AN EVALUATION OF AUDIO-VISUAL PERSON RECOGNITION ON THE XM2VTS CORPUS USING THE LAUSANNE PROTOCOLS*

Kevin Brady, Michael Brandstein, Thomas Quatieri, Bob Dunn

MIT Lincoln Laboratory, 244 Wood St., Lexington MA 02420-9185
{kbrady, msb, tfq, rbd}@ll.mit.edu

ABSTRACT

A multimodal person recognition architecture has been developed for the purpose of improving overall recognition performance and for addressing channel-specific performance shortfalls. This multimodal architecture includes the fusion of a face recognition system with the MIT/LL GMM/UBM speaker recognition architecture. This architecture exploits the complementary and redundant nature of the face and speech modalities. The resulting multimodal architecture has been evaluated on the XM2VTS corpus using the Lausanne open set verification protocols, and demonstrates excellent recognition performance. The multimodal architecture also exhibits strong recognition performance gains over the performance of the individual modalities.

Index Terms— Multimodal Person Recognition

1. INTRODUCTION

Multimodal audio-visual approaches have been investigated for a broad range of speech technology applications, including speech recognition [5], speaker recognition [6, 13], speech segmentation [7], and speech enhancement [8]. The allure of combining these specific modalities is strongly motivated by both the complementary and redundant natures of the visible articulators and resulting speech. This complementary nature is primarily through visual cues that can aid a listener in tracking the acoustical signal. An example of the redundant nature of these channels is the well known McGurk effect [9]. Reviews of person recognition using audio and visual modalities are available in [6, 13].

Visual person recognition approaches fall into one of two categories: model-based or appearance-based. In model-based approaches various aspects of the face (i.e. height and width of lips) are estimated and serve as a source for the feature set. These algorithms are typically limited by the quality of the feature estimates. Appearance-based approaches [15] implement a general representation of the face based on the original image. The inherent challenge of this approach is the "curse of dimensionality", necessitating methods for reducing facial image to a representative low dimensional space. This challenge was addressed by Turk and Pentland [1] in the 90's utilizing a Principal Component Analysis (PCA)-based approach popularly known as Eigenfaces. Increasingly impressive face recognition results have been demonstrated by Eigenfaces, as well as Fisherfaces [10] which is a Linear Discriminant Analysis (LDA) extension to Eigenfaces, as well as Kernel extensions to both of these methods [11].

While much work has been done in audio-visual person recognition, it is noteworthy that there has been little evaluation on open corpora using established testing criteria. This differs from the face recognition and speaker recognition communities where formal evaluations are common with standard testing protocols. Comparing results is challenging due to this lack of common testing standards.

The remaining sections of the paper are organized as follows. Section 2 provides an overview of the multimodal recognition task. Section 3 will provide a brief overview of the audio recognition task. Section 4 provides an overview of the visual recognition task. A brief overview of the XM2VTS Multimodal Corpus and corresponding Lausanne protocols are discussed in Section 5. While these protocols have been widely used for face verification, this is the first known instance of their use for multimodal verification. The experimental results are in Section 6 and concluding remarks are in Section 7.

2. MULTIMODAL RECOGNITION

Figure 1 illustrates a multimodal recognition engine that is composed of a preprocessing stage and a recognition engine.

* This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors, and are not necessarily endorsed by the United States Government.

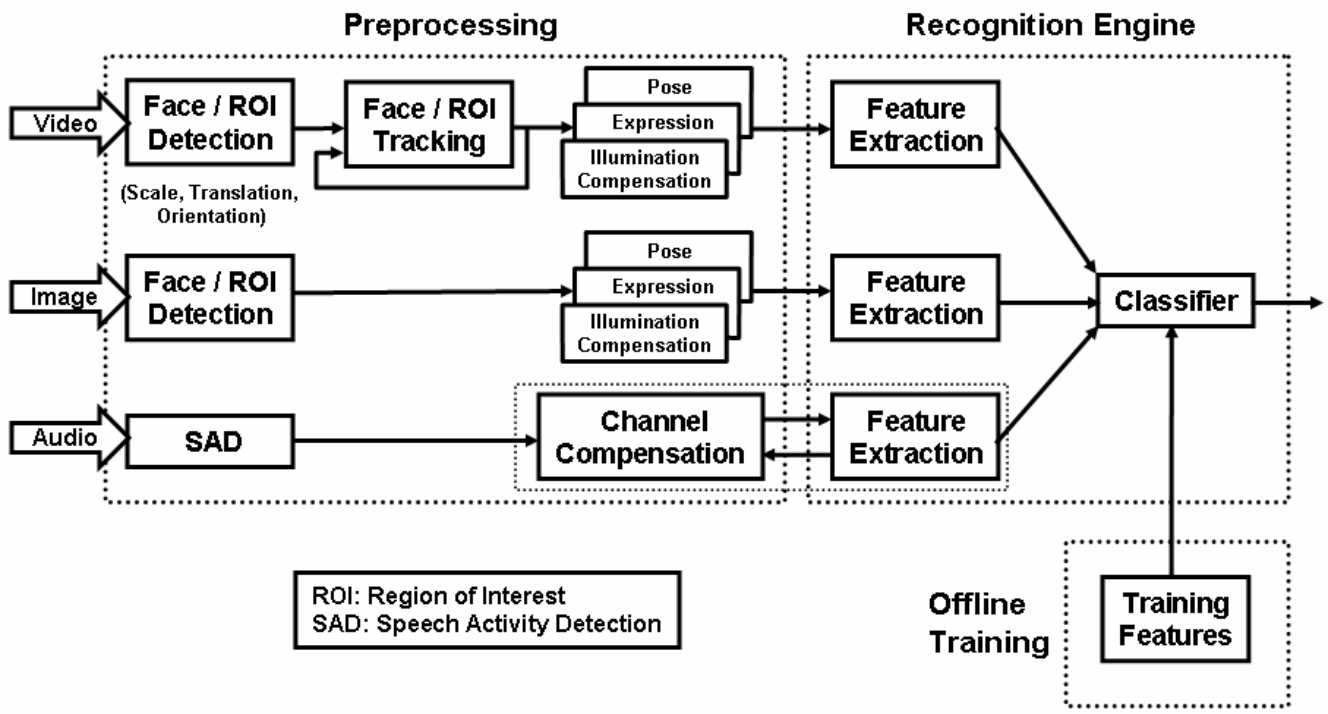

Figure 1 Multimodal Person Recognition Architecture

Only the Image and Audio modalities are evaluated in this paper. The preprocessing stage handles the detection and channel normalization for each of the respective input modalities: video, image, or audio. For audio files this includes speech activity detection (SAD). Audio channel normalization may occur either before (i.e. noise preprocessing) or after (i.e. RASTA) feature extraction as seen in Figure 1. For images and video this includes the detection [14] (and perhaps tracking) of the face as well as the normalization of the face relative to illumination, pose, and perhaps expression. The recognition engine extracts the relevant features from the respective channels and compares the features to an existing target model during the classification process. The model is generated in an offline process utilizing a set of training data.

A description of the feature sets used for each input modality will be discussed in the following sections. Separate scores are calculated for each mode and these are fused in a late integration process. Late integration has been shown to be highly successful for the speaker recognition task. The score fusion is calculated using a single stage perceptron with a log sigmoid output layer. Note that the output of such a network is well known to be a good representation of the posterior probability of target match.

3. AUDIO RECOGNITION

The MIT/LL speaker recognition architecture [4] utilizes a Gaussian Mixture Model (GMM) that statistically represents the features for each of the target speakers. The underlying features are cepstral features, as well as delta cepstral features. There is one GMM model per target and a Universal Background Model (UBM) that is used as an initial model from which each speaker model is trained by adapting the means of the Gaussians. The UBM is also used during testing as an alternate hypothesis whose score is compared to the target speaker model's score to create a likelihood ratio test. The UBM is typically generated with speech material collected across channel types so that all channels are represented.

4. VISUAL RECOGNITION

Two visual recognition engines are utilized in this paper: Eigenfaces [1] and Fisherfaces [10]. Both of these methods require a preprocessing stage wherein faces are detected and normalized. In the channel normalization step the images are uniformly resized, reoriented, and gray level histogram equalized to help address variations in pose and illumination.

The Eigenfaces approach, based on PCA, is an optimal linear space for representing faces. It vectorizes all of the normalized images (x_n) and calculates a mean face (x_0). This mean face is subtracted from each of the normalized faces ($A_n = x_n - x_0$) to form the matrix A. Normally, for PCA a sample covariance ($S = A A^T$) would be calculated at this point whose eigenvectors Φ would form the basis of the face space. Alternatively, Turk and Pentland suggest the formation of a lower dimensional sample covariance ($S = A^T A$) whose eigenvectors Φ are premultiplied ($\Phi_F = A\Phi$). This mathematical trick results in an equally effective basis for representing faces without the computational load of

doing an eigen-decomposition of the larger sample covariance. The columns of Φ_F form a basis for the Eigenfaces face space whose origin is the mean face Ψ. During the testing stage any normalized image can be projected into the face space as follows: $y_n = \Phi_F (x_n - x_0)$.

The Fishervectors approach, based on LDA, is an optimal linear space for discriminating faces. It utilizes the generation of the Eigenfaces of the respective images as described in the previous paragraph. The within class scatter matrix (S_w) and the between class scatter matrix (S_b) are calculated from the appropriate Eigenfaces. The goal is to solve a cost function for maximizing the feature space separation between classes relative to the separation within classes, which corresponds to solving the generalized eigenvalue problem ($S_b \times = S_w \times \lambda$). These Fishervectors can then be generated by projecting the appropriate Eigenvectors into the space defined by the eigenvector solution to the generalized eigenvalue problem.

Unlike the speaker recognition engine, there is no underlying statistical model for representing a target in this testing paradigm. Instead, there is an Eigenface / Fisherface vector corresponding to each image in the training set. Scores are obtained by directly comparing the estimated image vector to a target template. This results in multiple scores for each evaluated image (as many as there are images in the train set). Several popular scoring approaches for calculating these scores include the Euclidean metric, normalized correlation, and the Mahalanobis distance. The normalized correlation was used for our architecture.

5. XM2VTS MULTIMODAL CORPUS

The XM2VTS Audio Video Corpus [3] was collected at U. K. Surrey within the M2VTS (Multi-Modal Verification for Teleservices and Security Applications) project. It is a multimodal database consisting of four sessions on 295 subjects over a 4 month period. Each session consists of several audio-video sequences (A phonetically balanced sentence and several numeric sequences), a head rotation shot, and a static image shot. The audio is clean 16-bit 32 kHz, while the video is also clean and recorded at 25 Hz.

The XM2VTS Corpus has been widely used for evaluating recognition technology. This corpus consists of multimedia content for 295 subjects collected in 4 sessions over a period of 4 months. Two open set verification protocols [3], known as the Lausanne Protocols (I and II), were suggested for use on the XM2VTS Corpus, and have been utilized for several face verification competitions (ICPR 2000, AVBPA 2004, ICBA 2006). These verification protocols split the 295 XM2VTS subjects into 200 clients and 75 imposters that are further refined into train, test, and evaluation sets (as a function of session). These protocols will be used for the experimental results in the next section.

6. EXPERIMENT

The XM2VTS Corpus has been processed with this multimodal person recognition engine utilizing Lausanne Protocols I and II [3]. The audio material was extracted from the audio-video sequences and concatenated together resulting in about 20 seconds of data per subject per session. The visual material corresponded to the frontal static images collected for each session. The faces were manually registered.

The individual mode scores were fused via a single layer perceptron trained on the evaluation data. The results are illustrated with Detection Error Tradeoff (DET) plots in Figures 2 and 3 for Lausanne Protocols I and II respectively. The DET performance for the audio and image recognition approaches is quite similar. Note that the image recognition results are a result of fusing the Eigenface and Fisherface image recognition scores. What is quite dramatic is the recognition performance improvement obtained by fusing the scores from the visual and audio recognition approaches.

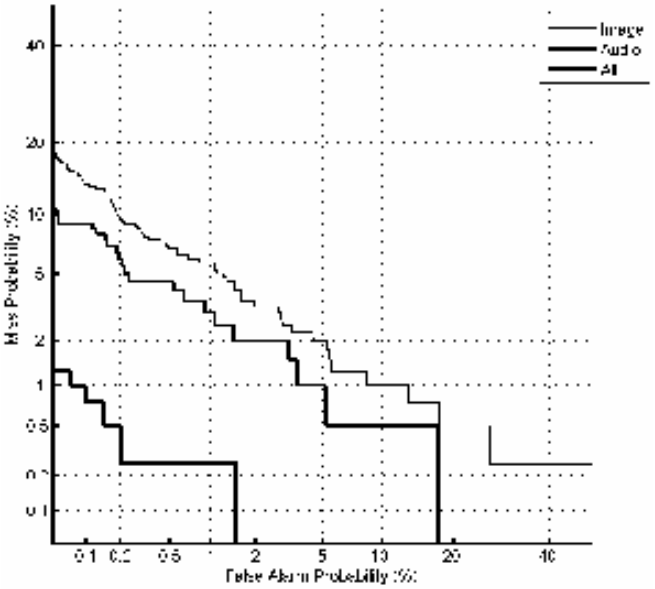

Figure 2 DET Results for XM2VTS Multimodal Corpus for Test Set of Lausanne Protocol I

The Equal Error Rates (EER) for each of the modes, including Eigenfaces and Fisherfaces alone, for the evaluation and test corpus are shown in Table 1. The EER describes the crossover point where the probability of miss and probability of false alarm are equal. As expected, Fisherfaces achieves a higher level of performance over the Eigenfaces approach. It is interesting to note, however, that

there is an obvious improvement in EER performance by fusing the Eigenfaces and Fisherfaces scores.

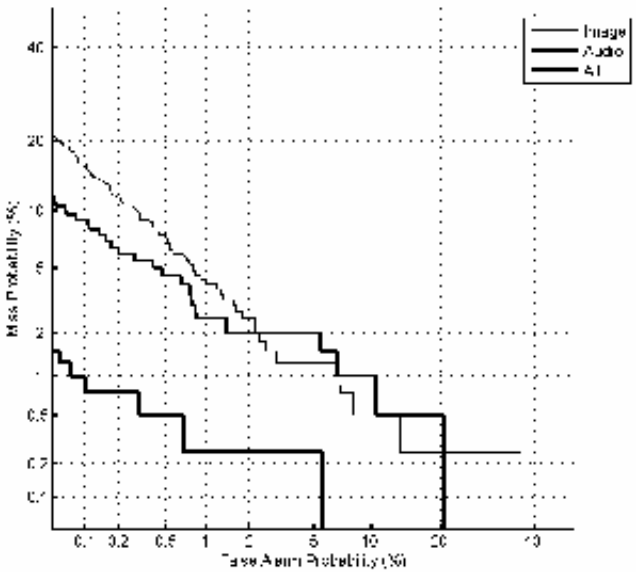

Figure 3 DET Results for XM2VTS Multimodal Corpus for Test Set of Lausanne Protocol II

	Lausanne I Evaluate (%)	Lausanne I Test (%)	Lausanne II Evaluate (%)	Lausanne II Test (%)
Eigenfaces	4.33	4.00	3.25	4.00
Fisherfaces	3.17	3.25	2.75	2.75
Image	3.00	2.81	2.50	2.23
Audio	1.67	2.00	1.57	2.00
Fusion	0.33	0.25	0.24	0.50

Table 1 EER Results for XM2VTS Multimodal Corpus

7. CONCLUDING REMARKS

A multimodal person recognition architecture has been presented and evaluated on the XM2VTS Corpus using the Lausanne Protocols. Significant recognition performance gains were obtained through the fusion of an audio and a visual modality over the performance of either modality alone.

In previous unpublished work, the authors successfully developed and evaluated an audio-video recognition engine using the same audio recognition engine and a GMM/UBM video recognition engine using the DCTs of the lip region. Future work will include the integration of this video recognition modality with the image and audio based recognition modalities. Based on the current work, an Eigenlip or Fisherlip approach to representing the lip or low face region would be a more effective option for representing the extracted features.

8. REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neurosciences, 3(1): 71-86, 1991.

[2] Tim Cootes, Chris Taylor, Hzizhuang Kang, Vladimir Petrovic, "Modeling Facial Shape and Appearance", Handbook of Race Recognition, Springer, 2005.

[3] K. Messer et al, "XM2VTSDB: The Extended M2VTS Database", AVBPA, 1999, Washington DC.

[4] D. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, Vol. 10, No 1-3, 2000, pp 19-41.

[5] G. Potamianos, C. Neti, G. Iyengar, Eric Helmuth, "Large-Vocabulary Audio-visual Speech Recognition by Machines and Humans", Proc. EUROSPEECH, Aalborg Denmark, Sept 2001.

[6] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-based Bimodal Recognition", IEEE Trans. On Multimedia, Vol 4, No 1, pp 23-37, March 2002.

[7] M. W. Mak, W. G. Allen, "Lip-motion Analysis for Speech Segmentation in Noise", Speech Communication, Vol. 4, pp 279-296, 1994.

[8] L. Girin, J. L. Schwartz, and G. Feng, "Audio-visual Enhancement of Speech in Noise", J. Acoust. Soc. Am., Vol. 109, #6, pp 3007-3020, June 2001.

[9] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices", Nature, pp 746-748, Dec 1976.

[10] V. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Trans. PAMI, 19 (7), pp 711-720, July 1997.

[11] M. H. Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition using Kernel Methods, Proc. Of IEEE Int. Conf. on Face and Gesture Recognition, pp 2150220, Washington DC USA, May 2002.

[12] C. C. Broun, X. Zhang, R. M. Mersereau, and M. A. Clements, "Automatic Speechreading with Application to Speaker Verification", ICASSP, Orlando USA, May 2002.

[13] C. Sanderson and K. K. Paliwal, "Identity Verification Using Speech and Face Information", Digital Signal Processing Journal, Vol 14, pp 449-480, 2004.

[14] S. Z. Li, "Face Detection", Handbook of Face Recognition, Springer, 2005.

[15] G Shaknarovich and B. Moghaddam, "Face Recognition in Subspaces" Handbook of Face Recognition, Springer 2005.