Robust Biometric Person Identification Using Automatic Classifier Fusion of Speech, Mouth, and Face Experts

Niall A. Fox, *Member, IEEE*, Ralph Gross, *Member, IEEE*, Jeffrey F. Cohn, *Member, IEEE*, and Richard B. Reilly, *Senior Member, IEEE*

***Abstract*—Information about person identity is multimodal. Yet, most person-recognition systems limit themselves to only a single modality, such as 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: audio, visual speech, and face. The system uses multimodal fusion in an automatic unsupervised manner, adapting to the local performance (at the transaction level) and output reliability of each of the three 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 acoustic babble noise and JPEG compression to degrade the audio and visual signals, respectively. Identification experiments were carried out on a 248-subject subset of the XM2VTS database. The multimodal expert system outperformed 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 16.1%, 48%, 75%, and 89.9%, respectively, representing a relative improvement of 19.9% over the best performing expert.**

***Index Terms*—Biometric fusion, expert reliability, hidden Markov models, image information loss, mouth features, multimodal, person recognition, robustness, tri-expert.**

I. Introduction

BIOMETRICS is a field of technology devoted to verification or identification of individuals using physiological or behavioural traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas identification, a multiclass problem, involves identifying a user from a set of enrolled subjects; and becomes more difficult as the number of enrollees increases. This paper deals with the task of closed-set person identification, thus, a reject option is also required to perform open-set identification. In audio-video processing, the video modality lends itself to two experts,[1] the face expert and the visual-speech expert. For this paper, the visual-speech expert is defined as a classifier utilizing a sequence of mouth images extracted from a video utterance, and will simply be referred to as the mouth expert. Person recognition systems based on the audio modality achieve high performance when the audio signal-to-noise ratio (SNR) is high. Yet, the performance degrades quickly as the test SNR decreases (referred to as a train/test mismatch), as shown in [1] and elsewhere [2], [3]. Similarly, face-based identification is susceptible to pose/illumination variation, occlusion, and also poor image quality [4]–[6]. Visual-speech-based person identification usually under-performs audio and face-based experts, and is generally treated as a supplement to audio and face-based recognition [1], [7].

To combat these limitations of unimodal audio and video-based experts, a multimodal fusion approach can be adopted, similarly to that for audio-visual speech recognition [8], [9], and both improve robustness and overall performance. The audio, face, and mouth modalities contain nonredundant, complementary information about person identity. In order to exploit this, 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[2] [10], [11]. The majority of studies are bimodal, employing either the audio and face modalities, [2], [12] or the audio and temporal mouth modalities (and ignoring face) [1], [3], [13], [14].

The benefits of audio-visual fusion for the purpose of speaker identification has been shown in [1], with the fusion method employing modality weightings found by supervised exhaustive search, such that the audio-visual accuracy was maximized.

Manuscript received May 10, 2005; revised December 23, 2005. This work was supported by the Informatics Research Initiative of Enterprise Ireland, by Enterprise Ireland's Informatics Advanced Research Technology Programme UCD-R8778, and in part by the U.S. Office of Naval Research under Contract N00014-00-1-0915. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhengyou Zhang.

N. A. Fox is with the MMSP Laboratory, School of Electrical, Electronic and Mechanical Engineering, University College Dublin, Belfield, Dublin 4, Ireland (e-mail: niall.fox@ee.ucd.ie).

R. Gross is with the Data Privacy Lab/SCS, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail rgross@cs.cmu.edu; ralph@ralphgross.com).

J. F. Cohn is with the University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: jeffcohn@pitt.edu).

R. B. Reilly is with the School of Electrical, Electronic and Mechanical Engineering, University College Dublin, Belfield, Dublin 4, Ireland (e-mail: richard.reilly@ucd.ie).

Color version of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2007.893339

[1]The term *expert* refers to a particular person identification classifier, which gives an opinion for each class, of the likelihood that class produced the test observation (see Section III).

[2]The term *speaker recognition* is used when recognition is based on speech experts, i.e., the audio and visual-speech experts. The more general term *person recognition* is used when nonspeech-based experts are considered, e.g., a face expert.

This highlights the potential of audio-visual fusion, yet it is not practical in a real-world scenario. An automated audio-visual speaker identification fusion approach was presented in [15]; however, the issue of an audio train/test mismatch was not addressed. In [3], audio-visual speaker verification experiments were carried out on 36 subjects; however, only experiments on an audio train/test mismatch were carried out. A visual train/test mismatch was not considered. In [16], robust audio-visual classifier fusion under both audio and visual train/test mismatch conditions was described. The adaptive fusion results were encouraging, with improved audio-visual accuracies over either modality alone. User-specific weighting schemes have also been investigated [17]; which are more suitable in the verification scenario. The audio, mouth, and face experts were combined in [10] and [11]; yet neither study employed expert weights that adapt automatically to local test conditions.

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 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. A cascade approach is adopted, where the video information (mouth and face) is first combined, and subsequently combined with the acoustic information.

This paper is organized as follows. Section II describes how person identification based on audio, mouth features, and face was performed. Section III investigates classifier fusion methods; specifically dealing with audio-visual-based fusion, and develops the proposed fusion strategy. In Section IV, the audio-visual corpus employed and its augmentation for the specific experiments is described. In Section V, we present results of extensive evaluations examining individual expert performance and fusion performance. The results are discussed in Section VI, and finally in Section VII, conclusions are drawn.

II. Audio- and Video-Based Identification

In this paper, the audio-video signal is divided into three separate modalities, namely, acoustic speech, visual speech (or mouth sequence), and face. We use the notation O_A, O_M, and O_F to respectively denote the observations arising from the three aforementioned modalities, where O_A and O_M take the form of a temporal sequence of features, whereas, O_F is a single still image.

A. Audio-Based Identification

Audio-based speaker identification is a mature topic, [18]. Standard acoustic methods were employed. The audio signal was divided into frames using a Hamming window of length 20 ms, with overlap of 10 ms to give an audio frame rate of 100 Hz. Mel-frequency cepstral coefficients (MFCCs) of dimension 16 were extracted from each frame [19]. The energy [19] of each frame was also calculated and used as a 17th 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 the available hidden Markov model (HMM) toolkit (HTK) functions [19] employing a W_D value of five frames. Cepstral mean normalization [19] was performed on the audio feature vectors (to each audio utterance).

A text-dependent speaker identification methodology was tested. For text-dependent modeling [20], the same utterance is spoken by the subject for both training and testing. It was employed, as opposed to text-independent modeling [18], due to its suitability to the database used in this study (see Section IV). The N subject classes are represented by N speaker HMMs [21] 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 of speech feature vectors or observations, $O_A = \{o_1, o_2, \ldots, o_t, \ldots o_{T_A}\}$, where o_t is the speech observation (frame) at time t and T_A denotes the number of observation vectors in the sentence. We obtain N class-conditional joint probabilities, $p(O_A|\lambda_n)$, that the observation sequence O_A was produced by the class (speaker model) λ_n. Assuming equal *prior* probabilities, then $p(O_A|\lambda_n)$ is referred to as the *likelihood* that O_A was caused by λ_n. For HMM classifiers, the scores are in *log-likelihood* form: $l(O_A|\lambda_n)$. The classification task (speaker identification) is to find the class with the maximum log-likelihood, i.e.,

$$\arg\max_n \{l(O_A|\lambda_n)\}, \quad 1 \le n \le N. \tag{1}$$

B. Mouth Features Expert

In tandem with audio-visual-speech processing, visual-speech feature analysis has also received much attention recently [22], [23]. It has been consistently shown in several visual-speech studies, that pixel-based features outperform geometric features [23], [24]. Geometric features/lip-contours require significantly more sophisticated mouth-tracking techniques compared to just locating the mouth region of interest (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 the salient speech features. Many types of transforms are examined in the literature, including the *discrete cosine transform* (DCT) [7], [23], *discrete wavelet transform* (DWT) [23], and *principal component analysis* (PCA) [24]. The DCT is one of most commonly employed image transforms. It has good de-correlation and energy compaction properties [25] and has been found to outperform other transforms [24].

For the mouth expert employed in this study, features derived from pixels were used to represent the visual information based on the DCT. The visual mouth features were extracted from the mouth ROI, which consists of a 49×49 color pixel block (see Fig. 4). 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.

A popular method of extracting the most important transform coefficients, consists of applying a mask to the transform coefficient matrix [23]. Considering that most of the information of an image is contained in the lower DCT spatial frequencies [25], the first 15 DCT coefficients were selected, using a mask that selects the coefficients in a tri-angular fashion (upper-left region of the transform matrix). Only 14 of these features are used for modeling since the first feature was zero valued due to the mean removal.

The visual sentences were modeled using HMMs as in Section II-A. The *static* features consist of the 14 DCT coefficients. Delta features were also calculated. Second order frame derivatives or *acceleration* features were also calculated from the W_A adjacent *delta* feature frames, where W_A is the acceleration window size. These were calculated using HTK, employing both W_D and W_A values of five frames. The three types of visual features were also concatenated to form a 42 dimensional feature vector. We have T_V visual observations (generally $T_A \approx 4 * T_V$) and the sequence of visual-speech feature vectors is denoted by $O_M = \{o_1, o_2, \ldots, o_t, \ldots o_{T_V}\}$. Each mouth expert HMM gives the *log-likelihood* $l(O_M|\lambda_n)$, that the observation sequence O_M was produced by the n^{th} mouth expert model λ_n.

C. 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 Kernel Methods [26], [27], linear discriminant analysis (LDA) [28], principal component analysis (PCA) [29], [30] and neural networks [31] 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. Pose variation and facial occlusion are often difficult to handle in these approaches [32].

The geometry feature-based methods analyze explicit local facial features, and their geometric relationships. Lanitis *et al.* have presented an active shape model in [33] extending the approach by Yuille [34]. Wiskott *et al.* developed an elastic bunch graph matching algorithm for face recognition in [35]. HMM methods [36] and Gaussian mixture model (GMM) methods [37] have also been examined. Penev and Atick [38] 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. Once extracted, the feature images are then encoded using PCA to obtain a compact description. The local features are then matched independently and fused at the score level to a combined matching score which is reported by the algorithm [39].

FaceIt was among the top-performing systems in a number of independent evaluations [5], [6], [40]. It has been shown to be robust against variations in lighting, facial expression and lower face occlusion. FaceIt can handle pose variations of up to 35 degrees from frontal. However, performance drops significantly for larger pose changes and for occlusion of the eyes (dark sunglasses) [6]. This suggests that the FaceIt algorithm places a lot of emphasis on the eye region and provides motivation for combining the face and mouth experts. 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*, denoted by $l(O_F|\lambda_n)$, rather than a log-likelihood. For FaceIt, the set of N templates, λ_n, $n = 1 \ldots N$, receives maximum and minimum scores of ten and zero, respectively, i.e., $l(O_F|\lambda_n) \in [0, 10]$.

III. Classifier Fusion

The fusion of classifiers is a research topic that predates work on audio-visual speech fusion [41]. A nonexhaustive list of levels at which fusion can take place includes: the *signal*, *feature*, *model*, *score*, and *decision* level. A *mapping* (*classifier* or *expert*), "transforms" the *feature-space* to give a hypothesis that the observation belongs to a specific class. The mapping output can take two forms: 1) a *decision* or 2) a *confidence/score*. A *classifier* outputs a decision (class label) with no associated confidence, whereas an *expert* outputs a confidence score. Confidence scores can take many forms, e.g., posterior probabilities, likelihood scores, and feature space similarity measures (e.g., the *Euclidean* distance).

At the higher levels of integration, it is easier to add additional experts to an existing system. Also, it becomes easier to drop or de-emphasize an existing expert that performs poorly for a particular classification test. Thus, higher-level integration strategies may be more robust if it is possible to account for the reliability of each expert, and this may compensate for the loss of information.

A. Levels of Fusion

The earliest level of fusion is signal fusion, followed by feature fusion. Both signal and feature fusion can be grouped into *early-integration* [20]. Feature fusion simply consists of concatenating the feature vectors into a larger dimensional feature vector, that has several disadvantages: 1) the "curse of dimensionality" [42]; 2) the difficulty to take the reliability of either modality into account (a corrupted modality can compromise the entire audio-visual feature vector and catastrophic fusion may occur; this has been demonstrated in [1]); and 3) inability to combine nonspeech-based modalities (e.g., a single face image). Also, the features from some experts may not be available due to proprietary issues. FaceIt, for instance, outputs confidence scores but no lower-level information. The next level available for fusion is at the mapping/modeling stage and is referred to as middle integration. Coupled HMMs in speech recognition are a common example of this approach. [8], [43], [44].

Multiple experts can be combined in the score-space and multiple classifiers combined in the decision-space. Both are grouped into *late-integration*. Integration can also take place at the post-classifier level, for example a secondary classifier employs the output scores from the primary classifiers as new features and performs a further classification [2]. Example decision combination rules include, the AND rule (all classifier decisions must agree), the OR rule (a decision is made if any classifier makes a decision), and the majority vote rule (a majority of the classifiers must agree). 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.

We will first consider the fusion of the expert output scores without the use of weights, and with the scores treated as probabilities [41]. Consider M experts operating in M feature spaces. An observation from the m^{th} feature space is denoted by O_m. Given N classes/models, λ_n, $1 \leq n \leq N$, the posterior probability that O_m was produced by the n^{th} class is $P(\lambda_n|O_m)$, which is formed using the probability density function $p(O_m|\lambda_n)$ and the class prior probability $P(\lambda_n)$. In order to attain a probability based on all M observations, we need to calculate $P(\lambda_n|O_1, O_2, \ldots, O_m, \ldots, O_M)$, where all M observations are considered simultaneously, implying that a joint probability density function must be determined, which generally is an intractable problem. It is more feasible to consider the output due to each observation O_m individually, and then combine them in some manner.

The *product rule* consists of multiplying the M posteriors together and is theoretically the statistically optimal method of classification. It is sensitive to expert errors; in the extreme case, if any single expert produces a close to zero posterior estimate for a specific class; the combined posterior for that class will be close to zero. The *sum rule* is defined as

$$P(\lambda_n|O_1, O_2, \ldots, O_m, \ldots, O_M) = \frac{1}{M} \sum_{m=1}^{M} P(\lambda_n|O_m) \quad (2)$$

and is also referred to as the mean rule. It 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 [41].

B. Existing Methods for AV Integration

Before the proposed method of fusion is described, we provide a brief review of existing methods that integrate audio with mouth, audio with face, and all three experts to carry out person recognition.

In [13], the audio- and visual-speech modalities were combined to perform person identification using a secondary classifier to determine the audio and visual weights. Robustness to visual degradation was not tested. Face and speech information was combined in [2], using secondary classifiers, yielding higher performance compared to the face and speech experts. Audio and visual speech was combined for person identification in [45] using audio and visual reliability measures. Face information was not considered.

The audio, visual speech, and face modalities were combined in [10] to perform person recognition. The fusion methods were basic, employing a choose two from three approach (agreement of any two experts), and an AND decision combination of all three experts. Due to fusion at the decision level, no individual expert reliability information could be considered. This system was again presented in [46]; this time the weighted sum rule was also employed. The weights could only be varied manually, and hence could not adapt automatically to changing testing conditions. The audio, visual speech, and face modalities were also combined in [11] to carry out verification of automatic person identification. The audio and mouth features were concatenated and jointly modeled using an audio-visual HMM. The audio-visual score was combined with the face expert score using weighted summation, thus giving an audio-visual-face score. The weights were global and set empirically, i.e., there was no adaptation to local variation of the signal reliability or expert confidence.

C. The Proposed Method

For the proposed method, the following design criteria were taken into account. Information from the audio, mouth, and face signals are to be combined to perform closed-set person identification. The fusion method should easily allow the addition of other experts. The system must be robust to mild through adverse test levels of both audio and visual (both face and mouth) noise. 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. The expert score weightings must be determined in an automatic unsupervised manner. Given these criteria, we chose score level late-integration based on the weighted sum rule. It should be noted here that when used to combine log-likelihoods, the sum rule is a variant of the product rule.

D. Score Normalization

Expert scores can take many forms such as posteriors, likelihoods, and distance measures. Nonnormalized scores cannot be integrated sensibly in their raw form, as it is impossible to fuse incomparable numerical scales. Example normalization methods include *min-max*, *Z-norm*, *decimal-scaling*, *Median-MAD*, and the $tanh$ transformation [47]. The min-max technique is the most basic form of score normalization, which shifts and scales the scores into the range [0,1]. The min-max norm is most suitable when the pre-normalized scores have known bounds (e.g., for the face expert employed $l(O_F|\lambda_n) \in ([0, 10])$, however, it can still be used otherwise but will be extremely sensitive to outlier scores. While being straightforward to implement, the min-max norm has been found to have comparable performance to more complicated normalization methods [47], hence, it was used for experiments reported here. The poor robustness of min-max score normalization to outlier scores can be circumvented (in the person identification scenario) by considering only the top K ranked class scores for normalization and integration. This omits the worst (outlier) expert scores.

A fusion strategy was first developed for fusing any two experts (e.g., audio with face). Then this general bi-expert fusion

strategy was cascaded to include an additional third expert. We use $l(O_m|\lambda_n)$ to denote the confidence score output from the m^{th} expert representing the general likelihood that the observation O_m was caused by the subject model/template λ_n, $n = 1 \ldots N$, where $m \in \{A, M, F\}$. Each expert provides a list of N scores $\{l(O_m|\lambda_n)\}_{n=1\ldots N}$ which are ranked into descending order and, by using the min-max rule (applied to only the top K ranked scores), are normalized to give $\{l'(O_m|\lambda_{n_k})\} \in [0, 1]$, where

$$l'(O_m|\lambda_{n_k}) = \begin{cases} \left(\frac{l(O_m|\lambda_{n_k}) - l_{\min}}{l_{\max} - l_{\min}}\right), & 1 \le k \le K \\ 0, & \mathrm{K} < k \le N \end{cases} \quad (3)$$

and λ_{n_k} denotes the k^{th} ranked subject model (based on the observation score). Using a high value for K may retain outlier scores, which could unfairly skew the distribution. A very low value, would result in loss of information, the limit being $K = 1$, where all confidence information has been lost. Tests showed that the system performance degraded for $K < 50$ and $K > 100$. A value for K of 75 was employed for this study,[3] although this value will depend on N. The set of *ranked* normalized scores is denoted by $\{l'(O_m|\lambda_{n_k})\}_{k=1\ldots N}$. The fusion module should take the local testing conditions into account and adapt the fusion parameters accordingly. Thus, we have the *weighted sum rule* (for the specific case of two experts; i.e., $M = 2$)

$$\begin{aligned} l(O_1, O_2|\lambda_n) &= \sum_{m=1}^{M} \alpha_m \cdot l'(O_m|\lambda_n) \\ &= \alpha_1 \cdot l'(O_1|\lambda_n) + \alpha_2 \cdot l'(O_2|\lambda_n) \end{aligned} \quad (4)$$

where $l(O_1, O_2|\lambda_n)$ represents the *nonnormalized* combined likelihood that the observations O_1 and O_2 were produced by the subject class λ_n; 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 $m = 1 \ldots M$. A reliability measure must be devised, which takes the confidence associated with each expert into account, and thus used to determine the α_m values.

E. Reliability Measures

Expert reliability parameters can be calculated at the signal or at the score level. Signal-based reliability measures are derived directly from signal observations prior to feature extraction. Audio examples include estimations of the *signal-to-noise* ratio [43] and the *degree of voicing* [48], and in [49], fingerprint image quality was employed. An audio only reliability measure is undesirable, as the integrity of the visual signal is not considered. Even if an observation signal is of high quality, the expert may still give a misclassification for two (nonexhaustive) reasons: 1) the correct subject class may be indistinguishable for the given expert, and may be consistently misclassified and 2) the model/template for the correct subject may be a poor representation. A signal-based reliability measure cannot take these scenarios 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* [48], *dispersion* [48], *variance* [3], *cross classifier coherence* [16], and *difference* [13]. For a test observation vector O_m, we have the set of ranked normalized scores $\{l'(O_m|\lambda_{n_k})\}$. We define ξ, the difference between the two highest ranked scores, normalized by the mean score

$$\xi_m = \frac{l'(O_m|\lambda_{n_1}) - l'(O_m|\lambda_{n_2})}{l'_{mean}} \quad (5)$$

where λ_{n_1} and λ_{n_2} are the subject classes achieving the highest and second highest ranks, respectively, m denotes the expert, and the mean is calculated over the first K ranked values of $\{l'(O_m|\lambda_{n_k})\}$. ξ was employed as the reliability measure for this study because it is computationally cheap and is not specific to any expert or noise type.

F. Reliability Mapping

A mapping between the reliability estimates and the expert weightings is required. In [7], [8], and [48] a sigmoidal mapping was used to map the reliability estimates to the fusion weights. The parameters of the sigmoid curve required training, which is difficult when the amount of audio-visual data is scarce. Also, these parameters may be specific to the noise type. In [43], an empirical regression was used to map the test SNR values to α_A. This is unsuitable here, as we need to consider the visual reliability also. Considering the small amount of audio-visual training data generally available, it was decided to use a nonlearned approach to map the reliability estimates to the α_m values. This was carried out as follows.

1) For each specific identification trial (user interaction), the system is presented with two expert observations, O_1 and O_2.
2) 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 ranked scores $\{l'(O_1|\lambda_n)\}$ and $\{l'(O_2|\lambda_n)\}$, and the reliability estimates ξ_1 and ξ_2 are calculated, using (5).
3) The fusion parameter α_2 is varied from 0 to 1 in steps of 0.05. For each α_2 value, the expert score lists $\{l'(O_1|\lambda_n)\}$ and $\{l'(O_2|\lambda_n)\}$ are combined using (4) (with $\alpha_1 = 1 - \alpha_2$), to give the set of N *nonnormalized* combined scores $\{l(O_1, O_2|\lambda_n)\}$.
4) The combined score set is subsequently normalized as before, to give $\{l'(O_1, O_2|\lambda_{n_k})\}$, and the combined score reliability estimate, denoted by ξ_{12}, is calculated, similarly to (5), using
$$\xi_{12} = \frac{l'(O_1, O_2|\lambda_{n_1}) - l'(O_1, O_2|\lambda_{n_2})}{l'_{mean}(O_1, O_2)}. \quad (6)$$
5) We choose the α_2 value that maximizes ξ_{12} for the given test according to (7), to give the fusion parameters $\alpha_{2\text{opt}}$ and $\alpha_{1\text{opt}} = 1 - \alpha_{2\text{opt}}$. The maximum ξ_{12} value should

[3]The overall performance did not vary significantly for $50 < K < 100$.

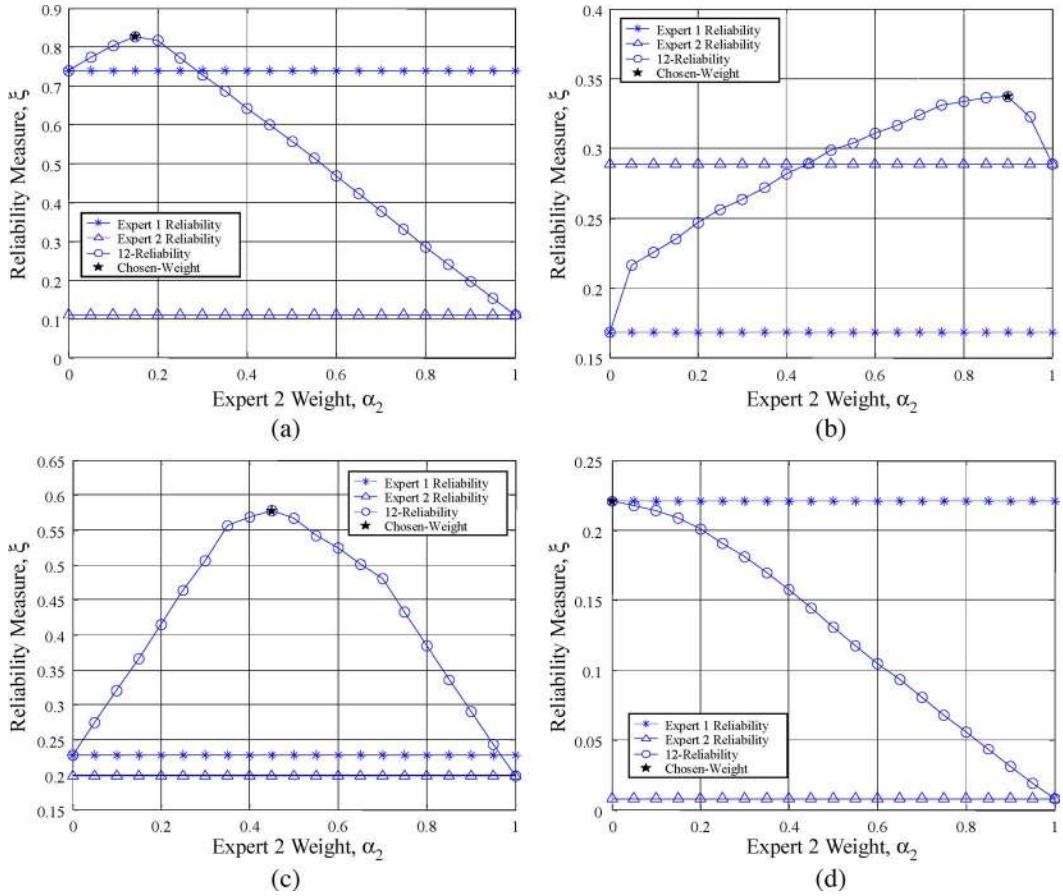

Fig. 1. Illustration of the weight selection procedure. The variation of the combined score reliability estimate w.r.t. α_2; and the individual expert reliability estimates are shown for four scenarios: (a) expert 1 is more reliable, (b) expert 2 is more reliable, (c) experts 1 and 2 have similar reliabilities, and (d) expert 2 has a very low reliability estimate, consequently $\alpha_{2\text{opt}} = 0$.

correspond to the combined scores of highest confidence, i.e., maximize the score separation between the highest ranked class and the other classes. Finally, we combine $\{l'(O_1|\lambda_n)\}$ and $\{l'(O_2|\lambda_n)\}$ using $\alpha_{1\text{opt}}$, $\alpha_{2\text{opt}}$, and (4), to form the combined score list $\{l(O_1, O_2|\lambda_n)\}_{opt}$ which is used to make the final identification decision.

$$\alpha_{2\text{opt}} = \arg \max_{\alpha_2 \in [0,1]} \{\xi_{12}|\alpha_2\}. \tag{7}$$

It should be noted that the above procedure is carried out for every identification transaction, and thus the fusion weights are determined online and automatically in an unsupervised manner. The weights can adapt to the local performance of each expert, i.e., the confidence of the score given by each expert for a given user transaction is considered. No assumptions are made here as to which experts are employed, i.e., O_1 and O_2 above can represent any of the face, mouth, or audio experts. To illustrate this procedure Fig. 1 gives four examples of the specific case of fusing the scores arising from audio and mouth test observations, i.e., $O_1 = O_A$ and $O_2 = O_M$. For Fig. 1(a), the audio and mouth score reliability estimate values, ξ_1 and ξ_2, are 0.74 and 0.11, respectively. The *audio-mouth* reliability estimate ξ_{12} reaches a maximum value at an α_2 value of 0.15. Thus, 0.15 and 0.85 (1–0.15) are chosen for α_2 and α_1, respectively, i.e., the scores of expert 1 are weighted more heavily. Similarly, for Fig. 1(b) the scores of expert 2 are weighted more heavily. In Fig. 1(c), due to the similarity of ξ_1 and ξ_2, the two experts receive approximately equal weightings. For Fig. 1(d), expert 2 has no contribution to the final decision. These four examples show that the weight selection procedure has the ability to adapt the weights to the reliability of each expert. Fig. 2 illustrates the fusion procedure described above, which is also for the specific case of fusing audio and mouth observations.

G. Fusion of the Three 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. 3, where "N *Score Integration*rdquo; refers to the general bi-expert fusion block as illustrated in Fig. 2. 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 be weighted highly to counteract this. The "*face-mouth*" scores are subsequently fused with the N audio scores to give the "*audio-face-mouth*" scores and a tri-expert

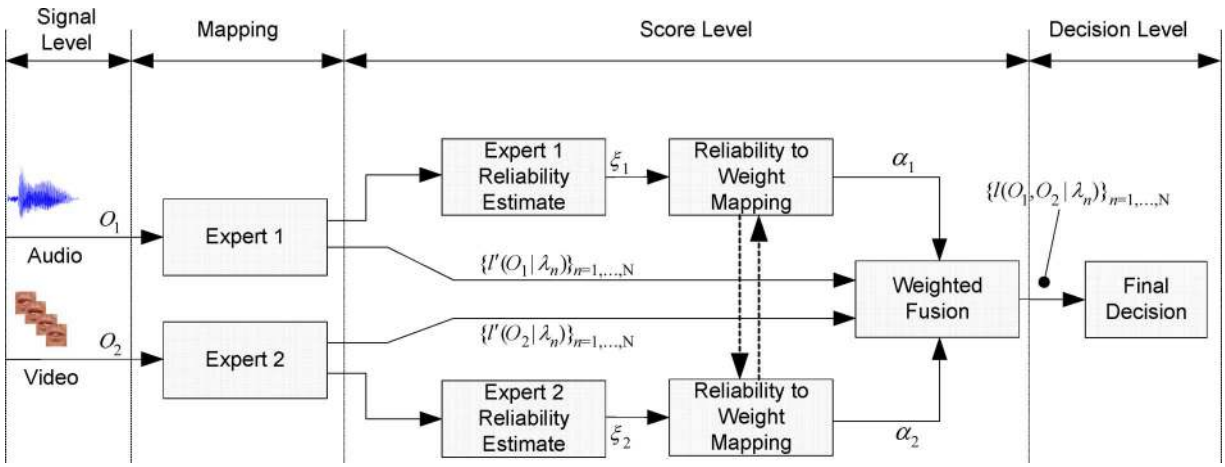

Fig. 2. Block diagram of the fusion strategy, for the specific case of audio and mouth observations.

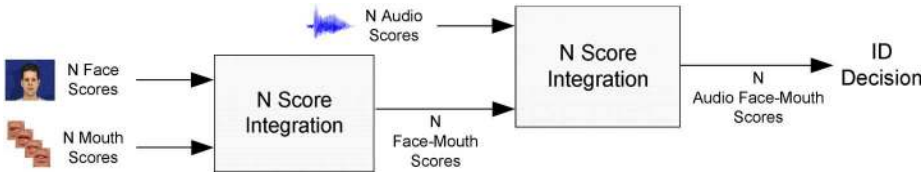

Fig. 3. Flow diagram for the fusion of all three experts.

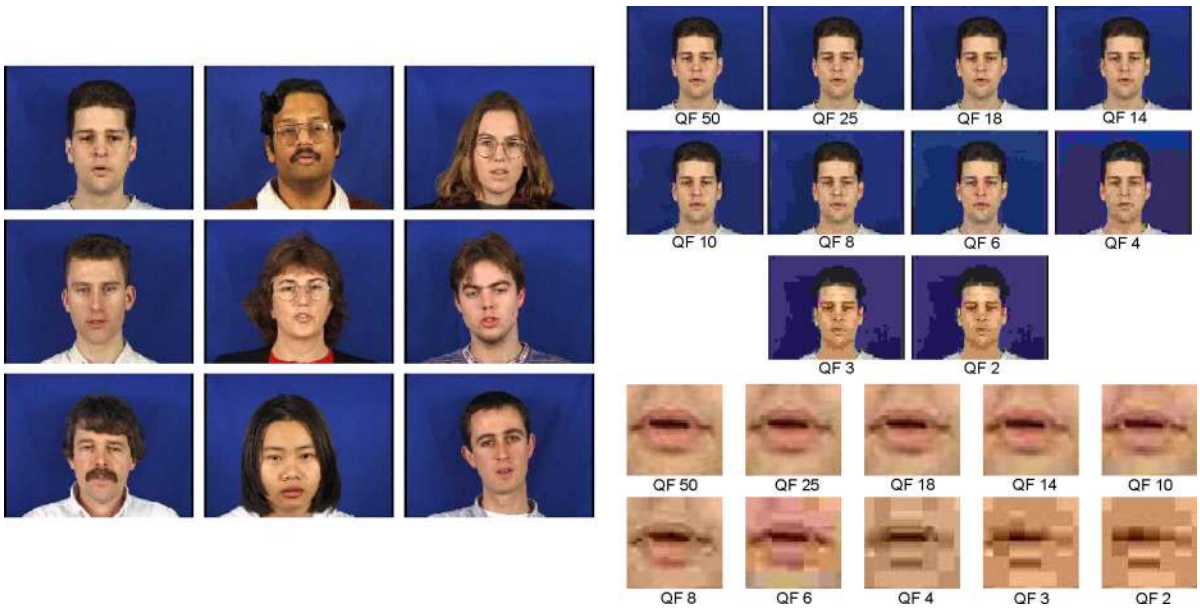

Fig. 4. Left: sample still images for nine subjects from the XM2VTS database. Right: ten levels of JPEG compression and corresponding mouth ROI images.

identification decision. We will now describe the fusion experiments that were carried out using the proposed method.

IV. Audio-Visual Corpus

A 248-subject subset of the XM2VTS audio–visual database [50] was used for the experiments described in this paper. The database consists of video data recorded from 295 subjects in four sessions, spaced monthly. The first recording per session of the phonetically balanced sentence ("*Joe took father's green shoe bench out*") was used. The original frame resolution of 720× 576 was downsampled to 360 × 288. Sample stills are shown in Fig. 4. The position of the mouth ROI was determined by manually labeling the left and right labial corners and taking the center point. This was carried out for every tenth frame only; the ROI positions for the other frames were interpolated.

In order to examine the robustness of the proposed system, both the audio and visual (face sequence) test signals were degraded to provide a train/test mismatch. Ten levels of audio and visual degradation were applied. Babble noise, taken from the NOISEX database [51], was applied to the clean audio data at SNR levels ranging from 24 to 6 dB in decrements of 2 dB. The video frame images were compressed

using JPEG compression, with ten levels of quality factor $(\mathrm{QF}) : QF \in \{50, 25, 18, 14, 10, 8, 6, 4, 3, 2\}$, where a QF of 100 represents the original image. The compression was applied to each video frame individually. The mouth ROI was then extracted from the compressed images. Mouth coordinates determined on the uncompressed images were employed for the compressed images, 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. 4. Blocking artefacts are evident at the lower QF levels.

V. Experiments and Results

The proposed tri-expert classifier fusion system was applied to the specific problem of closed-set person identification. It is not restricted to this application, and can also be applied to the more general problem of open-set person identification. For closed-set identification, there is prior knowledge that only enrolled subjects will access the system, i.e., the test subject will be identified as one of the enrollees.

A. Audio Expert Experiments and Results

The N HMMs were trained using the Baum Welch algorithm (with the maximum likelihood criterion) and tested using the Viterbi algorithm, both carried out using HTK [19], where $N = 248$ here. There was one background HMM. The first three sessions were used for training and the last session was used for testing. The background HMM was trained using three of the sessions for all N subjects, which was initialised using a prototype model; consisting of zero means, the unit matrix, equal mixtures, and left-to-right state transition probabilities of 0.5. This background model captures the audio speech variation over the entire database. The background model was used to initialise the training of the speaker models. All models were trained using the clean speech and tested using the various SNR levels. This provides for a mismatch between the audio testing and training conditions. The number of audio HMM states that maximized the audio accuracy was found empirically to be eleven, with a mix of two Gaussians per state. Fig. 5 and Table III show how the audio methodology performs w.r.t. the audio SNR. A maximum accuracy of 96.8% was achieved at 24 dB. At 6 dB, the accuracy dropped to 16.1%.

B. Mouth Expert Experiments and Results

The effect of the number of HMM states on the performance of the four visual feature types (*static (S)*, *delta (D)*, *acceleration (A)*, and *S-D-A*) was tested. These tests were carried using matched train/test data, i.e., the original images. For a HMM, each state is associated with a locally stationary section of the speech signal, whereas the state transitions model the temporal nature [21]. It would be expected that the *delta* and *acceleration* features would perform better using more states compared to the static visual features.

The number of states was increased from one, until a performance trend became apparent. In each case one Gaussian per state was used. The results of this are shown in Fig. 6. The number of states, that maximized the visual accuracies for each of the four feature types, are given in Table I. The *static* features performed best with just one state and decreased steadily with increasing number of states. The number of states, that maximized the accuracies for the *delta* and *acceleration* features, were 18 and 15, respectively. The concatenated *S-D-A* feature vector was modeled best using four states. For the visual experiments, the mouth expert HMMs were trained on the uncompressed images and tested on the degraded images. This provided for a visual train/test mismatch. The tests on the degraded mismatched visual data were carried out using the number of states, which maximized the accuracies, for each of the four visual feature types. Table I and Fig. 7 show how the visual features perform w.r.t. JPEG degradation.

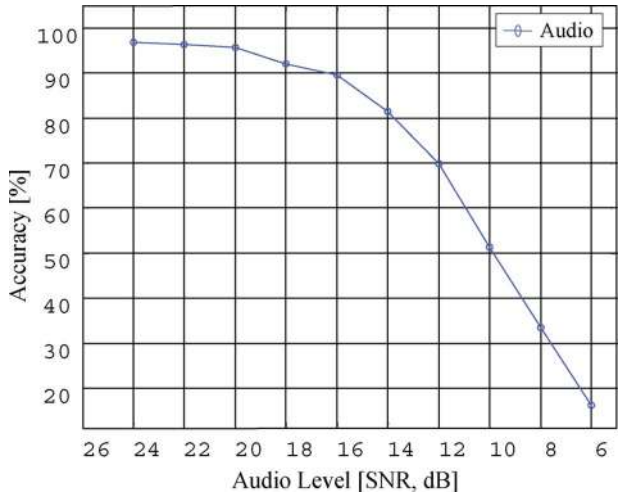

Fig. 5. Effects of audio degradation on person identification accuracy for "babble" noise.

C. Face Expert Experiments and Results

The face gallery set, comprising of three images, was formed by arbitrarily extracting the ninth image frame from each of the first three training sessions. These were used to form a face template for each of the N subjects. In all the face experiments, the probe images used for testing were obtained from the final (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 face (FaceIt) performance w.r.t. JPEG QF is given in Table II and Fig. 8(c).

D. Fusion Experiments and Results

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 (tri-expert fusion). By comparing the results of these four experiments, we can observe the performance gained with the addition of each expert. For comparison, the nonweighted sum rule was also employed for tri-expert fusion.

For the mouth expert, the *static* features exhibited the highest robustness to JPEG QF mismatch. Hence, the mouth expert using only the *static* features was employed for the fusion

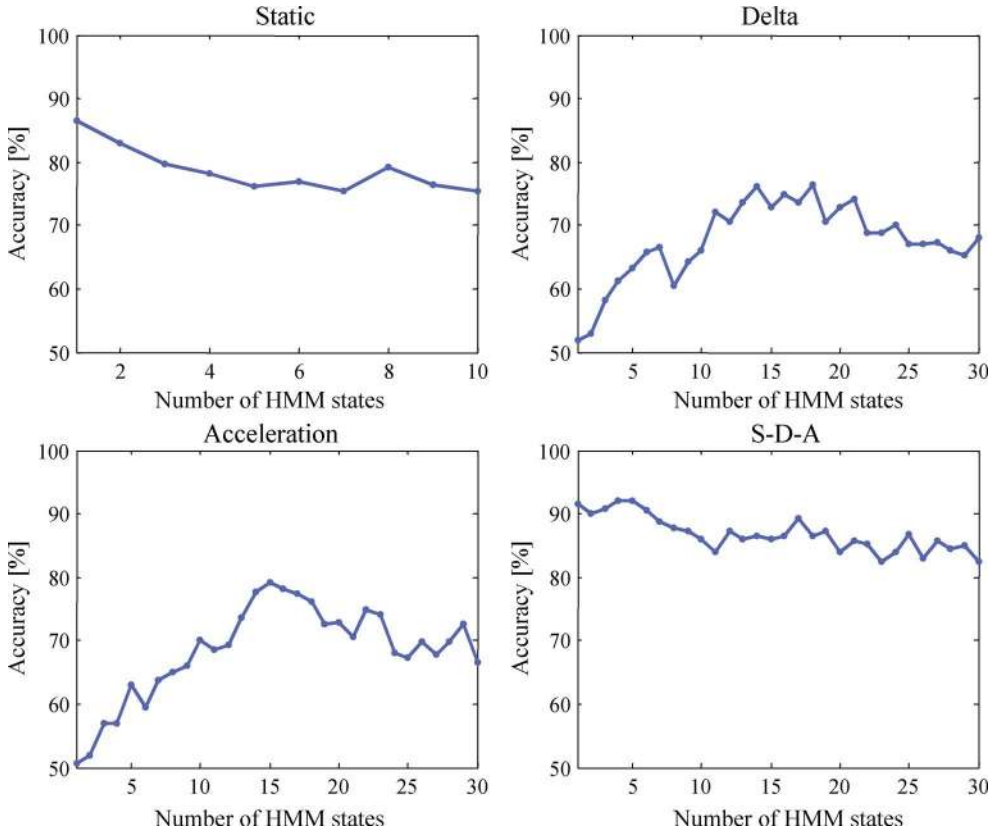

Fig. 6. Visual performance versus number of HMM states for each of the four types of mouth expert visual features, namely: static, delta, acceleration, and S-D-A.

TABLE I
NUMBER OF STATES THAT MAXIMIZED THE ACCURACY FOR EACH OF THE FOUR TYPES OF VISUAL FEATURES, AND THE PERFORMANCE ACROSS THE TEN LEVELS OF JPEG QF. ACCURACIES ARE PERCENTAGE CORRECT OF $N = 248$

Features	HMM States	Clean	QF 50	QF 25	QF 18	QF 14	QF 10	QF 8	QF 6	QF 4	QF 3	QF 2
Static	1	86.5	85.9	85.1	84.3	84.3	82.7	80.2	79.4	60.5	50.8	48.0
Delta	18	76.5	74.1	68.1	64.5	52.6	33.5	18.3	10.4	3.2	2.0	2.0
Accel.	15	79.3	77.3	70.1	67.7	56.2	38.2	20.7	13.1	3.2	1.6	2.8
S-D-A	4	92.0	90.4	89.2	88.0	87.6	83.3	75.7	67.3	42.6	35.9	34.7

experiments. The face, mouth, and *face-mouth* performance w.r.t. JPEG QF mismatch is given in Table II and Fig. 8(c).

The results for each of the three fusion experiments are presented in Fig. 8 and Table III, with the *audio-mouth* results in Fig. 8(a), the *audio-face* results in Fig. 8(b), and the *audio-face-mouth* (tri-expert), results in Fig. 8(d).

VI. 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, which can be seen in Fig. 5. 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 when using static features, rather, a simpler GMM approach [18] would be sufficient. Other person recognition studies based on the mouth ROI have ignored the temporal mouth information and modeled the statistical distribution of the mouth shape over the training utterances using GMMs [13], [52]. The higher number of HMM states required for the *delta* and *acceleration* features was expected due to the dynamic visual-speech information contained in these features. The fact that the *S-D-A* feature vector was modeled best using four states suggests a conflict between the static and dynamical features, with static features performing best with a single-state model whereas dynamical features perform better with a multistate model. Hence, a score-level integration-based approach of the S, D, and A scores, may yield higher *S-D-A* performance than the feature-concatenation fusion approach used here.

The best mouth expert performance of 92% is surprisingly high, considering that only mouth information was employed. While the *S-D-A* features outperform the *static* features for high QFs (*S-D-A* 90.4% versus *static* 85.7% at a QF of 50), the performance at a QF of 2 is 34.7%, which is poorer than the *static* performance of 48.2%. The dynamic features perform very poorly for low QF levels, both falling to around 2% at a QF of 2. It should be noted that in an applied scenario, where the ROI is automatically segmented, rather than manually,

TABLE II
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	98.0	93.1	87.5

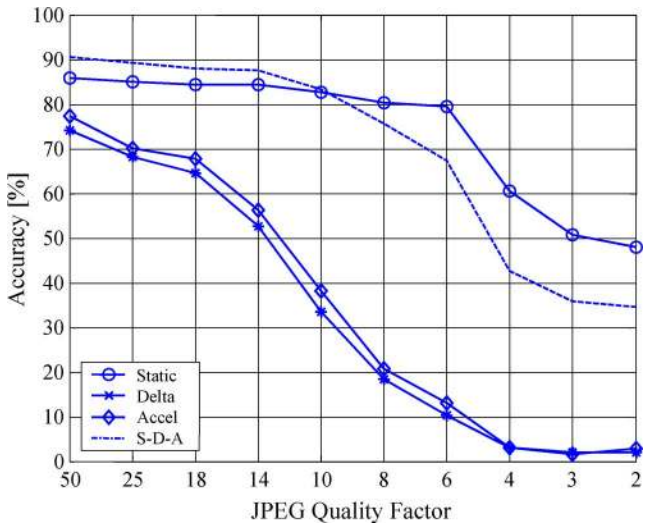

Fig. 7. Effects of visual degradation on the person identification accuracy for ten levels of JPEG QF and for each of the four types of mouth visual features examined.

poorer robustness to visual degradations would be expected. The results also show that the *static* features are more important and more robust than the *dynamic* features for person identification and also, that nontemporal GMM modeling may be more suitable than temporal HMM modeling.

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%, a relative difference of 56%. 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. Nonetheless, it is still interesting to examine if the combination of the FaceIt and mouth experts would yield any improvement in performance and robustness. The robustness of the face expert against JPEG compression is in line with results from the Face Recognition Vendor Test 2000 [5] 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: 1) the face expert emphasizes eye information and hence the mouth expert is complementary and 2) 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 the lower levels of audio and visual train/test mismatch, e.g., at the (16 dB, 8QF) operating point, the audio, mouth, and *audio-mouth* accuracies are 89.5%, 80.2%, and 98.4%, respectively. However, the fusion results are disappointing for the higher levels of mismatch, e.g., at the (6 dB, 8QF) operating point, the audio, mouth, and *audio-mouth* accuracies are 16.1%, 80.2%, and 65.7%, respectively. The *audio-face* results show an improvement over the individual experts at all levels of mismatch. At the (10 dB, 2QF) operating point, the audio, face, and *audio-face* accuracies are 51.2%, 75%, and 87.5%.

For the tri-expert experiments, perfect audio-face-mouth 100% accuracies were achieved at the majority of operating points. The tri-expert fusion attains a significant increase in robustness to both audio and visual degradations. This is evident from the flatness of the audio–visual surface in Fig. 8(d) compared to Fig. 8(a) and Fig. 8(b). From Fig. 8(d), 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 the (6 dB, 2QF) operating point, the audio, mouth, and face accuracies are 16.1%, 48%, 75%, respectively, whereas the audio-mouth, audio-face, and audio-face-mouth accuracies are 48.8%, 77.8%, and 89.9%, respectively. This exemplifies the robustness of our tri-expert fusion method to both audio and visual degradation. Importantly, integrating a highly mismatched scenario (e.g., audio 16.1% at 6 dB) with a "clean" test (e.g., face 75%, mouth 48% at QF2) does not result in catastrophic fusion (audio-face-mouth 89.9%), unlike for the audio-mouth only case and for the nonweighted sum rule (73%). These results were achieved with the tri-expert fusion block having no prior knowledge of the level or type of audio or visual degradation. Hence, we have a generalized fusion methodology, which should not be adversely affected by varying types of audio/video mismatch noise.

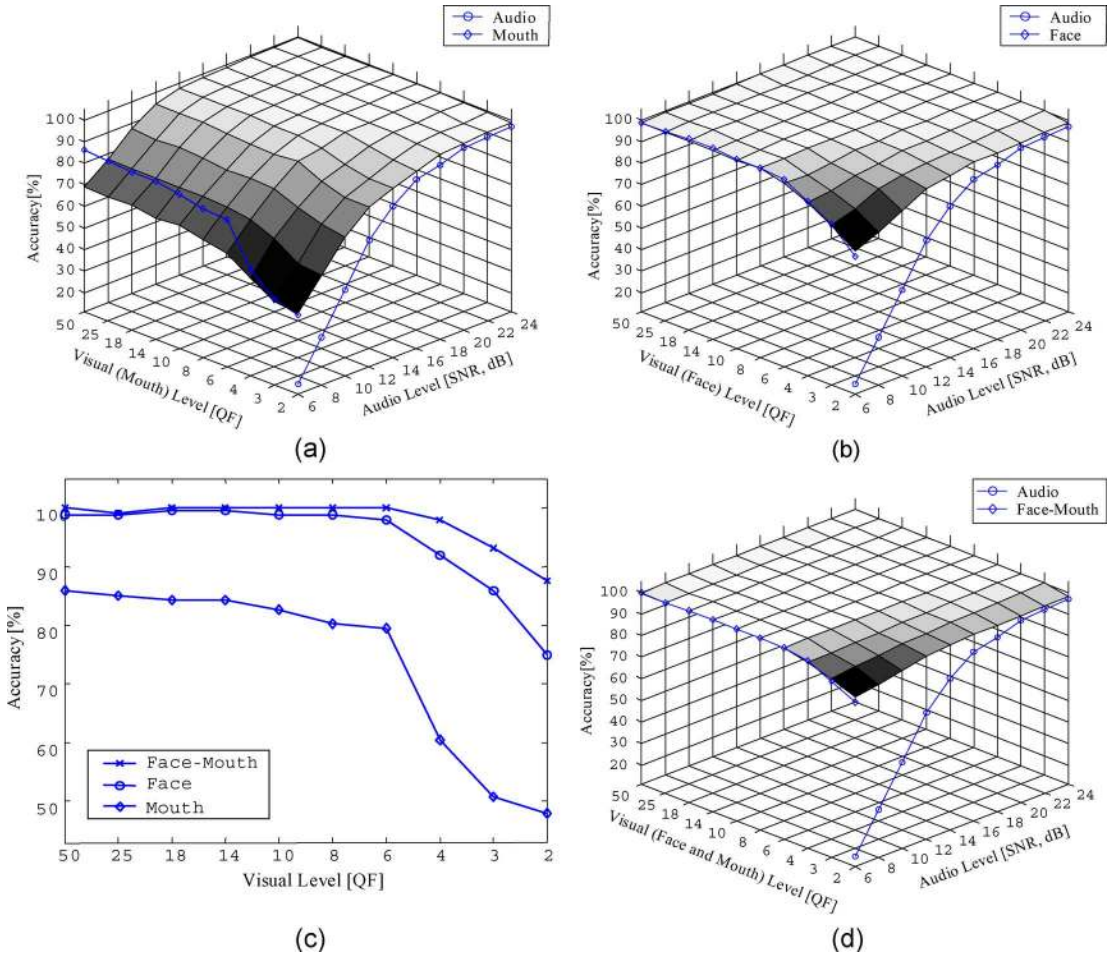

Fig. 8. Identification accuracies for the four fusion experiments carried out: (a) the audio and mouth, (b) the audio and face, (c) the face and mouth, and (d) the audio, face, and mouth (tri-expert fusion). For comparison, the results for the ten levels of visual (JPEG QF) and audio (dB) degradation are displayed.

TABLE III
PERSON IDENTIFICATION ACCURACIES (%) FOR THE MOUTH (M), FACE (F), AUDIO (A), AND THE FOUR FUSION EXPERIMENTS CARRIED OUT, NAMELY THE FUSION OF: (A) THE FACE AND MOUTH (FM), (B) THE AUDIO AND MOUTH (AM), (C) THE AUDIO AND FACE (AF), AND (D) THE AUDIO, FACE, AND MOUTH (AFM). RESULTS FOR BOTH AUTOMATIC FUSION (*AUTO*) AND THE NONWEIGHTED SUM RULE (*SUM*) ARE GIVEN

QF			dB	14		12		10		8		6	
			A	81.5		69.8		51.2		33.5		16.1	
				Auto	Sum	Auto	Sum	Auto	Sum	Auto	Sum	Auto	Sum
8	M	80.2	AM	96.8	96.8	92.3	91.5	85.9	85.5	76.6	73.8	65.7	60.5
	F	98.8	AF	99.6	99.2	99.6	98.4	99.6	94.8	99.2	88.7	98.8	80.6
	FM	100.0	AFM	100.0	99.6	100.0	99.2	100.0	94.8	100.0	90.3	100.0	86.3
6	M	79.4	AM	95.6	95.6	93.5	91.9	85.1	85.5	75.0	72.2	63.7	58.5
	F	98.0	AF	98.8	98.4	98.8	97.6	97.2	94.4	96.0	87.9	96.4	80.2
	FM	100.0	AFM	100.0	99.6	100.0	99.2	100.0	95.2	100.0	90.7	100.0	83.9
4	M	60.5	AM	92.3	91.9	88.7	89.5	80.6	81.0	70.6	66.9	56.5	53.6
	F	91.9	AF	98.4	97.6	96.8	95.6	95.6	91.5	92.7	84.3	91.5	74.6
	FM	98.0	AFM	98.8	98.8	98.4	98.0	98.0	94.4	97.6	88.3	97.6	82.7
3	M	50.8	AM	89.9	90.3	85.5	84.7	77.4	74.6	63.3	61.7	50.0	46.4
	F	85.9	AF	97.2	96.0	96.0	94.4	93.5	89.9	88.7	81.5	85.1	71.4
	FM	93.1	AFM	98.0	98.4	97.6	97.6	96.4	91.9	95.6	85.9	94.0	79.4
2	M	48.0	AM	89.5	89.5	85.1	83.5	75.4	73.8	62.1	60.9	48.8	44.8
	F	75.0	AF	95.6	95.2	93.5	91.5	87.5	83.9	82.3	75.4	77.8	62.9
	FM	87.5	AFM	96.8	97.6	96.0	96.0	93.5	90.3	91.5	81.5	89.9	73.0

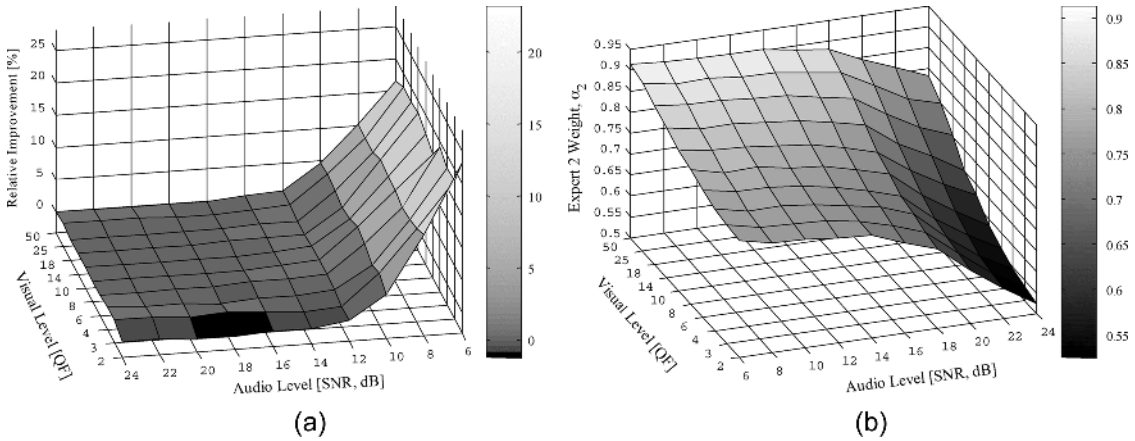

Fig. 9. (a) Shown is the relative percentage improvement of the automatic fusion scheme over the nonweighted sum rule, for tri-expert fusion, versus JPEG QF and audio SNR. (b) Automatically determined "visual" weights for tri-expert fusion versus JPEG QF and audio SNR.

Fig. 9(a) shows that, at the majority of operating points, the automatic fusion scheme yields a positive improvement compared to the nonweighted sum rule, with significant improvements (greater than 15% relative) attained at the lower QF and SNR levels. In Fig. 9(b), the general profile of the automatically determined "visual" weights is as expected, i.e., the visual scores receive a higher weighting when the QF is high and the SNR is low. The profile is biased towards weighting the "face-mouth" visual scores as all of the weights are greater than 0.5. This is due to the stronger performance of the "face-mouth" "expert" compared to the audio expert.

Further work includes dynamically varying the order of the fusion cascade and testing the performance of the fusion system described in this study using different types of audio and visual degradations. Since the reliability estimation is carried out at the score level, and not at the signal level, it is expected that varying the type of degradation causing the train/test mismatch will not adversely affect the fusion performance.

VII. Conclusion

A multiple-expert biometric person identification system has been presented, which combines information from three experts, namely: audio, visual speech, and face 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 (audio, face, and mouth) fusion studies use nonweighted 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. The results show improved fusion accuracies for the gamut of tested levels of audio and visual degradation, compared to the individual expert accuracies. Also, the automatic fusion scheme outperforms the nonweighted sum rule, particularly at the higher levels of audio and video degradation. The results highlight 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. These results as a whole are important for person recognition applications, where bandwidth is limited and uncontrolled acoustic noise is probable, such as, video telephony and online authentication; and demonstrate the utility of a multiple-expert person identification system based on automatic classifier fusion that is robust to both audio and visual train/test mismatch.

References

[1] N. A. Fox and R. B. Reilly, "Audio-visual speaker identification based on the use of dynamic audio and visual features," in *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication*, Guildford, U.K., Jun. 2003, pp. 743–751.

[2] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Dig. Signal Process.*, vol. 14, no. 5, pp. 449–480, Sep. 2004.

[3] T. J. Wark, S. Sridharan, and V. Chandran, "The use of speech and lip modalities for robust speaker verification under adverse conditions," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, Florence, Italy, Jun. 1999, vol. 1, pp. 812–816.

[4] C. Sanderson, S. Bengio, and Y. Gao, "On transforming statistical models for non-frontal face verification," *Pattern Recognit.*, vol. 39, no. 2, pp. 288–302, 2006.

[5] D. Blackburn, M. Bone, and P. J. Phillips, Facial Recognition Vendor Test 2000: Evaluation Report Feb. 2001 [Online]. Available: http://www.frvt.org, Tech. Rep.

[6] R. Gross, J. Shi, and J. F. Cohn, "Quo vadis face recognition," in *Proc. 3rd Workshop on Empirical Evaluation Methods in Computer Vision*, Kauai, HI, Dec. 2001, pp. 119–132.

[7] 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, Nov. 2003, pp. 25–32.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1324, Sep. 2003.

[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. 23–35, Mar. 2002.

[10] U. Dieckmann, P. Plankensteiner, and T. Wagner, "SESAM: A biometric person identification system using sensor fusion," *Pattern Recognit. Lett.*, vol. 18, no. 9, pp. 827–833, Sept. 1997.

[11] Y. Yemez, A. Kanak, E. Erzin, and A. M. Tekalp, "Multimodal speaker identification with audio-video processing," in *Proc. Int. Conf. Image Processing*, Barcelona, Spain, Sep. 2003, vol. 3, pp. 5–8.

[12] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1065–1074, May 1999.

[13] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Dig. Sig. Process.*, vol. 11, no. 3, pp. 169–186, July 2001.

[14] A. Kanak, E. Erzin, Y. Yemez, and A. M. Tekalp, "Joint audio-video processing for biometric speaker identification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, Apr. 2003, vol. 2, pp. 377–380.

[15] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 955–966, Oct. 1995.

[16] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "Adaptive classifier integration for robust pattern recognition," *IEEE Trans. Syst., Man, Cybern. B; Cybern.*, vol. 29, pp. 902–907, Dec. 1999.

[17] A. K. Jain and A. Ross, "Learning user-specific parameters in a multibiometric system," in *Proc. Int. Conf. Image Processing.*, Rochester, NY, Sep. 22–25, 2002, vol. 1, pp. 57–60.

[18] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[19] 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, U.K.: Cambridge Univ. Eng. Dept.: Microsoft Corporation, 2001.

[20] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: Applied to text dependent speaker recognition," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 495–506, Jun. 2005.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[22] I. Matthews, T. F. Cootes, J. A. Bangham, J. A. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[23] G. Potamianos, H. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998, vol. 3, pp. 173–177.

[24] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Tokyo, Japan, Aug. 2001, pp. 825–828.

[25] A. N. Netravali and B. G. Haskell, *Digital Pictures*. New York: Plenum, 1988.

[26] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

[27] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition*, Grenoble, France, Mar. 2000, pp. 300–305.

[28] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[29] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Amer. A*, vol. 4, no. 3, pp. 519–524, 1987.

[30] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[31] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, 1997.

[32] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[33] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 743–756, Jun. 1997.

[34] A. Yuille, "Deformable templates for face recognition," *J. Cog. Neurosci.*, vol. 3, no. 1, pp. 59–70, 1991.

[35] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[36] A. V. Nefian, L. H. Liang, T. Fu, and X. X. Liu, "A bayesian approach to audio-visual speaker identification," in *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication*, Guildford, U.K., Jun. 2003, pp. 761–769.

[37] F. Cardinaux, C. Sanderson, and S. Bengio, "Face verification using adapted generative models," in *Proc. 6th IEEE Int. Conf. Automatic Face and Gesture Recognition (AFGR)*, Seoul, 2004, pp. 825–830.

[38] P. Penev and J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Comput. in Neural Syst.*, vol. 7, no. 3, pp. 477–500, 1996.

[39] C. BenAbdelkader and P. Griffin, "Comparing and combining depth and texture cues for face recognition," *Image Vis. Comput.*, vol. 23, pp. 339–352, 2005.

[40] P. J. Phillips, P. Grother, P. Michaels, D. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," in *Proc. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures, (AMFG)*, Nice, France, Oct. 2003, p. 44.

[41] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[42] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[43] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[44] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, vol. 1, pp. 857–860.

[45] N. A. Fox and R. B. Reilly, "Robust multi-modal person identification with tolerance of facial expression," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, The Hague, The Netherlands, Oct. 10–13, 2004, vol. 1, pp. 580–585.

[46] R. W. Frischholz and U. Dieckmann, "BiolD: A multimodal biometric identification system," *Computer*, vol. 33, no. 2, pp. 64–68, 2000.

[47] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, 2005.

[48] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1260–1273, Nov. 2002.

[49] J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, and A. Jain, "Incorporating image quality in multi-algorithm fingerprint verification," in *Proc. Int. Conf. Biometrics (ICB)*, Hong Kong, China, 2006, pp. 213–220.

[50] K. Messer, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: "The extended M2VTS database"," in *Proc. 2nd Int. Conf. Audio and Video-based Biometric Person Authentication*, Washington, DC, Mar. 1999, pp. 72–77.

[51] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition Speech Res. Unit, Defence Research Agency, Malvern, U.K., 1992, Tech. Rep..

[52] T. Wark, S. Sridharan, and V. Chandran, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *Proc. 14th Int. Conf. Pattern Recognition*, Brisbane, Australia, Aug. 1998, vol. 1, pp. 123–125.

Niall A. Fox (M'04) received the B.E. degree in electronic engineering in 2001 and the Ph.D. degree for research in the area of digital signal processing in 2005, both from University College Dublin (UCD), Belfield, Dublin, Ireland.

His research interests include multimodal signal processing and classifier combination theory. He is currently a Research Engineer at BiancaMed, NovaUCD, Belfield, Dublin. More details of his research are available at http://www.niallfox.com/.

Ralph Gross (M'04) received the Diploma degree in computer science and mathematics from the University of Karlsruhe, Karlsruhe, Germany, in 1998. He is currently pursuing the Ph.D. degree at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, as a member of the Data Privacy Lab.

His research interests are in the areas of computer vision and machine learning, specifically in face modeling, tracking, and face recognition. More details of his research are available at http://www.ralphgross.com.

Mr. Gross is a member of the IEEE Computer Society.

Jeffrey F. Cohn (M'04) received the Ph.D. degree in clinical psychology from the University of Massachusetts, Amherst, and completed his Clinical Internship at the University of Maryland Medical Center.

He is a Professor of Psychology and Psychiatry at the University of Pittsburgh, Pittsburgh, PA, and Adjunct Faculty at the Robotics Institute, Carnegie Mellon University, Pittsburgh. For the past 20 years, he has conducted investigations in the theory and science of emotion, depression, and nonverbal communication. He has co-led interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis of facial expression and prosody; and applied these tools to research in human emotion, social development, nonverbal communication, psychopathology, biomedicine, and biometrics. His research has been supported by grants from the National Institute of Mental Health, the National Institute of Child Health and Human Development, the National Science Foundation, the Central Intelligence Agency, the Defense Advanced Research Projects Agency, and the Naval Research Laboratory.

Richard B. Reilly (M'92–SM'04) received the B.E., M.Eng.Sc., and Ph.D. degrees in 1987, 1989, and 1992, respectively, all in electronic engineering, from the National University of Ireland.

In 1988, he joined Space Technology Ireland and the Department de Recherche Spatiale (CNRS Group), Paris, France, developing DSP-based on-board experimentation for the NASA satellite WIND. In 1990, he joined the National Rehabilitation Hospital and in 1992 became a Post-Doctoral Research Fellow at University College Dublin (UCD), Dublin, Ireland, focusing on signal processing for speech and gesture recognition. Since 1996, he has been on the academic staff of the School of Electrical, Electronic and Mechanical Engineering, UCD. He is currently an Associate Professor conducting research in neurological signal processing and multimodal signal processing.

Dr. Reilly was the 1999/2001 Silvanus P. Thompson International Lecturer for the IEE. In 2004, he was awarded a U.S. Fulbright Award for research collaboration into multisensory integration with the Nathan Kline Institute for Psychiatric Research, New York. He is a member of the IEEE Engineering in Medicine and Biology Society and Signal Processing Society. He is the Republic of Ireland Representative on the Executive Committee of the IEEE U.K. and Republic of Ireland Section. He is an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and a Reviewer for IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, the JOURNAL OF APPLIED SIGNAL PROCESSING, *Signal Processing*, and *IEE Proceedings on Vision, Image, and Signal Processing*.