction error to reconstruct the *dependent* frames. Forward prediction, where the new frames are predicted from previous frames, or backward prediction, where the frames are predicted from future frames, or both can be used for estimation of the new frames in the middle. Many codecs use this technique for video compression.

The natural motion of the head while speaking is minimal. Moreover, it is not too difficult to acquire video of an arbitrary person uttering the 10 digits. Given such a sequence an attacker could proceed as discussed below.

Assuming the video is captured with a stationary camera, the background will be near-constant. Therefore, little information is lost or added between adjacent frames, such as teeth, mouth and eyes. That is, there is a high probability that part of a frame exist in another frame although translated to a different location. First the points in motion are extracted using absolute difference between the two frames. These two frames are extracted from the image sequences of the last frame of a digit and the first frame of the succeeding digit.

$$AD = |\mathbf{F}(k,l) - \tilde{\mathbf{F}}(k,l)| \tag{1}$$

Now that we know the points/blocks in motion, the motion vector (MV) is calculated only for these points. For each point or block in motion on frame $\mathbf{F}$, we look for its parallel pattern in frame $\tilde{\mathbf{F}}$ within a local neighborhood by using block matching algorithm.

3.1 Block Matching Algorithm

Block matching is a standard video compression techniques to encode motion in video sequences [14]. A review of block matching algorithms is given in [15], [14], [13], [19].

In our approach, equal sized non-overlapping blocks are created over the frame $\mathbf{F}$ (*the frame at time* t_0). Then, for those blocks on $\mathbf{F}$ containing points in motion, a search area is defined on frame $\tilde{\mathbf{F}}$ (*the frame at time* t_1). The search area is larger than the size of the block by *expected* displacement of the object in

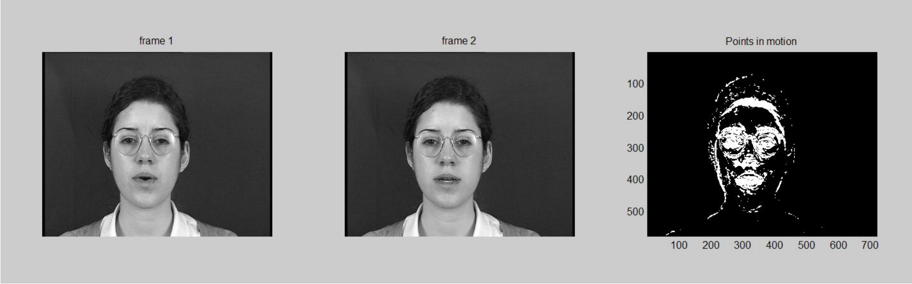

Fig. 2. Points in Motion between frame **F** and $\tilde{\mathbf{F}}$

motion e. Then we apply Schwartz inequality to find the most parallel pattern for the block in frame **F** from the search area in frame $\tilde{\mathbf{F}}$.

Let **f** and $\tilde{\mathbf{f}}$ be vector representations of patterns from frame **F** and $\tilde{\mathbf{F}}$ and $<, >$ be the scalar product defined over the vector space. We then have

$$| < \mathbf{f}, \tilde{\mathbf{f}} > | \leq \|\mathbf{f}\| \|\tilde{\mathbf{f}}\|$$

$$cos(\theta) = \frac{| < \mathbf{f}, \tilde{\mathbf{f}} > |}{\|\mathbf{f}\| \|\tilde{\mathbf{f}}\|} = \frac{|\mathbf{f}^T \tilde{\mathbf{f}}|}{\|\mathbf{f}\| \|\tilde{\mathbf{f}}\|} \leq 1 \qquad (2)$$

where $\mathbf{f} = (f_1, f_2, ..., f_{k-1}, f_k, f_{k+1}, ...)$ and $\tilde{\mathbf{f}} = (\tilde{f}_1, \tilde{f}_2, ..., \tilde{f}_{k-1}, \tilde{f}_k, \tilde{f}_{k+1}, ...)$ are vector forms of the 2D pattern **f** and $\tilde{\mathbf{f}}$ from frames **F** and $\tilde{\mathbf{F}}$ respectively and $cos(\theta) \in [0, 1]$ is the similarity measure between the patterns.

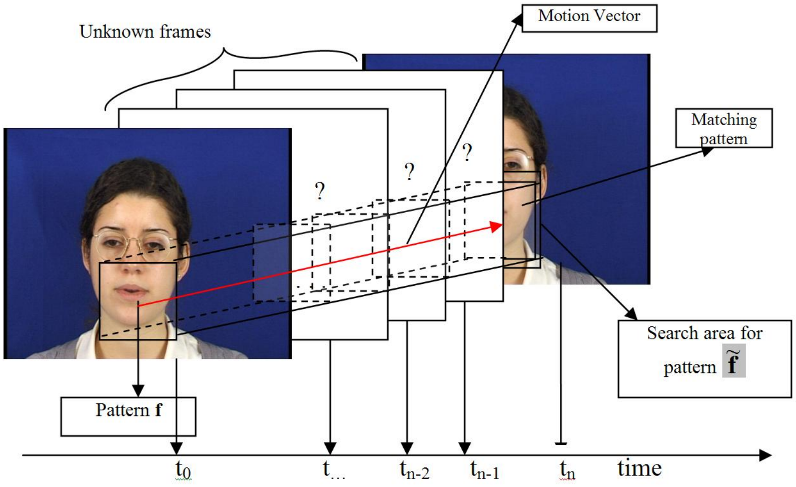

Fig. 3. Motion Vector and unknown frames in a sequence

The most parallel pattern $\tilde{\mathbf{f}}$ in the frame $\tilde{\mathbf{F}}$ is found by maximizing $cos(\theta)$. This can be done by repetitive scalar products. That is the pattern $\mathbf{f}$ from frame $\mathbf{F}$ is glided over an *expected* parallel local neighborhood of $\tilde{\mathbf{f}}$ in the frame $\tilde{\mathbf{F}}$ and the most similar pattern is selected as a match.

The motion vector for the point at the center of pattern $\mathbf{f}$ is calculated as the displacement between pattern $\mathbf{f}$ and pattern $\tilde{\mathbf{f}}$ (Fig. 3). That is:

$$MV(k,l) = x + iy \tag{3}$$

Where x and y are the horizontal and vertical displacements respectively of the block/pattern $\mathbf{f}$, *(k,l)* is the index for the center of pattern $\mathbf{f}$ and $i = \sqrt{-1}$.

3.2 Frame Prediction

Finally, the motion vector is used to predict the unknown frames between $\mathbf{F}$ and $\tilde{\mathbf{F}}$ (Fig. 3) if the image sequence is to be perceived realistic with minimum amount of discontinuity. The number of frames to be predicted depends on the norm of the motion vector and is determined at run-time. For a length 2 time units, the actual frame prediction is done by dividing the motion vector at point (*k*,*l*) by 2 and moving the block in frame $\tilde{\mathbf{F}}$ centered at (*k+x,l+y*) to the middle frame at (*k+x/2,l+y/2*). Then the block at the new location in frame $\tilde{\mathbf{F}}$ is moved back the same distance. That is, let $\mathbf{F}$ be the frame at t_0, $\tilde{\mathbf{F}}$ the frame at t_1 and $\mathbf{F}^{'}$ be the frame at $\frac{(t_1+t_0)}{2}$, then

$$\mathbf{F}^{'}(k + x/2, l + y/2) = \tilde{\mathbf{F}}(k + x, l + y) \tag{4}$$

$$\mathbf{F}^{'}(k, l) = \tilde{\mathbf{F}}(k + x/2, l + y/2) \tag{5}$$

where x and y are the real and imaginary parts of the motion vector MV at (*k*,*l*).

To avoid overlaps in the interpolation process, those blocks that are already been interpolated are flagged. Consecutive predictions are made in analogous manner. The motion vector is adjusted and a new frame is created as necessary between frames $\mathbf{F}$ and $\mathbf{F}^{'}$ as well as between $\mathbf{F}^{'}$ and $\tilde{\mathbf{F}}$. This process continues until all the necessary frames are created.

4 Video Synthesis

Audio signal is extracted from the input audio-video signal and forwarded to the speech recognizer. The recognizer uses the GMM models to return transcription file containing the start and end time of each digit spoken in the audio signal. A search for the prompted text is done against the transcription file and the time gap of each prompted digits within the video signal is captured (Fig. 4).

The discontinuity between the image sequences of each digit is compensated by predicting the unknown frames (Fig. 3) using the motion estimation technique

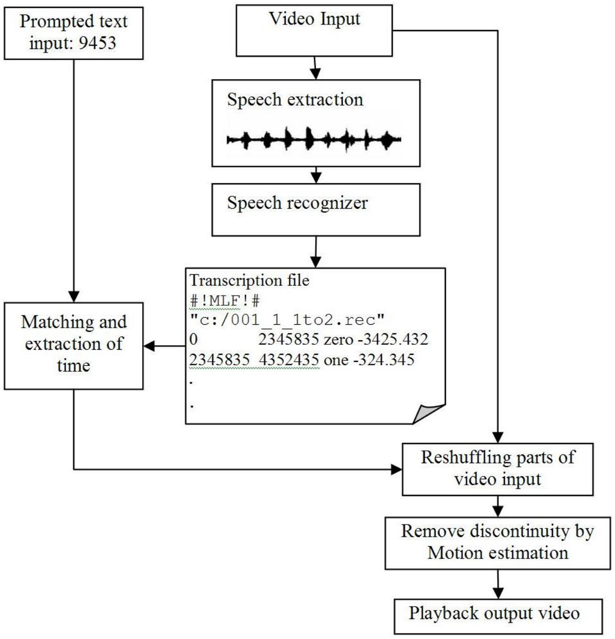
Prompted text input: 9453
Video Input
Speech extraction
Speech recognizer
Transcription file
#!MLF!#
"c:/001_1_1to2.rec"
0 2345835 zero -3425.432
2345835 4352435 one -324.345
.
.
Matching and extraction of time
Reshuffling parts of video input
Remove discontinuity by Motion estimation
Playback output video

Fig. 4. Process flowchart

summarized in section 3. The new predicted frames are attached to a silence sound and are inserted to their proper locations in the video signal to decrease the discontinuity of utterances. Finally, the video is played to represent the digit sequence prompted by the biometric system.

5 Experiment

Th experiments are conducted on all the digit speaking face videos of the XM2VTS database (295 persons). The accuracy of the reshuffled video signal is mainly dependent on the accuracy of the speech recognition system. The accuracy of our GMM based speech recognition system is 96%. The frame prediction algorithm works well when the block size is set to 3x3 pixels. When the block size is larger some visible deformations appear on the predicted frame mainly due to rotational effects. Therefore, we used 3x3 blocks to predict the unknown frames. The discontinuity of the reshuffled video signal is reduced significantly as evaluated by the human eye, the authors.

Biometric authentication and liveness detection systems that make use of motion information of zoomed in face, head, lip and text prompted audio-video are easy targets of such system attacks.

6 Conclusion

The risk of spoofing is forcing biometric systems to incorporate liveness detection. Assuring liveness especially on remotely controlled systems is a challenging task. The proposed method shows a way to produce play-back attacks against text-prompted systems using audio and video. The result shows that assuring liveness remotely by methods that rely on apparent motion can be bypassed. Our results suggest the need to increase the sophistication level of biometric systems to stand up against advanced play-back attacks.

Acknowledgment

This work has been sponsored by the Swedish International Development Agency (SIDA).

References

1. Jain, A., Ross, A., Prebhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics **14(1)** (January 2004)
2. Jain, A., Pankanti, S., Prabhakar, S., Hong, L., Ross, A.: Biometrics: A grand challenge. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004. Volume 2. (2004) 935–942
3. Ortega-Garcia, J., Bigun, J., Reynolds, D., Gonzalez-Rodriguez, J.: Authentication gets personal with biometrics. IEEE Signal Processing Magazine **21(2)** (March 2004) 50–62
4. Faundez-Zanuy, M.: Biometric security technology. IEEE Aerospace and Electronic Systems Magazine **21(6)** (2006) 15–26
5. Bigun, J., Fronthaler, H., Kollreider, K.: Assuring liveness in biometric identity authentication by real-time face tracking. In: IEEE international Conference on Computational Intelligence for Homeland Security and Personal Safety. Venice, Italy. (July 2004)
6. Stephanie, A., Schukers, C.: Spoofing and anti-spoofing measures. Information Security Technical Report (2002)
7. Ratha, N.K., Connell, J.H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal **40(2)** (2001) 614–634
8. Kollreider, K., Fronthaller, H., J., B.: Evaluating liveness by face images and the structure tensor. In: AutoID 2005: Fourth Workshop on Automatic Identification Advanced Technologies.IEEE Computer Society. (October 2005) 75–80
9. Li, J., Wang, Y., Tan, T., Jain, A.K.: Live face detection based on the analysis of fourier spectra. In Jain, A.K., Ratha, N.K., eds.: Biometric Technology for Human Identification. Proceedings of the SPIE, Volume 5404. (August 2004) 296–303
10. Faraj, M., Bigun, J.: Person verification by lip-motion. In: Computer Vision and Pattern Recognition Workshop (CVPRW). (June 2006) 37–45
11. Messer, K., Matas, J., Kitler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended M2VTS database. In: 2nd International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99). (1999) 72–77

12. Bigun, J.: Vision with Direction:A Systematic Introduction to Image Processing and Computer Vision. Springer-Verlag Berlin Heidlberg (2006)
13. Jain, J., Jain, A.K.: Displacement measurement and its application in interframe image coding. IEEE Transactions on Communication **COM 29** (December 1981) 1799–1808
14. Gyaourova, A., Kamath, C. Cheung, S.C.: Block matching for object tracking. Technical report UCRL-TR-200271. Laurence Livermore Technical Laboratory (Occtober 2003)
15. Cheng, K.W., Chan, S.C.: Fast block matching algorithms for motion estimation. In: ICASSP-96: IEEE International Conference on Acoustic Speech and Signal Processing. Volume 4(1). (May 1996) 2311–2314
16. Reynolds, D., Rose, R.: Robust text independent speaker identification using gaussian mixture models. IEEE Transactions on Speech and Audio Processing **3(1)** (January 1995) 72–83
17. Young, S., Evermann, G., Gales, M., Hein, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The htk book. for version 3.3. http://htk.eng.cam.ac.uk/docs/docs.shtml (April 2005)
18. Veeravalli, A.G., Pan, W., Adhami, R., Cox, P.G.: A tutorial on using hidden markov models for phoneme recognition. In: Thirty-Seventh Southeastern Symposiumon System Theory, SSST 2005. (2005)
19. Aly, S., Youssef, A.: Real-time motion based frame estimation in video lossy transmission. In: Symposium on Applications and the Internet. (January 2001) 139–146