University of Sheffield, Electronic Systems Group Report No. 95/47
Also appears in IEEE ICASSP, Atlanta, GA, May 1996.

VISUAL SPEECH RECOGNITION USING ACTIVE SHAPE MODELS AND HIDDEN MARKOV MODELS

Juergen Luettin, Neil A. Thacker and Steve W. Beet

Department of Electronic and Electrical Engineering
University of Sheffield, Sheffield, UK
J.Luettin@sheffield.ac.uk

ABSTRACT

This paper describes a novel approach for visual speech recognition. The shape of the mouth is modelled by an Active Shape Model which is derived from the statistics of a training set and used to locate, track and parameterise the speaker's lip movements. The extracted parameters representing the lip shape are modelled as continuous probability distributions and their temporal dependencies are modelled by Hidden Markov Models. We present recognition tests performed on a database of a broad variety of speakers and illumination conditions. The system achieved an accuracy of 85.42 % for a speaker independent recognition task of the first four digits using lip shape information only.

1. INTRODUCTION

It has been shown that the robustness and accuracy of automatic speech recognition can be improved by the use of visual information of the speaker's lip movements in addition to the acoustic speech signal [1]. The main difficulty in incorporating visual information into an acoustic speech recognition system is to find a robust and accurate method for extracting important visual speech features. The two main approaches for extracting speech information from image sequences are the image based approach [1, 2, 3] and the model based approach [4,5].

In the image based approach the image intensities are pre-processed and then used as the feature vector. Pre-processing normally consists of filtering and dimension reduction. The advantage of this approach is that no data is thrown away. The disadvantage is that it is left to the classifier to learn the nontrivial task of finding the generalisation for translation, scaling, rotation, illumination and linguistic variability. Another disadvantage is the high dimensionality and high redundancy of the feature vector.

In the model based approach a model of the visible speech articulators, mainly the lip contours, is built and its configuration is described by a small set of parameters. The advantage of the model based approach is that important features are represented in a low dimensional space and are normally invariant to translation, rotation, scale and illumination. A disadvantage is that a particular model may not consider all relevant speech information. The main difficulty in the model based approach is to build a model which represents the lip shape efficiently and which is able to locate and track the lip contours of different speakers and under different illumination conditions.

We describe a model based speechreading system where a model of the lips is constructed from a training set. The model is subsequently used to locate, track and parameterise lip contours in image sequences. We show how Hidden Markov Models (HMMs) can be used to model visual speech and describe recognition tests purely based on lip shape features.

2. LOCATING AND TRACKING LIPS

Deformable templates [6] have been proposed [4][5] to locate and track lip contours, but because deformation of the model is constrained by the initial choice of polynomials, representing the contour, they are often unable to represent various lip shapes in fine detail. "Snakes" [7] on the other hand are able to resolve fine contour details but shape constraints are difficult to incorporate [8] and one has to compromise between the degree of elasticity and the ability to resolve fine contour details. Image search for deformable templates and "snakes" is normally performed by fitting the model to the edges of the image, assuming strong edges along the lip contours. This assumption is often overestimated as lip edges vary across speakers and depend on illumination, visibility of teeth and mouth opening. Edges on the lower outer lip contour are particularly hard to distinguish and edges inside the mouth often originate from teeth.

We use an approach based on Active Shape Models (ASMs) [9] to model, locate and track lip contours, which is described in detail in [10]. These are flexible models which represent the boundary or other significant loca-

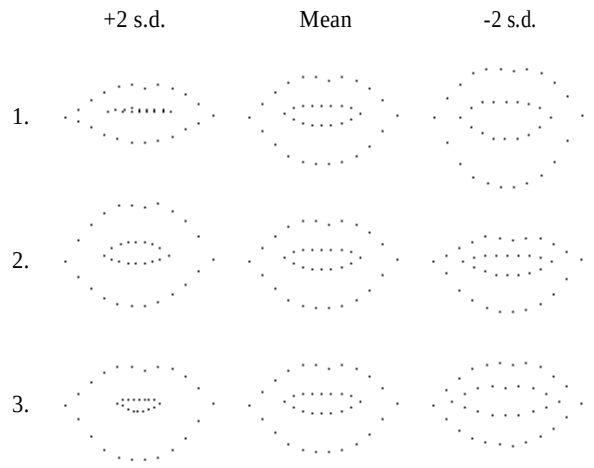

Figure 1: Mean shape and the first three principal modes of variation at ± 2 standard deviations.

tions of an object by a set of labelled points. ASMs use *a priori* knowledge about shape deformation from the statistics of a training set which was labelled by hand. The main modes of shape variation are projected into a linear subspace obtained by Principal Component Analysis (PCA). Any shape can therefore be approximated by a linear combination of the mean shape and the first few main modes of variation. No heuristic limits for shape deformation are used. Instead we constrain each shape parameter to lie within ± 3 standard deviations of the training set which accounts for about 99% of variation.

We built two models of the lips, one representing the outer lip contour and one describing the inner and outer lip contour. Figure 1 shows the first 3 principal modes of deformation captured in the training set for the double contour model.

The models are then used to locate and track lips in image sequences. During image search a cost function is used which measures the fit between the model and the image. We have found that image gradients are inappropriate for representing lip boundaries. Instead we use a profile model which learns typical intensity values around lip contours from the training set. We sample one-dimensional intensity profiles $\mathbf{g}_{ij}$ of length n, perpendicular to the contour and centred at model point i for each training image j, as described in [9], but we concatenate the profiles of all model points of a training image j to form a global profile vector $\mathbf{h}_j$. Similar to describing shape deformation, we constrain the main modes of profile variation, captured in the training set, to lie in a low-dimensional linear subspace which is obtained by PCA. Any profile in the training set can now be approximated by

$$\mathbf{h} = \bar{\mathbf{h}} + \mathbf{P}\mathbf{b} . \qquad (1)$$

where $\bar{\mathbf{h}}$ is the mean profile, $\mathbf{P}$ the matrix of the first column eigenvectors, corresponding to the largest eigenvalues and $\mathbf{b}$ a vector containing the weights for each eigenvector.

The motivation for this approach is to build a model which describes the mean intensity profile of the training set and its main modes of variation which originate from different speakers, different lighting conditions and different "mouth states". For example the profile inside the mouth contains large intensity variation and depends on the mouth opening and the visibility of teeth and tongue.

We use the Downhill Simplex Method [11] for image search, which performs a multi-dimensional minimisation process. The model is first placed at an initial position in the image, then the mean profile is aligned to the image profile $\mathbf{h}$ as closely as possible using the first few modes of profile variation. The profile weight vector is found using

$$\mathbf{b} = \mathbf{P}^T (\mathbf{h} - \bar{\mathbf{h}}) . \qquad (2)$$

The cost E at a particular location and shape is calculated as the mean square error (MSE) between the image profile and the aligned profile model using

$$E = (\mathbf{h} - \bar{\mathbf{h}})^T (\mathbf{h} - \bar{\mathbf{h}}) - \mathbf{b}^T \mathbf{b} . \qquad (3)$$

We assume equal prior probabilities of all shapes within the deformation constraints and therefore do not include a term for shape deformation in the cost function.

Locating the lips in the first frame of an image sequence is performed as described above. For subsequent frames the estimated position and shape of the lips in the previous frame are used as the initial estimate for the search algorithm.

3. VISUAL SPEECH FEATURE EXTRACTION

The parameters describing the shape of the lips are extracted at each time frame and used as visual speech feature vectors. The parameters are invariant to scale, rotation, translation and illumination and can directly be used by the recognition network. The translation and rotation parameters are not used for recognition because they are unlikely to provide speech information.

Much speech information is contained in the dynamics of the lip movements rather than the actual shape. Furthermore dynamics of lip movements might be less sensitive to linguistic variability. We therefore performed some recognition tests by including temporal differences of each feature (delta shapes). Scale might contain relevant speech information but absolute values are hard to estimate and may vary from speaker to speaker. We omitted absolute scale information but we performed

Fig. 2: Examples of image sequences with lip tracking results.

some recognition tests by including scale differences (delta scale).

4. VISUAL SPEECH MODELLING

Visual speech is modelled by representing each utterance as a sequence of visual speech vectors. Their emission probabilities are modelled by continuous Gaussian distributions and temporal changes are modelled by Hidden Markov Models. We used whole-word HMMs and trained one HMM for each word class to be recognised. The models are trained using the Baum-Welch re-estimation algorithm. Recognition is performed using the Viterbi algorithm, which estimates the likelihood for each HMM of having generated the observed sequence and the model with the highest likelihood is chosen as the recognised word. This is a standard approach used in acoustic speech recognition systems [12].

The shape features contain some information which contributes to class discriminability and some information which describes between- and within-speaker variability (linguistic variability). If we have sufficient training data we assume that the recognition network will learn which features contribute to class discriminability and which do not. Since the database we used was very small we performed a variety of recognition tests by using only the first few shape parameters, corresponding to the largest variances, assuming that these parameter estimates are more robust and contain most of the speech information.

Coefficients	Single Contour Model	Double Contour Model
sm	58.33 %	67.71 %
sm + dsm	68.75 %	79.17 %
sm + dsm + ds	80.21 %	85.42 %

Table 1: Word accuracy using one shape mode (sm) with optional delta shape mode (dsm) and delta scale (ds).

5. EXPERIMENTS

Experiments were performed using the Tulips1 database [3] which consists of grey level image sequences of the first four digits. Each digit was spoken twice by 12 individuals (9 males, 3 females). The database reflects a broad variety of speakers and illumination conditions.

Experiments for locating and tracking lips were individually evaluated and are described in detail in [13]. Figure 2 shows examples of lip tracking results using the double contour model. The examples demonstrate that the profile model has learned how the profile at the inner lip contour can change due to mouth opening and visibility of teeth and tongue. The second row also shows that the model is able to track lips which extend beyond image boundaries.

We performed speaker independent recognition tests, using different speakers for training and testing to see how well the system generalises for new speakers. Because of the small size of the database, recognition tests were performed using the ‘jack-knife’ or ‘leave-one-out’ method, i.e. 11 subjects were used for training and the 12th subject for testing. The whole procedure was repeated 12 times, each time leaving a different subject out for testing. The results were averaged over all speakers. A large variety of visual front ends and HMM architectures was used to evaluate the method.

6. RESULTS

Word accuracies of 80.21 % were achieved using the single contour model and 85.42 % using the double contour model. These results demonstrate that lip contours are a rich source of speech information. This is contrary to Bregler and Omohundro [14], who found the outer lip contour not distinctive enough to give reasonable recognition performance.

Best results were achieved with HMMs of 5 or 6 states and one diagonal covariance matrix. This suggests that the training set, consisting of 22 training instances for each class was not large enough to estimate the parameters of HMMs with a full covariance matrix or more than one diagonal mixture component.

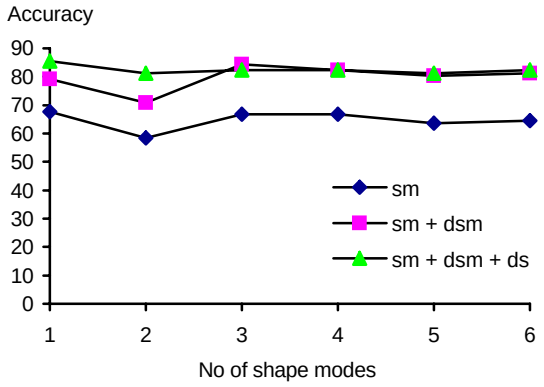

Fig. 3: Recognition accuracy for different numbers of shape modes using combinations of basic shape modes (sm), delta shape modes (dsm) and delta scale (ds).

Using only one shape parameter together with its delta coefficient and delta scale gave the best recognition rate. This might also indicate that the training set was not large enough to model more than the first main shape mode reliably. Table 1 shows results using one shape mode with optional "delta shape" and "delta scale" for 6 state HMMs with one diagonal mixture component. Figure 3 summarises results for different numbers of shape modes included in the feature vector.

7. CONCLUSIONS

We have described a new approach for visual speech recognition based on a data driven lip model and HMMs. Experiments have demonstrated high recognition performance using very low dimensional shape information only.

The recognition task described is relatively simple because it only consists of four word classes and only deals with isolated words. Nevertheless, recognition tests were speaker independent and have demonstrated high recognition accuracy and generalisation ability of the system. More extensive tests with more speakers and sub-word classes are necessary to estimate the discrimination ability of shape features for all phonemes.

Our results are not as good as the ones reported in [3] with 89.58% correct and which was about equivalent to the performance of untrained humans performing the same task. One reason for this might be the absence of additional intensity information particularly about the visibility of teeth and tongue. In the future we plan to extract this information from the profile weight vector and incorporate it in the visual feature vector.

The ability to locate and track lips accurately opens several other potential applications, as example model based image coding, facial animation, facial expression recognition and audio-visual person identification.

ACKNOWLEDGEMENTS

Juergen Luettin is funded by a University of Sheffield Scholarship and the German Academic Exchange Service (DAAD).

REFERENCES

[1] C. Bregler, H. Hild, S. Manke and A. Waibel, "Improved Connected Letter Recognition by Lipreading", Proc. IEEE ICASSP, pp. 557-560, 1993.

[2] B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski, "Integration of Acoustic and Visual Speech Signals using Neural Networks", IEEE Communications Magazine, pp. 75-81, 1989.

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

[4] M. E. Hennecke, K. V. Prasad and D. G. Stork, "Using Deformable Templates to Infer Visual Speech Dynamics", 28th Annual Asilomar Conference on Signals, Systems and Computers, 1994.

[5] R. R. Rao and R. M. Mersereau, "Lip Modeling for Visual Speech Recognition", 28th Annual Asilomar Conference on Signals, Systems and Computers, 1994.

[6] A. L. Yuille, P. Hallinan and D. S. Cohen, "Feature extraction from faces using deformable templates", Int. J. Computer Vision, Vol. 8, pp. 99-112, 1992.

[7] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: active contour models", Int. J. Computer Vision, pp. 321-331, 1988.

[8] C. Bregler and S. Omohundro, "Surface Learning with Applications to Lip-Reading", J.D.Cowan, G.Tesauro and J.Alspector (eds.), Advances in Neural Information Processing Systems 6, Morgan Kaufmann Publishers, 1994.

[9] 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.

[10] J. Luettin, N. A. Thacker and S. W. Beet, "Active Shape Models for Visual Speech Feature Extraction", D. G. Storck (Editor), Speechreading by Man and Machine: Models, Systems and Applications (NATO Advanced Study Institute), Springer Verlag, in press.

[11] J. A. Nelder and R. Mead, "A simplex method for function minimization", Comput. J. Vol. 7(4), pp. 308-313, 1965.

[12] S. J. Young, "HTK Version 1.4: User, Reference & Programmer Manual", Cambridge University Engineering Department, Cambridge, 1992.

[13] J. Luettin and N. A. Thacker, S. W. Beet, "Locating and Tracking Facial Speech Features", submitted to ECCV'96.

[14] C. Bregler and S. M. Omohundro, "Nonlinear Manifold Learning for Visual Speech Recognition", Pro. ICCV, 1995.