Investigations into the robustness of audio-visual gender classification to background noise and illumination effects

Stewart, D., Wang, H., Shen, J., & Miller, P. (2009). *Investigations into the robustness of audio-visual gender classification to background noise and illumination effects*. 168-174. Paper presented at Digtal Image Computing: Techniques and Applications, Australia. https://doi.org/10.1109/DICTA.2009.34

Investigations into the robustness of audio-visual gender classification to background noise and illumination effects

Darryl Stewart, Hongbin Wang, Jiali Shen, Paul Miller
ECIT
Queen's University Belfast
Belfast, Northern Ireland
{dw.stewart, h.wang, j.shen, p.miller}@qub.ac.uk

***Abstract*—In this paper we investigate the robustness of a multimodal gender profiling system which uses face and voice modalities. We use support vector machines combined with principal component analysis features to model faces, and Gaussian mixture models with Mel Frequency Cepstral Coefficients to model voices. Our results show that these approaches perform well individually in 'clean' training and testing conditions but that their performance can deteriorate substantially in the presence of audio or image corruptions such as additive acoustic noise and differing image illumination conditions. However, our results also show that a straightforward combination of these modalities can provide a gender classifier which is robust when tested in the presence of corruption in either modality. We also show that in most of the tested conditions the multimodal system can automatically perform on a par with whichever single modality is currently the most reliable.**

Keywords – Audio-Visual Fusion, Gender Classification

I. INTRODUCTION

Recently, there has been much work in the area of behavior analysis for video surveillance. However, an equally important issue, that has received relatively little attention thus far, is the ability to profile people in video data based on age and gender. Such profiling would allow future intelligent CCTV systems which could determine the intrinsic threat posed by certain individuals, or groups of individuals, to others. There are also important commercial applications of such systems, for instance in the dynamic provision of suitable gender-specific advertising in shops, transport systems or even the internet. In this work we will only investigate gender classification without consideration of age. Specifically, we will investigate gender classification using face images and voice samples.

Whilst there has been a large amount of work on face recognition in the past [1][2], there has been much less work aimed at gender recognition using faces. Moghaddam and Yang [3] investigated the use of nonlinear support vector machines (SVM) for gender classification with low-resolution "thumbnail" faces. The SVM performance (3.4% error) was shown to be superior to traditional classifiers such as linear, quadratic, and Fisher linear discriminate, and more modern techniques such as radial basis function classifiers and large-ensemble RBF networks. Furthermore, the SVM technique was shown to be stable and robust with respect to scale and degree of facial detail. In other work, Buchala et al [4] analyzed the importance of principal component analysis (PCA) order in classifying faces with respect to gender, ethnicity, age and identity. They found that gender, ethnicity and age could be encoded in a relatively few number of PCA components, and that these were to be found predominantly amongst the first few. Regarding gender, they found that the third component encodes information related to the complexion, length of nose and the presence or absence of hair on the forehead, and the fourth encodes information related to eyebrow thickness and the absence or presence of a smile. Using a Fisher discriminant classifier they achieved 86% correct gender classification. In further work [5], they replaced PCA by curvilinear component analysis (CCA) to account for the nonlinear nature of real data. Using CCA they were able to reduce the data to its intrinsic, or true, dimension. Similar error rates of 6% were achieved using an SVM classifier following CCA and PCA; however, the former used 14 coefficients whilst the latter required 273. They then investigated the importance of local to global features for gender recognition, and discovered that a combination of both gave significantly better classification than either alone [6]. In addition, they showed association between the errors made by the computational models and those made by humans for local features such as eyes and mouths. In other work, Abdelkader and Griffin [7] demonstrated superior performance for a local features based proprietary algorithm, FaceIt®, when compared with a global eigenface approach. Best performance obtained was 94%. The comparison was performed using the most extensive database, consisting of 13,000 faces, reported in the literature thus far. Most recently, Makinen and Raisamo [8] investigated the importance of automatic alignment of faces before gender classification for a number for different classifiers. They found alignment made no difference and achieved best classification accuracy of around 87% with an SVM. They also found that a face size of 36 × 36 pixels provided superior performance compared to 24 ×24 and 48 × 48. Within the area of psychology, studies have shown that reducing resolution and increasing noise leads to reduction in gender classification by human subjects [9].

978-0-7695-3866-2/09 $26.00 © 2009 IEEE
DOI 10.1109/DICTA.2009.34

Furthermore, this study showed that male faces are more efficiently categorised than female faces. Environmental conditions, such as the level of illumination, can have a significant impact on the performance of the aforementioned visual systems. Consequently, a lot of research has been performed aimed at improving robustness of face recognition under variable illumination conditions. Torre et al. [10] proposed filtered component analysis, instead of PCA, in their appearance models to increase robustness. Gupta et al. [11] introduced 3D image acquisition devices to acquire 3D facial data, which contains additional range data. The use of 3D data resulted in a lower misclassification rate for face recognition. The active appearance model is powerful for modeling deformable visual objects, such as faces, with different illumination and expressions. Saragih extended this model for monocular and stereo applications [12].

Similarly, whilst there has been much work in the area of speaker recognition, there has been little on gender recognition using speech. Although some work has shown that standard voice modeling approaches can provide highly accurate results for gender classification in noise-free conditions [13][14], there has been little research into the robustness of these approaches when operating in noisy conditions. Various approaches have been developed in the past to improve the robustness of speaker recognition in noisy conditions and it is feasible that these approaches would provide improved robustness for gender modeling [15].

An alternative approach to improving the robustness of a gender classification system for the types of corruption mentioned above is to combine different modalities which are unaffected by the same types and sources of corruption. For instance, image and audio information are affected by different types of corruption; hence, when one feature stream is corrupted, the other stream may still be clean and offer high accuracy. Numerous studies have shown the effectiveness of a multimodal approach to person identification [16][17][18]. In this work we take a similar approach to the analogous problem of gender classification. The modalities can be integrated by feature fusion or by score fusion. Feature fusion assumes relationships between the audio and video streams. Decision fusion is performed by combining the scores produced by the separate classifiers. We will be investigating a score fusion approach in this work.

The remainder of the paper is structured as follows: in Sections II and III, we describe the monomodal classifiers we will be using followed by a description in Section IV of the method we take for fusing these monomodal classifiers into a multimodal classifier. Section V then provides details of the experimental results we have obtained using these classifiers. Finally, Section VI draws some conclusions.

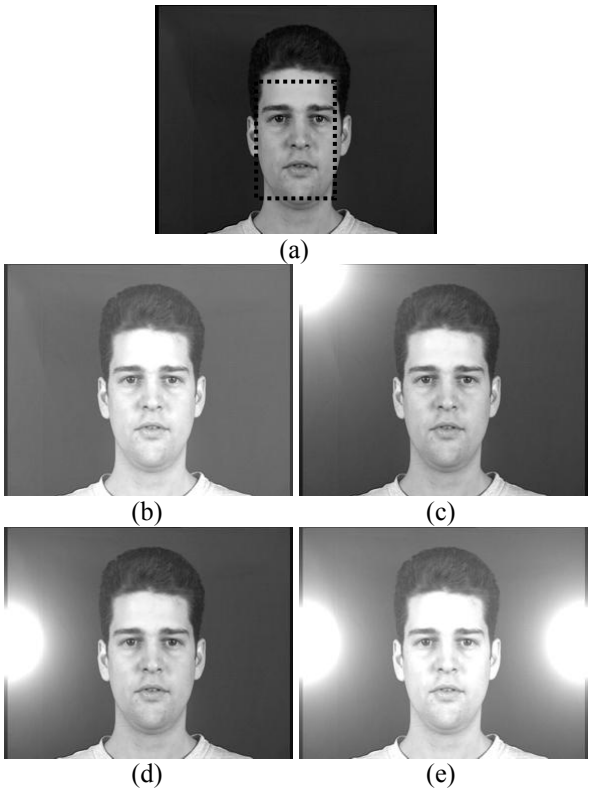

Figure 1. Different lighting effects: (a) Original, (b) uniform lighting change, (c) corner lighting, (d) side lighting, (e) lighting from both sides.

II. VISUAL CLASSIFIER

Given a training set of $m \times n$ images $\{A^t\}_{t=1}^{N_T}$, we can form a set of vectors $\{x^t\}$, where $x \in \Re^{N=mn}$, by lexicographic ordering of the pixels of each image. The basis functions for the eigenspace are obtained by solving

$$\Lambda = \Phi^T \sum \Phi \qquad (1)$$

where $\sum$ is the covariance matrix, Φ is the eigenvector matrix, and Λ is the corresponding diagonal matrix of eigenvalues. In PCA, a partial transform is performed, using a subset of eigenvectors with the largest eigenvalues, to obtain a principal component representation $y = \Phi_M^T \tilde{x}$, where $\tilde{x} = x - \bar{x}$ is the mean normalized image vector and Φ_M is a submatrix of Φ containing the principal eigenvectors. By ranking the eigenvectors with respect to their eigenvalues and selecting the first M principal components, we form an orthogonal decomposition of the vector space into two mutually exclusive and complementary subspaces: the principal subspace $F = \{\Phi_i\}_{i=1}^{M}$ containing the principal components, and its orthogonal complement $\bar{F} = \{\Phi_i\}_{i=M+1}^{MN}$. In PCA the residual reconstruction error is defined as

$$\boldsymbol{\varepsilon}^2(\mathbf{x}) = \sum_{i=M+1}^{N} \mathbf{y}_i^2 = \|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{M} \mathbf{y}_i^2 \quad (2)$$

The labeled PCA representations (y, l) where $l = -1,1$ are then input into a binary SVM which solves the following optimization problem

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=-1}^{1} \xi_i \quad (3)$$

subject to

$$l_i \left[w^T \phi(\mathbf{y}_i) + b \right] \geq 1 - \xi_i \quad (4)$$

where $\xi_i \geq 0$ is the error tolerance of the classifier. Here the training samples are mapped into a high dimensional space by the function ϕ. The SVM tries to find a separating hyperplane with the maximal margin in this high dimensional space. $C \geq 0$ is the error penalty parameter and

$$K(\mathbf{y}_i, \mathbf{y}_j) = \phi(\mathbf{y}_i)^T \phi(\mathbf{y}_j) \quad (5)$$

is called the kernel function. In this paper, the radial basis function kernel is used, which is defined as:

$$K(\mathbf{y}_i, \mathbf{y}_j) = \exp\left(-\gamma \|\mathbf{y}_i - \mathbf{y}_j\|^2\right), \quad \gamma > 0 \quad (6)$$

III. Audio Classifier

The audio features we use are Mel Frequency Cepstral Coefficients (MFCCs), which are a well known method for speech signal parametric representation. They are commonly employed in speaker and speech recognition systems [19], and have been previously shown to work well for gender classification in clean audio conditions [20]. To extract these features, the power spectrum from a 20ms hamming window of speech is pre-emphasized and processed by a mel-scaled filterbank. A DCT is then applied to the log-energy filter outputs to produce the cepstral coefficients. The zeroth order coefficient is not used in the feature vector. The hamming window is shifted 10ms each time in order to produce a feature vector for each 10ms of data. This is the period over which the audio signal is assumed to be reasonably stationary. In this work we extract 12 MFCCs derived from a 26-channel Mel-scale filter bank. This gives a twelve element feature vector for each 10ms of audio data. It is also common in speech and speaker recognition systems to augment these static MFCCs with their first-order delta (Δ) features and second-order delta delta (ΔΔ) features which capture some of the temporal dynamics of the speech signal. In this work we investigate if these additional features are also useful for gender classification.

The audio modeling approach adopted in this work is inspired by the approach taken to model individual speakers voices in many speaker identification or verifications systems. We model each gender using a Gaussian mixture model (GMM). A detailed description of GMMs and their application to acoustic modeling is provided in [21] and will not be repeated here. During the training phase the feature vectors in each voice sample are used to estimate a GMM for the corresponding gender. During the testing phase, the gender GMMs are used to calculate likelihoods for each acoustic frame for the unknown speech segment and the gender with the maximum mean likelihood for the speech segment is chosen. For this work we found through experimentation that 32 mixtures was sufficient to produce almost perfect classification results in clean conditions and no significant improvement was gained by increasing the dimensionality beyond this. Therefore 32 mixture GMMs where used for all of the experimental work described later.

IV. Multimodal Classifier

As described in [22], an important first step in combining classifiers is to normalize their scores in order to make them compatible for fusion and remove any bias caused by differing scales. In this work we found that *z-score* normalization provided slightly better results than alternatives such as *min-max* normalization and so this method was used in all the experimental work presented later. The z-scores for each gender from each classifier are calculated as follows:

$$z_g = \frac{s_g - \mu_g}{\sigma_g} \quad (7)$$

where s_g is the classifier output score for a specific gender, μ_g is the mean score output by the classifier for correct samples of this gender and σ_g is the corresponding standard deviation. The values of μ_g and σ_g are estimated from a set of tests using a held-out portion of the test data.

TABLE I. Summary of fusion techniques

Sum of Scores	$z_A + z_V$
Maximum Score	$\max(z_A, z_V)$
Where z_A is the z-score from the audio classifier and z_V is the z-score from the visual classifier	

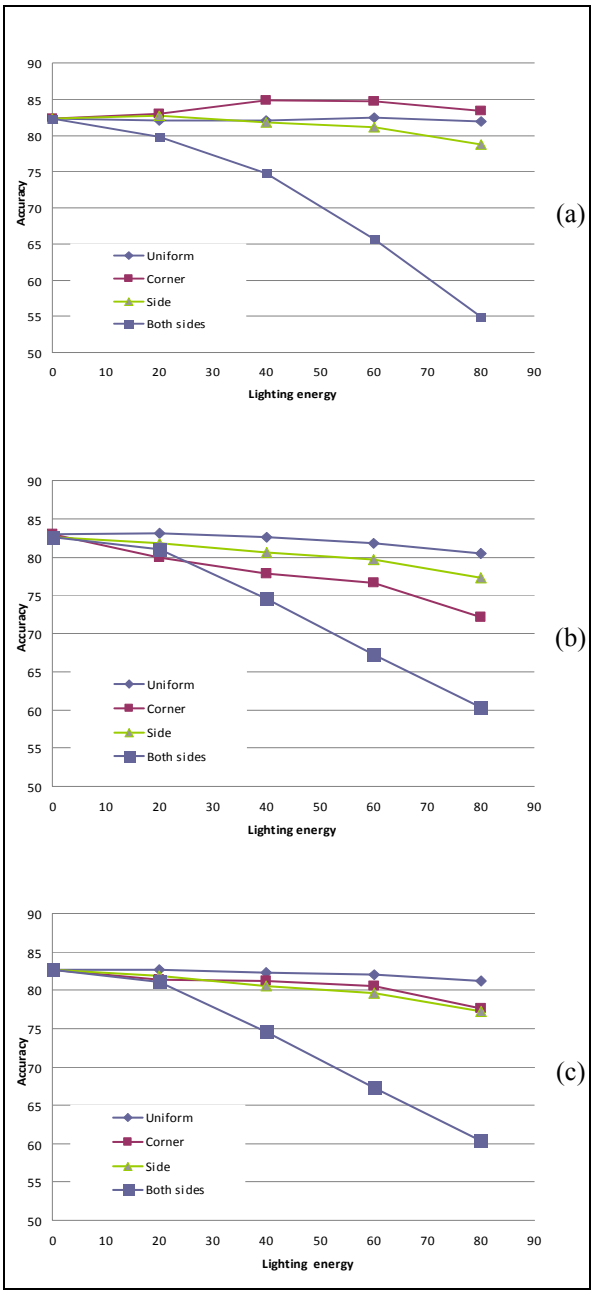

Figure 2. Percentage of correct visual-based gender classifications under 4 different illumination conditions for (a) Males, (b) Females, (c) Average for both Males and Females.

Having normalized the scores, there are a number of well-known alternative fusion techniques which can be used. The correct fusion algorithm for an application depends on the level of correlation between classifiers and also the extent to which the classifiers will be subject to noise and classification error. In [23] it was shown that when combining conditionally independent classifiers, as is the case in this work, the simple and intuitive *Sum* of classifier scores provided the most robust fusion method. The Sum method was demonstrated to be less sensitive to errors caused by noisy observations than other possible approaches, such as the Product. Therefore in this work we have used the Sum of Scores method for fusion and for comparison purposes we also used the Maximum Score as outlined in Table I.

V. EXPERIMENTS

In this section we describe three sets of experiments: visual classification, audio classification and multimodal classification.

A. Visual Classification

In visual classification, a database of 1841 female, and 1918 male face images was used. This was obtained online from [24]. Two thirds of the faces from each group were used for training and the remainder for testing under normal illumination conditions. For training and testing the face detection module extracts a normalized 24×24 template of each face region without hair as indicated in Figure 1. (a) by the area within the dotted rectangle. In addition, four different lighting conditions were applied to the test images; uniform, corner, single side and both sides. This was done by simulation enabling control of both direction and strength of illumination. Example images are shown in Figure 1. (b)-(e). We tested the visual classifier using images which have matching, Figure 1. (a), and mismatched, Figure 1. (b)-(e), illumination conditions to measure its robustness to these effects.

The results obtained under different lighting conditions are shown in Figure 2. , which shows that uniform lighting has almost no effect, other than a slight decrease in accuracy when the illumination starts to cause saturation. In the case of corner lighting, the female classification accuracy drops off significantly as the degree of illumination increases. Surprisingly, the male classification increases to a maximum with increasing illumination, before starting to drop off as the illumination further increases. Side illumination results in a slight decrease in accuracy with increasing illumination for both male and female. Lastly, illumination of both sides has the largest effect, with the classification accuracy dropping almost to chance with increasing illumination, particularly for male images.

B. Audio Classification

For this work we use the well known TIMIT database [25]. There are 630 distinct speakers in the database and each speaker has provided several phonetically rich utterances that have an average duration of approximately 3 seconds. In total this provides 1360 female and 3260 male utterances for training and 560 female and 1120 male utterances for testing. No speakers used in training the system are also used in the test data.

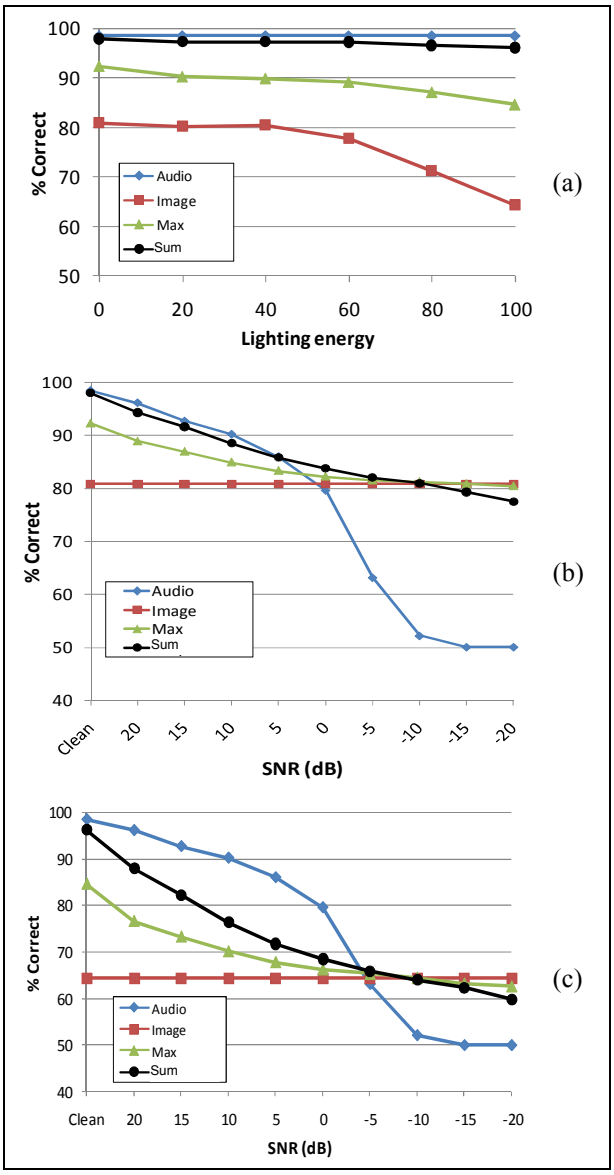

Figure 3. Mean percentage of correct classifications using monomodal classifiers and multimodal classifiers with different fusion techniques in 3 different types of corruption: (a) Clean audio data combined with images illuminated on both sides, (b) Clean image data combined with noisy audio data at various SNRs, (c) Noisy image data with illumination on both sides (lighting energy =100) combined with noisy audio data at various SNRs.

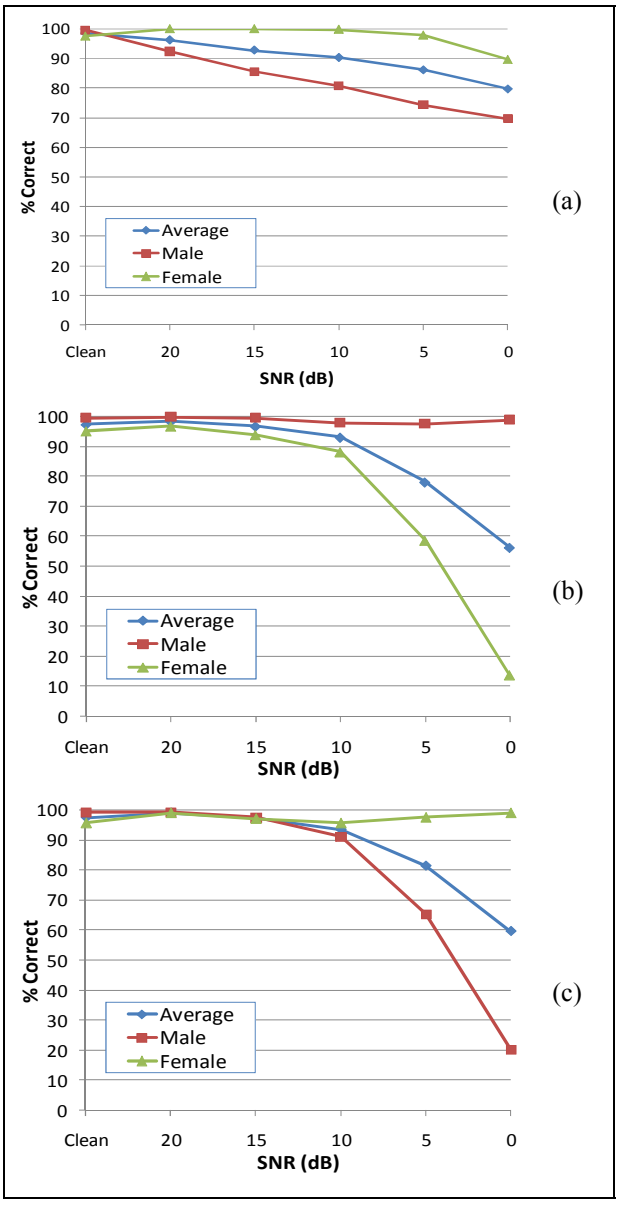

Figure 4. Percentage of correct audio-based gender classifications for clean and noisy speech data at various SNRs using 3 different feature vector types: (a) 12 mfcc, (b) 12 mfcc + Δ, (c) 12 mfcc + Δ + ΔΔ

In these experiments we are comparing the performance of three audio classifiers employing the following types of audio features described previously in section 3. These are: (a) 12mfcc, (b) 12mfcc + 12Δ, and (c) 12mfcc + 12Δ + 12ΔΔ. As with the visual modality we also tested the audio classifiers using both clean test data and also corrupted test data formed by adding full band white noise to the clean test data at a variety of Signal to Noise Ratios (SNRs).

The results are shown in Figure 3. It can be seen that for all three feature vector types the average classification rate under clean audio conditions is almost 100%. However, as the noise level increases the male classification rate for feature vector (a) drops off steadily as the SNR decreases whilst the female classification rate actually increases slightly for some low levels of noise before beginning to decrease after 5dB SNR. Feature vectors (b) and (c) appear to demonstrate an improvement in robustness for male classification compared to (a) at low levels of noise (above 10dB SNR). However, the average classification rate then decreases rapidly after 10dB SNR. Interestingly, it can be seen that the noise tends to have both a masculinising effect at low SNRs for feature vector (b) and a similarly feminizing effect on (c) under the same conditions.

It would be undesirable for a classifier to become strongly biased towards a particular class when noise is added as this would make fusion with other modalities less reliable. A more desirable response for a classifier to increasing noise levels would be for the classification rate for all classes to degrade together, i.e. the noise would have the effect of reducing the output scores for all classes. Feature vector (a) appears to behave in this manner more closely than either (b) or (c) and hence we will be using feature vector (a) in the multimodal classification system.

C. Multimodal Classification

Although there are a wealth of resources available which provide either face or voice samples for monomodal classification tests, there are very limited resources available which provide face and voice samples taken from the same individual. To overcome this problem we have made the reasonable assumption that the gender characteristics of a person's face and voice are conditionally independent given the gender of the person. This assumption allowed us to create a new large database for multimodal gender classification using the same resources we used for monomodal tests. We achieved this by systematically pairing each of the 1120 male voice samples from the TIMIT test data along with each of the 641 male face images from our face database described in section 2. This provided 717,920 male test cases. The same pairing process was carried out for the female test data producing 344,400 female test cases.

Pairing all of the speakers with all of the faces means that the system was tested in all possible cases which might arise. For instance the male faces which would score highly as masculine would be paired with male voices which would score highly as masculine, borderline as masculine and also low as masculine or even as feminine. The same extensive set of combinations would be tested for females. This provides a more complete set of tests than pairing random selections of faces and voices which could lead to favorably/unfavorably biased results.

Although we have tested the multimodal classifiers in all of the image and audio corruption conditions demonstrated earlier for the single modal systems, space determines that we can only present a subset of these results. The experimental results for this subset are illustrated in Figure 4. and demonstrate the nature of the multimodal system in three important conditions:

(a) clean audio data combined with noisy visual data,
(b) noisy audio data combined with clean visual data and
(c) noisy audio data combined with noisy visual data

It can be seen from Figure 4. (a) and (b) that the Sum fusion method produces results which are close to the performance of the best modality in each test condition. In (a) the results are very similar to the Audio-only results even though the visual modality is degrading. In (b) the SNR has been significantly reduced to -20dB in order to illustrate the response of the multimodal classifiers when the audio modality becomes less reliable than the visual modality. It can been that the Sum method tracks the performance of the audio modality until 5dB after which it tends to stay closer to the more reliable image modality. In some conditions (between 0dB and -5dB) the Sum and Max methods both slightly outperform either of the modalities on their own. In all but two of the tested conditions (-15dB and -20dB) the Sum method outperforms the Max method.

In (c) where the reliability of the two individual modalities is much poorer, the fusion of the results using either method is less impressive. This is because in many of the tested conditions the differences between the male and female scores produced are marginal due to the high level of corruption.

VI. Conclusion

An audio-visual gender classification system has been described in this paper, which uses SVMs along with PCA features for visual classification and GMMs along with MFCC features for audio classification. The scores from each classifier are normalized using z-score normalization before fusion using a Sum of Scores approach. In the experiments it was shown that in most of the tested conditions the multimodal classifier was able to automatically perform in a similar manner to whichever individual modality was currently the most reliable. Hence it would be able to adapt to changing environmental conditions in which either one of the modalities could become corrupted. Such a multimodal classifier would be useful in situations where the environmental conditions will change over time and are difficult to predict.

Acknowledgment

This work was supported by the UK EPSRC under Grant EP/E028640/1 ISIS.

References

[1] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation", IEEE Trans. Pattern Analysis & Machine Intelligence, vol.19, pp.696-710, 1997.

[2] Rahimi, B. Recht, and T. Darrell, "Learning Appearance Manifolds from Video", Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, Washington, DC, USA, pages 868—875, 2005.

[3] Moghaddam and M.-H. Yang, "Learning Gender with Support Faces", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pages 1-5, July 2002.

[4] Buchala, S., Davey, N., Gale, T. M., and Frank, R. J., "Analysis of Linear and Nonlinear Dimensionality Reduction Methods for Gender Classification of Face Images," International Journal of Systems Science, 2005.

[5] S. Buchala, N. Davey, T.M. Gale, and R.J. Frank. "Principal component analysis of gender, ethnicity, age, and identity of face images", In IEEE ICMI 2005, 2005.

[6] Buchala, S., Davey, N., Gale, T. M., and Frank, R. J. and Loomes, M., "The Role of Global and Feature Based Information in Gender

Classification of Faces: A Comparison of Human Performance and Computational Models", Int. Jour. Of Neural Sys., 15, pp. 1-8, 2005.

[7] BenAbdelkader, C., and Griffin, P., "A Local Region-based Approach to Gender Classification From Face Images", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[8] Makinen, E. and Raisamo, R., "Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces", IEEE Trans. PAMI,30, pp. 541-547, 2008.

[9] Cellerino, A., Borghetti, D., and Sartucci, F., "Sex differences in face gender recognition in humans", Brain Research Bulletin, 63, pp. 443-449, 2004.

[10] FD Torre, A. Collet, M Quero, JF Cohn, T Kanade, "Filtered Component Analysis to Increase Robustness to Local Minima in Appearance Models" , Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 665-672, June 2007.

[11] S Gupta, JK Aggarwal MK. Markey, AC Bovik, "3D Face Recognition Founded on the Structural Diversity of Human Faces" Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 643-649, June 2007.

[12] J Saragih, R Goecke, "Monocular and Stereo Methods for AAM Learning from Video" Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 680-687, June 2007.

[13] P. Moreno and S. Agarwal, "An Experimental Study of EM_Based Algorithms for Semi-Supervised Learning in Audio Classification", in Proc. of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, Washington DC, USA, 2003.

[14] L.Walawalkar, M. Yeasin, A. M. Narasimharmurthy and R. Sharma, "Support vector learning for gender classification using audio and visual cues: a comparison," Pattern Recognition with Support Vector Machines; LNCS 2388, eds. S.-W. Lee and A. Verri, Springer, 2002, pp. 144-155.

[15] J. Ming, D. Stewart, and S. Vaseghi, "Speaker identification in unknown noisy conditions - a universal compensation approach," Proc. of ICASSP, March 2005

[16] C.C. Chibelushi, F. Deravi, J.S. Mason "Audio-visual person recognition:an evaluation of data fusion strategies.", European Conference on Security and Detection, pp. 26-30, 1997.

[17] P.S. Aleksic, A.K. Katsaggelos, "Audio-visual Biometrics", Proceedings ofIEEE, Vol.94, No.11, pp. 2025-2044, 2006.

[18] R. Brunelli & D. Falavigna, "Person identification using multiple cues" IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 17, pp.995-966, Oct 1995.

[19] D. O'Shaughnessy, Speech Communications – Human and Machine, 2nd ed., IEEE Press, New York, 2000.

[20] H. Harb, Liming Chen, "Gender identification using a general audio classifier," IEEE International Conference on Multimedia and Expo - Volume 2 no. 2, pp. 733-736, (ICME '03), 2003.

[21] D.A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models" IEEE Transactions On Speech and Audio Processing, Vol. 3, No. 1, 1995.

[22] Snelick, R., Indovina, M., Yen, J., and Mink, A. "Multimodal biometrics: issues in design and testing", in Proceedings of the 5th international Conference on Multimodal interfaces 2003, ACM, New York, NY, 68-72.

[23] J. Kittler, M.Hatef, R.P. Duin, J.G. Matas, "On Combining Classifiers", IEEE Trans. on PAMI 20(3)(1998) 226-239.

[24] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features", in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1 2001, pages 1-511 – 1-518.

[25] John S. Garofolo, et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia, 1993