DICTA2002: Digital Image Computing Techniques and Applications, 21-22 January 2002, Melbourne, Australia

Robust Lip Tracking with Improved Active Shape Models

Suhuai Luo
CSIRO Telecommunications & Industrial Physics,
P.O.Box 76, Epping, NSW 2121, Australia
Email: Suhuai.Luo@tip.csiro.au

Abstract

This paper introduces a robust lip tracking algorithm based on improved active shape models. After introducing the principle of active shape models, our two main achievements in dealing with lip tracking are presented. One is that a robust model fitting procedure substantially differing from the original active shape model approach has been developed. The other is that tracking on each new frame is based on a priori knowledge about lip appearance and the resultant position and shape of the previous frame as well as the current image structure. This combination has secured a consistent and accurate global lip tracking. The algorithm has been tested on various real video sequences and has consistently demonstrated robust performance on tracking deformable shapes.

1. Introduction

Lip tracking in video sequences is a difficult problem in computer vision due to the inherent variability between individuals and non-rigid nature of the lip contour. Here the lip contour means the outer contour of the lip. The appearance of a lip in an image varies with different subject, content of speech and lighting conditions. Substantial research has been done on lip tracking recently. Major reported techniques for lip tracking are active contour [1], snake models [2] and deformable templates [3].

We use a top-down strategy in developing a model-based lip tracking algorithm. The first step is to build models describing non-rigid shapes of lips. Secondly, fitting these models to images to locate lips. Among various model-based image interpretation algorithms [1, 2, 3, 4, 5], the active shape model (ASM) approach introduced by Cootes et al. provides more favorable shape description properties. ASMs (also known as point distribution models) are statistical models of the shapes of objects which iteratively deform to lock on an instance of an object in an image. Here it is important to note that the locked instance may not necessarily be the real shape of the current object. However, all the resultant shapes are constrained by statistical shape models to vary only in ways occurring in the training set. The training set has previously been manually labelled.

This paper presents our two main achievements in dealing with lip tracking with ASMs. Firstly, a robust model fitting procedure substantially differing from the original approach has been developed. In our model fitting, a pattern matching criteria, instead of original Mahalanobis distance, has been used to improve convergence and accuracy. Secondly, tracking on each new frame is based on both the statistics of the training set and the resultant position and shape of the previous frame. This combination has secured a consistent and practical object tracking.

This paper is organized as follows. Section 2 introduces the active shape models. Section 3 describes our efforts on improving the performance of original ASMs. Section 4 gives details of the robust lip tracking algorithm. Experiment results on lip tracking are given in Section 5. Conclusions are made in Section 6.

2. Active Shape Models

2.1 Statistical shape models

The essentials of active shape models are statistical shape models (SSMs) which are built from analyzing the structures of labelled examples. We learn what are and what are not plausible structure variations with these statistical shape models, to find the best plausible structure in a test image.

Following three stages are needed to build SSMs when a set of shape examples is given.

Stage 1: Suppose there are *k* images in the training set, then each shape in the training set is represented by *n* labelled landmark points. Here the landmark points are so selected that they are consistent from one image to another. This selection guarantees that a given landmark point corresponds to a particular part of the object. For example, the left-most corner point is selected as a landmark in lip tracking.

Stage 2: All the labelled tracking examples are aligned into common coordinates in such a way that the sum of the distance of each shape to the mean of the training set is minimized. The alignment operation, symbolized as $A_{X_t, Y_t, \theta, s}$, includes translating (X_t, Y_t), rotating (θ) and scaling (s). After the alignment, a shape in 2D image can be represented by a *2*n* element vector as $\hat{x} = (x_1, \ldots, x_n; y_1, \ldots, y_n)$.

Stage 3: Since the training set forms a distribution in the *2*n* dimensional space, a parameterized model of the form $\hat{x} = F(\hat{b})$ can represent the training examples and generate plausible new examples similar to training set. Here F is a function of $\hat{b}$, and $\hat{b}$ is the shape parameter vector of the model. The model can be simply formulated by applying the principal component analysis (PCA) [6] on the training data. The resultant shape model can be expressed as

$$\hat{x} \approx \bar{x} + \hat{\Phi}\hat{b} \qquad (1)$$

where $\bar{x}$ is the mean of the $\hat{x}$, $\hat{\Phi}$ is a *2*n*t* dimension vector containing *t* eigenvectors of the covariance of $\hat{x}$. The selected *t* eigenvectors correspond to the *t* largest eigenvalues of the covariance matrix. (*t* can be considered as the number of models to be retained.) $\hat{b}$ is a *t* element vector of shape parameters given by

$$\hat{b} = \hat{\Phi}^T(\hat{x} - \bar{x}) \qquad (2)$$

Where *T* represents transpose operation.

It can be seen that new shapes can be generated by varying the shape parameters.

2.2 Active shape models

2.2.1 definition

When the statistical shape models described above are deformed iteratively to fit the shape of an object, they are considered active shape models. There are two essential components in the rules governing the deformation. One is that shapes are constrained to vary only in ways seen in the training set. Another is that a model instance and a real target are matched or fitted in an optimal way.

In dealing with matching or fitting a model instance $(\hat{x})$ and a real target $(\hat{t})$, both the shape parameter $(\hat{b})$ and the pose parameters (X_t, Y_t, θ, s) are involved. Under the assumption that a rough starting position is known, ASMs proceed in following three steps to iteratively fit to an image.

(1) finding local matching to model points: for each point $(\vec{X}_i = (x_i, y_i))$ on current shape $\hat{x}$, find its best local match $\vec{X}'_i = (x'_i, y'_i)$. These locally matched points form a new shape $\hat{x}'$.

(2) updating the shape and pose parameters $(\hat{b}, X_t, Y_t, \theta, s)$ to best fit the model instance to the newly found shape $\hat{x}'$.

(3) repeating (1) and (2) until convergence is achieved.

The details of steps (1) and (2) are given below.

2.2.2 algorithm

(1) finding local matching to model points

For any intermediate outcome of shape finding, the next possible position of each model point $\vec{X}_i$ can be deduced from its local statistical structure in the training set. A practical way of selecting the local structure is to consider the grey-scale values on the profile normal to the shape boundary on the model point. Figure 1. illustrates the consideration, where a profile consisting of five points (solid circles) is drawn perpendicular to the shape boundary.

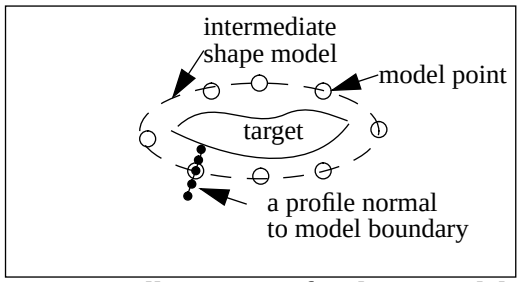


Figure 1. illustration of a shape model and a profile on a model point

In the active shape models introduced by Cootes et al. [4, 5, 7], grayscale derivatives at one centre point and k points either side of a model point are sampled along its profile. For each training image, the *2k+1* gray scale derivatives on the *i*th model point can be

represented with a vector $\vec{d}_i$. The distribution of the grayscale derivatives on the *i*th model point for all the training images can be looked upon as a multivariate Gaussian with mean $\bar{d}$ and covariance $\vec{S}_d$. This gives a statistical profile model for the model point. In testing, the fitness of the profile structure on the *i*th model point to the statistical profile model is given by the Mahalanobis distance (MD) [8] as

$$f(\vec{d}_i) = (\vec{d}_i - \bar{d})^T \vec{S}_d^{-1} (\vec{d}_i - \bar{d}) \qquad (3)$$

To find the best fitting (i.e., minimal $f(\vec{d}_i)$), the fitness test along the profile is done *2(m-k)+1* times with the centre of $\vec{d}_i$ shifted one pixel for every following test. Here *m* (which is always greater than *k*) decides the range of sampling along the profile.

After the above matching process has been done for all the model points, the new position of the shape model ($\vec{X}' = \{\vec{X}_i', i = 1, \ldots, n\}$) is given.

(2) updating shape and pose parameters

With a new model position $\vec{X}'$, the procedure of updating the current shape parameters $\vec{b}$ and the pose parameters (X_t, Y_t, θ, s) to match a model instance to the model position can be described as below.

- generate current model instance as $\vec{X} \approx \bar{x} + \vec{\Phi}\vec{b}$.
- find the pose parameters (X_t, Y_t, θ, s) that best align $\vec{X}$ to $\vec{X}'$ by minimizing the sum of their square distances.
- project $\vec{X}'$ into the model coordinates by using the inversion alignment as $\vec{y} = A_{X_t, Y_t, \theta, s}^{-1}(\vec{X}')$.
- update the shape parameters as $\vec{b} = \vec{\Phi}^T(\vec{y} - \bar{x})$.
- apply plausible constraints on $\vec{b}$ as $p(\vec{b}) > p_t$, where p_t is a suitable threshold on the probability distribution function.

(3) repeating of steps (1) and (2)

Steps (1) and (2) are repeated until convergence is achieved or some pre-defined times of iteration is reached. This will result a model instance $\vec{y}$ that is the best match or fit to the test image.

3. Improved Active Shape Models

As discussed in the above section, in fitting a model to a test image, the local fitness of every model point is judged by the matching between its profile structure $\vec{d}_i$ and its statistical profile structure which is specified by $\bar{d}$ and $\vec{S}_d$. In the approach described by Cootes et al., Mahalanobis distance is adopted to judge the goodness of fit.

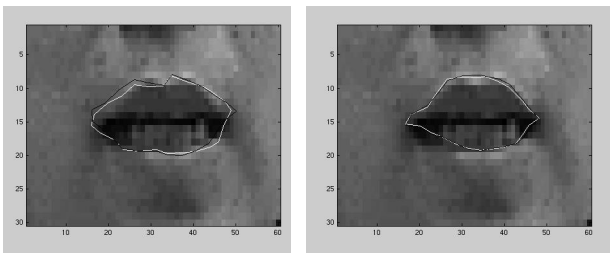
Through a detailed study on the performance of the MD, we have found that it does not give consistent goodness of fit on two shapes. This can be illustrated by a simple example. For two vectors $\vec{\alpha} = (1, 2, 3, 4, 5)$ and $\vec{\beta} = (5, 4, 3, 2, 1)$, though they both have $\bar{d} = 3$ and $\vec{S}_d = 2.5$, they are two different structures when representing sample values in a 2D image.

In order to introduce a more accurate fitting scheme, we consider the local fitting as a process of profile pattern matching. Suppose function $f(t)$ and $g(t)$ are two discrete functions representing the profile of a current model point and its corresponding statistical profile respectively. Then their degree of matching $M(f, g)$ is defined as

$$M(f, g) = \sum_i f(t + i) \bullet g(i) \qquad (4)$$

For the illustrated vectors $\vec{\alpha}$ and $\vec{\beta}$, we have $M(\vec{\alpha}, \vec{\beta}) = 35$ and $M(\vec{\alpha}, \vec{\alpha}) = 55$. This illustrates that $M(f, g)$ gives a good indication of goodness of match.

To further illustrate the performance improvement by introducing $M(f, g)$, two real examples of locating lips using the original ASM approach and our improved active shape models are given in Figure 2 (a) and Figure 2(b) respectively.



(a) **(b)**

Figure 2. Lip locating performance comparison between (a) original ASM approach and (b) our improved ASM approach

In Figure 2 (a), the white curve is the located shape after 15 iterations, and the dark curve is that of the *14th* iteration. It is apparent that divergence exists on model points. However, when the improved ASM approach was applied (see Figure 2 (b)), convergence was achieved after 3 iterations. Convergence can be

seen by the fact that the white curve of the 3rd iteration overlaps with the black curve of the 2nd iteration. Because it has converged, the improved ASM approach yields a more reasonable and correct lip shape.

4. Robust Lip Tracking

To develop a robust lip tracking algorithm, we have considered making use of both a priori knowledge of lip appearance and the resultant position and shape of previous frame. The a priori knowledge is reflected by the statistics of the training set. The use of the statistics of the training set results in a shape that is plausible to the shapes seen in the training set. The consideration of the resultant position and shape of the previous frame will compensate for the lack of globally optimized fitting of ASMs. This combination results in a consistent and practical object tracking.

The tracking algorithm starts with an input video sequence and ASMs corresponding to a specified training set. Based on the assumption that the lip centre of the first frame is known (which is done manually), the algorithm first creates a model instance for every frame. For each frame except the first one, the algorithm adapts the instance's initial position according to the located lip on the previous frame. The 3-step ASM fitting process is then iteratively applied to locate current lip contour. The iteration will stop either when convergence is reached or when a maximal iteration operation has been done. Figure 3 is the block diagram of the algorithm.

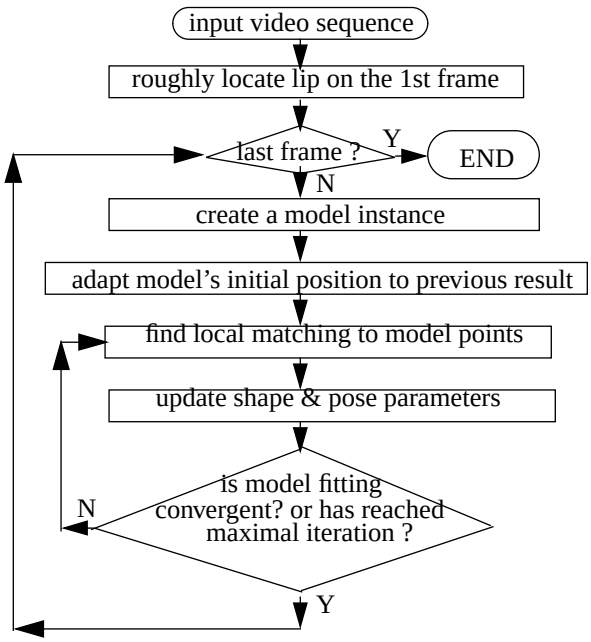


Figure 3. Block diagram of the proposed lip tracking algorithm

5. Experiments and Results

5.1 data collecting

In our lip tracking experiments, face-and-shoulder video sequences of multiple subjects have been processed. There were altogether 16 video sequences corresponding to 4 subjects. For each subject, 4 sequences were recorded corresponding to 4 different scenes. In scene one, the subject did not speak and barely moved; in scene two, the subject barely moved head but talked naturally; in scene three, the subject naturally talked and moved; in scene four, the subject kept silent but the head moved naturally. Each sequence contains about 33 frame images covering about 6 second period. Each image contains 240*320 pixels.

In selecting the training and test sets, complete separation has been chosen. More specifically, 12 sequences of 3 subjects were chosen as the training set, 4 sequences of the fourth subject were chosen as the test set. Figure 4 illustrates some mouth area images of the subjects. It can be seen that the appearances of individual frames are different and the change of lip shape is non-rigid.

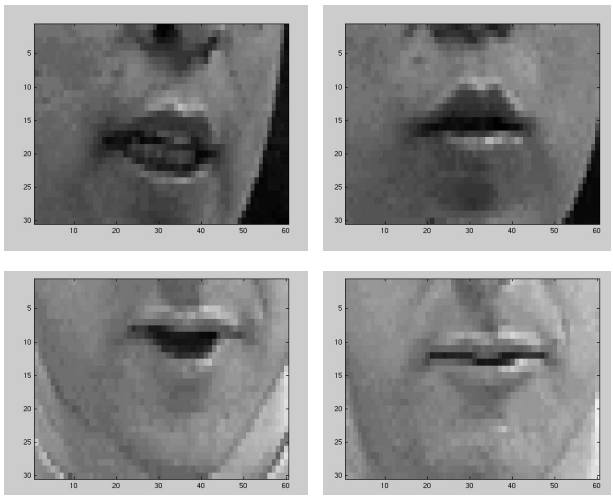


Figure 4. Examples of mouth area images

All the lip contours in the training images were manually labelled. Model parameters were derived

from these images.

5.2 performance

Subjective evaluation has been chosen to investigate the performance of the algorithm. For all the test sequences the algorithm tracked the lip correctly from the first frame up to the last frame. Figure 5 shows results of lip tracking on scene three, where both the

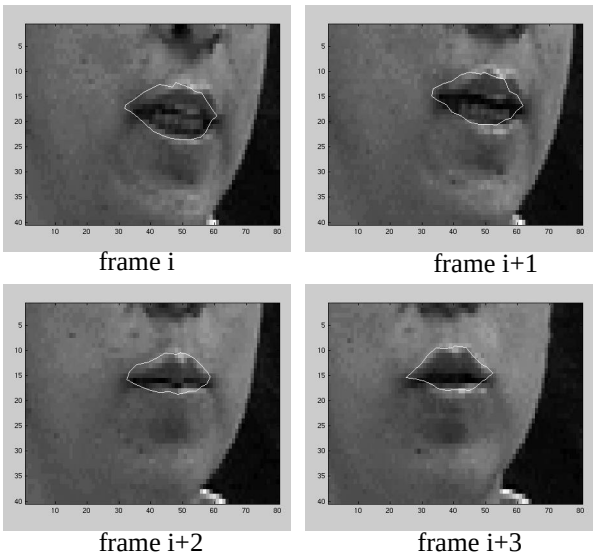


Figure 5. Results of lip tracking on 4 consecutive frames

head motion and lip motion took place. In the figure, the change of lip shape is apparent. The existence of head motion can be seen by changes in the dark background.

From Figure 5 it can be seen that the lip tracking is robust and accurate. Other examples have shown similar results. There is no restriction on lip shape due to speech or head motion.

6. Conclusions

This paper introduces a robust lip tracking algorithm. Two major contributions have been discussed in detail. One is that a robust model fitting procedure substantially differing from the original active shape model approach has been developed. The other is that tracking on each new frame is based on a priori knowledge about lip appearance and the resultant position and shape of the previous frame as well as the current image. This combination has secured a consistent and accurate global lip tracking. Experiments have shown that the proposed algorithm performed well with no restriction on lip shape due to speech or head motion.

Possible future work and improvements may include the introduction of an automatic lip locator on the initial frame, more complicated performance evaluation and its application on other non-rigid shape analysis.

Acknowledgment

The author wish to acknowledge the contributions of Stephen Brown and Michael Dadd to this paper.

References

[1] Kaucic, R. and Blake, A., *Accurate real-time, unadorned lip tracking*, Proc. 6th International Conference on Computer Vision, pp. 370-375, 1998.

[2] Kass, M., Witkin, A., and Terzopoulus, D., *Snakes: active contour models*, International Journal of Computer Vision, vol. 1, pp.321-331, 1987.

[3] Yuille, A. L., Hallinan, P., and Cohen, D. S., *Feature extraction from faces using deformable templates*, International Journal of Computer Vision, vol. 6, pp. 99-112, 1992.

[4] Cootes, T. F., Hill, A., Taylor, C. J., and Haslam, J., *Use of active shape models for locating structures in medical images*, Image, Vision Computing, vol. 12, no. 6, pp. 355-366, 1994.

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

[6] Jolliffe, I. T., *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[7] Cootes, T. F., and Taylor, C. J., *Statistical Models of Appearance for Computer Vision*, Report, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, University of Manchester, U.K., http://www.wiau.man.ac.uk, Feb. 28, 2001.

[8] Mahalanobis, P. C., *On tests and measures of groups divergence*, Journal of the Asiatic Society of Benagal, 26:541, 1930.