To appear in Proceedings of the International Conference on Pattern Recognition, Vienna, Austria, 1996.

Learning to Recognise Talking Faces

Juergen Luettin[1,2], Neil A. Thacker[1], Steve W. Beet[1]

[1]Dept. of Electronic and Electrical Engineering
University of Sheffield
Sheffield S1 3JD, UK
{N.Thacker, S.Beet}@shef.ac.uk

[2]IDIAP
CP 592, 1920 Martigny
Switzerland
Luettin@idiap.ch

Abstract

An approach for person identification is described based on spatio-temporal analysis of the talking face. A person is represented by a parametric model of the visible speech articulators and their temporal characteristics during speech production. The model consists of shape parameters, representing the lip contour and intensity parameters representing the grey level distribution in the mouth region. The model is used to track lips in image sequences where the model parameters are recovered from the tracking results. While some of these parameters relate to speech information, others are intuitively related to different persons and we show that models based on these features enable successful person identification. We model the shape and intensity parameters as mixtures of Gaussians and their temporal dependencies by Hidden Markov Models. Identifying a talking person is performed by estimating the likelihood of each model for having generated the observed sequence of features and the model with the highest likelihood is chosen as the identified person.

1. Introduction

Recognising persons is a task which humans perform with remarkable accuracy but which still remains a very challenging problem for computers. Two of the main approaches for person recognition by machine are face recognition [1 , 2 , 3] and speaker recognition [4 , 5 , 6]. Both of these approaches have mainly been treated independently.

The appearance of a face can change considerably during speech and due to facial expressions. In particular, the mouth is subject to fundamental changes but at the same time it is one of the most distinctive parts of a face. Face recognition research has largely ignored appearance changes of the face and focused on static images with neutral facial expressions. Approaches for facial expression analysis have been proposed in [7 , 8 , 9 , 10 , 11]. A face recognition system which accounts for these changes is likely to be more robust to talking faces and facial expressions.

Although visual speech information is well known to provide information, complementary to the acoustic speech signal which can improve speech perception [12], speaker recognition has mainly concentrated on the acoustic signal and ignored the multi-modal nature of speech.

Person recognition combining face recognition and speaker recognition have recently been proposed in [13][14]. In this case, face recognition was performed on static images with neutral expressions and speaker recognition was based on acoustic analysis of isolated digits uttered by the subject.

We describe a new approach for person identification based on spatio-temporal information extracted from the speaking face. We show that these parameters provide important speaker dependent information, which could be incorporated in visual (face recognition), acoustic (speaker recognition) and audio-visual person recognition systems to increase their performance and robustness to impostors.

2. Feature Extraction

Face recognition and facial expression recognition requires detailed analysis of the whole face but most facial motion during speech occurs around the mouth area. We are interested in facial changes due to speech and therefore analyse the mouth region of the talking face. We assume that much distinct information of a given speaker is contained in the lip contours and intensity values around the mouth area. During speech production the lip shape varies smoothly but still

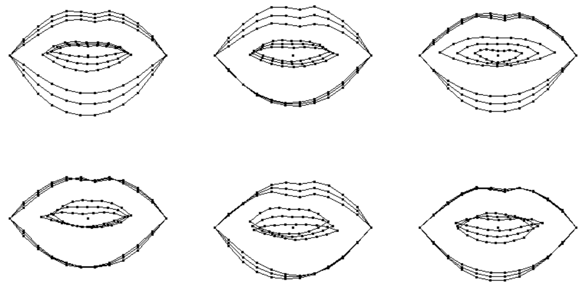

Figure 1: First six principal modes of shape variation captured in the training set across all subjects and over all word sequences.

contains speaker dependent information. We try to exploit this fact by building a spatio-temporal model for each speaker which describes the mouth of the speaker and its temporal changes during speech production. For recognition, the likelihood of these models for having generated the observed mouth features is estimated and the model with the highest likelihood is chosen to be the recognised person.

We use a shape model to describe the lip contours and a profile model to describe intensity values around the lip contours. An approach based on active shape models [15] is used to locate and track the lips over an image sequence. These are deformable models which represent an object by a set of labelled points, in our case the lip contours. The principal modes of deformation are obtained by performing principal component analysis on a labelled training set. Any shape can be approximated by a linear combination of the mean shape and the first few principal modes of deformation.

We use a grey level model for representing the intensity values around the lip contours. It is used for image search to describe the fit between the model and the image. The model describes one dimensional intensity profiles centred at the model points and perpendicular to the contour, similar to the one described in [15]. The difference of our approach is that we concatenate the profiles of all model points to form a global profile vector. Principal component analysis is performed to obtain the main modes of profile variation. Any profile of the training set can be approximated by a linear combination of the mean profile and the first few modes of profile variation. The profile model deforms with the contour model and therefore always represents the same object features.

This method enables robust tracking of the inner and outer lip contour for different persons and lighting conditions. The weights for the principal shape modes and profile modes are used as shape parameters and

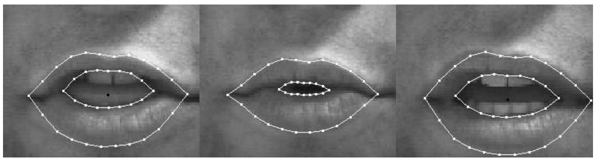

Figure 2: Example images from a sequence of the word "three" with tracking results.

profile parameters, respectively. They are obtained from the tracking results and serve as features for the recognition system. We have described the detailed feature extraction method elsewhere [16][17].

The first few modes of shape variation are shown in Figure 1. The modes account for variation between speakers and variation due to speech production. Figure 2 shows example images of a tracking sequence for a person saying the word "three".

3. Modelling Talking Faces

We follow the approach where the face, or in this work part of the face, is represented in 2D, with no explicit 3D information. This approach is also motivated by psychophysical experiments, which suggest that humans may represent faces in a caricature like manner by considering shape information of individual parts and their spatial relationship [18]. We use the model parameters and their temporal dependencies during speech production for representing a particular talking person.

The shape and intensity parameters are extracted at each time frame to form a frame dependent feature vector. They are invariant to scale, translation, and rotation. The shape parameters are also invariant to illumination. Although scale might contain important speaker dependent information, we did not use it as feature since it was not possible to estimate absolute scale values from the database we used.

The shape and intensity parameters contain both, speech relevant information and speaker dependent information (intensity parameters also contain illumination information). We have previously shown that these features provide important information for visual speech recognition (lipreading) [19]. Here, for speaker identification, the training method has to learn which features contain speaker dependent information rather than speech dependent information.

3.1 Representation

A visual observation $\mathbf{O}$ of a speaker is represented by a sequence of feature vectors

$$\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots \mathbf{o}_T \quad (1)$$

where $\mathbf{o}_t$ is the feature vector extracted at time t. We assume that the feature vectors of a person follow continuous probability distributions which we model by mixtures of Gaussians. We further assume that temporal changes during speech are piece-wise stationary and follow a first-order Markov process. Thus each Hidden Markov Model (HMM) state represents several consecutive feature vectors. These assumptions are not strictly true, but are also often made in the acoustic domain. They can be improved by increasing the number of states which decreases the number of times an observation frame remains at a particular state. Similar modelling techniques have been used for acoustic speaker modelling [6].

A HMM representing a particular person is defined by the parameter set

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi). \tag{2}$$

$\mathbf{A} = \{a_{ij}\}$ is the matrix of state transition probabilities from state i to state j, $\mathbf{B}$ the matrix of observation probabilities $b_j(\mathbf{o})$ for state j and π the vector with probabilities π_i of entering the model at state i. The observation probabilities are modelled as mixtures of Gaussian distributions:

$$b_i(\mathbf{o}) = \sum_{m=1}^{M} c_{im} N(\mathbf{o}, \mu_{im}, \Sigma_{im}) \tag{3}$$

where c_{im} is the mixture weight for state i and mixture m and $N(\mathbf{o}, \mu, \Sigma)$ a multivariate Gaussian with mean μ and covariance matrix Σ. A model of a talking person in shown in Figure 3.

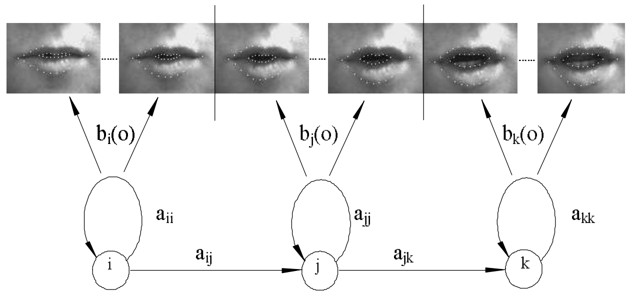

Figure 3: Spatio-temporal model of a talking face for a 3 state HMM with observation probabilities *b*(o) and transition probabilities *a*.

3.2 Training

Speaker recognition tests can be classified into text dependent (TD) and text independent (TI) tasks. For text dependent tasks the test utterance is known while for text independent tasks it is not known. We performed experiments for both TI and TD mode, where TI mode here is restricted by the size of the vocabulary. For the TD mode, we built one HMM per word class and speaker while for TI mode, only one HMM was built per speaker, representing all word classes. In TD mode, the spoken word is known and only HMMs of that word class are used for identification. In TI mode, the spoken word is not known, thus text independent HMMs representing all word classes are used for identification.

We trained HMMs which only allow self-loops and sequential transitions between the current and the next state. The initial state probabilities are set to zero for all states but the first. The remaining parameters are estimated from the extracted model parameters of the training set. Each HMM is initialised by linear segmentation of the training vectors onto the HMM states followed by iterative segmental k-means clustering and Viterbi alignment. The models are further re-estimated using the Baum-Welch procedure, which maximises the likelihood of model λ for having generated the observed sequence $\mathbf{O}$. This is a common training procedure used in acoustic speech modelling [20].

3.3 Identification

Speaker identification is performed using the Viterbi algorithm which calculates the most likely state sequence for each HMM of having generated the observed sequence. Classification is performed by estimating the maximum *a posteriori* probability (MAP)

$$\arg\max_i P(\lambda_i|\mathbf{O}) \tag{4}$$

where λ_i represents the model of the identified person i and $\mathbf{O}$ the observation sequence. The *a posteriori* probability can be obtained using Bayes rule

$$P(\lambda_i|\mathbf{O}) = \frac{p(\mathbf{O}|\lambda_i)P(\lambda_i)}{p(\mathbf{O})}, \tag{5}$$

where $P(\lambda_i)$ represents the prior probability of subject i and $P(\mathbf{O}|\lambda_i)$ the probability distribution of the feature sequence $\mathbf{O}$ for model λ_i. The terms $P(\lambda_i)$, which are assumed to be equal for all subjects, and $p(\mathbf{O})$ are constant for all speakers and can therefore be ignored in the MAP calculation. $P(\mathbf{O}|\lambda_i)$ is simply the product of the transition probabilities a and the output probabilities $b(y)$ of the most likely state sequence.

4. Experiments

Experiments were performed using the Tulips 1 database [21]. It consists of 96 grey-level image sequences of 12 speakers (9 male, 3 female) each saying the first four digits in English twice. We used the first utterance of each word and each speaker as the training set for the HMMs and the second instances as the test set. All images were Gaussian filtered and scaled to compensate for global illumination differences.

Speaker recognition performance is often evaluated as a function of training duration and test duration and typically performance increases if either of those periods is increased. Typical periods used in speaker recognition are 10-30 seconds for training and 3-10 seconds for testing. In our experiments, average training duration was 0.3 seconds for TD tests and 1.3 seconds for TI tests. Average test duration was 0.3 seconds for both tasks. Both periods were therefore considerably shorter than typical periods used in the acoustic domain. Particularly for TD mode, the small size of the database, providing only one example of each word for a given speaker, caused difficulties in estimating the HMM parameters. We tried to alleviate these problems by applying different training and parameter tying methods [22].

We performed experiments for different feature vectors and different HMM architectures. Best results were obtained with HMMs of 4 states and 3 mixture components. The performance generally increased by using more mixture components, but the database only permitted to train about 3 or 4 components. The feature vector either contained shape parameters or intensity parameters or both. 10 parameters were used for describing the shape and 20 parameters for describing the profile.

4.1 Text Dependent Test (TD)

For TD recognition we built a separate HMM for each subject and each word class, resulting in a total of 48 models. Due to the small size of the database we used a sequential training procedure, where re-estimation is based on the models trained in the previous step:

I. Estimating variances for one *global model* using all training data.

II. Re-estimating means, mixture weights and transition probabilities for *subject independent word models*.

III. Re-estimating the mean and mixture weights for *subject dependent word models*.

All HMMs have therefore the same variances and the transition probabilities of any word class are tied for all subjects. Only the means and mixture weights are estimated individually for each class and each subject.

Identification is based on speaker dependent models of the spoken word. The likelihood for each speaker is estimated and the speaker with the highest likelihood is chosen as the identified person. Results for TD tests are summarised in Table 1. Best performance was achieved by using both, shape and intensity parameters.

	Shape	Intensity	Shape + Intensity
TD	72.9 %	89.6 %	91.7 %
TI	83.3 %	95.8 %	97.9 %

Table 1: Accuracy for text dependent (TD) and text independent (TI) person identification tests using shape and intensity parameters.

4.2 Text Independent Test (TI)

For text independent person recognition we built one HMM for each subject, representing all utterances. The motivation behind this approach is to construct one model which represents different word classes by different mixture components. Parameter estimation was not as critical as in text dependent mode and was performed as follows:

I. Estimating variances, means and mixture weights for one *global model*.

II. Re-estimating mean, mixture weights and transition probabilities for a *text independent speaker model*.

Only the variances are therefore tied for all models. Table 1 shows the results for text independent person identification. Best performance was also obtained by using both, shape and intensity parameters. Although the performance for TI mode is generally worse than for TD mode in the acoustic domain, our system performed better for TI mode. But this is likely to be due to the very small training set for TD mode and the constrained training procedure needed to train the models. For both tasks, performance was higher for intensity parameters than for shape parameters. This could however be due to the first few intensity modes which probably account for different illumination and therefore bias recognition results. The use of a larger database with different lighting conditions could be used to reduce this effect.

5. Conclusions

We have described a new approach for person identification based on spatial and temporal analysis of the talking face. An important property of the extracted facial parameters is their very low dimension and their invariance to scale, translation and rotation. Although we achieved high performance by using these features, they only provide partial information of a talking face. Other speech related information might be contained in the texture of the lips and the shape of the teeth.

Because of the novelty of the approach we were only able to perform experiments on a small database. Considering the small training and test duration, results are very encouraging and demonstrate that lip information is an important cue for person identification which might be used to enhance the accuracy and robustness of current acoustic or visual person verification systems.

Acknowledgements

This work was funded by the University of Sheffield, the German Academic Exchange Service (DAAD) and the European ACTS-M2VTS project.

References

[1] R. Chellappa, C. L. Wilson and S Sirohey, "Human and Machine Recognition of Faces: A Survey", Proc IEEE, vol. 83, no. 5, pp. 705-740, 1995.

[2] A. Samal and P. Iyengar, "Automatic Recofnition and Analysis of Human Faces and Facial Expressions: A Survey", Pattern Recognition, Vol. 25, No. 1, pp. 65-77, 1992.

[3] D. Valentin, H. Abdi, A. O'Toole and G. W. Cottrell, "Connectionist Models of Face Processing: A Survey", Pattern Recognition, Vol. 27, No. 9, pp. 1209-1230, 1994.

[4] S. Furui, "Cepstrum analysis techniques for automatic speaker verification", IEEE Trans. Acoustic, Speech and Signal Processing, vol. 29, no. 1, pp. 254-272, 1981.

[5] G. R. Doddington, "Speaker recognition, identifying people by their voices", Proc. IEEE, vol. 73, no. 11, 1985.

[6] T. Matsui and S Furui, "Concatenated phoneme models for text-variable speaker recognition", Proc. Int. Conf. Acoustics, Speech and Signal Processing, pp. 391-394, 1993.

[7] K. Mase, "Recognition of facial expression from optical flow", IEICE Transactions, vol. E 74, no. 10, pp. 3474-3483, 1991.

[8] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 6, 1993.

[9] Y. Yacoob and L. S. Davis, "Computing Spatio-Temporal Representations of Human Faces", Pro. IEEE Computer Vision and Pattern Recognition, 70-75, 1994.

[10] I. A. Essa and A. P. Pentland, "Facial Expression Recognition using a Dynamic Model and Motion Energy", Proc. Int. Conf. Computer Vision, pp. 20-23, 1995.

[11] A. Lanitis, C. J. Taylor and T. F. Cootes, "A Unified Approach To Coding and Interpreting Face Images", Proc. Int. Conf Computer Vision, 1995.

[12] K. W. Grant and L. D. Braida, "Evaluating the articulation index for audio-visual input", J. Acoustic Soc. Am., Vol. 89, pp. 2952-60, 1991.

[13] R. Brunelli, D. Falavigna, L. Stringa and T. Poggio, "Automatic Person Recognition by Using Acoustic and Geometric Features", Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994.

[14] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 10, pp. 955-966, 1995.

[15] T.F.Cootes, A.Hill, C.J.Taylor and J.Haslam, "Use of active shape models for locating structures in medical images", Image and Vision Computing, Vol. 12, No. 6, pp. 355-365, 1994.

[16] J. Luettin, N. A. Thacker and S. W. Beet, "Active Shape Models for Visual Speech Feature Extraction", D. G. Stork and M. E. Hennecke (eds.), Speechreading by Humans and Machine, NATO ASI Series, Berlin, Springer Verlag, pp. 383-390, 1995.

[17] J. Luettin, N. A. Thacker and S. W. Beet, "Locating and Tracking Facial Speech Features", Proc. Int. Conf. on Pattern Recognition", 1996.

[18] G. Rhodes, S. E. Brennan and S. Carey, "Identification and rating of caricatures: implications for mental representations of faces", Cognitive Psychology, vol. 19, pp. 473-497, 1987.

[19] J. Luettin, N. A. Thacker and S. W. Beet, "Visual Speech Recognition using Active Shape Models and Hidden Markov Models", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1996.

[20] L. R. Rabiner and B. H. Juang, " Fundamentals of Speech Recognition", Prentice Hall, New Jersey, 1993.

[21] J.R.Movellan, "Visual Speech Recognition with Stochastic Networks", G.Tesauro, D.Touretzky, T.Leen (eds.) Advances in Neural Information Processing Systems. Volume 7, MIT Press Cambridge, 1995.

[22] L. R. Bahl, F. Jelinek and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE Trans. Pattern Anal. Machine Intell., Vol. 5, pp. 179-190, 1983.