Gender Classification via Lips: Static and Dynamic Features

Stewart, D., Pass, A., & Zhang, J. (2013). Gender Classification via Lips: Static and Dynamic Features. IET Biometrics, 2(1), 28-34. DOI: 10.1049/iet-bmt.2012.0021

Published in:
IET Biometrics

Document Version:
Peer reviewed version

Gender Classification via Lips: Static and Dynamic Features

Darryl Stewart, Adrian Pass, Jianguo Zhang

Abstract

Automatic gender classification has many security and commercial applications. Various modalities have been investigated for gender classification with face-based classification being the most popular. In some real-world scenarios the face may be partially occluded. In these circumstances a classification based on individual parts of the face known as local features must be adopted. We investigate gender classification using lip movements. We show for the first time that important gender specific information can be obtained from the way in which a person moves their lips during speech. Furthermore our study indicates that the lip dynamics during speech provide greater gender discriminative information than simply lip appearance. We also show that the lip dynamics and appearance contain complementary gender information such that a model which captures both traits gives the highest overall classification result. We use Discrete Cosine Transform based features and Gaussian Mixture Modelling to model lip appearance and dynamics and employ the XM2VTS database for our experiments. Our experiments show that a model which captures lip dynamics along with appearance can improve gender classification rates by between 16-21% compared to models of only lip appearance.

I. Introduction

The ability to automatically classify an individual according to gender holds potential for a wide range of commercial applications. For example it may be desirable to tailor automatic advertisements according to the individuals gender or perhaps restrict access to gender specific facilities. It may also be used as part of an automatic census system or in security/surveillance related applications. For instance CCTV systems have become widely deployed in the monitoring of passengers on travel networks. Despite this increase in the use of CCTV, the impact on anti-social and criminal behavior has been minimal. Passenger assaults on trains and buses are still a major problem for transport operators. This is largely due to the fact that there is too much CCTV footage to be monitored in real-time by human security operatives. Automatic gender profiling is seen as one of the fundamental tasks for intelligent surveillance on a CCTV system as it may help an automatic system determine the potential threat posed to or by certain individuals or groups of individuals.

With these applications in mind and with improvements in computational ability and classification algorithms there has been a good deal of recent interest in visual based automatic gender classification systems. Much of the work has been concerned with full faces [1] [2] [3] [4], though promising results have also been obtained using alternative representations such as gait [5] or full body images [6] and even hands [7]. It is understood however that individuals under surveillance by CCTV are not always cooperative (knowingly or unknowingly) and it may be difficult to capture certain traits cleanly in video footage. For instance a person's face may be partially occluded by sunglasses and/or a hat. In such circumstances it would be inadvisable to base the classification on their whole face and instead only the unoccluded parts of the face should be used. Indeed, in many real-world scenarios it may be most effective to use a fusion of various classifiers using different traits as they are captured over time as in [8]. For such a system to work there needs to be an understanding of how effective the various parts of the face can be in determining a persons gender.

A. Gender classification from face-parts

While there have been numerous studies on face recognition there have been many fewer studies focused on gender recognition by using faces and even fewer studies on using individual face parts for gender recognition. One comprehensive study into the efficacy of the static appearance of various facial features for gender classification can be found in [9]. In that work, the authors compare the performance of the mouth, chin, nose, eyes, full face and inner/outer faces using still images from both the FERET [10] and XM2VTS databases for a number of different classifiers. Dimensionality reduction was performed using Principal Component Analysis (PCA). Although the work reports conflicting results as to the superior individual facial feature, it is nonetheless the mouth and eyes that perform consistently well providing results comparable to that of the global facial features. Although no dynamic information is considered, the work clearly demonstrates that mouth *appearance* contains significant discriminatory information. In [11] the authors combine head and mouth movement with facial appearance using a bespoke video dataset in a what we classify as a *speaker-dependent* gender classification task. By speaker-dependent we do not mean that the same data has been used for training and testing the system, we mean that the system has been trained using data for a specific set of speakers and then tested using new data for the same set of speakers. Head and mouth movement is accounted for using normalised tracking coordinates along with mouth width and height parameters, whilst facial appearance is encoded using a PCA based *eigenface* approach. Modelling is performed using Gaussian Mixture Models (GMMs) for the individual subsystems followed by a score fusion step. No results are given for mouth features alone, though it is clearly demonstrated that they provide additional information complementary to the other modalities. It is to be noted that due to the way in which the GMM handles multiple frames or samples, the final classification decision can be considered a form of *majority voting* of the individual scores from each static video frame when used in this way. In contrast, the authors of [12] attempt to capture the dynamics of the speakers entire face by extracting spatio-temporal, Local Binary Pattern (LBP) features which make use of width X, height Y and time T such that the feature transformation is applied over three orthogonal planes XY, XT and YT. Using Support Vector Machines (SVMs) the authors compare the gender classification performance of this spatio-temporal approach to a spatial only LBP approach with majority voting, for both speaker dependent and independent paradigms. They show that full facial dynamic information can be beneficial to *speaker-dependent* gender classification. However they also conclude that such features can be detrimental in the speaker-independent scenario, and that static only features may be superior when considering the full face.

In this work we wish to investigate whether the dynamic movements of the mouth hold gender specific information which could benefit a gender classification system beyond the mere static appearance of the lips. Unlike the work in [11] we are focusing on a *speaker-independent* task where the people being classified are not used in the training data. This is a significantly different and more challenging problem which is more closely related to the real-world scenarios discussed above where an unknown individual needs to be classified. Also unlike the work in [12] we are focusing purely on the mouth region of the face rather than the full face as again we are interested in the efficacy of the mouth region for applications where the face is partially occluded.

To be very clear, in this work it is not our intention to propose the lips as a better form of gender biometric than the full face or even other parts of the face. Instead we are specifically focusing on the utility of dynamic lip movements in conjunction with static lip appearance for applications where the face may be partially occluded. As far as we are aware this is the first attempt at performing gender classification using *solely* lip dynamics for a speaker-independent task. Therefore one of the important outcomes of this work is a baseline for comparison by others who perform similar experiments on the widely used XM2VTS database [13].

The rest of the paper is organised as follows. In Section II we describe the methods which have been proposed for modelling lip movements in other related works and describe the approach taken in this work in detail. In Section III we present the experimental results and finally a summary and concluding remarks are given in Section IV.

II. Lip Modelling Framework

Petajan [14] showed that visual information derived from a speakers lip movements may be used as an additional modality in Automatic Speech Recognition (ASR) systems, improving robustness to the effects of noise corruption in the audio. It has since been found that complementary speaker specific information also exists in these lip movements, allowing the creation of robust multi-modal speaker identification/verification systems [15] [16]. Additional dynamic modalities such as this make it much more difficult for an impersonator attempting to fool the system, whilst the use of individual facial features can improve robustness to partial facial occlusions such as sunglasses, or a shadow cast by headgear. Speech and speaker recognition using lips are analogous problems to the one we are investigating in this paper and so we can learn a great deal about how to capture and model the appearance and dynamics of the lips by examining the features and model types used for those problems.

A. Lip Features

Lip features are usually extracted from the video frames using a process similar to that shown in Figure 1. Depending on the content of the video (i.e., does it contain more than one speakers face), it may be necessary to start with a face detection stage which returns the most likely location of the speakers face in the video frame. The consecutive stages of face localization and mouth localization provide a cropped image of the speakers mouth.

The lip parameterization stage may be geometric based or image transform based. Petajan's original system [14] is an example of geometric based feature extraction which used simple thresholding of the mouth image to highlight the lip area, and then measurements of mouth height, width and area were

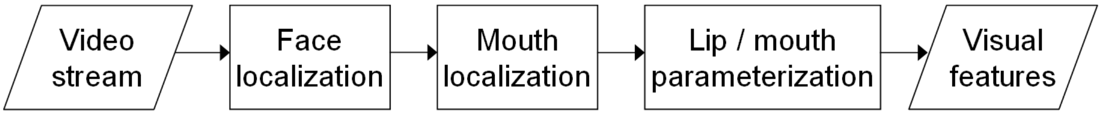

Fig. 1. The general process of visual feature extraction.

taken from that. Since then many approaches have been developed which exploit our knowledge of the shape of a human mouth to fit more complex models to speaker's mouths [17]–[19].

Whereas geometric methods utilize knowledge of the structure of the human mouth to extract features which describe its shape, image transform methods attempt to transform the image pixel values of each video frame into a new lower-dimensional space, which removes redundant information and provides better class discrimination. As with geometric-based approaches, there have also been numerous studies using different image transform methods. These methods include Discrete Cosine Transform (DCT) [20]–[22], Discrete Wavelet Transform (DWT) [23], Principal Component Analysis (PCA) [20], Linear Discriminant Analysis (LDA) [24].

In [20] Potamianos et al. give a comparison of DCT, DWT, Walsh, Karhunen-Loève transform (KLT) and PCA transforms and concludes that the DWT and DCT transforms are preferable to other transforms such as PCA which require training. They also tested the features under several noisy video conditions including video field rate decimation, additive white noise and JPEG image compression and showed that image transform based features are quite robust to these conditions. Other similar studies [25] have drawn the same conclusion on the effectiveness of DCT features for modelling the appearance of the lips and based on this we will be applying the DCT transform to extract features of the static lip appearance in each video frame. A common and widely accepted approach for estimating dynamic features of the lips has been to calculate the first and second order derivatives of the static features which correspond to the velocity and acceleration of the DCT components and these will be the dynamic features we will be applying in our work.

We follow a standard Discrete Cosine Transform (DCT) based feature extraction process which has been shown to be state of the art for visual speech recognition [25]. Following mouth region of interest (ROI) cropping using the mouth tracking coordinates supplied with the dataset detailed in section III-A,

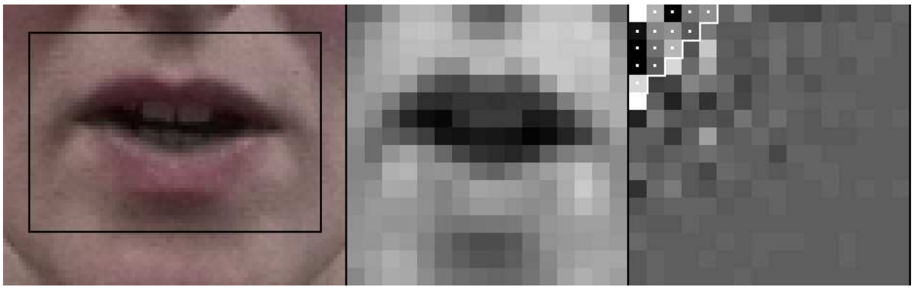

Fig. 2. From left to right: original lip image, sub sampled 16×16 ROI, DCT output showing 5×5 triangle coefficient selection.

the video frames were then converted to greyscale, sub-sampled to 16 by 16 pixels and a 2D DCT applied. The top 15 high energy coefficients were taken for each frame in a zigzag pattern from the top left of the DCT to create the per frame static feature vector. Figure 2 illustrates this approach.

At this point, where used, 1st and/or 2nd order derivative features were calculated across each session, corresponding to velocity and acceleration of DCT components, and concatenated to make the total feature vector. In this work we consider all 7 concatenations of static and 1st/2nd order derivatives. Finally the features were mean and variance normalised across each session individually.

B. Lip Models

In the visual *speech* recognition domain it is common to employ *Hidden Markov Models* (HMMs) to model each unit of speech (i.e., the words or sub-word units known as visemes). Each HMM uses a number of states to model temporal changes in the signal, and each state uses a *Gaussian mixture model* (GMM) with a small number of mixtures to model visual variation in the features [25]. In the visual *speaker* recognition domain, where the aim is to model the speaker independently of the text which is spoken, each individual speaker can be modeled using a single GMM with typically a very large number of mixtures [26]. Our intention in this work is to model purely the characteristics of the two genders independent of the specific content of the speech. In that way the system will be able to classify the persons gender regardless of the words they speak. Therefore, our problem can be viewed as analogous to the speaker recognition problem where in this case we have only two identities to model, i.e. male and female. Therefore we will model each identity, i.e. gender, using a single GMM to capture all the variation in the features for that gender.

We adopt the following GMM approach to modelling the two gender classes, with a likelihood function of the form;

$$p(o|\lambda) = \prod_{t} \sum_{k=1}^{K} w_{tk} g_{tk}(o_t) \tag{1}$$

The summation is over all k mixture components of Gaussian g with corresponding weights w, whilst the product over time t allows for variable length observation sequences. The classified gender of a particular sequence o can then be found as the model emitting the highest accumulative log-likelihood, i.e.

$$\underset{\lambda \in \{\lambda_{male}, \lambda_{female}\}}{\operatorname{argmax}} \quad p(o|\lambda) \tag{2}$$

All GMMs in this work used 64 mixture components, this being found optimal during preliminary testing, with a diagonal covariance matrix. Models were initialized by uniformly segmenting training utterances, followed by training via Expectation-Maximisation (EM).

III. Experimental Setup

A. Database

In order to set a baseline for comparison by others, this work utilises the Lausanne protocol of the publicly available XM2VTS dataset [13]. The dataset consists of 295 subjects split into 158 males and 137 females, uttering the digits 0 ('zero') to 9, with each subject uttering 20 digits per session over 8 sessions, providing 160 digits per subject in total. 109 males (872 sessions) and 91 females (728 sessions) were used for training, with the remaining 49 males (392 sessions) and 46 females (368 sessions) used for testing. Some sample images form the database can be seen in Figure 3 For the baseline results, each session of 20 digits is considered to be a single training or testing sample. However we also provide results using individual digits in order to investigate the efficacy of different mouth movements for gender classification. Audio Hidden Markov Models (HMMs) were trained for each individual digit using TIDGITS [27] audio data and the Hidden Markov Toolkit (HTK) [28], enabling forced alignment to be performed on the audio from XM2VTS. Where gaps between digits were calculated, start and end points of digit boundaries were extended to fill these gaps in order to capture visual lip transitions between digits. The process is illustrated in Figure 4

Fig. 3. Example full head shot video frames from XM2VTS dataset

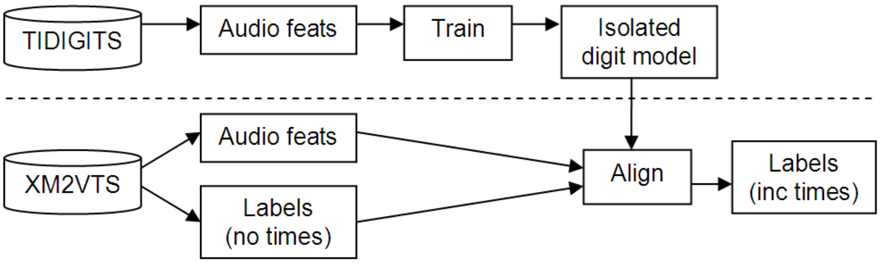

Fig. 4. Process used to isolate individual digits in the XM2VTS database

1) Facial hair: We felt it important in our experiments to isolate and investigate the effect that facial hair had on the performance of the classifier. Therefore we manually inspected and annotated each male video session in which the speaker had visible facial hair. We found there to be 216 sessions in the training data and 148 sessions in the testing data, containing facial hair. Figure 4 shows some example video frames from each set of data. To ensure that our experiments ascertained the gender classification performance of lip movement alone, a second train-test split was also created by omitting these utterances leaving 88 different males for training and 32 for testing, spanning 656 and 244 sessions respectively.

B. Speaker Identification/Speaker-Dependent Gender Classification

Prior to testing our modelling approach for the speaker-*independent* gender classification problem we initially ran some preliminary experiments to verify that the DCT lip features and GMM models were capable of capturing the lip appearance effectively. To do this we carried out some speaker identification experiments using the same features and models. We trained a GMM for each speaker in the XM2VTS database using 6 sessions and then used the remaining 2 for test purposes. We repeated these experiments 6 times using different combinations of static DCT features and 1st/2nd order derivatives.

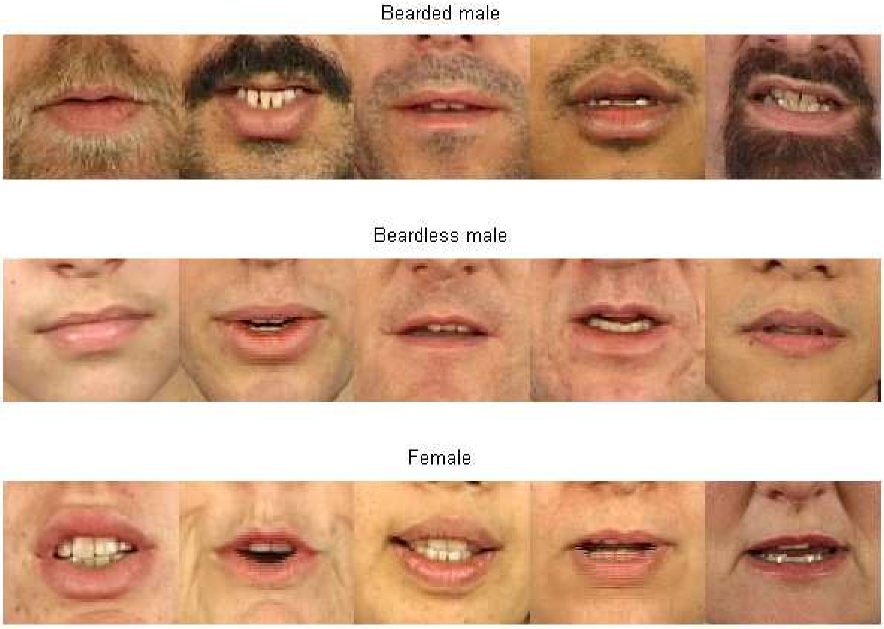

Fig. 5. Example video frames from XM2VTS dataset. Top: Males annotated with facial hair. Middle: Males annotated without facial hair. Bottom: Females.

The results are shown in Table I. It can be seen that in all cases the identification rates are very high which verifies that the DCT features and GMM modelling approach is effective in capturing the variations in lip appearance for each speaker. It can also be seen that the addition of dynamic features improved recognition accuracy significantly. Adding Δ features reduced the error rate by approximately 50%. Adding $\Delta\Delta$ features improved it further but the improvement was much smaller. It is worth noting that Δ features on their own outperformed static features on their own. This suggests that most of the useful discriminatory information is coming from the actual lip movements rather than the static appearance of the lips and mouth.

When the specific errors made by the system were examined it was found that none of the misclassifications crossed gender boundaries, i.e. no males were misreocgnised as females or vice versa. Therefore, if we consider these tests as speaker-dependent gender classification tests then a 100% gender classification accuracy was achieved. Again this is further verification that DCT features and GMMs are suitable for this application.

TABLE I
SPEAKER IDENTIFICATION ERROR RATE OF DCT FEATURES USING DIFFERENT COMBINATIONS OF STATIC AND DYNAMIC FEATURES.

Static	Δ	ΔΔ	Identification Rate(%)
●			97.59
	●		98.80
		●	95.70
●	●		98.80
	●	●	97.77
●	●	●	98.97

Given that this was on a set of 295 speakers (158 males/137 females) these results compare very favourably with other similar speaker-dependent studies such as in [11] where 96.2% accuracy was achieved on a considerably smaller data set consiting of 13 speakers.

C. Speaker-Independent Gender Identification

In this section and in all the remaining sections we report results for the speaker-independent gender classification task which we view as being most useful for real-world applications, as explained in Section I. For this, we trained GMM models for each gender as described in Section II,B.

Tables II and III show the gender identification rates achieved for the full dataset and the subset omitting facial hair respectively, using all 7 combinations of static, 1st and 2nd order derivative (marked Δ and ΔΔ) features. The results give both individual male/wfemale identification rates along with the averaged identification rate. Averaging the score in this way removes any bias resulting from any uneven split of male/female test utterances.

Comparison of the two tables clearly highlights the influence of facial hair on classification scores, particularly in the case of static and Δ features where the male recall rates are most affected by the presence or absence of facial hair. Most of the male specific information appears to be contained within the appearance based static features, even in the absence of facial hair. In contrast, the Δ only features appear to provide more female specific gender information, i.e. complementary to the static features, and so bias classification the other way. It is the ΔΔ features that provide the best average accuracy when comparing individual features in each case, also giving the most equal balance between male and female recall rates. Interestingly the two orders of dynamic features appear to provide conflicting information

TABLE II
IDENTIFICATION RATES (%) FOR ALL FEATURE COMBINATIONS ON FULL DATASET

Static	Δ	$\Delta\Delta$	**Male**	**Female**	**Avg**
●			83.93	73.70	78.81
	●		79.59	81.64	80.62
		●	80.36	81.92	81.14
●	●		84.69	77.26	80.98
	●	●	81.63	76.71	79.17
●		●	85.46	78.36	81.91
●	●	●	85.20	79.18	82.19

TABLE III
IDENTIFICATION RATES (%) FOR ALL FEATURE COMBINATIONS ON SUBSET DATA (NO FACIAL HAIR)

Static	Δ	$\Delta\Delta$	**Male**	**Female**	**Avg**
●			77.46	69.04	73.25
	●		71.72	77.53	74.63
		●	77.05	75.07	76.06
●	●		78.28	76.44	77.36
	●	●	78.28	71.23	74.76
●		●	81.97	74.79	78.38
●	●	●	83.20	75.07	79.13

when combined, as shown by the reduction in female recall rates. This would appear to indicate that some form of decision fusion or feature weighting may be more appropriate than straightforward concatenation of features. Nonetheless, it is the combination of static and dynamic features that provides the best overall classification score of 82.19% for the full dataset.

To further illustrate the effect of lip movement on gender classification, figure 6 shows the gender classification performance when classification is performed based on shorter utterances containing only one digit and using combined static and Δ features. Results are given as raw accuracies and were obtained using two different models. Firstly the GMMs from the previous experiments which were trained using utterances containing all digits were used and secondly, new GMMs trained only on the same corresponding digit were also used, i.e. a GMM was trained on utterances only containing the digit *'one'* and then tested on utterances containing only the digit *'one'*. Although the latter approach appears

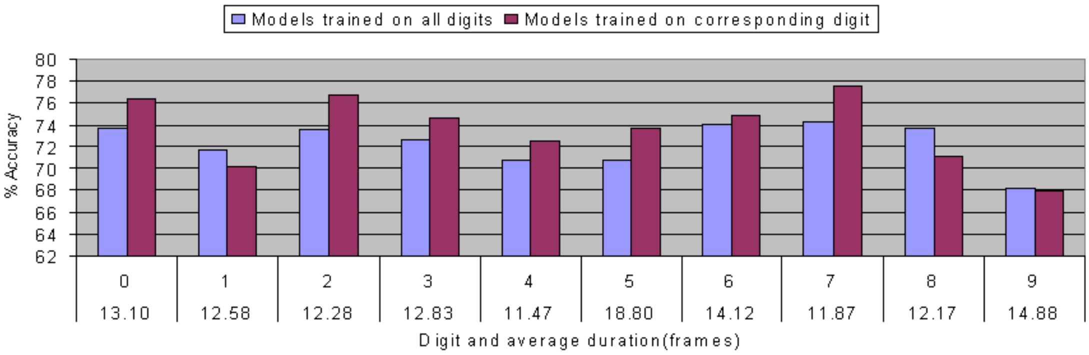

Fig. 6. Gender classification performance of individual digits, using all digits model and individual digits' models, showing average duration of each digit (video frames)

superior, the general trends across digits remain the same and it is the digits that require the most extensive lip/mouth movements that provide the highest overall accuracies. In particular digits *'zero'* and *'seven'* which each contain 2 syllables, along with the digit *'two'* which generally requires a significant pursing of the lips appear to contain the greatest gender specific content. In contrast the digit *'nine'* which is mostly articulated within the mouth cavity, thus producing a predominantly neutral mouth shape, shows the poorest classification performance. It is also worth noting from figure 6 that the duration of an utterance bears very little correlation with classification accuracy, further suggesting dependence upon the *content* of the utterance.

D. Speaker-Independent Gender Verification

As a second performance metric we also report results based on a gender *verification* task using the Receiver Operating Characteristic (ROC) curve of each system. The log-likelihood scores of each test utterance o belonging to the male (λ_{male}) and female (λ_{female}) models was obtained, and then a normalised score was calculated by subtracting the female from the male scores;

$$P(o_{norm}) = P(o|\lambda_{male}) - P(o|\lambda_{female}) \tag{3}$$

The gender classification then depends on a threshold value such that values of $P(o_{norm})$ exceeding this threshold correspond to a male classification and values below correspond to female. From these normalised scores we generated ROC curves depicting the True Positive Rate (TPR) and False Positive

Rate (FPR) for a range of threshold values, and used the Equal Error Rate (EER) and Area Under the Curve (AUC) for each curve as the performance measure. The EER is defined as the point on a curve where the accept and reject error rates are equal, i.e. where it passes through the line FPR = 1-TPR where 1-TPR = False Negative Rate (FNR). A lower EER indicates a system with higher accuracy. The AUC gives a measure of the discriminative ability of the system, i.e. the probability that a random positive sample will produce a higher score than a random negative sample.

Figure 7 shows the ROC curves for the 3 individual feature types and the combination of all 3 using the full dataset (incl. beards). For clarity only the shoulder of the curve is shown. In line with the recall rates the static features alone give the lowest performance, providing the highest EER here of 21.1%. The EER threshold indicates a bias toward male model log-likelihoods which is backed up by the higher male idntification rates in both tables 2 and 3. In addition, the AUC appears to suggest that there may be more of an overlap between male and female scores than those of the dynamic features giving lower discriminative ability. The combination of all 3 feature types gives the lowest EER of 18.36%, however the EER threshold again indicates a bias toward the male model log-likelihoods. The $\Delta\Delta$ features on the other hand achieve a comparable EER of 18.63% at a much lower threshold value, whilst also giving the highest AUC. This suggests that the dynamic features alone provide the most even distribution of male/female scores with the most pronounced separation, thus potentially providing higher discriminative ability than the static lip appearance and indeed the combination of features.

In order to provide some further insight into this last observation, we analysed the kurtosis of the distributions for individual feature components from male and female feature vectors separately. This gave us a measure of the peakedness of their distributions. A higher kurtosis equates to a distribution with a sharper peak about the mean, with most of the variance being caused by infrequent extreme values, in contrast to a lower kurtosis where the variance lies in more frequent and modest values. The difference between the kurtosis of corresponding feature components was calculated across male and female classes, i.e. the female kurtosis was subtracted from the male kurtosis. Therefore a positive difference corresponds to a higher male component kurtosis and vice versa. The results are shown in figure 8 for both static and Δ features separately. Although more prominent in the Δ features, in the majority of cases the difference is positive, showing that the male feature components generally form distributions with a higher kurtosis than those of the females. Furthermore, given that all the speakers in the tests were saying exactly the

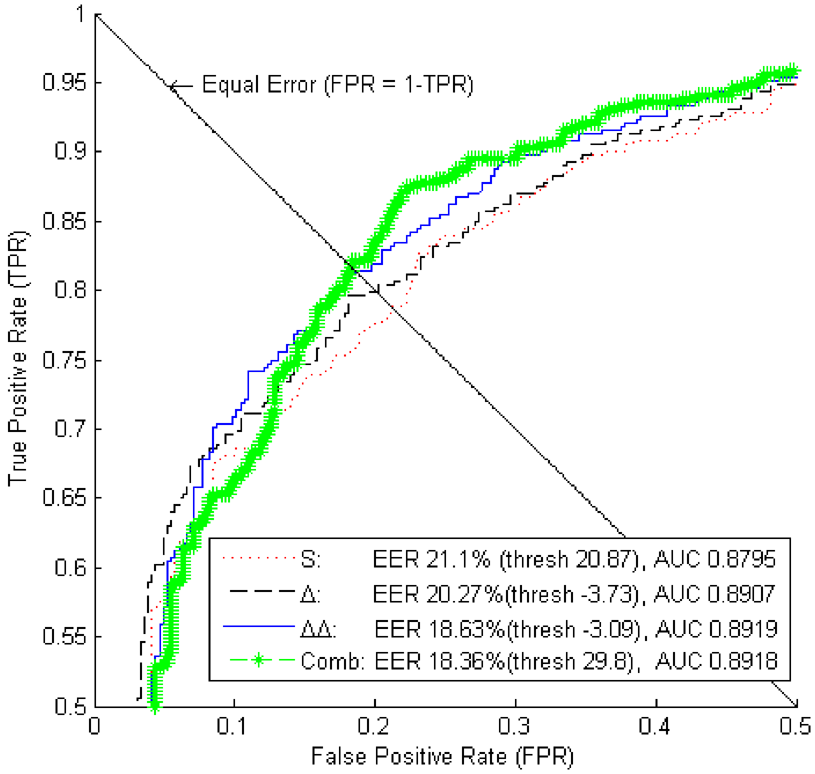

Fig. 7. ROC curves (shoulder only) for individual static/dynamic features and all 3 combined using full dataset (incl. beards). Showing EERs, the thresholds at which EER occurs and AUCs.

same utterances, from the plots in figure 8 it appears that the male subjects generally exhibit a smaller degree of lip movement and velocity during speech than the females with the exception of a few extremes.

IV. Conclusion & Future Work

Automatic gender classification of unknown individuals has a variety of potential commercial and security related applications. In some applications where a person's face may be partially occluded by sunglasses or headgear a gender classification system which uses the full face may not be appropriate. For these challenging applications it is likely that a classification framework which combines the outputs of a series of different classifiers would be used, for instance classifiers based on body shape and any unoccluded facial parts. With these applications in mind, in this paper we focus on the problem of gender classification of unknown people using only the mouth region of the face. The mouth region has

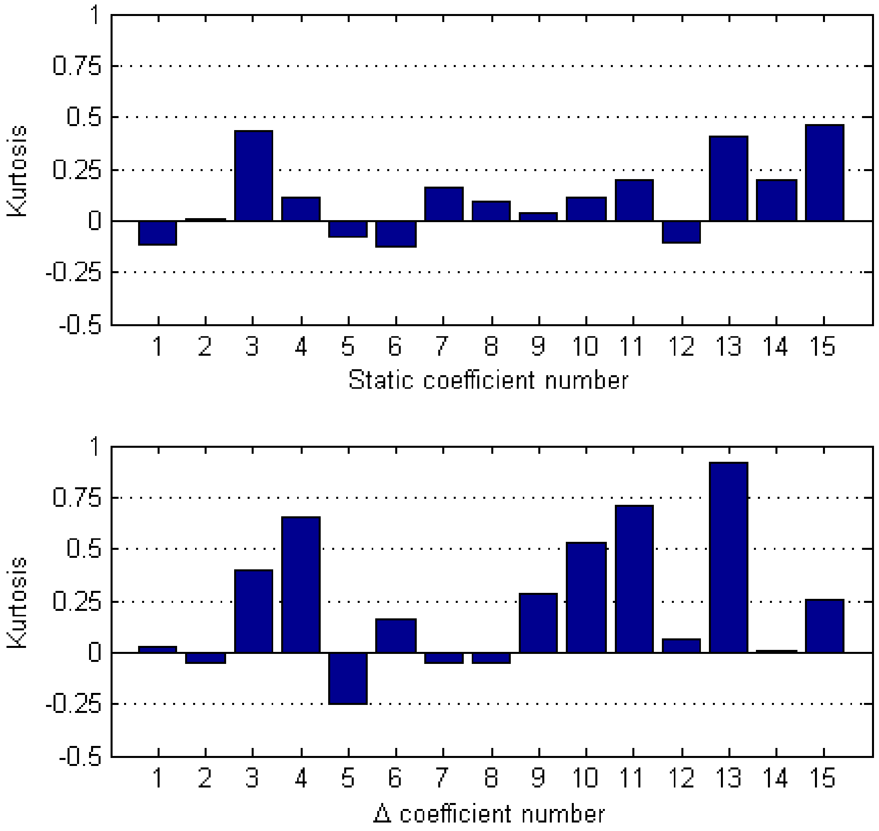

Fig. 8. Difference between Kurtosis of DCT coefficients between male and female features. Top: static features, Bottom: Δ features

the advantage that it is often a part of the face which is left uncovered in order to aid communication. We presented a lip modelling framework based on Gaussian Mixture Models and Discrete Cosine Transforms which captures both the lip appearance and dynamics for males and females. This modelling approach was shown to be highly effective in speaker-dependent gender classification experiments, giving 100% accuracy on the large XM2VTS database. In our speaker-independent experiments it was shown that the dynamics of speakers' lips during speech provide more gender specific information than the static appearance of the lips alone. This has been shown both through analysis of the features and the use of sequences rich in lip movements. We have also shown dynamic and static features to be complementary to one another in terms of the gender specific information they represent, and that the highest overall classification rates are achieved through their combination. To the best of our knowledge this is the first demonstration of the efficacy of lip dynamics themselves for gender classification on unknown speakers. These results demonstrate that lip appearence and dynamics could be a useful additional modality for automatic gender classification, particularly under conditions of partial facial occlusion. Possible extensions to this work which we will be investigating in the future include an examination of how the system performs under different illumination conditions and changes in the speaker's pose. We also will be investigating ways in which this work can be integrated with other modalities such as audio-based gender classification for applications where both the video and audio stream is present.

ACKNOWLEDGMENT

This work was supported by DEL and the UK EPSRC under Grant EP/E02-8640/1 ISIS.

REFERENCES

[1] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *PAMI*, vol. 30, pp. 541–547, March 2008.

[2] D. Stewart, H. Wang, J. Shen, and P. Miller, "Investigations into the robustness of audio-visual gender classification to background noise and illumination effects," in *Proceedings of the 2009 Digital Image Computing: Techniques and Applications*, DICTA '09, (Washington, DC, USA), pp. 168–174, IEEE Computer Society, 2009.

[3] B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Trans.Pattern Anal.Mach.Intell.*, vol. 24, no. 5, pp. 707–711, 2002.

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

[5] C. Shan, S. Gong, and P. W. McOwan, "Learning gender from human gaits and faces," in *AVSS '07: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, (Washington, DC, USA), pp. 505–510, IEEE Computer Society, 2007.

[6] M. Collins, J. G. Zhang, P. Miller, and H. B. Wang, "Full body image feature representations for gender profiling," in *VS09*, pp. 1235–1242, 2009.

[7] G. Amayeh, G. Bebis, and M. Nicolescu, "Gender classification from hand shape," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, p. 1, 23-28 2008.

[8] J. Ma, W. Liu, and P. Miller, "An evidential improvement for gender profiling.," in *Belief Functions* (T. Denoeux and M.-H. Masson, eds.), vol. 164 of *Advances in Soft Computing*, pp. 29–36, Springer, 2012.

[9] Y. Andreu and R. A. Mollineda, "The role of face parts in gender recognition," in *ICIAR '08: Proceedings of the 5th international conference on Image Analysis and Recognition*, (Berlin, Heidelberg), pp. 945–954, Springer-Verlag, 2008.

[10] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1104, 2000.

[11] F. Matta, U. Saeed, C. Mallauran, and J.-L. Dugelay, “Facial gender recognition using multiple sources of visual information,” in *International Workshop on Multimedia Signal Processing, MMSP*, pp. 785–790, 2008. DBLP:conf/mmsp/2008.

[12] A. Hadid and M. Pietikainen, “Combining motion and appearance for gender classification from video sequences,” in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, p. 1, 8-11 2008.

[13] K. Messer, J. Matas, J. Kittler, J. Lttin, and G. Maitre, “Xm2vtsdb: The extended m2vts database,” in *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pp. 72–77, 1999.

[14] E. D. Petajan, *Automatic lipreading to enhance speech recognition (speech reading)*. PhD thesis, Champaign, IL, USA, 1984.

[15] T. Wark, S. Sridharan, and V. Chandran, “The use of speech and lip modalities for robust speaker verification under adverse conditions,” in *ICMCS ’99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, (Washington, DC, USA), pp. 812–816, IEEE Computer Society, 1999.

[16] A. G. de la Cuesta, J. Zhang, and P. Miller, “Biometric identification using motion history images of a speaker’s lip movements,” in *IMVIP ’08: Proceedings of the 2008 International Machine Vision and Image Processing Conference*, (Washington, DC, USA), pp. 83–88, IEEE Computer Society, 2008.

[17] R. Kaucic, B. Dalton, and A. Blake, “Real-time lip tracking for audio-visual speech recognition applications,” in *Proc. European Conf. on Computer Vision*, (Cambridge, UK), pp. 376–387, 1996.

[18] M. Gordan, C. Kotropoulos, and I. Pitas, “Pseudoautomatic lip contour detection based on edge direction patterns,” in *Proc. of 2nd IEEE R8 - EURASIP Symposium on Image and Signal Processing and Analysis*, (Pula, Croatia), pp. 138–143, June 2001.

[19] R. Goecke, J. B. Millar, A. Zelinsky, and J. Robert-Ribes, “A detailed description of the AVOZES data corpus,” in *In Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing*, (Salt Lake City, USA), pp. 486–491, May 2001.

[20] G. Potamianos, H. P. Graf, and E. Cosatto, “An image transform approach for HMM based automatic lipreading,” in *Proc. of Int’l Conf. on Image Processing*, vol. 3, (Chicago), pp. 173–177, 1998.

[21] P. Císař, M. Železný, J. Zelinka, and J. Trojanová, “Development and testing of new combined visual speech parameterization,” in *Proc. of the Intl Conf. on Auditory-Visual Speech Processing (AVSP 2007)*, (Hilvarenbeek, Netherland), 2007.

[22] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, “DCT-based video features for audio-visual speech recognition,” in *Proc. of International Conference on Spoken Language Processing*, (Denver, Colorado), pp. 1925–1928, September 2002.

[23] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, “A comparison of model and transform-based visual features for audio-visual LVCSR,” in *Proc. International Conference on Multimedia and Expo*, (Tokyo, Japan), p. 210, 2001.

[24] G. Potamianos and H. Graf, “Linear discriminant analysis for speechreading,” in *Proc. Works. Multimedia Signal Process.*, (Los Angeles), pp. 221–226, 1998.

[25] R. Seymour, D. Stewart, and J. Ming, “Comparison of image transform based features for visual speech recognition in clean and corrupted videos,” *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–9, 2008.

[26] D. Dean, S. Sridharan, and T. Wark, “Audio-visual speaker verification using continuous fused hmms,” in *Proceedings of the HCSNet workshop on Use of vision in human-computer interaction - Volume 56*, VisHCI ’06, (Darlinghurst, Australia,

Australia), pp. 87–92, Australian Computer Society, Inc., 2006.

[27] R. Leonard, “A database for speaker-independent digit recognition,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’84.*, vol. 9, pp. 328–331, 1984.

[28] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen, *The HTK Book for HTK V2.0*. Cambridge University Press, Cambridge, UK, 1995.