Robust Automatic Human Identification Using Face, Mouth, and Acoustic Information

Niall A. Fox[1], Ralph Gross[2], Jeffrey F. Cohn[2], and Richard B. Reilly[1]

[1] Dept. of Electronic and Electrical Engineering,
University College Dublin, Belfield, Dublin 4, Ireland
niall.fox@ee.ucd.ie, richard.reilly@ucd.ie
http://ee.ucd.ie/mmsp/
[2] Robotics Institute, Carnegie Mellon University,
5000 Forbes Ave, Pittsburgh, PA 15213
rgross@cs.cmu.edu, jeffcohn+@cs.cmu.edu

Abstract. Discriminatory information about person identity is multimodal. Yet, most person recognition systems are unimodal, e.g. the use of facial appearance. With a view to exploiting the complementary nature of different modes of information and increasing pattern recognition robustness to test signal degradation, we developed a multiple expert biometric person identification system that combines information from three experts: face, visual speech, and audio. The system uses multimodal fusion in an automatic unsupervised manner, adapting to the local performance and output reliability of each of the experts. The expert weightings are chosen automatically such that the reliability measure of the combined scores is maximized. To test system robustness to train/test mismatch, we used a broad range of Gaussian noise and JPEG compression to degrade the audio and visual signals, respectively. Experiments were carried out on the XM2VTS database. The multimodal expert system out performed each of the single experts in all comparisons. At severe audio and visual mismatch levels tested, the audio, mouth, face, and tri-expert fusion accuracies were 37.1%, 48%, 75%, and 92.7% respectively, representing a relative improvement of 23.6% over the best performing expert.

1 Introduction

Biometrics is a field of technology devoted to verification or identification of individuals using physiological or behavioral traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas identification, a multi-class problem, involves identifying a user from a set of enrolled subjects; and becomes more difficult as the number of enrollees increases. In audio-video processing, the video modality lends itself to two experts, the face expert and the visual speech expert (referred to as the mouth expert here).

Deployed person recognition systems are generally unimodal. Face based identification is susceptible to pose/illumination variation, occlusion, and poor image quality [1], [2]. Audio-based identification achieves high performance when the signal-to-noise ratio (SNR) is high. Yet, the performance degrades quickly as the test

W. Zhao, S. Gong, and X. Tang (Eds.): AMFG 2005, LNCS 3723, pp. 264 – 278, 2005.

SNR decreases (referred to as a train/test mismatch), as shown in [3] and elsewhere. Visual speech based person identification under performs audio and face based experts, and is not thought of as a stand-alone person recognition expert.

To combat these limitations of unimodal audio-video based experts, a multimodal fusion approach can be adopted. This can both improve robustness and overall performance. The audio, face, and mouth modalities contain non-redundant, complementary information about person identity. For example, it was reported in [2] that the performance of the FaceIt face recognizer [4] is extremely sensitive to eye occlusion (dark sunglasses), yet the effect of mouth occlusion (scarf) was significantly lower. This provides motivation for combining the FaceIt and mouth experts, i.e. combining an expert emphasizing eye information with an expert emphasizing mouth information. Also, it is expected that, for person identity, audio and video information are complementary.

In order to exploit this complementary information, issues arise, such as how to account for the reliabilities of the modalities and at what level to carry out the fusion. Only a few studies have investigated the combination of audio, face, and temporal mouth information for the purpose of person recognition [5], [6], [7]. The majority of studies are bi-modal, employing either the audio and face modalities, [8], or, the audio and temporal mouth modalities (ignoring face) [3], [9], [10].

The audio, mouth, and face experts were combined for person recognition in [5], [6], [7]; yet none of these studies employed expert weights that adapt automatically to local test conditions. In [5], fusion was carried out at the decision level, thus no individual expert reliability information could be considered. The weighted sum rule was employed in [7], however, the weights could only be varied using manual supervision. In [6], the weights were global and set empirically. To the best knowledge of the authors, no person recognition system exists, that combines the audio, mouth, and face experts in an automatic unsupervised manner, while adapting to the local performance of each expert.

The aim of this study was to develop a tri-expert person recognition fusion system, combining audio, mouth sequence, and face information in an automatic unsupervised manner. Specifically the tri-expert information was to be combined, such that the fused system provided improved performance beyond existing systems, exhibiting higher robustness to mild through adverse test levels of both audio and visual (face and mouth) noise (train/test mismatch). Therefore, to fully fulfill the aims of this study, the contribution from each source of information to the final decision must be weighted dynamically by taking the current reliability of each source into account.

This paper is organized as follows. Sections 2 and 3 describe how person identification based on audio, mouth features, and face was performed. Section 4 investigates classifier fusion and develops the proposed fusion strategy. In Section 5, the audio-video corpus employed and its' augmentation for the specific experiments is described. In Section 6, we present results of extensive evaluations examining individual expert performance and fusion performance. The results are discussed in Section 7 and finally in Section 8, conclusions from the results are drawn.

2 Audio Identification

Audio based speaker identification is a mature topic, [11]. Standard acoustic methods are employed here. For the feature extraction, the audio signal was divided into frames using a Hamming window of length 20 ms, with an overlap of 10 ms. Mel-frequency cepstral coefficients (MFCCs) of dimension 16 were extracted from each frame [12]. The energy [12] of each frame was also calculated and used as a 17^{th} static feature. Static features refer to features extracted from individual audio frames that do not depend on other frames. Seventeen first order derivatives or delta features were calculated using W_D adjacent static frames, where W_D is the delta window size. The delta frames were appended to the static audio features to give an audio feature vector of dimension 34. These are calculated using HTK [12], employing a W_D value of five frames. Cepstral mean normalization [12] was performed on the audio feature vectors (to each audio utterance).

A text dependent speaker identification methodology was tested. For text dependent modeling [13], the same utterance is spoken by the subject for both training and testing. It was employed, as opposed to text independent modeling [11], due to its suitability to the database used in this study (see Section 5). The N subject classes S_n, $n=1,2,\ldots,N$, are represented by N speaker hidden Markov models (HMMs) denoted by λ_n, $n=1,2,\ldots,N$. The speaker utterance that is to be classified is a sentence, which is represented by a sequence, O_A, of feature vectors or observations denoted by,

$$O_A = \{o_1, o_2, \ldots o_t, \ldots, o_{T_A}\}, \tag{1}$$

where o_t is the speech frame at time t and T_A denotes the number of observations. For HMMs, the output scores are in *log-likelihood* form, denoted by $ll(O_A|\lambda_n)$.

3 Video Based Identification

Visual speech based speaker recognition differs from face recognition in two major ways. Firstly, face recognition employs the entire face area, conversely, visual based speaker recognition employs a region of interest about the speakers' mouth, where most of the speech information is contained. Secondly, for face recognition, a gallery of *static* face images forms a template, whereas for visual based speaker recognition, it is attempted to model the temporal characteristics of the visual speech signal.

3.1 Mouth Features Expert

It has been consistently shown in several visual speech studies, that pixel based features outperform geometric features [14], [15]. Geometric features/lip-contours require significantly more sophisticated mouth-tracking techniques compared to just locating the mouth region of interested (ROI) for pixel-based features. This may be difficult, particularly when the visual conditions are poor. Pixel based features employ linear transforms to map the image ROI into a lower dimensional space, removing the redundant information while retaining pertinent speech information. Many types of transforms are examined in the literature, including the *discrete cosine transform* (DCT) [14], [16], *discrete wavelet transform* (DWT) [14], and *principal component*

analysis (PCA) [15]. The DCT is one of most commonly employed image transforms. It has good de-correlation and energy compaction properties and has been found to outperform other transforms [15]. The visual mouth features were extracted from the mouth ROI, which consists of a 49×49 color pixel block (see Fig. 3). To account for varying illumination conditions across sessions, the gray scale ROI was histogram equalized and the mean pixel value was subtracted. The two dimensional DCT was applied to the pre-processed gray scale pixel blocks.

Considering that most of the information of an image is contained in the lower DCT spatial frequencies, the first 15 non-zero DCT coefficients were selected, using a mask that selects the coefficients in a tri-angular fashion (upper-left region of the transform matrix) [14]. This gives the static features. The visual sentences were modelled using the same HMM methodology as described for the audio sentences. Dynamic features (frame derivatives of the static features) were employed in previous studies, but exhibited very poor robustness to video degradation, compared to using just static features [17], and were not employed here. We have T_V visual observations (generally $T_A \approx 4xT_V$) and a sequence, O_M, of visual mouth speech feature vectors or observations denoted by,

$$O_M = \{o_1, o_2, ..., o_t, ..., o_{T_V}\}. \tag{2}$$

Each mouth expert HMM gives the *log-likelihood* $ll(O_M|\lambda_n)$, that the observation sequence O_M was produced by the n^{th} mouth expert model λ_n.

3.2 Face Expert

Most current face recognition algorithms can be categorized into two classes, image template-based or geometry feature-based. The template-based methods compute the correlation between a face and one or more model templates to estimate the face identity. Statistical tools such as Support Vector Machines (SVM) [18], Linear Discriminant Analysis (LDA) [19], [20], Principal Component Analysis (PCA) [21], [22], Kernel Methods [23], and Neural Networks [24] have been used to construct a suitable set of face templates. While these templates can be viewed as features, they mostly capture global features of the face images. Facial occlusion is often difficult to handle in these approaches.

The geometry feature-based methods analyze explicit local facial features, and their geometric relationships. Cootes et al. have presented an active shape model in [25] extending the approach by Yuille [26]. Wiskott et al. developed an elastic bunch graph matching algorithm for face recognition in [27]. Penev et. al [28] developed PCA into Local Feature Analysis (LFA) which is the basis for the commercial face recognition system FaceIt. LFA addresses two major problems of PCA. The application of PCA to a set of images yields a global representation of the image features that is not robust to variability due to localized changes in the input. Furthermore the PCA representation is non topographic, so nearby values in the feature representation do not necessarily correspond to nearby values in the input. LFA overcomes these problems by using localized image features in form of multi-scale filters. The feature images are then encoded using PCA to obtain a compact description.

FaceIt was among the top performing systems in a number of independent evaluations [1], [2], [29]. It has been shown to be robust against variations in lighting, facial expression and lower face occlusion. Each of the registered N subjects is represented by a face template λ_n. Unlike for the audio and mouth experts employed here, FaceIt gives a *confidence score*, rather than a log-likelihood, denoted here by $l(O_F|\lambda_n)$, i.e. the likelihood that the face observation O_F belongs to the n^{th} face template λ_n. For FaceIt, the set of N templates, λ_n, $n=1...N$, receives maximum and minimum scores of ten and zero respectively, i.e. $l(O_F|\lambda_n) \in [0,10]$.

4 Classifier Fusion

The fusion of audio and video information falls into two broad categories, *early integration* and *late integration* [13]. Early-integration consists of concatenating the feature vectors, from the different modalities, to give a combined larger dimensional feature vector. This has the disadvantage of high dimensionality and the inability to take the reliability of the individual modalities into account. Furthermore, features from some experts may not be suitable or even available for fusion with speech-based features, e.g. the FaceIt face recognizer.

Late integration can occur at the *score* level or at the *decision* level; and has several advantages: a) late integration involves lower data dimensions than early integration, b) early integration is less robust to sensor failure, c) for late integration, it is more straightforward to add new experts, d) late integration allows the fusion of modalities possessing different temporal synchrony e.g. face and audio.

A significant amount of information is lost when the expert confidence scores are mapped to the class labels (decisions). This is why that, if the individual expert reliabilities are to be considered, fusion should occur at the *score* and not the *decision* level, as the score level information is crucial for discerning the reliability of each expert. For *decision* fusion, the number of classifiers should be higher than the number of classes. This is reasonable for person verification. For person identification, the number of classes is large, rendering decision fusion unsuitable.

Two typical methods of combining the output scores from the N_E experts are the *product* and the *sum* rules [30]. The *product rule* consists of multiplying the N_E scores together. It is sensitive to expert errors; in the extreme case, if any single expert produces a close to zero score for a specific class; the combined score for that class will be close to zero. The *sum rule* is less sensitive to expert errors and will outperform the product rule when the expert errors are large. The robustness of the sum rule to expert errors was shown theoretically and verified experimentally in [30].

Experts scores can take many forms such as posteriors, likelihoods, and distance measures. Non-normalized scores cannot be integrated sensibly in their raw form, as it is impossible to fuse incomparable numerical scales. The min-max technique shifts and scales the scores into the range $[0,1]$. Given a set or a list of N scores $\{S_n\}_{n=1...N}$ corresponding to N class labels the normalized score is calculated as:

$$S_n^{'} = \left(\frac{S_n - S_{\min}}{S_{\max} - S_{\min}} \right), \quad S_n^{'} \in [0,1], \tag{3}$$

where S_{max} and S_{min} are the maximum and minimum scores from the set $\{S_n\}$. While been straightforward to implement the min-max norm, has been found to have comparable performance to more complicated methods [31], hence, it was used for experiments reported here. Its' poor robustness to outlier scores can be circumvented (in the person identification scenario) by considering only the top M ranked scores for normalization. This omits the worst (outlier) expert scores.

4.1 The Proposed Method

The fusion strategy was first developed for fusing any two experts, and was then extended to include an additional third expert. Each expert provides a list of N likelihoods: $\{l(O_m|\lambda_n)\}_{n=1...N}$ with $m \in \{A, M, F\}$. These are ranked into descending order and normalized into the range $[0,1]$ using Eqn. (3), applied to only the top M scores. Using a high value for M may retain the worst s(outlier) scores, which could unfairly skew the distribution. A very low value, would result in information loss; the limit been $M=1$, where all confidence information has been lost. Tests showed that the system performance degraded for $M<50$ and $M>100$. A value for M of 75 was employed for this study[1]. This value may depend on N, the number of classes. The set of M *ranked* normalized scores is denoted by $\{S(O_m|\lambda_i)\}_{i=1...M}$. We have the *weighted sum rule* (for the specific case of two experts):

$$S(O_1, O_2 \mid \lambda_i) = \sum_{m=1}^{N_E} \alpha_m \cdot S(O_m \mid \lambda_i) = \alpha_1 \cdot S(O_1 \mid \lambda_i) + \alpha_2 \cdot S(O_2 \mid \lambda_i), \tag{4}$$

where $S(O_1,O_2|\lambda_i)$ represents the combined likelihood that the observations O_1 and O_2 were produced by the subject class λ_i; and α_m is the weight of the m^{th} expert, subject to the constraints that $\sum\alpha_m=1$ and $0 \le \alpha_m \le 1$ for $m=1...N_E$. Given that the weights α_m are variable, some sort of reliability measure must be devised, which takes the confidence associated with each expert into account, and is used to determine the α_m values.

Expert reliability parameters can be calculated at the *signal* or at the *score* level. Signal based reliability measures are generally acoustic based [32] which have the disadvantage of having no corresponding video reliability measure. Even if an observation signal is of high quality, the expert may still give a misclassification for two (non-exhaustive) reasons: 1) the correct subject class may be indistinguishable for the given expert, and may be consistently misclassified, 2) the model/template for the correct subject may be a poor representation. A signal based reliability measure cannot take these into account. The distribution of the set of expert confidence scores contains information not only about the integrity of the observation signal, but also the reliability of that experts' decision. Taking these points into account, it is better to calculate the reliability measure based on the expert scores.

If the highest ranked class receives a high score and all of the other classes receive relatively low scores, then the confidence level is high. Conversely, if all the classes receive similar scores, the confidence is low. Various metrics exist, which can be used to capture this confidence information. Examples include, score *entropy* [32], *dispersion* [32], *variance* [9], and *difference* [9]. For a test observation vector O_m, we

[1] The overall performance did not vary significantly for $50<M<100$.

have the set of M ranked normalized scores $\{S(O_m|\lambda_i)\}_{i=1...M}$. The difference, ξ, between the two highest ranked confidence scores is calculated as

$$\xi_m = S(O_m \mid \lambda_1) - S(O_m \mid \lambda_2), \tag{5}$$

where λ_1 and λ_2 are the subject classes achieving the highest and second highest ranks respectively, and m denotes the expert. This metric was employed for this study.

A mapping between the reliability estimates and the expert weightings is required. In [16], [32] a sigmoidal mapping was used to map the reliability estimates to the fusion weights. The sigmoidal parameters require training, which is difficult when the amount of audio-visual data is scare, and may be specific to the noise type. Another option is to form bins of evaluation reliability values and the corresponding α_m values (found by exhaustive search), effectively a lookup table, but again this requires extensive training. Considering the small amount of audio-visual training data generally available, it was decided to use a non-learned approach to map the reliability estimates to the α_m values. This was carried out as follows:

For each specific identification trial (user interaction), the system is presented with two expert observations, O_1 and O_2.

1. The two experts each generate a set of N match scores, $\{l(O_1|\lambda_n)\}$ and $\{l(O_2|\lambda_n)\}$, which are normalized to give the sets of M ranked scores $\{S(O_1|\lambda_i)\}_{i=1...M}$ and $\{S(O_2|\lambda_i)\}_{i=1...M}$.
2. The fusion parameter α_2 is varied from 0 to 1 in steps of 0.05. For each of these α_2 values, the expert score lists $\{S(O_1|\lambda_i)\}$ and $\{S(O_2|\lambda_i)\}$ are combined using Eqn. (4) (with $\alpha_1=1-\alpha_2$), to give the combined set of N scores $\{S_{12,n}\}=\{S(O_1,O_2|\lambda_n)\}_{n=1...N}$. We have N, not M, S_{12} scores here because the sets of M normalized scores arising from experts 1 and 2 will in general correspond to different sets of M subject classes; some of the N S_{12} scores will be zeroed valued.
3. The combined score set is subsequently normalized as before, to give $\{S(O_1,O_2|\lambda_i)\}_{i=1...M}$, and the combined score reliability estimate, denoted by ξ_{12}, is calculated, as in Eqn. (5). ξ_{12} can be thought of as a linear weighted combination of the individual expert reliabilities ξ_1 and ξ_2 because

$$\begin{aligned} \xi_{12} &= S(O_1, O_2 \mid \lambda_1) - S(O_1, O_2 \mid \lambda_2) \\ &= \alpha_1 \cdot S(O_1 \mid \lambda_1) + \alpha_2 \cdot S(O_2 \mid \lambda_1) - \alpha_1 \cdot S(O_1 \mid \lambda_2) - \alpha_2 \cdot S(O_2 \mid \lambda_2) \\ &= \alpha_1 \cdot \xi_1 + \alpha_2 \cdot \xi_2, \end{aligned} \tag{6}$$

   where λ_1 and λ_2 are the subject classes achieving the highest and second highest ranks respectively, as before. However, the ξ_{12} value is calculated using Eqn. (5) and not Eqn. (6) because the set of scores $\{S_{12,n}\}$ is normalized, and hence Eqn. (6) does not hold exactly.
4. We choose the α_2 value that maximizes ξ_{12} for the given test according to Eqn. (7), to give the fusion parameters α_{2opt} and $\alpha_{1opt} = 1-\alpha_{2opt}$. The maximum ξ_{12} value should correspond to the combined scores of highest confidence, i.e. maximizes the score separation between the highest ranked class and the other classes. Finally, we combine $\{S(O_1|\lambda_i)\}$ and $\{S(O_2|\lambda_i)\}$ as in Eqn. (4) (using α_{1opt} and α_{2opt}), to form the combined score list $\{S_{12,n}\}_{opt,\ n=1,...,N}$ which is used to make the final identification decision.

$$\alpha_{2_{opt}} = \arg \max_{\alpha_2 \in [0,1]} \{\xi_{12} | \alpha_2\}. \quad (7)$$

It should be noted that the above procedure is carried out for every identification trial, and thus the fusion weights are determined online and automatically in an unsupervised manner. Also, O_1/O_2 above can represent any of $m \in \{A, M, F\}$. For illustration, Fig. 1 gives four examples of the specific case of fusing the scores arising from audio and mouth observations. The four examples show that the weight selection procedure has the ability to adapt the weights to the reliability of each expert.

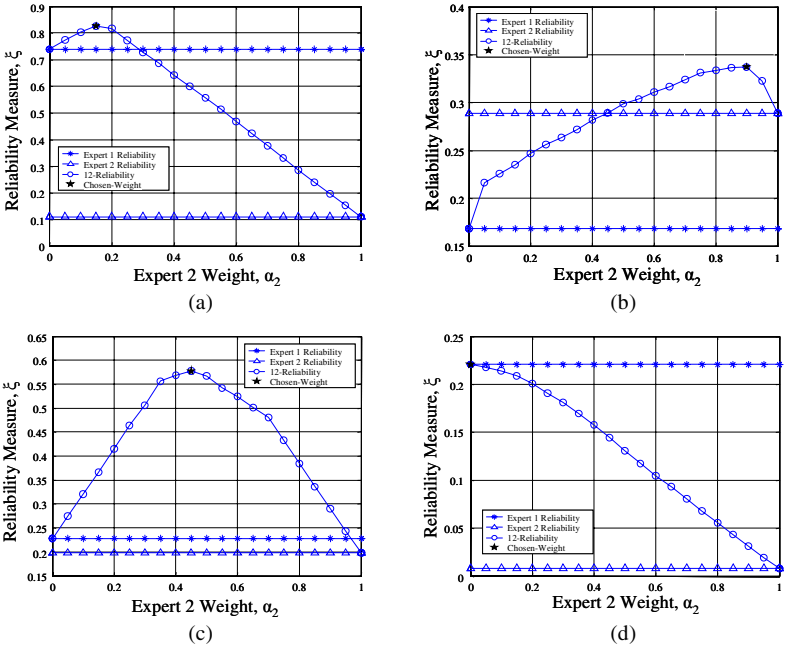

Fig. 1. The variation of the combined score reliability estimate w.r.t. α_2; and the individual expert reliability estimates are shown for four scenarios: (a) expert 1 is more reliable (selected α_{2opt} = 0.15), (b) expert 2 is more reliable (α_{2opt} = 0.9), (c) experts 1 and 2 have similar reliabilities (α_{2opt} = 0.5), and (d) expert 2 has a very low reliability (ξ_2 = 8x10-4), and α_{2opt} = 0

4.2 Fusion of the Three Experts

The bi-expert fusion method developed above can be employed to combine the output scores from any pair of person identification experts. In order to carry out tri-expert fusion of the audio, mouth, and face experts, a cascade approach is employed. Firstly, the two visual based experts (face and mouth) are combined, thus giving *N "face-mouth"* scores. This is shown in the first block of Fig. 2, where *"N Score Integration"* refers to the general bi-expert fusion block as described above. It is

intuitive to fuse the two visual experts initially, as a noisy visual observation signal is likely to affect both the face and mouth experts; in which case, the audio scores can still be weighted highly to counteract this. The *"face-mouth"* scores are subsequently fused with the *N* audio scores to give a tri-expert identification decision. We will now describe the fusion experiments that were carried out using the proposed method.

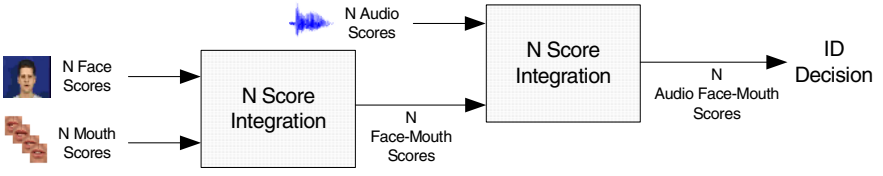

Fig. 2. Flow diagram for the fusion of all three experts

5 Audio-Visual Corpus

The XM2VTS audio-visual database [19] was used for the experiments, and consists of video data recorded from 295 subjects in four sessions, spaced monthly. The first recording per session of the phonetically balanced third sentence (*"Joe took father's green shoe bench out"*) was used. Some sentence recordings were clipped. Due to this and other errors, only 248 subjects were used for the experiments. The position of the mouth ROI was determined by manually labeling the left and right labial corners and taking the center point. Frames were manually labeled for every 10^{th} frame only; the ROI positions for the other frames were interpolated.

To test the robustness of the proposed system, both the audio and video (face sequence) test signals were degraded to provide a train/test mismatch. Ten levels of audio and visual degradation were applied; emulating mild to adverse train/test mismatch noise levels, which may be encountered in a realistic operating environment. Additive white Gaussian noise was applied to the clean audio at SNR levels ranging from 48 dB to 21 dB in 3 dB decrements. In [14], an image transform based approach was used to carry out visual word recognition. The system demonstrated robustness to JPEG compression, with no significant drop in performance until JPEG quality factors (QF) levels fell below 10. For our study, in order to account for practical video conditions, the video frame images were compressed using JPEG compression. We tested ten levels of JPEG QF, i.e.

$$QF \in \{50, 25, 18, 14, 10, 8, 6, 4, 3, 2\}, \tag{8}$$

where a QF of 100 represents the original uncompressed image. The compression was applied to each video frame individually. The mouth ROI was then extracted from the compressed images. Manually labeled mouth coordinates were employed, so that any drop in performance would be due to mismatched testing rather than poorer mouth tracking. The variation of the face and corresponding mouth ROI images w.r.t. JPEG QF is shown in Fig. 3. JPEG blocking artifacts are evident at the lower QF levels.

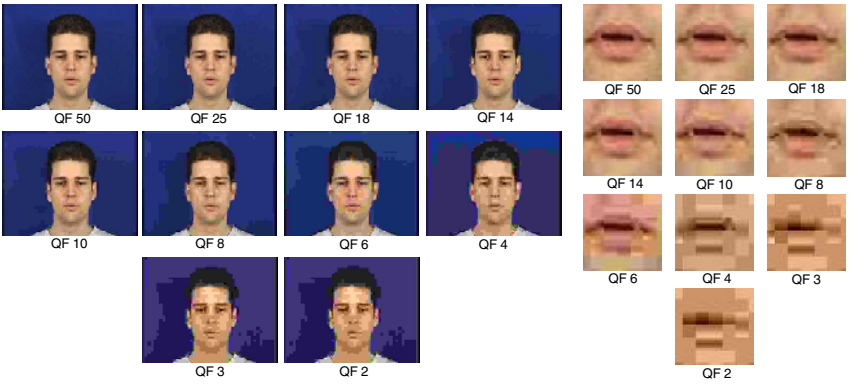

Fig. 3. Ten levels of JPEG compression and corresponding mouth ROI images

6 Experiments and Results

The proposed tri-expert system was applied to closed-set person identification. It can also be applied to the more general problem of open-set person recognition.

Audio Expert: The HMMs were trained/tested using HTK [12]. The first three sessions were used for training and the last for testing. A prototype HMM consists of the initial parameters. Since there are only three training utterances per subject, there was insufficient training data to train a speaker HMM directly from a prototype model. For this reason, a background HMM was trained using three of the sessions for all N subjects, and was used to initialize the training of the speaker models. All models were trained on the clean speech and tested on the various SNR levels. This provides for an audio train/test mismatch. The num+ber of HMM states that maximized the audio accuracy was found empirically to be eleven (with a mix of two Gaussians per state). Fig. 4 shows how the audio expert performs w.r.t. audio degradation. A maximum accuracy of 97.6% was achieved at 48dB, with dropped to 37.1% at 21 dB.

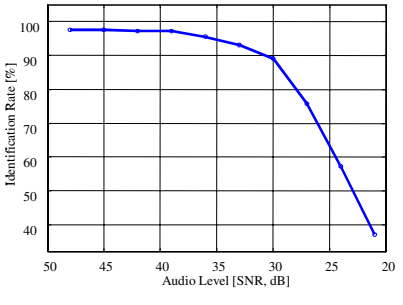

Fig. 4a. Audio expert performance versus audio degradation level

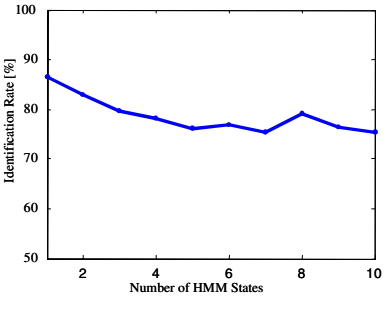

Fig. 4b. Mouth expert performance versus number of HMM states

Mouth Expert: The effect of the number of HMM states on the performance of the mouth expert was initially tested. One Gaussian per state was used. The result of

this is shown in Fig. 4b. The mouth expert performed best with just one state and decreased steadily with increasing number of states. For the visual degradation experiments the mouth expert HMMs were trained on the "clean" (uncompressed) visual images and tested on the degraded visual images. This provided for a visual train/test mismatch. The results are given in Table 1 and Fig. 5c.

Face Expert: The face gallery set, comprising of three images, was formed by arbitrarily extracting the 9th image frame from the first three sessions. The probe images used for testing were obtained from the fourth session (again, the 9th frame). The gallery sets consisted of the original uncompressed images and the probe sets consisted of degraded images at the ten levels of JPEG compression. This provided for a gallery/probe mismatch. The results are given in Table 1 and Fig. 5c.

Fusion Experiments: Four fusion experiments were carried out using the proposed fusion method: 1) the face and mouth experts, 2) the audio and mouth experts, 3) the audio and face experts, and 4) the audio, face, and mouth experts. The face, mouth, and *face-mouth* fusion performance w.r.t. JPEG QF mismatch is given in Table 1 and Fig. 5c. For the three audio-visual fusion experiments, ten levels of both visual (JPEG QF) and audio (dB) degradation were examined. The results for these experiments are given in Fig. 5 and Table 2, with the *audio-mouth* results in Fig. 5a, the *audio-face* results in Fig. 5b, and the *audio-face-mouth,* results in Fig. 5d.

Table 1. The mouth, face, and face-mouth fusion accuracies for the ten levels of JPEG QF

JPEG QF	50	25	18	14	10	8	6	4	3	2
Mouth [%]	85.9	85.1	84.3	84.3	82.7	80.2	79.4	60.5	50.8	48.0
FaceIt [%]	98.8	98.8	99.6	99.6	98.8	98.8	98.0	91.9	85.9	75.0
Mouth-FaceIt [%]	100.0	99.2	100.0	100.0	100.0	100.0	100.0	98.4	92.7	87.5

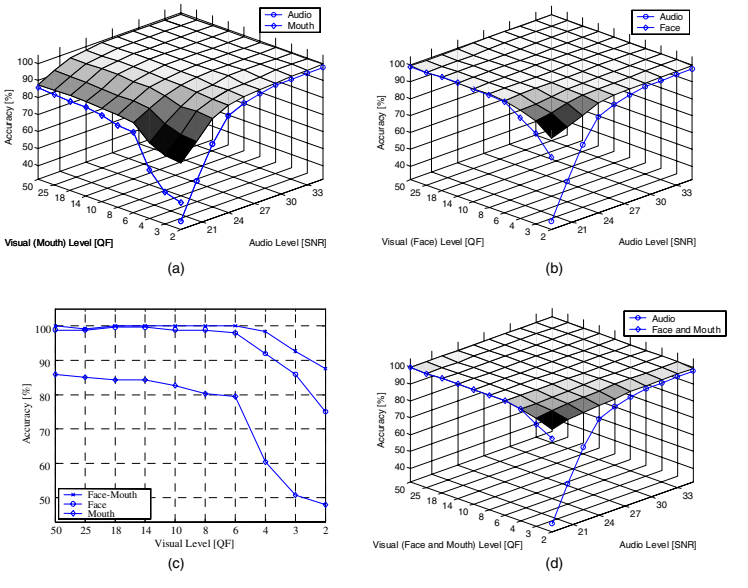

Fig. 5. The accuracies for the fusion of: (a) the audio and mouth experts, (b) the audio and face experts, (c) the face and mouth experts, and (d) the audio, face, and mouth experts

Table 2. The accuracies for the mouth (M), face (F), audio (A), and the fusion of: (a) the face and mouth experts (FM), (b) the audio and mouth experts (AM), (c) the audio and face experts (AF), and (d) the audio, face, and mouth experts (AFM)

QF			dB	33	30	27	24	21
			A	93.1	89.1	75.8	57.3	37.1
8	M	80.2	AM	98.8	97.6	96.0	91.1	86.3
	F	98.8	AF	99.6	99.6	99.6	99.6	99.6
	FM	100.0	AFM	100.0	100.0	100.0	100.0	100.0
6	M	79.4	AM	98.8	97.2	94.4	89.9	85.1
	F	98.0	AF	99.6	99.6	99.6	99.2	98.8
	FM	100.0	AFM	100.0	100.0	100.0	100.0	100.0
4	M	60.5	AM	97.6	96.4	91.1	84.7	76.6
	F	91.9	AF	99.2	99.2	98.0	96.8	96.4
	FM	98.4	AFM	99.6	99.2	98.8	98.4	98.4
3	M	50.8	AM	97.6	95.6	91.9	81.5	72.6
	F	85.9	AF	99.2	98.8	97.2	94.8	93.1
	FM	92.7	AFM	99.2	98.8	98.4	97.6	96.4
2	M	48.0	AM	97.2	95.2	91.1	80.6	71.4
	F	75.0	AF	98.8	97.2	93.1	89.9	86.3
	FM	87.5	AFM	97.6	97.2	96.4	95.6	92.7

7 Discussion

With regard to the specific experiments, the audio expert performed very well under near "clean" testing conditions, however the accuracy roll off w.r.t. SNR is very high. For the mouth expert experiments, the fact that the static visual features performed best with just one state indicates that HMMs may not be required to model visual speech, rather, a Gaussian mixture model (GMM) approach [11] would be sufficient. The best mouth expert accuracy is 85.9%. A reasonable level of robustness to video degradation is exhibited; with an accuracy of 48.2% at a QF of 2.

It was expected that FaceIt, a commercial system, employing features located throughout the entire face would outperform an expert employing features extracted from just the mouth ROI. The face expert outperformed the mouth expert at all levels of train/test mismatch. The highest face expert accuracy of 98.8% is 15% higher (relative) than the highest mouth expert accuracy of 85.9%. The face expert also exhibits higher robustness to JPEG compression, when compared to the mouth expert, with accuracies exceeding 98%, for all test mismatch levels exceeding a QF of 4. At the highest mismatch QF level of 2, the face expert accuracy was 75%, and the mouth expert accuracy was 48%. The superior performance of FaceIt is more impressive when considering that the FaceIt gallery consists of only three images, whereas the mouth expert model has the advantage of *"seeing"* three sequences of video frames and hence more variation in the subjects' appearance. The robustness of the face expert against JPEG compression is in line with results from the Face Recognition Vendor Test 2000 [1], where similar observations were made.

For the fusion of the face and mouth experts, a perfect *face-mouth* accuracy of 100% is achieved at several levels of JPEG QF mismatch. Also, the *face-mouth* accuracies are higher than either of the face or mouth expert accuracies for all levels of JPEG QF mismatch, i.e. enhancing fusion. The most significant improvements are yielded for the higher levels of mismatch, for example at the lowest QF level of 2, the *face-mouth*, face, and mouth accuracies are, 87.5%, 75%, and 48% respectively, representing a 17% relative improvement over the face expert alone. The improved *face-mouth* performance indicates that the mouth features complement the facial

features that the FaceIt engine employs. The improvement may be due to two factors. a) the face expert emphasizes eye information and hence the mouth expert is complementary, b) the fact that the mouth expert can capture the variation of the mouth ROI over the training video frame sequences.

The *audio-mouth* accuracies represent an improvement over the individual audio and mouth expert accuracies at all tested levels of audio and visual train/test mismatch. At the (21dB, 2QF) operating point, the audio, mouth, and *audio-mouth* accuracies are 37.1%, 48%, and 71.4% respectively, representing a relative improvement of 49% over the mouth expert. The *audio-face* results also show an improvement over the individual experts. At the (21dB, 2QF) operating point, the audio, face, and *audio-face* accuracies are 37.1%, 75%, and 86.3% respectively, representing a 21% relative improvement over the *audio-mouth* accuracy.

For the tri-expert experiments, perfect audio-face-mouth 100% accuracies were achieved at the majority of operating points. From Fig. 5 it is evident that the tri-expert performance exceeds the performance of either the audio-mouth or audio-face fusion. The improvements in performance were most significant, at the highest levels of train/test mismatch. At 21dB, the audio accuracy is 37.1% and at a JPEG QF of 2, the face and mouth accuracies are 75% and 48% respectively. At the (21dB, 2QF) operating point, the audio-mouth, audio-face, and audio-face-mouth accuracies are 71.4%, 86.3%, and 92.7% respectively. Improvements over the face-mouth accuracies were also achieved, particularly at the (21dB, 2QF) operating point, where an accuracy of 92.7% outperforms the face-mouth accuracy of 87.5% at a QF of 2. This highlights the increased robustness of the tri-expert fusion over bi-expert fusion and exemplifies the robustness of our tri-expert fusion method to both audio and visual degradation. Importantly, integrating a highly mismatched scenario (e.g. audio 37.1% at 21dB) with a "clean" test (e.g. face 75%, mouth 48% at QF2) does not result in catastrophic fusion (audio-face-mouth 92.7%). These results were achieved with the tri-expert fusion block having no prior knowledge of the level or type of audio or visual degradation. The fusion method is not computationally expensive as only $1+1/0.05 = 21$ fusion-parameter steps are carried out to determine the best fusion weight and also, the reliability measure is computed with a basic subtraction.

Further work includes testing the performance of the fusion system using different types of audio and visual degradations, and examining other reliability measures.

8 Conclusion

A multiple expert biometric person identification system has been presented, which combines information from three experts, namely: face, audio, and visual speech information in an automatic unsupervised fusion, adapting to the local performance of each expert, and taking into account the output-score based reliability estimates of each of the experts. Previous tri-expert (face, mouth, and audio) fusion studies employ un-weighted fusion or else fixed weights; expert reliability information is not considered. A benefit of the approach described is that audio-visual training data is not required to tune the fusion process. Importantly, no assumption has been made about the type of audio or visual noise that may cause an expert to perform poorly. The results show improved fusion accuracies for the gamut of tested levels of audio

and visual degradation, compared to the individual expert accuracies. This highlights the complementary nature of the mouth and face experts under clean and noisy test conditions, and in turn, the complementary nature of audio and video based information. The deployment tri-expert information should be robust to varying facial expressions, which may deform the eye or mouth region. These results are important for remote authentication applications, where bandwidth is limited and uncontrolled acoustic noise is probable, such as, video telephony and online authentication.

Acknowledgements

This work was supported by Enterprise Ireland's Informatics Advanced Research Technology Programme UCD-R8778 and in part by the U.S. Office of Naval Research contract N00014-00-1-0915.

References

[1] D. Blackburn , M. Bone, and P. J. Phillips, "Facial Recognition Vendor Test 2000," Evaluation report 2000.

[2] R. Gross, J. Shi, and J. F. Cohn, "Quo Vadis Face Recognition," in *Third Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

[3] N. A. Fox and R. B. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features," in *Proc. of the fourth Int'l Conf. on Audio- and Video-Based Biometric Person Authentication*, Guildford, UK, pp. 743-751, 2003.

[4] www.identix.com, "Identix Corp.," 5600 Rowland Road, Minnetonka, MN 55343.

[5] U. Dieckmann, P. Plankensteiner, and T. Wagner, "SESAM: A biometric person identification system using sensor fusion," *Pattern Recognition Letters*, vol. 18, pp. 827-833, 1997.

[6] Y. Yemez, A. Kanak, E. Erzin, and A. M. Tekalp, "Multimodal Speaker Identification with Audio-video Processing," in *Proc. of the Int'l Conf. on Image Processing,*, vol. 3, pp. 5-8, 2003.

[7] R. W. Frischholz and U. Dieckmann, "BiolD: a multimodal biometric identification system," *Computer*, vol. 33, pp. 64-68, 2000.

[8] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, pp. 449-480, 2004.

[9] T. Wark and S. Sridharan, "Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification," *Digital Signal Processing*, vol. 11, pp. 169-186, 2001.

[10] N. A. Fox and R. B. Reilly, "Robust Multi-modal Person Identification with Tolerance of Facial Expression," in *in the Proc. of the IEEE Int'l Conf. on Systems, Man and Cybernetics,*, The Hague, The Netherlands, vol. 1, pp. 580-585, 2004.

[11] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Tran. on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[12] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.1)*. Cambridge University Engineering Department: Microsoft Corporation, 2001.

[13] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: Applied to text dependent speaker recognition," *to appear in the IEEE Transactions on Multimedia*, vol. 7, 2005.

[14] G. Potamianos, H. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," in *Proc. of the IEEE Int'l Conf. Image Processing*, Chicago, vol. 3, pp. 173-177, 1998.
[15] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A Comparison of Model and Transform-based Visual Features for Audio-Visual LVCSR," in *proc. of the IEEE Int'l Conf. on Multimedia and Expo*, pp. 825-828, 2001.
[16] N. A. Fox, R. Gross, P. de Chazal, J. F. Cohn, and R. B. Reilly, "Person Identification Using Automatic Integration of Speech, Lip, and Face Experts," in *ACM SIGMM workshop on Biometrics Methods and Applications*, Berkley, CA., pp. 25-32, 2003.
[17] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "Audio-Visual Speaker Identification via Automatic Fusion using Reliability Estimates of both Modalities," in *to appear in the Proc. of the 5th Int'l Conf. on Audio- and Video-Based Biometric Person Authentication*, Rye Brook, NY, 2005.
[18] V. Vapnik, *The nature of statistical learning theory*: Springer Verlag, 1995.
[19] K. Messer, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: "The Extended M2VTS Database"," in *The Proc. of the Second Int'l Conf. on Audio and Video-based Biometric Person Authentication*, Washington D.C., pp. 72-77, 1999.
[20] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 711-720, 1997.
[21] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A. 4*, pp. 519-524, 1987.
[22] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.
[23] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. of the Fourth IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pp. 300-305, 2000.
[24] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *Neural Networks, IEEE Tran. on*, vol. 8, pp. 98-113, 1997.
[25] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 743-756, 1997.
[26] A. Yuille, "Deformable Templates for Face Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 59-70, 1991.
[27] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 775-779, 1997.
[28] P. Penev and J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Computation in Neural Systems*, vol. 7, pp. 477-500, 1996.
[29] P. J. Phillips, P. Grother, P. Michaels, D. Blackburn , E. Tabassi, and M. Bone, "Face Recognition Vendor Test 2002," Evaluation report 2002.
[30] [30]J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226-239, 1998.
[31] A. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Biometric Systems," *to appear in Pattern Recognition*, 2005.
[32] M. Heckmann, F. Berthommier, and K. Kristian, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1260-1273, 2002.