

**This is the author-manuscript version of this work - accessed from
<http://eprints.qut.edu.au>**

**Lucey, Patrick J. and Dean, David B. and Sridharan, Sridha (2005)
Problems Associated with Current Area-Based Visual Speech
Feature Extraction Techniques. In Proceedings International
Conference on Auditory-Visual Speech Processing (AVSP), pages pp.
73-78, Vancouver Island, British Columbia, Canada.**

Copyright 2005 (please consult author)

PROBLEMS ASSOCIATED WITH CURRENT AREA-BASED VISUAL SPEECH FEATURE EXTRACTION TECHNIQUES

Patrick Lucey, David Dean and Sridha Sridharan

Queensland University of Technology
Speech, Audio, Image and Video Research Laboratory
GPO Box 2424, Brisbane 4001, Australia
{p.lucey, d.dean, s.sridharan}@qut.edu.au

ABSTRACT

Techniques such as principle component analysis (PCA), linear discriminant analysis (LDA) and the discrete cosine transform (DCT) have all been used to good effect in face recognition. As these techniques are able to compactly represent a set of features, researchers have sought to use these methods to extract the visual speech content for audio-visual speech recognition (AVSR). In this paper, we expose the problems of employing such techniques in AVSR by running some visual-only speech recognition experiments. The results of these experiments illustrate that current area-based feature extraction techniques are heavily dependent on the visual front-end, as well as being ineffective in decoupling adequate speech content from a speaker's mouth. As a potential solution, we introduce the concept of a free-parts representation, which may be able to circumvent many of these problems experienced by current area-based techniques.

1. INTRODUCTION

It is largely agreed upon that the majority of visual speech information comes from a speaker's mouth [1]. As a result, a large proportion of the work that has been conducted in *audio-visual speech recognition* (AVSR) has been towards the goal of finding a suitable mouth representation for recognition purposes. The more discriminant and compact the mouth representation, generally the easier the recognition task is.

Mouth features can be categorized into two types, namely: area, and contour based representations. Area-based representations are concerned with transforming the whole *region of interest* (ROI) mouth pixel intensity image into a meaningful feature vector. Contour based representations, are concerned with parametrically atomising the mouth, based on a priori knowledge of the components of the mouth (i.e. outer and inner labial contour, tongue, teeth, etc.) [2]. In a paper by Potamianos et al. [3], a review was conducted between area and contour features for the tasks of visual-only

speech recognition. In this work, it was shown that area representations obtained superior performance as well as being more robust to visual noise and compression artifacts.

Much work performed in area-based AVSR visual feature extraction has closely paralleled work done previously in face recognition. For example, after Turk and Pentland's [4] ground breaking paper on *Eigenfaces* was written, a similar strategy for compactly representing the mouth's major modes of variation using principal component analysis (PCA) was formulated by Bregler [5], which they referred to as *Eigenlips*. Similarly, after some innovative work by Belhumeur et al. [6] demonstrating considerable benefit in employing linear discriminant analysis (LDA) for recognising faces, similar benefit was cited by Duchnowski et al. [7] employing an LDA strategy for representing the mouth. Since then, LDA in combination with some post-processing strategies of maximum likelihood linear transform (MLLT) [8, 3] and mean-subtraction [8, 3] have been used to good effect and are considered the best mouth feature set to date. However, in a review paper by Chibelushi et al. [9], it was reported that no significant difference in speech classification accuracy was obtained between PCA and LDA features. As a result, recent research in this field [10, 11, 12, 13] has looked towards using the computationally efficient discrete cosine transform (DCT), which enables real-time automatic systems to be developed [8]. The DCT has gained its popularity in AVSR, through its ability to compactly represent visual speech just as effectively as PCA and LDA without requiring supervision or massive amounts of training data.

Even though these area-based techniques have been used to reasonable effect for AVSR, they are still plagued by many problems. Firstly, they are susceptible to pose, camera and lighting variations [8]. They also heavily rely on the visual front-end to detect and track a speaker's mouth with extreme precision. However, their biggest problem appears to be that they are not capable of capturing enough speech content from the visual modality to reliably recognise speech. In this paper, we expose these latter problems, and as a potential solution, introduce the concept of a free-

parts area-based representation which may be able to circumvent many of these current problems.

The rest of this paper is organised as follows. In Section 2, an evaluation of some of the current area-based feature extraction techniques are given. In Section 3, the importance of the visual front-end in an AVSR system is explained and its specific importance to the visual speech feature extraction process. Section 4, details the setup and the methods used in the experiments. Section 5, gives the results from the experiments and highlights the problems associated with the current feature extraction techniques. Section 6, introduces the concept of a free-parts representation of the ROI and explains how this can be used to counteract these problems. A summary follows in Section 7.

2. AREA-BASED VISUAL SPEECH FEATURES

Current area-based visual speech feature extraction techniques such as PCA, LDA and DCT are all based on *monolithic* representations [14] of the mouth. The term monolith is used to describe the holistic vectorised representation of the mouth based purely on pixel values within an image array. For the purposes of this study, only the PCA and DCT feature extraction techniques were used due to their natural ability to bring a priori knowledge of the mouth to the representation as well as them being unsupervised processes.

Generally, an accurate measure of the quality of visual features is indicative of how well it performs in the task it is being used for, which in this case is visual-only speech recognition. For purposes of notation the mouth image matrix $I(x, y)$ is also expressed as the vectorised column vector $y = \text{vec}(I)$. A description of the visual features is as follows:

PCA: was used to create a twenty dimensional subspace Φ_{PCA} preserving the highest 20 linear modes of mouth variation. Delta features were also taken, resulting in a feature vector of 40.

MRPCA: is where the mean removed mouth sub-image y^* is calculated from a given temporal mouth sub-image sequence $Y = \{y_1, \dots, y_T\}$ such that,

$$y_t^* = y_t - \bar{y}, \text{ where } \bar{y} = \frac{1}{T} \sum_{t=1}^T y_t \quad (1)$$

This approach is very similar to cepstral mean subtraction used on acoustic cepstral features to improve recognition performance by providing some invariance to unwanted variations. In the visual scenario, this unwanted variation usually stems from the subject's appearance. Mean-removal PCA (MRPCA) uses these newly adjusted y^* mouth sub-images to create a new twenty dimensional subspace Φ_{MRPCA} preserving the highest modes of mean removed mouth variation. This approach was first proposed by Potamianos et al.

[15] for improved visual speech recognition performance. The delta features were also taken giving a feature vector of 40.

DCT: is where a 2-D DCT is performed on each mouth image matrix $I(x, y)$, and the top 20 DCT coefficients according to a zig-zag scan are retained. Delta features were also used.

MRDCT: is similar to MRPCA, with a 2-D DCT being applied to the mean-removed mouth images y^* . The top 20 DCT coefficients according to a zig-zag scan were retained with delta features being used.

3. VISUAL FRONT-END

Before the visual speech features can be extracted, the ROI has to be detected and tracked. In an AVSR system, this is performed by the visual front-end. For AVSR to be effective, it is essential that the visual front-end be highly accurate, otherwise these errors will cascade throughout the system and have a large effect on the ability of the final AVSR system to reliably recognise speech. This is known as the *front-end effect*.

In this study, the visual front-end consisted of three stages; face location, eye location and lip location. As shown in Figure 1, each stage was used to help form a search region for the next stage.

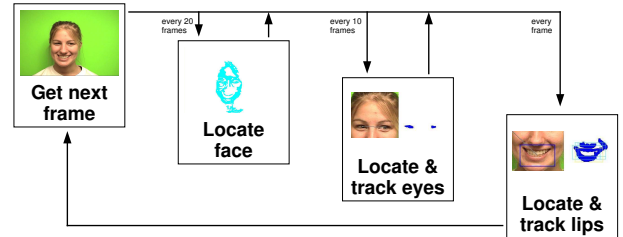


Figure 1: Overview of lip tracking system.

3.1. Face Location

Before face location was performed on the videos, 10 manually selected skin points for each speaker are used to form thresholds for the red, green and blue (r, g, b) values in colour-space for skin segmentation. The thresholds for each colour-space were calculated from the skin points as

$$\mu_c - \sigma_c \leq p_c \leq \mu_c + \sigma_c, \quad (2)$$

Where $c \in \{r, g, b\}$, μ_c and σ_c are the mean and standard deviation of the 10 points in colour-space c , and p_c is the value of the pixel being thresholded in colour-space c .

Once the thresholds were calculated, they were used for skin segmentation of the video to generate a bounding

box of the face region within the frames every 20 frames, and this face location was remembered in the intermediate frames.

3.2. Eye Location and Tracking

When transformed into $YCbCr$ space, the eye region of face images exhibit a high concentration of blue-chrominance, and a low concentration of red-chrominance. Therefore eye detection can be done in the $Cr - Cb$ space with reasonable results. However, eyebrows often appear as false positives and can degrade results. To remove the influence of eyebrows the $Cr - Cb$ image can be shifted vertically and subtracted from the original $Cr - Cb$ image. This will cancel the eyebrow minima by subtracting the eye minima, whereas the eye minima will be subtracted by the high values in the skin region and receive a large negative value suitable for thresholding [16].

To locate the eyes from the face region from the previous stage, the top half of the face region was designated as the eye search-area, which was then searched using the shifted $Cr - Cb$ algorithm for the eye locations. The possible eye candidates were searched for two points that were not too large, too close horizontally, and not too distant vertically. Finally the two candidates which had the largest horizontal distance were chosen to be the eye locations. This process was performed every 10 frames, and the locations were remembered in the intermediate frames.

3.3. Lip Location and Tracking

Once the eye locations have been found, they are used to calculate a lip search region, as shown in Figure 2. The lip search region is then rotation-normalised, converted to R/G colour-space, and thresholded. The lip candidates from the thresholding are examined to remove unlikely lip locations (eg. too small, wrong shape). A search-window of 125×75 pixels is then scanned over the lip candidate image to find the windows with the highest concentration of lip candidate regions. The final lip ROI is chosen as the lowest, most central of these windows. Once the ROI was correctly located, it was rescaled to 30×18 pixels for the experiments.

4. EXPERIMENT SETUP AND METHODS

Training and evaluation visual speech was taken from the Clemson University, *CUAVE*, audio-visual database [17]. The *CUAVE* database was selected as it is presently the only common audio-visual database which is available for all universities to use. This is important for benchmarking and comparison purposes. Even though the *XM2VTS* database [18] is another database which is available to researchers for the same purposes, the *CUAVE* database was chosen due to the fact that it is freely available. The *CUAVE*

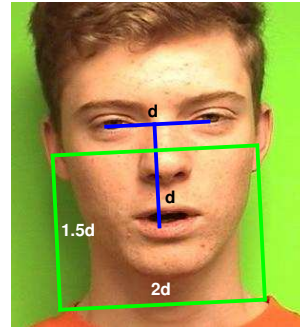


Figure 2: Calculating lip search region from eye locations.

database consists of two major sections, one of individual speakers and one of speakers pairs. For this study, only the individual speakers were used.

The normal stationary-speech sections of the *CUAVE* database were used for these experiments. The normal section of the database consisted of each of the 36 individual speakers uttering the isolated digits “zero” to “nine” a total of 5 times each. For each speaker, 4 of the isolated sequences were chosen as the training set, with 1 sequence used for testing. In these experiments, each of the digits were modelled using 3 state and 3 mixture Hidden Markov models (HMMs), using HTK [19]. This HMM topology was used as previous work in this field showed that this topology gave superior results [20]. The models were tested using both speaker independent and speaker dependent models. The speaker dependent models were developed using per-speaker maximum likelihood linear regression (MLLR) adaptation of the HMMs.

5. EXPERIMENTAL RESULTS AND PROBLEMS WITH CURRENT TECHNIQUES

The visual speech recognition results are given in Table 1. As can be seen in the non-adapted results, the PCA and DCT features gave comparable results suggesting that the DCT is more conducive to AVSR, as it is more computationally efficient in terms of processing and is not data dependent. The results for these two techniques, however, were not good, with both achieving word error rates (WERs) around 50%. This suggests that they did not extract sufficient speech content from the video sequences. A possible reason for this is illustrated in Figure 3, in which the second and third DCT coefficients for three digits are compared for two speakers. As can be seen, there is no speech class separation between the words, however, there is clear speaker class separation. This result was observed across all speakers for all digits. This may give a hint toward the fact that these feature extraction techniques are not able to adequately decouple the useful speech content from the speaker information. Intu-

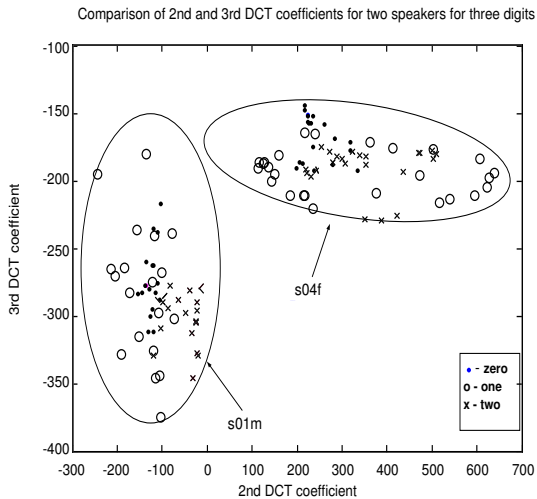


Figure 3: Plot showing the class separation between two different speakers using the 2nd and 3rd DCT coefficients

itively, this result makes sense, as most of the information contained in these images essentially consist of redundant speaker skin and shape information, with little change happening between adjacent frames.

Table 1: Experimental results showing the Word Error Rates (WER) percentages of the various area-based features extraction techniques. (NA = speaker independent models, SA = speaker dependent models using MLLR adaptation)

Feature Extraction	WER (%)	
	NA	SA
PCA	52.52	37.82
MRPCA	44.19	24.46
DCT	49.66	25.92
MRDCT	42.74	21.03

For speech recognition, speaker information is not important, but the dynamic speech content is. This is shown by the fact that mean-removed performance of the DCT and PCA features improved by around 6-8%. This result suggests that the dynamic features of speech, including the delta features are beneficial for speech recognition. As shown in Figure 4, the speaker separation is essentially lost just leaving the speech information. However, as it can be seen in this plot, speech separation is still not evident, which is shown also in the relatively poor results.

As already mentioned, the results obtained in these experiments are not impressive. However, these experimental results are around the same level as other similar work as reported in [10, 11]. The results can be improved greatly by employing speaker adaptation, and by doing this it can

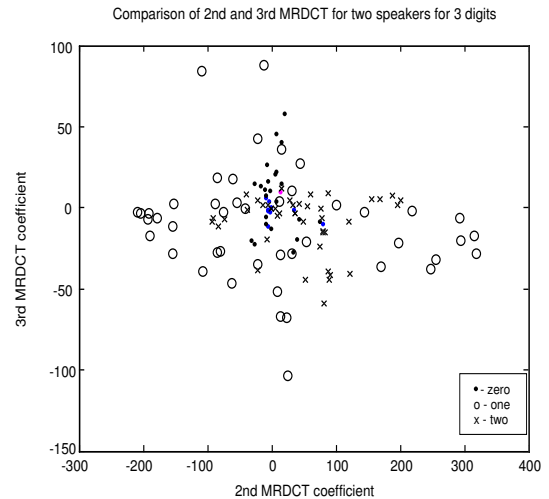


Figure 4: Plot showing the class separation between two different speakers using the 2nd and 3rd MRDCT coefficients

improve the WERs by over 20% as shown in Table 1. However, for AVSR to be employed in “real-world” scenarios in the near future, it is expected that these applications would be restricted to a small vocabulary task such as kiosk type application (e.g. train station ticket machine). In a small vocabulary task, the AVSR system would not have sufficient data to adapt the models to the various speakers, so it is really not a viable solution.

There are basically two main reasons for the poor results. The first one being that the monolithic techniques used depend heavily on a speaker’s ROI being in the same position in every frame. However, in practice this does not occur and as such, we get small variations in speaker mouth orientation and pose. These small variations in a speaker’s ROI have a large effect on the eventual features for each frame. As these monolithic techniques only rely on the pixel intensity values, these slight variations create errors in the new compact representations, thus greatly affecting the final recognition results. This was experienced in this case, as the visual front-end used in this experiment had the tendency to allow small variations in the position of the speaker’s ROI. This example highlights that the current area-based techniques are not robust to the effects of variations in the visual front-end. The second problem stems from the lack of quality training data. It is well known and widely accepted that in the field of AVSR, the lack of a large audio-visual database is a major limiting factor to the success of this technology. This study also exhibits this problem. Having only about half-an hour of available visual speech data is a major reason why the visual speech recognition rates of these experiments are an order of magnitude away from their acoustic counterpart. However, with the recent improvement in



Figure 5: Comparison of ROI tracking variabilities, the top being tracked correctly and the bottom not.

computer technology, this is becoming less of a concern.

6. FREE-PARTS REPRESENTATION OF THE MOUTH AREA

As seen in the previous section, current AVSR area-based feature extraction techniques are not robust enough to deal with “real-world” scenarios. However, in [14], a novel feature extraction technique in face recognition has been developed that is able to circumvent existing problems by obtaining features that lead to robust, generalisable discriminant classifiers via the use of a free-parts representation.

The term free-parts is employed to denote a representation of the mouth that can be considered as an ensemble of image patches of the mouth image array where the position/structure of these patches within the image can be relaxed. The relaxation of structure in the mouth has the major benefit of obtaining a “distribution” instead of a “point” for each visual speech frame. By utilising a distribution instead of a point structure, the dependence on the detection and tracking of the face is greatly reduced. For example, in Figure 5 there are two identical ROI’s, with one being correctly tracked and the other not. With the current monolithic techniques being used, the two would be recognised as totally different as their respective pixel values or projected image spaces would not correlate. However, the respective distributions of the two would be much similar and therefore the speech would be more likely to correspond to each other, thus making the free-parts representation more robust to the visual front-end.

Another example which illustrates the robustness of the free-parts representation is given in the toy example shown in Figure 6. In this figure, the image is partially corrupted by some type of visual noise. In a monolithic representation, this image would be pretty useless as some of the



Figure 6: Example of an image being partially corrupted.

pixel intensities have been corrupted, and thus an image transform on this particular image would produce an erroneous output. However, for a free-parts representation, the image would be broken up into blocks, and by averaging over all the blocks, the corrupted data would be effectively smoothed out in the distribution, thus minimising the effect the corrupted data has on the visual speech.

It has been widely reported that visual speech classifiers [20] are generally under trained in comparison to acoustic classifiers due to the unavailability of training observations. A free-parts representation will go some way to lessening this problem through the natural creation of current monolithic area-based features are that the features are too data dependent, in that the features generated from a development set of one set of speakers do not generalise well to an evaluation set containing another set of speakers. This can largely be attributed to feature extraction processes like PCA and LDA being highly data dependent. Free-parts representations will be able to circumvent some of these problems as the feature extraction process is largely data independent.

Implementation of the free-parts representation for visual speech is as follows. Firstly, each image is broken up into patches or blocks, with the blocks being possibly overlapping. Features for each block are then obtained using techniques such as the DCT. Instead of modelling each utterance with the features of the entire image, the features of each block are used to train the HMMs. The difficulty of employing this technique is being able to use multiple feature vectors for each observation.

7. SUMMARY AND CONCLUSION

In this paper, the problems of current area-based feature extraction techniques were investigated. As a result, the ability of the current techniques such as DCT and PCA were shown to inadequately decouple the useful speech content from the redundant speaker information. It was shown that by removing the static speaker information by subtracting the mean image, greater speech intelligibility can be found

by utilising the dynamic information. The reliance on the visual front-end for correct tracking and detecting speaker's ROI became evident, with even the smallest variations causing errors. As these variations are indicative of the types of problems an AVSR system would encounter in a "real-world" environment, a novel technique of using a free-parts representation instead of the monolithic representation was introduced. The free-parts representation has been used to good effect in face recognition and its potential in the field of AVSR is quite tantalising with its robustness against face pose and orientation variation and its ability to overcome problems of inadequate training data. The implementation of the free-parts technique was also described and the results pertaining to this will hopefully be published soon.

8. ACKNOWLEDGEMENTS

We would like to thank Clemson University for freely supplying us their CUAVE audio-visual database for our research.

9. REFERENCES

- [1] F. Lavagetto, "Converting speech into lip movements: a multimedia telephone for hard of hearing people," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no. 1, pp. 90–102, 1995.
- [2] T. Wark and S. Sridharan, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *International Conference on Pattern Recognition*, 1998, vol. 1, pp. 123–125.
- [3] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *Proceedings of IPCIP'98 International Conference on Image Processing*, Chicago, IL, USA, 1998, vol. 3, pp. 173–177, IEEE Comput. Soc.
- [4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.
- [5] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, 1994, vol. 2, pp. 669–672.
- [6] D. J. Kriegman P. N. Belhumeur, J. P. Hespanha, "Eigenfaces vs fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711–720, 1997.
- [7] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip reading," in *International Conference on Spoken Language and Processing*, Yokohama, Japan, 1994, pp. 547–550.
- [8] G. Potamianos, C. Neti, J. Luetin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. MIT Press, Boston, 2004.
- [9] C.C. Chibelushi, F. Deravi, and J.S.D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 23–37, 2002.
- [10] G. Potamianos and C. Neti, "Improved roi and within frame discriminant features for lipreading," in *Proceedings 2001 International Conference on Image Processing*, Thessaloniki, Greece, 2001, vol. 3, pp. 250–253, IEEE.
- [11] M. Heckmann, F. Berthommier, C. Savariaux, and K. Kroschel, "Effects of image distortions on audio-visual speech recognition," in *AVSP*, 2003, pp. 163–168.
- [12] J. N. Gowdy E. K. Patterson, "An audio-visual approach to simultaneous-speaker speech recognition," in *ICASSP '03*, 2003, pp. 780–783.
- [13] P. Motlicek, L. Burget, J. Cernocky, and I. Potucek, "Phoneme recognition of meetings using audio-visual data," in *AMI Workshop*, Martigny, 2004.
- [14] S. Lucey, "The symbiotic relationship of parts and monolithic face representations in verification," in *International Workshop on Face Processing in Video (FPIV)*, Washington D.C., USA, 2004.
- [15] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *Proceedings of IPCIP'98 International Conference on Image Processing*, Chicago, IL, USA, 1998, vol. vol.3, pp. 173–177, IEEE Comput. Soc.
- [16] D. Butler, C. McCool, M. McKay, S. Lowther, V. Chandran, and S. Sridharan, "Robust face localisation using motion, colour and fusion," in *Seventh International Conference on Digital Image Computing: Techniques and Applications*, C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, Eds., Macquarie University, Sydney, Australia, 2003, CSIRO Publishing.
- [17] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: a new audio-visual database for multimodal human-computer interface research," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, 2002.
- [18] K. Messer, J. Matas, J. Kittler, J. Luetin, and G. Maitre, "Xm2vts: The extended m2vts database," in *Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication*, Washington D.C., 1999, pp. 72–76, IEEE.
- [19] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Entropic Ltd, 2002.
- [20] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in *International Conference of Audio- and Video-Based Person Authentication (AVBPA)*, Guildford, U.K., 2003, pp. 260–267.