MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Recognizing Talking Faces From Acoustic Doppler Reflections

Kaustubh Kalgaonkar, Bhiksha Raj

TR2008-080 December 2008

Abstract

Face recognition algorithms typically deal with the classification of static images of faces that are obtained using a camera. In this paper we propose a new sensing mechanism based on the Doppler effect to capture the patterns of motion of talking faces. We incident an ultrasonic tone on subjects' faces and capture the reflected signal. When the subject talks, different parts of their face move with different velocities in a characteristic manner. Each of these velocities imparts a different Doppler shift to the reflected ultrasonic signal. Thus, the set of frequencies in the reflected ultrasonic signal is characteristic of the subject. We show that even using a simple feature computation scheme to characterize the spectrum of the reflected signal, and a simple GMM based Bayesian classifier, we are able to recognize talkers with an accuracy of over 90%. Interestingly, we are also able to identify the gender of the talker with an accuracy of over 90%.

IEEE International Conference on Face and Gesture Recognition 2008

Recognizing Talking Faces From Acoustic Doppler Reflections

Kaustubh Kalgaonkar
School of Electrical and Computer Engineering,
Georgia Institute of Technology.
Atlanta GA USA 30332
kaustubh@ece.gatech.edu

Bhiksha Raj
Mitsubishi Electric Research Labs.
Cambridge, MA USA 02139
bhiksha@merl.com

Abstract

Face recognition algorithms typically deal with the classification of static images of faces that are obtained using a camera. In this paper we propose a new sensing mechanism based on the Doppler effect to capture the patterns of motion of talking faces*. We incident an ultrasonic tone on subjects' faces and capture the reflected signal. When the subject talks, different parts of their face move with different velocities in a characteristic manner. Each of these velocities imparts a different Doppler shift to the reflected ultrasonic signal. Thus, the set of frequencies in the reflected ultrasonic signal is characteristic of the subject. We show that even using a simple feature computation scheme to characterize the spectrum of the reflected signal, and a simple GMM based Bayesian classifier, we are able to recognize talkers with an accuracy of over 90%. Interestingly, we are also able to identify the* gender *of the talker with an accuracy of over 90%.*

1. Introduction

In this paper we address the topic of automatic recognition of talking faces.

Automatic recognition of faces has usually been treated as a problem of visual processing. Nearly all methods for automatic face recognition begin with *images* taken with a camera. Faces may then be segmented out of the images using a variety of techniques [1], features of various kinds measured from them [2], and classification performed with a variety of classifiers [3, 4, 5]. The focus of research has primarily been on improved segmentation of faces out of the images, improved features and improved classifiers, retaining the assumption about the visual nature of the basic measurements captured by the sensor, *i.e.* the camera.

This paper proposes to use an entirely different sensing paradigm for the recognition of faces: an acoustic Doppler sonar (ADS). We incident ultrasonic sound waves on the subject's faces and capture the reflected signals. The energy patterns and the Doppler frequency shifts in the reflected signal are characteristic of the subject, particularly when they are talking, and are used to identify the subject. Since the Doppler frequency shifts resulting from facial movements related to talking are key, the approach is geared primarily towards recognition of *talking* faces.

Ultrasound measurements have commonly been used for imaging, particularly as a diagnostic tool (although we are not aware of any prior work on the use of ultrasound imaging for facial recognition). They have, however, been used chiefly as *imaging* tools (as noted above) that scan the target to recreate *images* of the target from the energy and spectral patterns of the reflected signal; further processing if any is performed on the inferred images. The final representation derived is thus still visual. In our work however, the sensor is static and performs no scan; we do not attempt to infer an image of the target. Instead, it is our contention that the information relating to the target is encoded in the reflected signal itself and it can hence be processed directly for classification, without resorting to an intermediate visual representation.

ADS sensors have also previously been shown to be useful sources of primary or secondary measurements for voice-activity detection [6], gait [7], and even speaker identification [8] (where Doppler measurements were used to augment the information in a speaker's voice); however this paper, to the best of our knowledge, is the first reported use of Doppler sonars as primary sources of information for recognizing faces.

Our ADS sensor is an inexpensive device that consists of a low-frequency ultrasound emitter and an acoustic transducer that is tuned to the transmitted frequency. An ultrasound tone output by the emitter is reflected from the subject's face and undergoes a Doppler frequency shift that is proportional to normal velocity of the portion of the face that it is reflected by. The reflected "Doppler" signal thus contains an spectrum of frequencies that represent the motion of the subject's facial features such as the cheeks, lips,

978-1-4244-2154-1/08/$25.00

tongue, etc. The pattern of movements of facial muscles while speaking is typical of the subject. By characterizing the velocities of these movements, the Doppler signal thus represents a signature that is quite specific to the person. Although the energy is the reflected signal also contains information about the physiognomy of the speaker's face, energy variations in the reflected signal due to modulation by the subject's physiognomy are indistinguishable from those arising simply from changing the distance of the subject to the sensor. However, the *temporal* variations in the reflected energy remains characteristic of the subject as it represents the typical movements of the subject's head.

It must be pointed out that the type of information captured by the Doppler sensor is fundamentally different from that captured by a camera. The camera primarily captures static images. Movements, such as those of a talking face are captured chiefly as the difference is subsequent snapshots in a series of images such as in a video. The signal captures by the Doppler sensor on the other hand actually represents a characterization of the *dynamics* of the face, and it may in fact not be possible to compute a static image of the face from it.

In our work the signals captured by the sensor are parameterized through a simple feature computation scheme and classification is performed using a very simple Gaussian classifier. Nevertheless, using only these very simple mechanisms we are able to achieve accuracies exceeding 90% in recognizing faces from our collection. As we argue in the concluding section of the paper, we have reason to believe that this performance could be improved further by better characterization of the signal and better modelling of distributions. Furthermore, we believe that as complimentary sources of information, cameras and Doppler sensors may, in fact, be used together to achieve better classification than either of them could get by themselves.

In Section 2 we briefly present some background on facial movement as a cue to a person's identity. In Section 3 we describe the basic hardware setup of the ADS. Our setup, built with off-the shelf components, costs only a few dollars (US); if replicated on a large scale it can be made far cheaper. In Section 4 we briefly discuss the Doppler principle that accounts for the information in the measurements. In Section 5 we describe the signal processing we employ to extract features from the Doppler signal for classification.

In Section 6 we describe the classification mechanism that we employ to recognize faces using the Doppler signal. We use a simple Bayesian mechanism within which we combine the likelihoods of features derived from the Doppler signal for this purpose. We describe experiments in Section 7 which demonstrate the effectiveness of our method. Finally in Section 8 we present our conclusions.

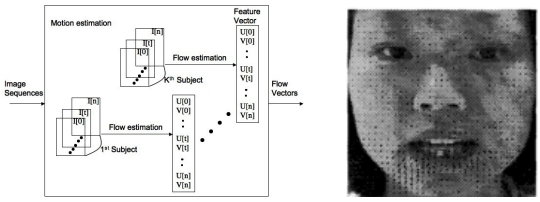


Figure 1. Left panel: Procedure proposed by Chao et. al. [12] to compute facial motion flow from a sequence of images. Right panel: An example of facial flow measurements from Chao et. al. [12]. Flow measurements such as these, that have been obtained from video, have previously been successfully used to classify talking faces.

2. FACIAL MOTION AND IDENTITY

A person's face is the primary cue to their identity – in fact it is believed that humans may have evolved specialized abilities to recognize faces. Moreover, and key to this paper, there is considerable evidence obtained both from studies of human subjects and inference from computer algorithms that the *movements* of a person's face, including facial gestures and the motion of facial structures that occurs while speaking also carry significant information about the identity of the person. A well-known study by Berry [9] demonstrated that both children and adults are able to identify the gender of a speaker through visualization of point-light displays of their faces as they conversed, clearly suggesting that information about the speaker's gender, at least, was present in their patterns of facial motion. Similarly, Knappmeyer, Thornton and Buelthoff [10] argue using another study that combines computer animation with psychophysical methods that facial movement carries information about a variety of characteristics of the subject such as their age, gender, emotion and identity.

Needless to say, talking faces produce speech sounds. It may be argued that the identity of the speaker lies primarily in this *speech signal*, and that the facial movement is only a secondary phenomenon that accompanies it and only presents an alternative characterization of information that is already present in the speech signal itself. Munhall and Buchan [11] provide contradicting evidence through a study where they show that even when the facial movements of the images in a video of talking faces corresponded to a *different* utterance than the one played out in the audio channel, the combination of video and sound results in improved identification of the talker, demonstrating that the facial movement has distinct cues about the identity of the talker that are independent of the accompanying audio.

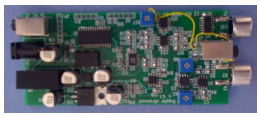
Other evidence about the cues to speaker identity in patterns of facial motion is derived through interpretation of results obtained computationally by various researchers. Some of this evidence is rather direct: Chao, Liao and Lin [12] show that features characterizing facial movement de-

rived from a video are very effective for identifying the talker. Other evidence is indirect: audio-visual speaker recognition algorithms attempt to identify speakers using a combination of the audio signals and the accompanying video [13]. Several of these methods augment static measurements from video with motion features that are computed through difference operations on adjacent frames, as this is observed to result in improved speaker recognition. The motion features in these methods effectively capture patterns of facial motion.

The work reported in this paper is based on the premise drawn from all the above that facial movement carries information about the identity of the talker. However, unlike prior work that characterizes such motion through differences in features derived from video snapshots, we characterize it directly in terms of the patterns of *velocity* of facial structures, as we explain in Section . One of the drawbacks of our approach is that our sensor integrates information from different facial components that all move with the same velocity. This results in a loss of resolution; nevertheless our results show that the approach is promising.

3. THE ACOUSTIC DOPPLER SENSOR

(a) The Doppler sensor used in our experiments. An ultrasonic emitter and a corresponding receiver were taped on either side of a long-barreled microphone. Signals from the receiver were captured by a high-end A/D converter and sampled at 96kHz.



(b) A newer version of the ADC. Captured ultrasonic signals are heterodyned down by 36kHz on the device itself and can be recorded through the microphone jack of a PC at 16k samples per second.

Figure 2. Doppler devices

Figure 2(a) shows our acoustic Doppler sonar setup for recognizing talking faces. It has two main components. The tiny pillbox-shaped object to the left is an ultra-sound emitter that emits a 40 kHz tone. The pillbox to the right is an ultra-sound receiver that is tuned to capture signals in a narrow band of frequencies centered at 40 kHz. The barrel-shaped device in the center is a high-quality microphone that we have also included in our setup to capture the speech uttered by the speaker; however we do not use this signal in any manner for the work reported in this paper and we shall not refer to it hereafter.

The sensor is arranged to point directly at subject's faces. Both the emitter and receiver in our setup have a diameter that is approximately equal to the wavelength of the emitted 40kHz tone, and thus have a beamwidth of about 60^o, making them quite directional. Signals emitted by the 40Khz transmitter are reflected by the subject's face and captured by the receiver. It must be noted that the receiver also captures high-frequency harmonics from the actual speech being uttered and any background noise; however these are significantly attenuated with respect to the level of the reflected Doppler signal in most standard operating conditions and can be safely ignored. The cost of the entire setup shown in the Figure (not including the microphone) is minimal: the high-frequency transmitter and receiver both cost less than $10 when bought singly and much lesser if bought in bulk. The signal captured by the receiver is digitized prior to further processing. Since the high-frequency transducer is highly tuned and has a bandwidth of only about 4Khz, the principle of band-pass sampling may be applied, and the signal need not be sampled at more than 12Khz (although in our experiments we have sampled the signal at 96Khz and down-shifted the frequencies in the signal algorithmically).

4. DOPPLER EFFECT ON SIGNALS REFLECTED BY A TALKING FACE

The Doppler sonar operates on the Doppler's effect, whereby the frequency perceived by a listener who is in motion relative to the signal emitter is different from that emitted by the source. In particular if the source emits a frequency f that is reflected towards a receiver by an object moving with velocity v with respect to the receiver, then a reflected signal sensed at the receiver $\hat{f}$ is given by

$$\hat{f} = \frac{v_s + v}{v_s - v} f \tag{1}$$

were v_s is the velocity of the sound in the medium. When the receiver is collocated with the transmitter, as it is for our ADS, f in the above equation also refers to the velocity of the object with respect to the transmitter. If the signal is reflected by multiple objects moving at different velocities then multiple frequencies will be sensed at the receiver.

The human face is an articulated object with multiple components capable of moving at different velocities. When a person speaks all components of the face including but not limited to the lips, tongue, jaw cheeks etc. move with velocities that depend on facial construction and are typical of the talker. The ultrasonic signal reflected off the face of a subject has multiple frequencies each associated with one of the moving components. This reflected signal can be mathematically modeled as

$$d(t) = \sum_{i=1}^{N} a_i(t) cos(2\pi f_i(t) + \phi_i) + \Psi_{person} \tag{2}$$

where f_i is the frequency of the reflected signal from the i^{th} moving component, which is dependent on its velocity

v_i. f_c is the transmitted ultrasonic frequency. $a_i(t)$ is a time-varying reflection coefficient that is related to the distance of the i^{th} facial component from the sensor. ϕ_i is an component-specific phase correction term. The term within the summation in Equation 2 thus represents the sum of a number of frequency modulated signals, where the modulating signals $f_i(t)$ are the velocity functions of all moving parts of the face . We do not, however, attempt to resolve the individual velocity functions via demodulation. The quantity Ψ_{person} is a person-specific term that accounts for the baseline reflection from the talker's face. It represents a crude zeroth order characterization of the bumps and valleys in the face and is not related to motion. Figure 3 shows a typical Doppler signal captured by the receiver on our Doppler sensor. The overall characteristics of this signal may be assumed to be typical of the talker.

5. SIGNAL PROCESSING

The received "Doppler signal" is initially sampled at 96 kHz. The ultrasonic sensor is highly frequency selective with a 3 dB bandwidth of only about 4 kHz; at 40 kHz $\pm$ 4 kHz the signal is attenuated by more than 12 dB. Moreover, the frequencies in the received signal rarely wander outside of this range (since facial features do not move fast enough). It can hence be safely assumed that the effective bandwidth of the Doppler signal is less than 8kHz. We therefore heterodyne the signal from the Doppler channel down by 36 kHz so that the signal is now centered at 4 kHz and resample it to 16 kHz. While we currently perform the heterodyning and resampling digitally, in a more recent version of our device (shown in Figure 2(b)), the analog Doppler signal is heterodyned down to have a center frequency of 4 kHz onboard, and the signal from it only need be sampled at 16 kHz, with no further resampling required.

The frequency characteristics of the Doppler signal vary slowly, since the articulators that modulate its frequency are relatively slow-moving. To capture the frequency characteristics of the Doppler signal we segment it into relatively long analysis frames of 40 ms. Adjacent frames overlap by 75%, such that 100 such frames are obtained every second. Each frame is Hamming windowed, a 1024-point Fourier transform computed from it, and the power in all the unique spectral terms in the resulting transform computed, to obtain a 513-point power spectral vector. The power spectrum is logarithmically compressed and a Discrete Cosine Transform (DCT) is applied to it. The first 40 DCT coefficents are retained to obtain a 40-dimensional cepstral vector. Each cepstral vector is then augmented by a difference vector as follows:

$$\begin{aligned} \Delta C^d[n] &= C^d[n+2] - C^d[n-2] \\ c^d[n] &= [C^d[n]^T \Delta C^d[n]^T]^T \end{aligned} \quad (3)$$

where $C^d[n]$ represents the cepstral vector of the n^{th} analysis frame, $\Delta C^d[n]$ is the corresponding difference vector and $c^d[n]$ is the augmented 80-dimensional cepstral vector. The dimensions of the feature vector are reduced to 20 using PCA. The 20-dimensional vectors are finally used for classification. We note here that the entire processing is very similar to that used to process audio signals for classification and is computational complexity is very low.

Figure 3 doppler signal acquired by the ultrasonic receiver.

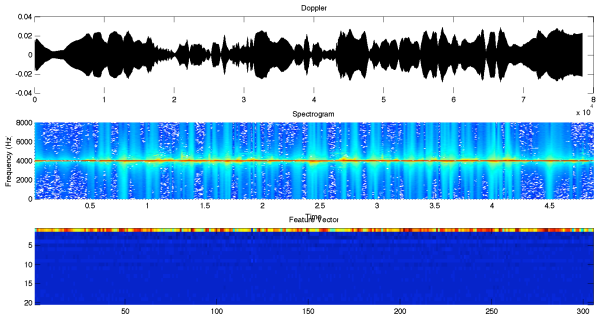


Figure 3. Doppler signal, spectrogram, and features from a talking face.

6. CLASSIFIER

We use a simple Bayesian formulation for recognizing talking faces. For each subject, we learn a separate distribution for the feature vectors from of the Doppler features computed from a set of training recordings. For the purpose of modelling these distributions, we assume that the sequence of feature vectors from any channel to be IID. Specifically, we assume that the distribution of the Doppler feature vectors for any subject w is a Gaussian mixture of the form:

$$P(\mathbf{d}|w) = \sum_i c_{w,i} \mathcal{N}(\mathbf{d}; \mu_{w,i}, R_{w,i}) \quad (4)$$

where $\mathbf{d}$ represents a random feature vector derived from the Doppler signal. $\mathcal{N}(X; \mu, R)$ represents the value of a multivariate Gaussian with mean μ and covariance R at a point X. $\mu_{w,i}$, $R_{w,i}$ and $c_{w,i}$ represent the mean, covariance matrix and mixture weight respectively of the i^{th} Gaussian in the distribution of Doppler feature vectors for subject w. All parameters of the distribution for any subject are learned from a small amount of training Doppler recordings from that subject.

Classification is performed using a simple Bayesian classifier. Let $\{\mathbf{D}\}$ represent the set of all doppler feature vectors obtained from any test recording. The recording is recognized as having come from a subject $\hat{w}$ according to the rule:

$$\widehat{w} = \underset{w}{\operatorname{argmax}}\, P(w) \prod_{\mathbf{d}\in\mathbf{D}} P(\mathbf{d}|w) \tag{5}$$

where $P(w)$ represents the *a priori* probability of the subject w. We assume the *a priori* probability to be uniform for all the subjects.

7. EXPERIMENTS

Experiments were conducted to evaluate the effectiveness of the ultrasonic Doppler sensing as a mechanism for recognition of talking faces. All experiments were conducted on a corpus of Doppler recordings collected at Mitsubishi Electric Research Labs. A total of 50 subjects were made to record 75 sentences each from the TIMIT corpus. Each sentence was treated as a separate recording; we thus has 75 recordings per subject. The subjects included people of both gender, including men with facial hair.

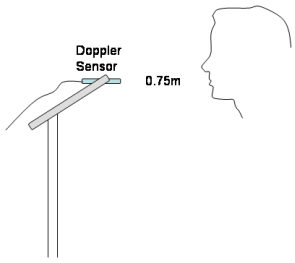


Figure 4. Experimental setup: Subjects spoke facing the Doppler sensor. Subjects typically sat a distance of 0.75m from the sensor.

For the recording, subjects were seated in a soundproofed room (since the audio data from the spoken utterances were also collected) facing the Doppler-augmented microphone setup of Figure 2. Before the experiments they were given a small demonstration on how to use the recording setup (*i.e.* how to use the keyboard/mouse to operate the recording setup). They were then instructed to read sentences which were displayed on a screen adjacent to the microphone naturally, without attempting to restrict the motion of their faces and heads in any manner. They were also instructed not to make any unnatural movements (*i.e.* malicious motions) in front of the setup, or to block their face in any manner during recording. No additional instructions were given. Subjects were not interrupted once recording began, nor were their actions corrected or modified during the recordings. All data from a subject were recorded in a single session, although they were allowed to take breaks.

The recorded data for each speaker were divided into two sets, a training set of 37 utterances and a test set of 38 utterances. Gaussian mixture densities with different numbers of Gaussians per density were trained for each subject. Table 1 shows the results obtained.

Table 1. Talker classification accuracy vs. number of Gaussians in the GMMs

# Gaussians	4	10	20	40	50
% Accuracy	81	85	87	90	90

Table 2. Percent accuracy in classifying the gender of the talker. Rows represent the test data, columns the ground truth. The overall accuracy is 91.25%.

%	Male	Female
Male	82.5	17.5
Female	0	100

We note that using GMMs composed of 40 or more Gaussians we are able to achieve an accuracy of over 90% in recognizing faces.

The facial structures of male and female humans is known to be different. It may hence be inferred that the patterns of facial motion in both genders is also different. This hypothesis is also corroborated by Berry [9]. To test this hypothesis we ran an alternate experiment, where we attempted to identify only the gender of the speaker. For this experiment we separated the set of male and female subjects into a training and a test set, retaining only a total 360 recordings each from males and females in our test set. The training set for each gender was also trimmed to 360 recordings to maintain balance. A separate GMM was trained for each gender, and the 720 test recordings were classified with these GMMs. Table 2 reports these results.

We note from the table that we are able to identify the gender of the speaker with an accuracy of over 91%. The results indicate that speech-related facial movements have statistically different patterns for male and female subjects. This relation of facial motion to gender has hithertofore not been quantified or studied to the best of our knowledge. Interestingly, males are far more likely to be misrecognized as female than the other way around.

8. CONCLUSION AND DISUCSSION

Our experiments indicate that talking-related facial movements are indeed unique and may be used to recognize faces. Importantly, the manner in which our measurements are taken make them fundamentally different from video, and the two may in fact be used together for further improved recognition. The results may also be interpreted as computational corroboration of human studies that show that facial movements may be distinctive cues for determining the gender and identity of subjects. Our approach for signal characterization and classification is not optimal either. We do not actually attempt to characterize the temporal nature of the data in any manner. The features derived from the signal are also very simplistic and make no

attempt to distinguish between different facial features that move with the same instantaneous velocity at a given instant (it may be possible to resolve these to some extent by analysis of modulation patterns of the energy in different frequency bands). We also believe that better classification may be achieved through the use of discriminative classifiers. These and other avenues remain to be explored.

Even the current set of experiments may, at best, be considered to demonstrate promise in the proposed approach. The data we have gathered are not actually representative of realistic patterns of facial motion that may be obtained in spontaneous talking. Rather, they only characterize the type of movements during more controlled *reading*. In order to obtain a more realistic measurement of the effectiveness of the proposed system, we may need to collect data from subjects talking in more conversational scenarios. Since such situations typically do not permit mounting of Doppler sensors in locations from which reliable measurements may be made, collection of such data is a difficult task.

In all of our recordings talkers faced the Doppler sensor. In more realistic scenarios, we may not have such control over the direction of the talker's face. In such situations we may need to use multiple, or arrays of receivers to capture the reflected signal, in order to achieve a more complete, multi-directional characterization of the talker's face. We are currently addressing these issues.

References

[1] M. H. Yang, D. J. D. J. Kriegman, and N Ahuja, "Detecting faces in images: a survey," *EEE Trans on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[2] C. Sanderson and S Bengio, "Robust features for frontal face authentication in difficult image conditions.," *Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pp. 495–504, 2003.

[3] P Viola and M. Jones, "Robust real-time object detection.," *Int. J of Computer Vision*, 2002.

[4] P. J. Phillips, "Support vector machines applied to face recognition," *conference on Advances in neural information processing systems*, pp. 803–809, 1999.

[5] B. Moghaddam, W. Wahid, and A Pentland, "Beyond eigenfaces: Probabilistic matching for face recognition," *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition*, 1998.

[6] K. Kalgaonkar, Rongquiang Hu, and B. Raj, "Ultrasonic doppler sensor for voice activity detection," *Signal Processing Letters, IEEE*, vol. 14, no. 10, pp. 754–757, Oct. 2007.

[7] K. Kalgaonkar and R Bhiksha, "Acoustic doppler sonar for gait recogination," *IEEE Int. Conf. of AVSS*, 2007.

[8] K. Kalgaonkar and R Bhiksha, "Ultrasonic doppler sensor for speaker recognition," *IEEE Int. Conf. of ICASSP*, 2008.

[9] D. S. Berry, "Child and adult sensitivity to gender information," *J of Ecological Psychology*, vol. 3, no. 4, pp. 349–366, 1991.

[10] B. K Knappmeyer, I. M. Thornton, and H. H. Buelthoff, "The use of facial motion and facial form during processing of identity," *Vision Research*, vol. 43, no. 18, pp. 1921–1936, 2003.

[11] K. G. Munhall and J. N. Buchan, "Something in the way she moves," *Trends in Cognitive sciences*, pp. 51–53, 2004.

[12] L. F. Chen, H. Liao, and J. C. Lin, "Person identification using facial motion," *Proc of Internation conf. on Image Prco. Proc of Internation conf. on Image Prco.*, , no. 677-680, 2001.

[13] "Audio and video based biometric person authentication," *Proc and Lecture Notes in Computer Science*, 2003.