Person Identification Using Automatic Integration of Speech, Lip, and Face Experts

Niall Fox
Dept. of Electronic and Electrical Engineering
University College Dublin, Belfield, Dublin 4, Ireland
Tel.: 353-1-7161914
niall.fox@ee.ucd.ie

Ralph Gross
Robotics Institute, Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
Tel.: 1 412 268 - 2078
rgross@cs.cmu.edu

Philip de Chazal
Dept. of Electronic and Electrical Engineering
University College Dublin, Belfield, Dublin 4, Ireland
Tel.: 353-1-7161959
philip@ee.ucd.ie

Jeffery F. Cohn
Robotics Institute, Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
Tel.:1 412 624 8825
jeffcohn+@cs.cmu.edu

Richard B. Reilly
Dept. of Electronic and Electrical Engineering
University College Dublin, Belfield, Dublin 4, Ireland
Tel.: 353-1-7161960
richard.reilly@ucd.ie

ABSTRACT

This paper presents a multi-expert person identification system based on the integration of three separate systems employing audio features, static face images and lip motion features respectively. Audio person identification was carried out using a text dependent Hidden Markov Model methodology. Modeling of the lip motion was carried out using Gaussian probability density functions. The static image based identification was carried out using the FaceIt system. Experiments were conducted with 251 subjects from the XM2VTS audio-visual database. Late integration using automatic weights was employed to combine the three experts. The integration strategy adapts automatically to the audio noise conditions. It was found that the integration of the three experts improved the person identification accuracies for both clean and noisy audio conditions compared with the audio only case. For audio, FaceIt, lip motion, and tri-expert identification, maximum accuracies achieved were 98%, 93.22%, 86.37% and 100% respectively. Maximum bi-expert integration of the two visual experts achieved an identification accuracy of 96.8% which is comparable to the best audio accuracy of 98%.

Categories and Subject Descriptors

I.5.2 [**Computing Methodologies**]: Pattern Recognition, Design Methodology– *Classifier design and evaluation.*

General Terms

Algorithms, Reliability, Experimentation, Security.

Keywords

Person identification, multi-expert, audio, face, lips, late integration, automatic weighting.

1. INTRODUCTION

Biometrics is a field of technology devoted to verification or identification of individuals using biological 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 subjects. Hence person identification is inherently a more difficult task, particularly when the number of registered subjects is large.

Person identification systems based on the analysis of audio signals achieve high performance when the signal to noise ratio (SNR) of the audio signal is high. However the performance degrades quickly with decreasing SNR values. It is expected that the integration of the audio expert (in the context of this paper, the term expert refers to a particular audio or visual based person identification system) with a visual expert will improve upon the audio scores for clean audio conditions and increase robustness to the presence of audio noise. In [5] and [24] the audio expert was fused with the visual expert to improve the robustness of speech recognition to audio noise. In [4] multi-expert person identification experiments showed that the integration of visual experts, incorporating visual features from the eyes, nose and mouth, with the audio expert, significantly improved the scores. In [13] and [2] person identification was found to be more robust to audio noise when the audio and lip motion experts are integrated.

Previous studies have shown that the FaceIt [9] person identification system places a lot of emphasis on the eye region.

In [7] FaceIt experiments were carried out for different face occlusions and it was shown that the performance of FaceIt degrades significantly for a high level of eye occlusion. The lip motion system described in this paper uses lip features extracted from a sequence of lip region of interest (ROI) images. It is expected that the integration of these two visual person identification systems, one emphasizing the eyes and the other emphasizing the lips will lead to a synergistic improvement.

Most systems for person identification use a single expert, such as speech or facial features. In this paper we investigated the relative performance of three approaches to person identification – speech, lip motion, and facial features – and the benefits or improvements possible by combining them in a multi-expert classifier under conditions of varying time-delays between training and testing sets, number of training sessions, and audio noise. Results are then presented for the integration of these experts in four experiments, namely the integration of: 1) FaceIt and the lip motion experts, 2) FaceIt and audio experts, 3) lip motion and audio experts, and 4) all three experts.

2. THE XM2VTS AV DATABASE

The XM2VTS Audio-Visual (AV) database [14], [16] was employed 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 third sentence ("Joe took fathers green shoe bench out") was used for this research. The start and end of some sentences were clipped. Due to this and other errors in the sentences, only 251 out of a possible 295 subjects were used for our experiments. Lip motion features were extracted from the mouth ROI. The ROI was identified manually for every 10th frame, and the ROI for the intermediate frames determined by interpolation. To identify the ROI manually, the midpoint between the two labial corners was identified and a 98×98 pixel block centered on this midpoint was extracted as the ROI. The FaceIt system processes an entire visual frame taken from the video sequence.

3. FACE RECOGNITION: FACEIT

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) [26], Linear Discriminant Analysis (LDA) [16, 1], Principal Component Analysis (PCA) [23, 25], Kernel Methods [12], and Neural Networks [11] 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 [10] extending the approach by Yuille [29]. Wiskott et al. developed an elastic bunch graph matching algorithm for face recognition in [27]. Penev et. al [18] 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 [7, 3, 19]. 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) [7].

4. AUDIO PERSON IDENTIFICATION

The audio signal was first pre-emphasized to increase the acoustic power at higher frequencies using the filter $H(z) = 1/(1-0.97z^{-1})$. The pre-emphasized signal was divided into frames using a Hamming window of length 20 ms, with overlap of 10 ms to give an audio frame rate, F_A, of 100 Hz. Mel-frequency cepstral coefficients (MFCC's) [6] of dimension 16 were extracted from each frame. The energy [28] of each frame was also calculated and used as a 17^{th} static feature. Seventeen first order differences or *delta features* were calculated between adjacent frames and appended to the static audio features to give an audio feature vector of dimension 34. The number of MFCC's employed was determined empirically to give the best performance. *Cepstral mean normalization* was performed on the audio feature vector [28] in order to compensate for long term spectral effects of the audio channel.

A text dependent person ID expert was used. For text dependent modeling [13], the subject says the same utterance for both training and testing. It was employed, as opposed to text independent modeling [21], because it was suited to the database used in this study. Also, text dependence has been found to out-perform text independence [15].

Subject S_i , $i = 1 \ldots N$, was modeled by a single audio sentence subject dependent HMM, where $N = 251$ here. There was one background HMM. Three sessions were used for training and one session for testing. The background HMM was trained using three of the sessions for all N subjects. This background model captures the audio speech variation over the entire database. Since there were only three training utterances per subject, there was insufficient training data to train a subject dependent HMM, which was initialized with a prototype model. Hence the background model was used to initialize the training of the subject dependent models.

A sentence observation, O, was tested against all N subjects, S_i, and the subject that gave the maximum score was chosen as the identified subject. To score an observation O against subject S_i, $P(S_i/O)$ is calculated and is normalized by dividing by F, the number of frames in the sentence observation.

5. IDENTIFICATION BY LIP MOTION

Transform based features were used to represent the visual information based on the Discrete Cosine Transform (DCT) which was used because of its high energy compaction [17]. The 98×98 color pixel blocks were converted to gray scale values. The gray scale ROI was then histogram equalized and the mean pixel value

was subtracted. This image pre-processing was carried out to account for varying lighting conditions across sessions and subjects. The DCT was applied to the gray scale pixel blocks. The first 15 coefficients were used, taken in a zigzag pattern to form the visual frame observation feature vector. However only 14 of these features were used for modeling since the first feature corresponds to the mean of the ROI, and due to the mean removal, was zero valued for the feature vector of each frame.

A Gaussian model consisting of a single probability density function was used to model the lip motion of each subject. The entire visual sentence was modeled by the Gaussian model. The mean feature vector and diagonal covariance matrix was calculated from the training data. The log likelihood probability was calculated for each frame of the test sentence and these scores were summed over the entire sentence. The summation score was then normalized by dividing by the number of frames in the test sentence. This was done for the *N* subject models and the model giving the highest score chosen as the identified subject.

6. LATE INTEGRATION

The various "experts" (audio, FaceIt, lip motion) were integrated using late integration (LI). The advantages of LI include the ability to account for the expert reliabilities, small feature vector dimensions per expert and ease of adding other experts to the system. For LI the expert scores are weighted to account for the reliability of each mode. The scores may be integrated via addition or multiplication as shown in Equations (1a) and (1b) respectively, for bi-expert LI. Both LI methods were investigated and it was found that the results achieved for both were similar. Hence the results for additive integration only, are presented in this paper. Prior to LI, all expert scores were normalized to fall into the range of 0 to 1.

$$P(S_i \mid x_A, x_V) = \alpha.P(S_i \mid x_A) + \beta.P(S_i \mid x_V). \qquad (1a)$$

$$P(S_i \mid x_A, x_V) = (P(S_i \mid x_A))^{\alpha} \times (P(S_i \mid x_V))^{\beta}. \qquad (1b)$$

where:

$$\alpha = \begin{cases} 0, & c \le -1, \\ 1+c, & -1 < c < 0, \\ 1, & c \ge 0, \end{cases} \qquad (2)$$

$$\beta = \begin{cases} 1, & c \le 0, \\ 1-c, & 0 < c < 1, \\ 0, & c \ge 1, \end{cases}$$

The fusion parameter *c* varies with the audio SNR. Figure 1 shows how the expert weights, α and β, depend on the fusion parameter *c*. Higher values of *c* (*>0*) place more emphasis on the audio expert whereas lower values (*<0*) place more emphasis on the visual expert. For $c \ge 1$, $\alpha = 1$ and $\beta = 0$, hence the decision is based entirely on the audio expert, whereas, for $c \le -1$, $\alpha = 0$ and $\beta = 1$, hence the decision is based entirely on the visual expert.

For automatic integration a mapping, $c(\rho)$, between the audio reliability measure, ρ, and the fusion parameter *c* is employed. The reliability measure employed is the sum of the difference between the top two highest scores and the difference between the second and third highest scores for each person test. As the audio SNR decreases, this reliability measure decreases because the audio scores become less discriminatory. To determine the mapping $c(\rho)$, the values of *c* which provided for optimum fusion, c_{opt}, were found by exhaustive search for the *N* tests at each SNR value. The mean reliability measure, ρ_{mean}, across the *N* tests at each SNR value was also found. A sigmoidal function

$$c(\rho) = c_{os} + \frac{h}{1 + \exp[d.(\rho + \rho_{os})]}, \qquad (3)$$

was employed to provide a mapping between c_{opt} and ρ_{mean}, where the parameters c_{os}, h, d and ρ_{os} were determined empirically to give the best performance. Hence, for automatic integration, ρ is calculated from the *N* scores for each test and *c* is determined using $c = c(\rho)$. This integration approach is similar to that used in [8].

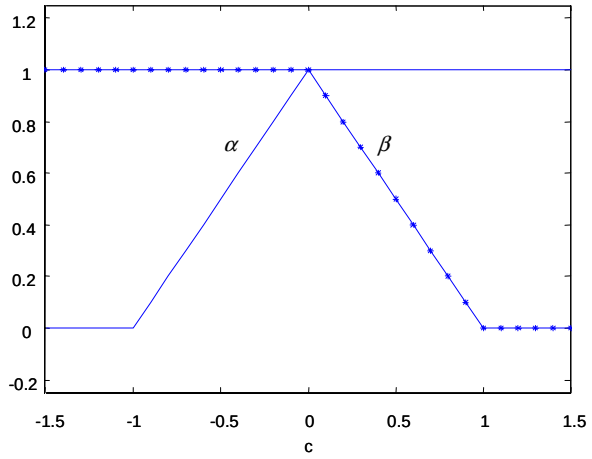

Figure 1. Variation of the expert weigths α and β, with the fusion parameter *c*.

7. EXPERIMENTS

In all the experiments, the probe images used for testing were obtained from the final (fourth) session. The FaceIt galleries used for training were formed from the first three sessions according to Table 1. Five trials were constructed to test how the performance of FaceIt varied when 1) the time difference between the gallery and the probe set varied between one and three months and 2) multiple sessions were used to form the gallery. Testing of the lip motion feature person identification system was also carried out for the five trials in Table 1.

Table 1. Outline of the five visual trials performed

Trial #	Gallery\Train	Probe\Test
1	3	4
2	1	4
3	2	4
4	2,3	4
5	1,2,3	4

The audio models were tested according to trial 5 of Table 1. Additive white Gaussian noise was applied to the clean audio at signal-to-noise ratios (SNR) ranging from 0dB to 48dB in steps of 4dB. All audio models were trained using clean speech and tested using speech containing noise.

The evaluation of the FaceIt system on the selected subjects of the XM2VTS database proceeded as follows. For trials 1 through 3 a single gallery image was chosen at random from the image sequence of each subject and compared to a randomly chosen probe image of each subject. FaceIt produces a matching score between 0.0 and 10.0 for each gallery/probe image pair. For each probe image the gallery image with the highest score was selected as recognition result. For trials 4 and 5 FaceIt was given two and three randomly chosen gallery images respectively that where internally combined by FaceIt to determine a matching score. In all cases original, unprocessed images were used. In a separate experiment it was verified that FaceIt's face finding module was able to reliably locate the face in each image.

Four integration experiments were carried out using the three experts.

7.1 Integrating FaceIt with Lip motion

The two visual experts were integrated to test if their fusion led to a synergistic improvement. The fusion was carried out according to Equation (1a) with fusion parameter $c = 0$, i.e. equal weighting of the two experts.

7.2 Integrating FaceIt with Audio

Under noise-free audio conditions an audio person identification system can achieve high accuracies. However in the presence of audio noise the performance can degrade significantly. The scores of FaceIt trial 5 were integrated with the audio scores at various audio SNR's in order to investigate if the audio accuracies can be improved upon at low SNR's. The expert weighting was determined automatically as described in Section 6.

7.3 Integrating Lip motion with Audio

The same tests described in Section 7.2 were carried out using the lip motion trial 5 scores in place of the FaceIt scores.

7.4 Integrating all Three Experts

All three experts were integrated. The visual weight, β, was divided equally between the two visual experts as shown in Equation (4).

$$P(S_i \mid x_A, x_{V_1}, x_{V_2}) = \alpha.P(S_i \mid x_A) + \frac{\beta}{2}.P(S_i \mid x_{V_1}) + \frac{\beta}{2}.P(S_i \mid x_{V_2}). \quad (4)$$

8. RESULTS

Left to right HMM's with a twelve state, two mixture topology were used in the audio classification experiments. The audio models were trained using the Baum Welch algorithm and tested using the Viterbi algorithm [20], implemented using the HMM toolkit, HTK [24]. The audio features were calculated using HTK. The background models were trained using three sessions. This gave *3*N (753)* training examples per background model. This HMM topology, was found by exhaustive search to give the best result. The audio results versus SNR are presented in Table 2.

Table 2. Audio, audio with FaceIT, audio with lip motion and tri-expert scores (trial 5)

SNR	Audio (%)	Audio & FaceIt (%)	Audio & Lip Motion (%)	3 Experts (%)
48	98.01	100.00	98.80	100.00
44	98.41	100.00	98.80	100.00
40	97.61	99.20	98.41	100.00
36	95.22	99.20	98.01	100.00
32	91.63	98.41	96.02	98.41
28	79.28	97.61	93.23	98.41
24	54.58	96.81	90.04	96.81
20	27.49	94.82	88.84	95.62
16	10.36	93.23	86.06	94.02
12	5.98	93.63	86.06	92.83
8	1.99	94.02	86.06	96.41
4	1.99	93.23	86.06	96.41
0	1.20	93.23	86.06	96.81

The results of the five FaceIt and lip motion trials are presented in Table 3. The results are presented as percentage accuracy and number of correctly identified subjects out of the possible *251* subjects.

Table 3. FaceIt, lip motion and FaceIt integrated with lip motion scores for the 5 trials

Trial #	FaceIT (%)	Lip Motion (%)	FaceIT & Lip Motion (%)
1	84.86	47.41	90.04
2	79.28	29.08	82.87
3	84.06	25.10	85.66
4	90.44	75.30	93.23
5	93.23	86.06	96.81
Mean	**86.37**	**52.59**	**89.72**

8.1 Integrating FaceIt with Lip motion

The results of late integration of the FaceIt and lip motion experts are presented in Table 3. The integrated scores presented are for equal weights.

8.2 Integrating FaceIt with Audio

The results of late integrating FaceIt trial 5 with the audio expert for the thirteen audio noise levels using automatic weights are presented in Table 2. The same results are also shown in Figure 3. The sigmoidal fit between c_{opt} and the reliability measure is shown in Figure 2.

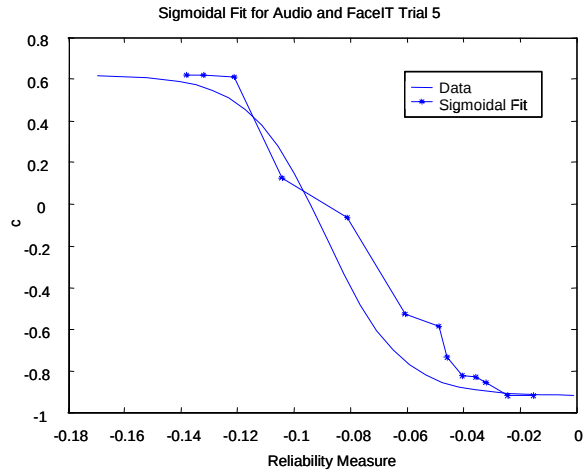

Figure 2. Sigmoidal fit for audio and FaceIt trial 5.

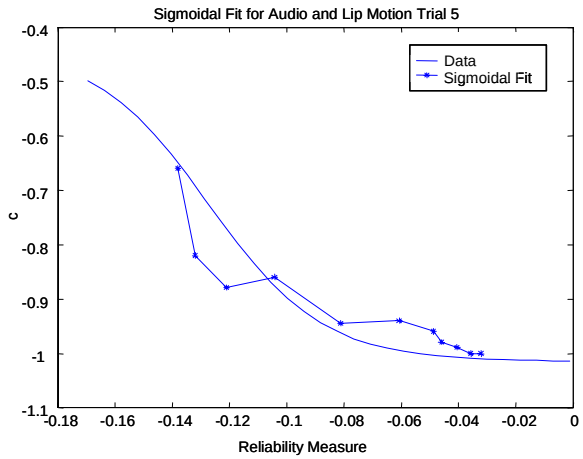

Figure 4. Sigmoidal fit for audio and lip motion trial 5.

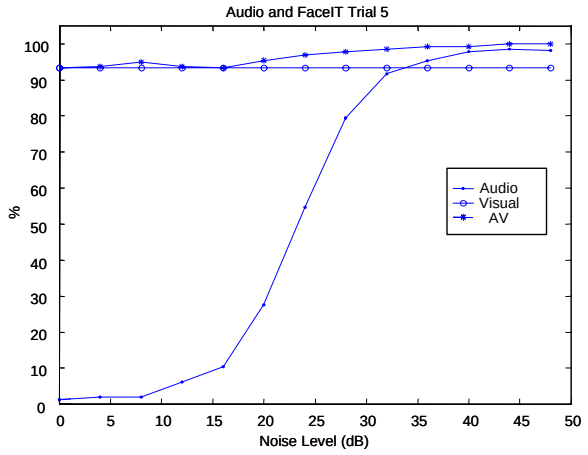

Figure 3. Integration of FaceIt trial 5 with audio versus SNR.

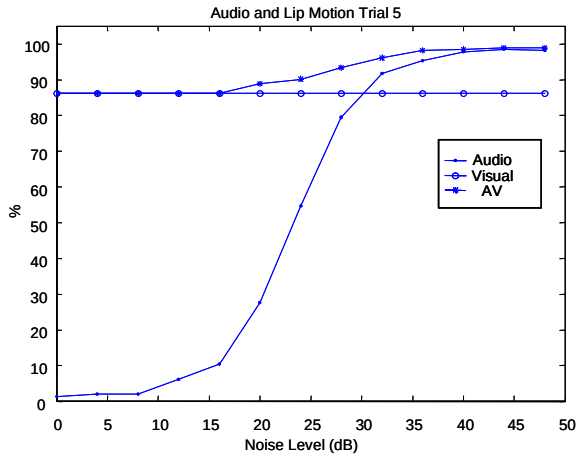

Figure 5. Integration of lip motion trial 5 with audio versus SNR.

8.3 Integrating Lip motion with Audio

The results of late integrating lip motion trial 5 with the audio expert for the thirteen audio noise levels using automatic weights are presented in Table 2. The same results are also shown in Figure 5. The sigmoidal fit between c_{opt} and the reliability measure is shown in Figure 4.

8.4 Integrating all Three Experts

The results of tri-expert late integration of trial 5 for the thirteen audio noise levels using automatic weights are presented in Table 2 and Figure 7. The sigmoidal fit between c_{opt} and the reliability measure is shown in Figure 6.

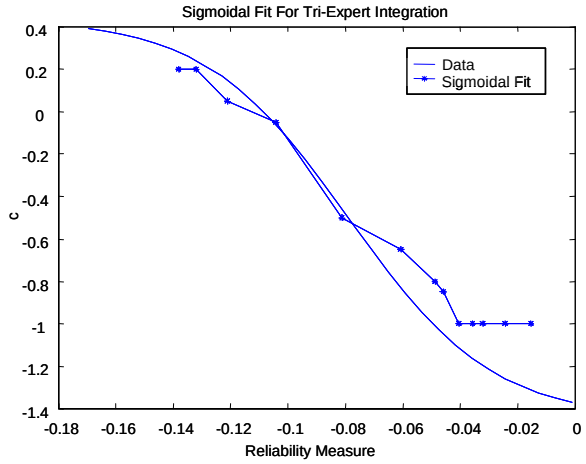

Figure 6. Sigmoidal fit for tri-expert integration.

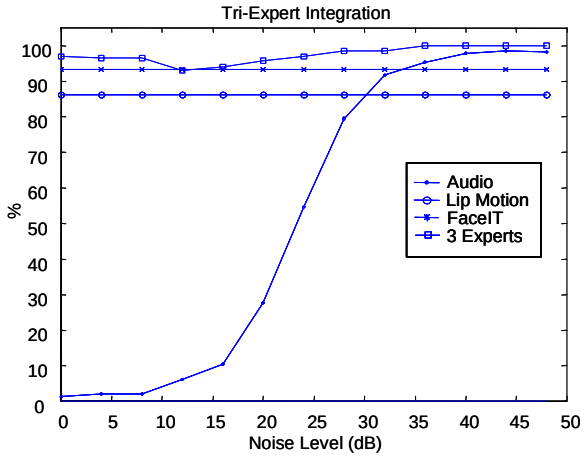

Figure 7. Integration of the 3 experts versus SNR.

9. DISCUSSION

The person recognition rates using the audio expert only (Table 2) were very high for high SNR's. The highest recognition rate was 98% attained at a SNR of 48dB. As the SNR decreased, the recognition rate decreased. There was a large change in recognition rate between a SNR of 32db (91.6%) and a SNR of 24 dB (54.6%). At a SNR of 0dB and lower the recognition rate was equal to recognition rate of random guessing (0.4%). The steep roll off of the recognition performance with respect to SNR was due to the mismatched audio testing conditions, i.e. training on noise-free audio and testing on audio of a lower SNR. It is expected that the roll off would be less steep if matched testing was employed, i.e. training and testing using audio of the same SNR.

Table 2 also shows that the lip motion expert only performs well when all three training sessions are used (trial 5), giving a recognition rate of 86%. This rate was obtained using manual ROI determination, which may have positively biased the results. An automatic ROI detection system will need to be used in order to test for such a bias. The lip motion scores for trials 1 to 3, 47.4%, 29%, 25% respectively, show a large variance. This may indicate that the Gaussian models were poorly trained or that test session 3 used in trial 1 provided for more person discrimination. However, trial 1 had the least time difference between training and testing and this may be the reason why it performed better than trials 2 and 3. Trial 1 also was the best of the FaceIt trials 1 to 3. However the score variance was less than that for the lip motion trials.

The FaceIt recognition rate for trials 1 to 3 decreased from 84.86% to 79.3% as the time difference increased between the training and testing data. This decrease in recognition rate may become more apparent in a practical system when the time difference between the training phase and the identification phase exceeds three months. Addition of more gallery sessions improved the FaceIt scores, with the use of all three sessions in trial 5 giving a score of 93.23%. Adding up to 5 additional gallery images from session 1 to trial 2 improved the recognition accuracy to 84.4% (from 79.28%), but falls far short of the 93.23% accuracy achieved in trial 5. We therefore conclude that the increase in performance is mostly due to the availability of gallery images over time.

9.1 Integrating FaceIt with Lip motion

Combining the two visual experts resulted in comparable performance to the audio expert. As shown in Table 2 the best result for combining the two image experts resulted in a recognition rate of 96.81% which was only 1.19% lower than the best recognition rate from the audio expert. This suggests that the visual experts are of a comparable importance as the audio expert for person identification. It is worth noting that the XM2VTS visual data is of extremely high quality with little variation in illumination, pose and emotion during and across the four recording sessions. A practical system may not produce such high quality visual data and hence a person recognition system would have a greater reliance on the audio expert.

For the five trials combining the two visual experts, employing equal weightings, resulted in an improvement in recognition rate over using either visual expert alone.

9.2 Integrating FaceIt with Audio

The integration of FaceIt and audio shown in Table 2 resulted in a perfect recognition rate (100%) for SNR greater than 40dB. The recognition rate did not decrease as rapidly with respect to SNR as the audio only expert (see Table 2). The integrated system performed better for all noise levels and is less sensitive to audio noise.

9.3 Integrating Lip motion with Audio

The integration of the lip motion expert and the audio expert shown in Table 2 was not as successful as the integration in Section 9.2. However, an improvement on the audio only scores was achieved for all noise levels. The recognition rate decreased more rapidly with respect to SNR than the equivalent rates for Section 9.2.

9.4 Integrating all Three Experts

A perfect recognition rate was achieved when the three experts were integrated when the SNR of the audio signal exceeded 32dB. The tri-expert integration also outperformed the bi-expert and uni-

expert recognition rates for most noise levels except at 12dB. The tri-modal scores between 25dB and 10dB decline and rise again to the level achieved by purely integrating the two visual experts. The poor performance between 25dB and 10dB is due to the sigmoidal fit (shown in Figure 6) which is trained using the reliability parameters ρ_{opt} . The ρ_{opt} values are determined globally across all *N* person tests (see Section 6), and may have a large variance. The sigmoidal fit does not take this variance about ρ_{opt} into account. Hence a different curve fit or a different reliability parameter exhibiting a lower variance, may improve upon the tri-expert scores of Figure 6.

As further work we aim to investigate the effect of image degradations on the identification performance.

9.5 Automatic determination of integration weights

It is important to note that the automatic weight sigmoidal mapping functions are trained and tested on the same data and this may have resulted in slightly optimistically biased performance results. Ideally the sigmoidal mapping functions should be trained and tested on the different data sets. It should also be noted that the reliability measure employed in this paper may behave differently w.r.t. SNR for different noise types. It would be interesting to carry out the same experiments using different noise types other than additive white Gaussian noise use here, and also employ other reliability measures such as dispersion, score variance, entropy and voicing index [8].

The automatic weights employed here do not take the reliability of the visual experts into account. This was not a major issue for the experiments carried out in this study, since the visual data was of a constant high quality. However, in the presence of varying visual degradations, the expert weights should depend on some visual expert reliability measure.

10. CONCLUSION

The integration of the two visual experts resulted in higher identification performance than the performance obtained using either visual expert alone. These results show that the two visual experts provide complementary information and hence are emphasizing different visual cues. Visual experiments, such as the performance of the system in the presence of eye occlusion would provide further insight.

The results showed that a system integrating the audio and either visual expert was more accurate for noise-free audio and is more robust to audio noise compared to the performance of the audio only system.

The use of automatically determined experts weights leads to synergistic integration. The integration method is efficient and not computationally expensive to carry out.

The tri-expert results show that the integration of all available experts leads to the best recognition rates. An advantage of a multi-expert system is that subjects that are difficult to identify with one expert may be more easily identified with another expert.

11. ACKNOWLEDGEMENTS

The research described in this paper was supported by the Informatics Research Initiative of Enterprise Ireland and in part by the U.S. Office of Naval Research contract N00014-00-1-0915.

REFERENCES

[1] P.Belhumeur and J. Hespanha and D.Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[2] Fox, N., Reilly, R.B., "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features", *Proc. 4th International Conference on Audio and Video Based Biometric Person Authentication*, June 2003.

[3] D.Blackburn and M.Bone and P.J.Philips, “Facial Recognition Vendor Test 2000”, *Evaluation report*, 2000.

[4] Brunelli, R. and Falavigna, D., “Person identification using multiple cues”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955-966, Oct. 1995.

[5] Chen, T., “Audiovisual Speech Processing”, *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9-21, Jan. 2001.

[6] Davis, S. and Mermelstein, P., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

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

[8] Heckmann,M.; Berthommier,F.; Kristian,K., “Noise Adaptive Stream Weigting in Audio-Visual Speech Recognition”, *EURASIP Journal on Applied Signal Processing, Special Issue on Joint Audio-Visual Speech Processing*”, vol. 2002, no. 11, pp. 1260-1273, Nov. 2002.

[9] Identix corp., *www.identix.com*, 5600 Rowland Road, Minnetonka, MN 55343.

[10] A.Lanitis and C.Taylor and T.Cootes, “Automatic Interpretation and Coding of Face Images Using Flexible Models”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743-756, 1997.

[11] S.Lawrence and C.Giles and A.Tsoi and A.Back. “Face Recognition: A Convolutional Neural Network Approach”, *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98-113, 1998.

[12] Y.Li and S.Gong and H.Liddell, “Support vector regression and classification based multi-view face detection and recognition”, *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[13] Lucey, S., *Audio-Visual Speech Processing. PhD thesis*, Queensland University of Technology, Brisbane, Australia, Apr.2002.

[14] Luettin, J. and Maitre, G., “Evaluation Protocol for the XM2VTSDB Database (Lausanne Protocol)”, *In IDIAP*

Communication 98-05, IDIAP, Martigny, Switzerland, Oct.1998.

[15] Luettin J., "Speaker verification experiments on the XM2VTS database", *In IDIAP Communication 98-02, IDIAP*, Martigny, Switzerland, Aug.1999.

[16] Messer, K., Matas, J., Kittler, J., Luettin J., and Maitre, G.: XM2VTSDB: "The Extended M2VTS Database", *The Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, Washington D.C., pp. 72-77, Mar.1999.

[17] Netravali, A. N. and Haskell, B. G., *Digital Pictures*, Plenum Press, pp. 408-416, 1998.

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

[19] P.J. Phillips and P. Grother and R. Michaels and D. Blackburn and E.Tabassi and M. Bone, "Face Recognition Vendor Test 2002", *Evaluation Report*.

[20] Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, Feb.1989.

[21] Reynolds, D. A. and Rose, R. C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, Jan.1995.

[22] B.D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.

[23] 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.

[24] Scanlon, P. and Reilly, R., "Visual Feature Analysis For Automatic Speechreading", *DSP Research Group*, UCD, Dublin, Ireland, 2001.

[25] M.Turk and A.Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol. 3, no.1, pp. 71-86, 1991.

[26] Vapnik, V, *The nature of statistical learning theory*, Springer Verlag, 1995.

[27] L.Wiskott and J-M.Fellous and N. Krueger and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775—779, 1997.

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

[29] A.Yuille, "Deformable Templates for Face Recognition", *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 59-70, 1991.