Lip-Tracking for Speech-Assisted Video-Processing

Juergen Luettin

IDIAP, CP 592, 1920 Martigny, Switzerland
luettin@idiap.ch

Abstract

This paper describes a robust method for facial feature extraction based on an appearance based model. Both the shape of the lips and the texture of the mouth is decomposed into a sum of weighted basis shapes and basis textures, using a Karhunen-Loéve decomposition. The model can be used to locate and track lip movements and to recover the model's parameters to be used for speech or image processing applications. Feature extraction experiments are described on two audio-visual databases, the Tulips1 isolated digit database of 12 subjects and the M2VTS continuous digit database of 37 subjects.

1 Introduction

The extraction of facial features from image sequences has become an important issue in both automatic speech processing and automatic image processing. Potential applications include speech-assisted image coding, speech-driven talking heads, lip synchronisation, lip animation for film dubbing, audio-visual speech recognition, and audio-visual person authentication. Robust facial feature analysis is a difficult object recognition problem because of appearance differences across subjects, variability during speech production, varying pose, and varying illumination. Several previous approaches have therefore simplified the lip tracking process by painting the subject's lips with a reflective marker, by performing experiments on one subject only, or by using controlled recording environments. Most applications however require the image analysis to be performed on natural images and under different environmental conditions.

2 Previous Approaches

Yuille et al. introduced the use of *deformable templates* for facial feature extraction [14] in which the outline of facial parts is modelled by a set of hand coded polynomials. Since the deformation of deformable templates is constrained by the initial choice of polynomials, they are often not able to resolve fine contour details. Image search is performed by fitting the template to image gradients, assuming strong edges at the feature contours. This assumption is however often violated, especially for lip contours.

Kass et al. [6] have described active contour models, so-called *snakes*, for lip-tracking. These 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 resolve fine contour details. Bregler et al. [1] described a method based on *snakes* for tracking the outer lip contour, but where the contour was constrained to lie in a sub-space learned from a training set. Similar to deformable templates, the approach assumes that image gradients are well suited to represent the lip contours.

An approach based on *splines* and Kalman filters has been described in [4]. Shape constraints were imposed on the deforming template by limiting the number of degrees of freedom. This was performed by fitting the spline to two extreme mouth shapes. Image search was done by searching for high contrast edges. However, tracking was only stable when a lip-stick was worn to enhance the contrast around the lips. Ramos Sanchez et al. [13] combined a spline based lip model with the use of colour information. Lip boundary search was formulated as a two-class ('lip' or 'skin') classification problem. They found however the lip chromaticity not to be adequately represented by a unimodal distribution. They therefore constructed individual chromaticity models for each subject, which prevents the use of the model for multi-speaker or speaker independent feature extraction.

Most of these methods have relied on heuristics about shape deformation and contour representation. The method described here aims to avoid the use of such assumptions by learning such knowledge from examples.

3 Lip Model

The method described here [8] is based on a point distribution model (PDM), also called active shape model (ASM) when used in image search [2], to model the shape of the lips. PDMs 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 average shape and the principal modes of variation are captured from the statistics of a labelled training set.

3.1 Shape Modelling

Shape deformation is modelled by decomposing a shape into a weighted sum of basis shapes using a Karhunen-Loéve (K-L) expansion. The basis shapes are obtained from the statistics of a representative training set. This formulation of shape deformation constrains the shape model to only deform to shapes similar to the ones seen in the training set. The training set is labelled by placing model points around the boundary of the lips to allow the statistical analysis of their coordinates. The i^{th} shape in the training set $(i = 1 \dots N_e)$ is described by a vector $\mathbf{x}_i$ with

$$\mathbf{x}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \dots, x_{iN_s-1}, y_{iN_s-1})^T \tag{1}$$

where (x_{ij}, y_{ij}) are the normalised coordinates of the j^{th} point $(j = 0 \dots N_s - 1)$ of the i^{th} shape. Normalisation is performed with respect to scale, rotation and translation.

A shape can be approximated by a weighted sum of basis shapes which are obtained by a K-L expansion. Given a set of N_e normalised labelled shapes, we can calculate the mean shape $\overline{\mathbf{x}}$ and the covariance matrix $\mathbf{S}_s$. The eigenvectors and eigenvalues of the covariance matrix are obtained by principal component analysis (PCA). The eigenvectors with the largest eigenvalues describe the most significant modes of variation, i.e. the variance described by an eigenvector is equal to its corresponding eigenvalue. A normalised shape can now be approximated by

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P_s b_s} \tag{2}$$

where $\mathbf{P_s} = (\mathbf{p}_{s1}, \mathbf{p}_{s2}, \dots, \mathbf{p}_{sT_s})$ is the matrix of the first T_s $(T_s < N_s)$ column eigenvectors which correspond to the largest eigenvalues and where $\mathbf{b_s} = (b_{s1}, b_{s2}, \dots, b_{sT_s})$ is a vector containing the weights for the eigenvectors. The number of eigenvectors used to describe the main modes of variation is normally much smaller than the number of variables of a shape instance.

The approach assumes that the principal modes are linearly independent, although there might be non-linear dependencies present. For objects with non-linear behaviour, linear models reduce the specificity of the model and can generate implausible shapes, which lead to less robust image search. They also require more modes of variation than the true number of degrees of freedom of the object. The specificity of a model can be improved by a nonlinear process, e.g. by nonlinear PCA.

3.2 Intensity Modelling

Intensity modelling is used here as a mean for robust image representation to define a distance measure between the model and the image. The recovered intensity parameters can however also be used as features for subsequent applications. In order to use PDMs for image search, we would like to have a cost function, which measures the fit between the model and the image.

The most common approach for representing contours is to use edges or gradients. It is however well known that the representation of object boundaries by edges are ill-posed and therefore do not have unique solutions. Particularly for lips, the appearance of contours is highly variable – even for the same person. The gradient values at the outer lip contour are often strong at the upper lip and week at the lower lip. The gradients at the inner contour are highly dependent on mouth opening and the visibility of teeth and tongue. Gradients are also dependent on the speaker (make up, facial hair, ethnic origin) and illumination. Furthermore, edges of the lip contours are often confused with other gradients, which can originate from specularity, visibility of teeth and tongue, shadows, or facial hair.

In analogy to the statistical description of the lip deformation we want to avoid the use of heuristics for image search and rather learn the grey-level appearance at the contour from a training set. Assuming that grey-level changes are not only important at each contour point but also in regions around each point, the statistics of the actual grey-level appearance are captured around each model point and their main modes of variation are estimated from a training set.

Following [3], one dimensional profiles $\mathbf{g}_{ij}$ of length N_p are sampled 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, the profiles of all model points are concatenated to construct a global profile vector $\mathbf{h}_i = (\mathbf{g}_{i0}, \mathbf{g}_{i1}, \ldots, \mathbf{g}_{iN_s-1})^T$ for each training image i. The global mean profile $\overline{\mathbf{h}}$ of dimension $N_i = N_s N_p$ and the global covariance matrix $\mathbf{S}_i$ is then calculated, and the eigenvectors and eigenvalues of the covariance matrix $\mathbf{S}_i$ are obtained by PCA. The eigenvectors with the largest corresponding eigenvalues describe the main modes of grey-level variation seen in the training set. Any profile can be approximated using

$$\mathbf{h} = \overline{\mathbf{h}} + \mathbf{P}_i \mathbf{b}_i \tag{3}$$

where $\mathbf{P}_i = (\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{iT_i})$ is the $N_i \times T_i$ matrix of the first T_i ($T_i < N_i$) column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}_i$ a vector containing the weights for each eigenvector.

The intensity weight vector $\mathbf{b}_i$ represents the intensity around the lip contour and accounts for features like the intensity of the oral cavity, visibility of teeth and tongue, and finer details like protrusion. These parameters therefore represent important features with respect to speech recognition and person identification.

3.3 Image Search

The task of image search is to localise and track the lips in the image. This is performed by matching the model to the image at different locations and with different shapes and by choosing the hypothesis with the lowest cost. The search method described here considers both shape variability and grey-level variability in the calculation of the cost E_p^2. To account for grey level variation captured in the training set, the model profile is first aligned to the image profile $\mathbf{h}$ as closely as possible, using the mean intensity and a weighted sum of the first few basis intensities. As the eigenvectors in $\mathbf{P}_i$ are orthogonal, the parameter vector $\mathbf{b}_i$ describing the weights for the basis images can be found using

$$\mathbf{b}_i = \mathbf{P_i}^T(\mathbf{h} - \overline{\mathbf{h}}). \tag{4}$$

To measure how well a model fits the image, the sum of square errors E_i^2 between the image profile and the aligned model profile is calculated using

$$E_i^2 = (\mathbf{h} - \overline{\mathbf{h}})^T(\mathbf{h} - \overline{\mathbf{h}}) - \mathbf{b}_i^T\mathbf{b}_i. \tag{5}$$

The Downhill Simplex Method (DSM) [11] is used to find a minimum of the cost function. The intensity weight parameter vector $\mathbf{b}_i$ can be obtained from the cost function and can be used for further processing.

4 Experiments

Experiments were performed on the Tulips1 database [10] and on the M2VTS database [12]. Tulips1 consists of 96 grey-level image sequences of 12 persons, each saying the first four digits in English twice. The M2VTS database consists of 185 image sequences of 37 persons. Each sequence contains the continuously pronounced French digits from zero to nine. This database contains a total of over 27,000 images. The original colour images of the M2VTS database were converted to grey-level images which were used for all experiments.

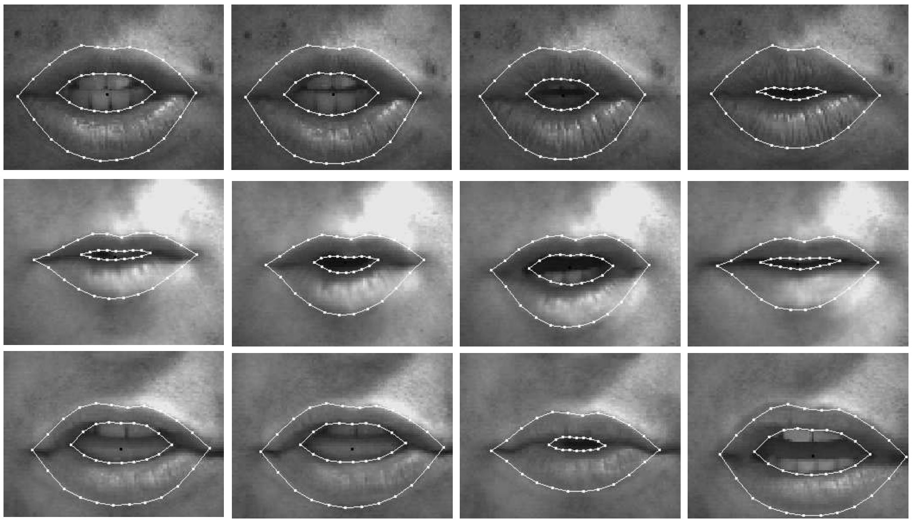

Figure 1: Examples of lip tracking results on the Tulips1 database. The tracking results are very accurate across subjects despite large appearance differences and varying visibility of teeth and tongue.

In order to compare different algorithms it is highly desirable to evaluate their performance. The evaluation of feature extraction algorithms is however difficult since measures about the quality of features are often not available.

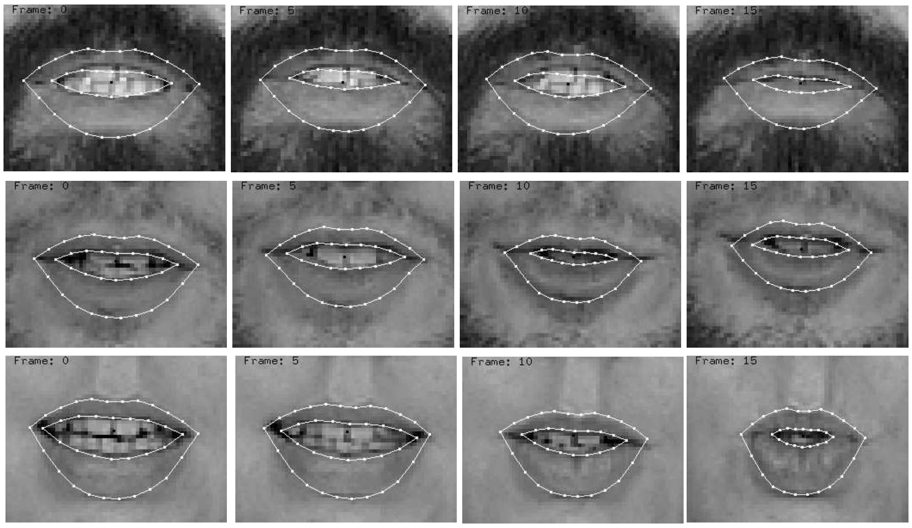

Figure 2: Examples of lip tracking results on the M2VTS database. The first row demonstrates the robustness of the algorithm for subjects with beard despite the reverse contrast between lips and skin.

We argue that the direct evaluation of features is an ill-posed problem and propose that such algorithms should be evaluated on the performance of the application for which the features are being used. We have therefore chosen to evaluate the performance of the lip tracker on the performance of a visual speech recognition system and a visual speaker recognition system which are described in [8, 9, 7, 5]. The algorithms were found to perform well for both databases and for both tasks. Examples of lip tracking results for the Tulips1 database are displayed in Fig. 1 and for the M2VTS database in Fig 2.

5 Conclusion

An appearance based lip model has been described for lip tracking and feature extraction. Both shape deformation and intensity variability are learned from training examples which avoids the use of heuristics about shape deformation and the determination of dominant features. The model parameters can be recovered from the lip tracking results and can be used for subsequent applications. Lip tracking experiments were performed on two publicly available databases and the extracted features were successfully used for the the application of speech recognition and speaker recognition.

References

[1] C. Bregler and S. M. Omohundro. Surface learning with applications to lipreading. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, 1994.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, Jan 1995.

[3] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proceedings of the International Conference on Computer Vision*, pages 242–246, 1993.

[4] B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 373–382. Springer Verlag, Berlin, 1996.

[5] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 319–326. Springer Verlag, 1997.

[6] M. Kass, A. Witkin, and Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[7] J. Luettin, N. A. Thacker, and S. W. Beet. Learning to recognise talking faces. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96)*, volume IV, pages 55–59. IAPR, 1996.

[8] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, February 1997.

[9] J. Luettin. Towards speaker independent continuous speechreading. In *Proceedings of the European Conference on Speech Communication and Technology*, 1997. to appear.

[10] J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT press, Cambridge, 1995.

[11] J. A. Nelder and R. Mead. A simplex method for function optimization. *Computing Journal*, 7(4):308–313, 65.

[12] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database. In *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[13] M. U. Ramos Sanchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with B-splines. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 69–76. Springer Verlag, 1997.

[14] A. L. Yuille, P. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8:99–112, 1992.