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

Locating and Tracking Facial Speech Features

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

This paper describes a robust method for extracting visual speech information from the shape of lips to be used for an automatic speechreading (lipreading) systems. Lip deformation is modelled by a statistically based deformable contour model which learns typical lip deformation from a training set. The main difficulty in locating and tracking lips consists of finding dominant image features for representing the lip contours. We describe the use of a statistical profile model which learns dominant image features from a training set. The model captures global intensity variation due to different illumination and different skin reflectance as well as intensity changes at the inner lip contour due to mouth opening and visibility of teeth and tongue. The method is validated for locating and tracking lip movements on a database of a broad variety of speakers.

1. Introduction

Several researchers have demonstrated the use of visual speech information of the speaker's face, mainly the lip movements, for automatic speech recognition (see for example [1 , 2 , 3 , 4]). However, robust and accurate visual feature analysis is a difficult object recognition problem, because of the large variation between subjects and the changing appearance of a specific subject due to pose, lighting, specularity, makeup and mouth opening. Most previous approaches have constrained or circumvented the feature extraction problem by marking the subject's lips with colour or a reflective marker, by recording the lip movements with a head mounted camera, by using one subject only, by hand segmenting the lip region, by tracking only the outer lip contour or by using very controlled lighting conditions.

We are interested in developing a system which is capable of extracting visual speech information reliably from a broad range of subjects and without the use of artificial aids to enable its operation in real world applications.

This paper describes a method for robust detection, tracking and parameterisation of visual speech information based on active shape models (ASM) [5 , 6 , 7]. We use a grey level profile model for image search, which learns global grey level distribution around the lip contours from a training set. During image search this profile model is used to describe the fit between the image and the model at a particular location in the image. We demonstrate the robustness and accuracy of the method by locating and tracking lip movements on a database consisting of a broad variety of speakers and various lighting conditions. Finally, we describe the extraction of important features which have successfully been used for speechreading and person identification.

2. Related Work

Deformable templates [8] have been used to locate lip contours, where the outline of the lips is modelled by a set of hand coded polynomials. These are matched onto the outline of the lips, which are represented by the image gradient. Since the deformation of deformable templates is constrained by the initial choice of polynomials, they are often not able to resolve fine contour details. The image search is performed by fitting the template to image gradients, assuming strong edges at the lip contours. This assumption is often violated since the gradient along the contour is dependent on the speaker, illumination, makeup, facial hair, visibility of teeth and mouth opening (Figure 1).

Another common approach for shape modelling is based on active contour models or so called snakes [9]. These were used to track facial features which were highlighted by a colour pen [10]. Snakes are able to resolve fine contour details but shape constraints are difficult to incorporate and one has to compromise between the degree of elasticity and the ability to

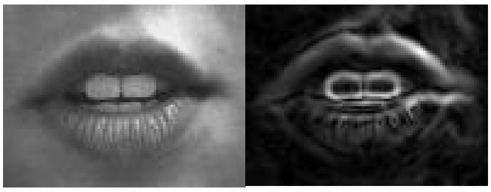

Figure 1: Example image and its gradient image.

resolve fine contour details. Bregler [11] has described a contour tracking method based on snakes, where the contour is constrained to lie in a sub-space learned from a training set. A gradient based image search is performed, similar to the one for deformable templates.

A method based on b-splines and Kalman filters has been described in [12]. A stochastic dynamic model is learned from example sequences which enhances the tracking speed and robustness to distractions. But tracking lips in frontal view was only reliable when a lip high-lighter was worn.

3. Active Shape Models

Our approach for modelling the lip contours is based on active shape models. These are flexible models which represent an object by a set of labelled points. The points describe the boundary or other significant locations of an object. The principal modes of shape variation are learned from a training set which is labelled by hand.

The training examples need to be labelled in a consistent manner in order to be able to compare equivalent points from different shapes. We choose to use the two corner points of the lips as reference points. Their distance is used as scale, their orientation to the horizontal as the angle and the centre between them as the origin. The other points are placed at equal horizontal distance. This definition enables us to map different examples of lip shapes to each other in a consistent manner. We built two different models of the lips: Model 1 describes the outer lip contour and Model 2 describes the outer and inner lip contour.

3.1 Shape Modelling

By using active shape models, we try to avoid the use of heuristic assumptions about legal shape deformation. Instead, *a priori* knowledge about shape deformation is obtained by examining a representative training set. This leads to a compact description of local and global deformation with a small set of parameters. The ith shape in the training set (i = 1 .. N) is described by a vector $\mathbf{x}_i$ with

$$\mathbf{x}_i = (X_{i0} Y_{i0} X_{i1} Y_{i1} \ldots X_{in-1} Y_{in-1})^T \qquad (1)$$

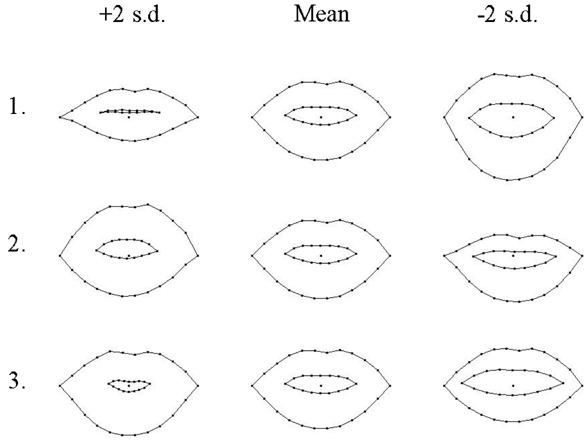

Figure 2: Mean shape and first three modes of variation of Model 2. The first three modes seem to be symmetric and mainly describe the mouth opening. Subsequent modes account for pose and finer contour details.

where (X_{ij} Y_{ij}) is the jth point (j = 0 .. n-1) of the ith shape. The training shapes are normalised by scaling to unit width, zero translation and zero rotation. The average shape $\overline{\mathbf{x}}$ is calculated and used to estimate the covariance matrix of the training shapes. Principal component analysis (PCA) is used to calculate the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors with the largest eigenvalues describe the most significant modes of variation. A normalised shape $\mathbf{x}$ can now be approximated by

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{Pb} \qquad (2)$$

where $\mathbf{P}=(\mathbf{p}_1,\mathbf{p}_2, \ldots \mathbf{p}_t)$ is the matrix of the first t ($t < n$) column eigenvectors corresponding to the largest eigenvalues and **b** a vector containing the weights for each eigenvector. This approach assumes that the principal modes are linearly independent although there might be non-linear dependencies present. Figure 2 shows the mean shape of Model 2 and the first three modes of variation by +/- 2 standard deviation (s.d.). Subsequent shape modes seem to account for pose and finer contour details.

3.2 Intensity Modelling

In order to use Active Shape Models for image search, we would like to have a cost function which measures the fit between the model and the image. We therefore need to find a way of representing dominant image features of the lip contours. The most common approach for representing contours is to use edges or gradients. However, the lip contour can appear in many different ways. The gradient has different values along the contour and is also dependent on speaker, illumination,

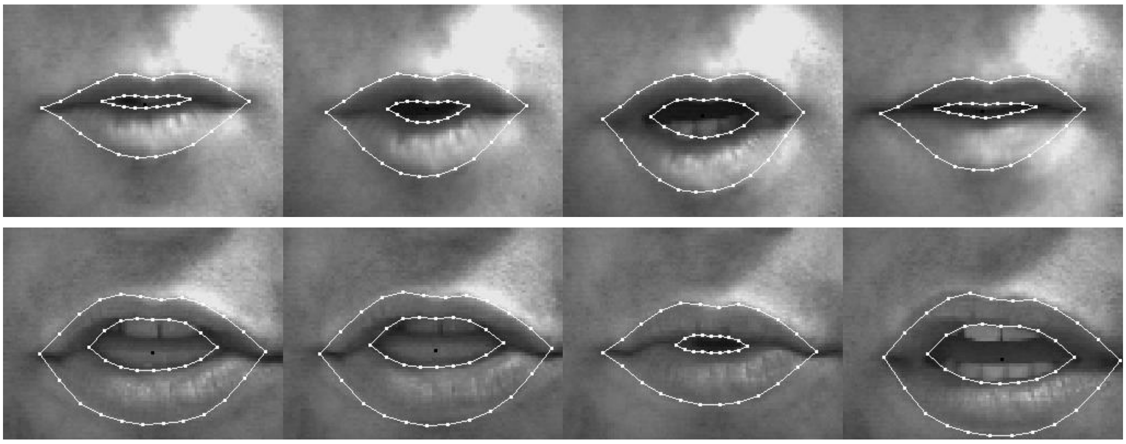

Figure 3: Tracking lips with Model 2. The examples demonstrate that the model finds the outer contour despite low contrast and is robust to intensity changes inside the mouth.

lighting direction, specularity, facial hair, visibility of teeth and mouth opening. The gradient can also change within an utterance, particularly for the inner contour. Figure 1 shows an example image and its gradient image after Gaussian smoothing.

In analogy to the statistical description of lip deformation we want to avoid the use of heuristics for image search and rather use *a priori* knowledge of the actual grey level appearance at the contour as well as its global variation across different examples. Assuming that grey-level changes are not only important at each contour point but also in regions around each point, we capture the statistics of the actual grey level appearance around each model point and estimate their main modes of variation within a training set.

Following [5] we choose to sample one dimensional profiles $\mathbf{g}_{ij}$ of length n_p perpendicular to the contour and centred at point j for each training image i. But instead of calculating individual mean profiles and covariance matrices for each model point, we concatenate the profiles of each model point to construct a global profile vector $\mathbf{h}_j$ for each training image using

$$\mathbf{h}_i = (\mathbf{g}_{i0}\mathbf{g}_{i1}\cdots\mathbf{g}_{in-1})^T . \tag{3}$$

We then calculate a global mean profile $\bar{\mathbf{h}}$ and the covariance matrix from the training set. The eigenvectors and eigenvalues of the covariance matrix are found by PCA.

This approach assumes that the variations of profiles at each model point are correlated as they are expected to be due to illumination effects and different skin and lip reflectance values. The eigenvectors with the largest corresponding eigenvalues describe the main modes of grey level variation seen in the training set and therefore enable the model to describe different lighting, speakers and mouth states. Any profile $\mathbf{h}$ can be approximated using

$$\mathbf{h} = \bar{\mathbf{h}} + \mathbf{P}_g\mathbf{b}_g \tag{4}$$

where $\mathbf{P}_\mathrm{g} = (\mathbf{p}_{g1}\mathbf{p}_{g2}\ldots\mathbf{p}_{gt})$ is the $(n*n_p)\times t$ matrix of the first t column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}_\mathrm{g}$ a vector containing the weights for each eigenvector. An example image with overlaid shape and profile model is shown in Figure 4.

3.3 Cost Function

The cost for the model at a particular location and shape is calculated as the mean square error (MSE) between the image profile and the aligned model profile. The mean profile $\bar{\mathbf{h}}$ is aligned to the image profile $\mathbf{h}$ as closely as possible by calculating the elements of the weight vector $\mathbf{b}_\mathrm{g}$. Since the eigenvectors in $\mathbf{P}_\mathrm{g}$ are orthogonal, the weight vector is found using

$$\mathbf{b}_g = \mathbf{P}_g{}^T(\mathbf{h}-\bar{\mathbf{h}}). \tag{5}$$

The cost E can then be obtained using

$$E = (\mathbf{h}-\bar{\mathbf{h}})^T(\mathbf{h}-\bar{\mathbf{h}}) - \mathbf{b}_g{}^T\mathbf{b}_g. \tag{6}$$

During image search, each shape mode is constrained to lie within +/- 3 s.d. of the training set which accounts for about 99% of variation. We assume equal prior probability for each shape mode within these limits and therefore do not include a term for shape deformation in the cost function.

3.4 Locating and Tracking Lips

For locating lips we assume that a rough estimate of the region of interest (ROI) containing the lips is known from another image processing algorithm. The model is initialised with the mean shape and placed at a random

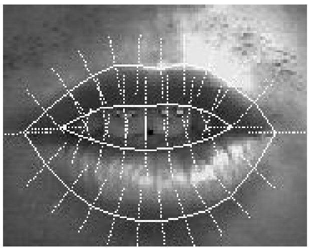

Figure 4: Shape model with profile vectors.

location in the ROI. We use the Downhill Simplex Method (DSM) [13] to find a minimum during image search. The algorithm uses the translation parameters, scale, rotation and the first few shape parameters as variables for the multidimensional optimisation process. The algorithm is initialised with the initial estimate of the model parameters and starts the search by perturbing each parameter by 2 s.d..

For tracking lips over an image sequence, the lips are located as described above for the first frame. For consecutive frames the previous frame is used as the initial estimate of the lip position and the search is performed by DSM. Although constraints could be introduced to limit the search to stay within certain limits during tracking, for simplicity we used the same constraints as for locating the lips. Figure 3 shows tracking examples using Model 2.

4. Experiments

We used the Tulips 1 database [14] for our experiments. It consists of 96 grey level image sequences of 12 speakers (9 male, 3 female) each saying the first four digits in English twice (later described as set 1 and set 2). The use of this particular database makes our experiments very challenging because it contains images of a variety of speakers with different ethnic background, different lip shape, make-up, facial hair and illumination.

The shape and profile models were built using a representative sample of frames of each sequences of set 1. This larger and more balanced training set has lead to higher performance levels then the ones reported in [7], where only the first image of each training sequence was used. Model 1 was represented by 22 points, Model 2 by 38 points and the dimension of the grey level profiles was 19. We used 8 shape modes for Model 1 and 10 shape modes for Model 2. All tests for locating and tracking were performed on set 2.

5. Results

Whereas most previous approaches have avoided performance evaluation of their algorithms, we think that this is important for characterising and comparing the accuracy and robustness of different methods. We choose to judge the performance by visual inspection on the following criteria: A search result was classified as *Good* if the lip contour was found within about one quarter of the lip thickness deviation. It was classified as *Adequate* if the outline of the contour was found between one quarter and half the lip thickness deviation and it was classified as a *Miss* otherwise.

Table 1 : Results for locating lips.

	Good (%)	Adequate (%)	Miss (%)
Model 1	97.9	0	2.1
Model 2	97.9	0	2.1

Table 2: Results for tracking lips.

	Good (%)	Adequate (%)	Miss (%)
Model 1	95.8	2.1	2.1
Model 2	91.7	6.2	2.1

Best results were achieved with about 12 profile modes for Model 1 and about 20 modes for Model 2. Table 1 shows the results for locating and Table 2 the results for tracking. For Model 1, results for tracking are similar to results for locating but for Model 2, results for tracking are worse than for locating. Inaccurate tracking results were mainly due to the model aligning to the teeth instead of the inner lip contour, the lips were never missed completely.

6. Feature Extraction and Applications

The parameters describing the shape of the lips can be extracted at each time frame from the tracking results and used as feature vector for a speechreading system [15]. The advantage of the shape features is that they are invariant to scale, translation, rotation and illumination.

Intensity information can be extracted from the profile weight vector $\mathbf{b}_g$, which describes the intensity around the mouth area. Some of the features account for intensity variation caused by different lighting conditions and different speakers while others account for the visibility of teeth and tongue, which provides speech related information [16]. The profile vectors, forming the sample space for the PCA, deform with the lip contours and therefore always represent the same object features. A similar approach for intensity modelling has been described in [6]. The modelling of these features for speech recognition and its combination with shape parameters is described in [17]. The extracted features do not only contain speech related information, they also contain subject related information. A person identification system purely based on spatio-temporal analysis of these features is reported in [18].

7. Conclusions

An approach for locating and tracking facial speech features has been described which uses a learned profile model to enable robust lip tracking for different subjects. High performance for locating and tracking was achieved without the need of high-lighting the lips with lipstick or a reflective marker. The model enables the extraction of shape and intensity information embedded in a low dimensional space which can be used for speechreading and person identification.

The assumptions of linear independence and uni-modal distribution seem to be appropriate for shape modelling but are likely to be inadequate for profile modelling. This is particularly the case for the profiles at the inner contour, which can take on intensity values of the teeth, tongue, lips and oral cavity. Future work will therefore concentrate on improved intensity modelling and on discriminating speaker dependent information from speech dependent information.

Acknowledgements

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

References

[1] E. Petajan, "Automatic Lip Reading to Enhance Speech Recognition", Proc. IEEE Computer Vision and Pattern Recognition, pp. 44-47, 1985.

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

[3] K. Mase and A. Pentland, "Automatic Lipreading by Optical-Flow Analysis", MIT Media Lab, Perceptual Computing Group, Technical Report #117, 1990.

[4] C. Bregler and S. Omohundro, "Nonlinear Manifold Learning for Visual Speech Recognition", Proc. Int. Conf. Computer Vision, 1995.

[5] 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.

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

[7] 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.

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

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

[10] 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.

[11] 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. San Francisco, CA: Morgan Kaufmann Publishers, 1994.

[12] R. Kaucic, B. Dalton and A. Blake, "Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications, Proc. European Conf. Comp. Vision, 1996.

[13] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, "Numerical Recipes in C", Cambridge University Press, Cambridge, 1992.

[14] 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.

[15] 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.

[16] Q. Summerfield, "Lipreading and audio-visual speech perception", Phil. Trans. R. Soc. Lond. B 335, pp. 71-78, 1992.

[17] J. Luettin, N. A. Thacker and S. W. Beet, "Speechreading Using Shape and Intensity Information", Proc. Int. Conf. on Spoken Language Processing, 1996.

[18] J. Luettin, N. A. Thacker and S. W. Beet, "Learning to Recognise Talking Faces", Proc. Int. Conf. on Pattern Recognition", 1996.