International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:2, No:4, 2008

Realtime Lip Contour Tracking For Audio-Visual Speech Recognition Applications

Mehran Yazdi Mehdi Seyfi Amirhossein Rafati Meghdad Asadi

***Abstract*—**Detection and tracking of the lip contour is an important issue in speechreading. While there are solutions for lip tracking once a good contour initialization in the first frame is available, the problem of finding such a good initialization is not yet solved automatically, but done manually. We have developed a new tracking solution for lip contour detection using only few landmarks (15 to 25) and applying the well known Active Shape Models (ASM). The proposed method is a new LMS-like adaptive scheme based on an Auto regressive (AR) model that has been fit on the landmark variations in successive video frames. Moreover, we propose an extra motion compensation model to address more general cases in lip tracking. Computer simulations demonstrate a fair match between the true and the estimated spatial pixels. Significant improvements related to the well known LMS approach has been obtained via a defined Frobenius norm index.

***Keywords*—**Lip contour, Tracking, LMS-Like

I. Introduction

A relatively large class of lipreading algorithms has been proposed based on lip contour analysis. Examples of such algorithms can be found in [1, 2]. In these algorithms, lip contour extraction is needed as the first step. By lip contour extraction, we usually refer to the process of lip contour detection in the first frames of an audio-visual image sequence. Obtaining the lip contour in subsequent frames is usually referred as lip tracking. In lip contour tracking the initial estimates are not always available for the first frames, they have to be produced by some means. Different authors tried different procedures to solve the extraction of a good lip contour in the initial frame(s). Of course, the goal would be to solve this task automatically and approaches based on region-based image segmentation and on edge detection have been successfully proposed. In the frontal images without any marking of the lips, the above-mentioned techniques unfortunately fail, and these images are most frequently used for speechreading. In such cases, the solution adopted is based on marking manually more or less points on the lip contour (or even on drawing manually the entire lip contour). When a small number of points (e.g. 15 to 25 points) are marked on the lip contour, these points are used to derive some model parameters such as the widely used ellipsoidal model [3] for the lip contour. In the latter case, the accuracy of lip contour extraction is limited by the fitness of the model to the real lip contour. For example, in the case of an asymmetric mouth image (due to a displacement of the video camera), the model-based lip contour representation might be different from the real lip contour. In this paper we propose a new approach to the problem of lip contour tracking in either gray level or RGB images. The proposed approach is based on tracking of some initially considered landmarks (15 landmarks here) in the lip region using a new adaptive LMS-like approach. In this paper some assumptions are made: some initial landmarks that could be found somehow automatically are given, in some sufficiently distant frames the corresponding landmarks, called the training landmarks, could also be found automatically, this is for updating the adaptive tracking procedure. Using the position coordinates of these landmarks via a normalizing process, extraneous movements of the face in sequential frames are compensated. Based on our observation that the landmarks sequential variations in successive frames obey an auto-regressive (AR) model, we propose an LMS-like adaptive scheme to estimate the landmark positions between the video frames where no landmark trainings for contour estimation are available. First the adaptive weight vector which is the AR model coefficients is initially set to track the landmarks variations, and next at each updating frame the weight vector is re-optimized due to new data. Finally, using the obtained landmark positions we utilize Active Shape Models [4] to extract the lip contour in each image. At the end, with the use of computer simulations, we study the efficiency of our proposed lip contour tracking for a practical video segment. Throughout the present work, matrices are indicated by capital bold face letters and vectors are presented by lower case bold face letters.

Manuscript received December 2007. M. Yazdi is with the Department of Electrical Engineering, School of Engineering, Shiraz University, Shiraz, Iran, e-mail: yazdi@shirazu.ac.ir

II. Extra motion compensation model

In order to track the lip motion, some important issues must be taken into consideration. First, to inject more generality to the procedure without any special limitation for any special gestures a person might have, we need to eliminate extra motions in the tracking process. These motions are generally due to any vertical or horizontal movement of face along the main axes, and due to rotational movements around the horizontal axis. To deal with the problem, two separate spots with least possible variations during facial expression are chosen as references. These spots are properly chosen on the nose and the middle point between these two references is considered as the center of rotation (see Fig. 1).

$$x_{origin,old} = (x^{right}_{nose,old} + x^{left}_{nose,old})/2$$
$$y_{origin,old} = (y^{right}_{nose,old} + y^{left}_{nose,old})/2 \quad (1)$$

After calculating the origin coordinates, the relative intermediate coordinate system is simply obtained as

$$x_{int} = x_{old} - x_{origin,old}$$
$$y_{int} = y_{old} - y_{origin,old} \quad (2)$$

As the movements along the main vertical and horizontal axes are compensated, it is now desired to compensate additional rotations relative to the horizontal axes. To overcome the problem it is needed to calculate the angle of rotation in each video frame. Hence according to new coordinates of the specified nasal references (see Fig. 2) the rotation can be obtained as

$$\theta = tan^{-1} \frac{y^{right}_{nose,int} - y^{left}_{nose,int}}{x^{right}_{nose,int} - x^{left}_{nose,int}} \quad (3)$$

By computing the rotation angle, we define new axis which differs with the old axis only in a rotation angle. With this new coordinate system, the landmarks' positions have new coordinates which must be computed as follow

$$\gamma = tan^{-1} \frac{y_{int}}{x_{int}} \quad (4)$$

$$r = \sqrt{x^2_{int} + y^2_{int}} \quad (5)$$

$$\alpha = \gamma - \theta \quad (6)$$

$$x = r.sin\alpha \quad y = r.cos\alpha \quad (7)$$

Considering N different positioned landmarks in the lip region it may be possible to represent the x and y components of the desired landmarks in the xy coordinate system as

$$[\mathbf{X}]_{rs} = x_{rs}$$
$$[\mathbf{Y}]_{rs} = y_{rs}$$

Where $r \times s$ = N.

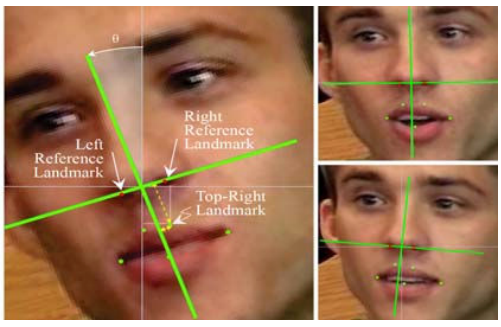

Fig. 1. Nasal reference points and the Coordinates calculation system.

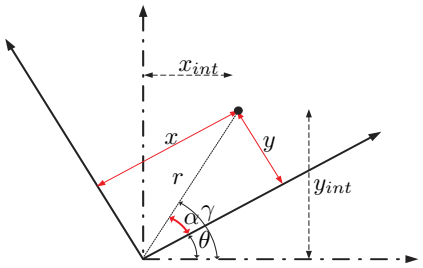

Fig. 2. Nasal reference points and the Coordinates calculations system.

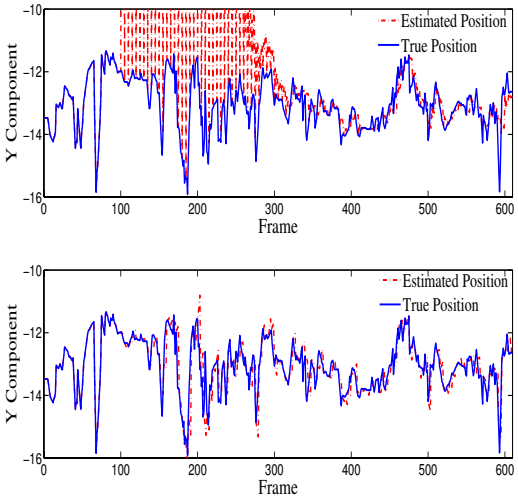

Fig. 3. Comparison between the LMS method (top) and the proposed method (bottom) for data recovery,$\mu = .9$.

III. LMS-Like AR Modelling For The Random Landmark Sequence

Let $x(t)$ be a continuous sample function of a zero-mean real process, and

$$x_n \stackrel{\text{def}}{=} x(nT_s)$$

be the discrete version of this function with sampling frequency $f_s = 1/T_s$. Let us also suppose that x_{n_m} is the m^{th} known point of the sequence x_n. Depending on the statistical properties of the signal it may be possible to fit an Auto Regressive (AR) model to this signal so that,

$$x_{n_m} = \sum_{i=1}^{p} a_i x_{n_m - i} + v_{n_m} \quad (8)$$

where v_{n_m} is the modelling error that is usually assumed to be zero mean, a_i s for $i = 1, \ldots, p$ are the so called AR parameters (also known as prediction coefficients), and p is the order of the model. Using such a model, x_{n_m} can be estimated from p previously known samples, i.e.

$$\hat{x}_{n_m} = \sum_{i=1}^{p} a_i x_{n_m - i} \quad (9)$$

where

$$x_{n_m-i} = \begin{cases} x_{n_m-i} & \text{if the sample is not lost} \\ \hat{x}_{n_m-i} & \text{if the sample is lost} \end{cases} \quad (10)$$

For such a model, we compute the prediction coefficient vector $\mathbf{a} = [a_1, \ldots, a_p]^T$ such that the following time dependent cost function is minimized,

$$J_k = \frac{1}{k}\sum_{m=1}^{k} \mu^{k-m}(x_{n_m} - \hat{x}_{n_m})^2 = \frac{1}{k}\sum_{m=1}^{k} \mu^{k-m} v_{n_m}^2 \quad (11)$$

where $0 < \mu < 1$ is a constant known as the forgetting factor.

In this paper we would like to estimate x_{n_m-i} using the AR model (9) when x_{n_m-i}'s are unknown, and where x_{n_m} is perfectly known for $m = 1, 2, \ldots k$.

Let us assume that the true x_n is known at instances $n = n_m$, the prediction vector $\mathbf{a}$ has to be found from (11) [5].

In this scenario, x_{n_m} can be estimated from (9).

Based on the above discussion the algorithm for estimation of unknown x_ns has to be computed as [5],

Step 1:Computing $\hat{x}_{n_m}$, using (9).

Step 2:Computing $\mathbf{a} = [a_1, \ldots, a_p]$ using (11).

Step 3:Prediction of missing samples between instances $n = n_m$ and $n = n_{m+1}$ for $m = 1, 2, \ldots$ using the last estimated AR parameters.

At step 2, one has to compute the gradient of the cost function $\partial J_k/\partial \mathbf{a}$, putting it equal to zero and compute the proper $\mathbf{a}$. A detailed description of the algorithm can be found in [5]. It is interesting to note that the gradient can be updated recursively at each instant $n = n_m$ for $m = 1, 2, \ldots$.

IV. The Proposed Tracking Scheme

Based on the previous discussion, in a lip landmark system with N different separated landmarks, one will be able to track the x component and the y component of elements of landmark matrices $\mathbf{X}$ and $\mathbf{Y}$ using the proposed algorithm. The landmark matrices at time instant k can be expressed in terms of p previous landmark matrices as,

$$\mathbf{X}_k = \sum_{i=1}^{p} \mathbf{A}_i \odot \mathbf{X}_{k-i} + \mathbf{V}_k$$
$$\mathbf{Y}_k = \sum_{i=1}^{p} \mathbf{B}_i \odot \mathbf{Y}_{k-i} + \mathbf{W}_k \quad (12)$$

Here, $\odot$ stands for the Hadamard product and $\mathbf{V}_k$ and $\mathbf{W}_k$ show the modelling error matrices being uncorrelated with each other. We call $\mathbf{A}_i$ and $\mathbf{B}_i$ for $i = 1, \ldots, p$ the *prediction matrices* where their rs^{th} element shows the i^{th} prediction coefficient for landmark position elements x_{rs} and y_{rs}.

In a similar way as before, the landmark matrices $\mathbf{X}_n$ and $\mathbf{Y}_n$ can be estimated using the previously known/estimated landmark matrices as,

$$\hat{\mathbf{X}}_k = \sum_{i=1}^{p} \mathbf{A}_i \odot \hat{\mathbf{X}}_{k-i}$$
$$\hat{\mathbf{Y}}_k = \sum_{i=1}^{p} \mathbf{B}_i \odot \hat{\mathbf{Y}}_{k-i} \quad (13)$$

Calculation of the elements of $\mathbf{A}_i$'s and $\mathbf{B}_i$'s is similar to what we explained in the previous section.

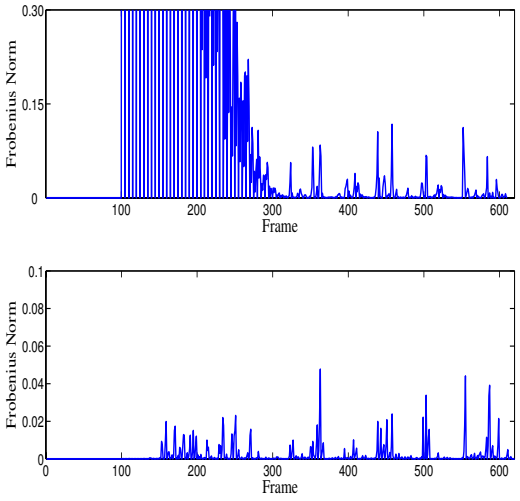

Fig. 4. Frobenius norm comparison of the LMS method (top) and the proposed method (bottom) for data recovery.

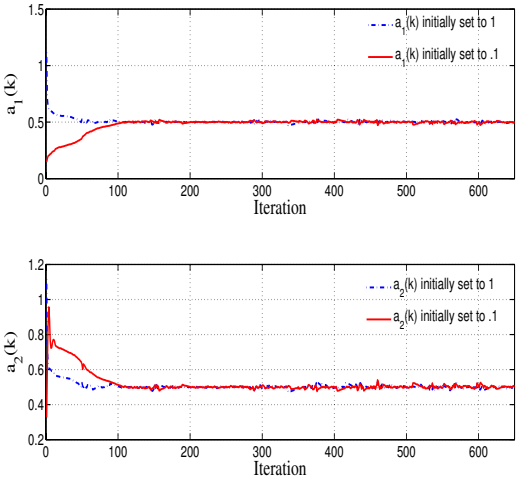

Fig. 5. Adaptive weight vector convergence with different initial conditions.

V. Computer Simulations

For the computer simulations, a practical video segment was utilized. The nasal reference points along with 15 more landmarks are chosen in each frame. This manual process is used for 50 initial video frames. Extracting the landmarks positions in successive frames and normalizing their reciprocal and rotational movements, the landmarks positions are tracked based on the proposed LMS-like procedure. Fig. 3 shows the tracked Y-position of a sample landmark where updating procedure is performed every 50 frames. The tracking algorithm is compared with that of LMS algorithm. As depicted in the figure the proposed algorithm shows a pretty good recovery

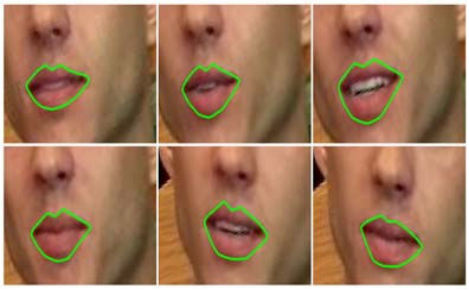

Fig. 6. Final Contour detection via ASM.

in comparison with the LMS algorithm.

Let us define the instantaneous relative landmark position squared estimation error with,

$$\varepsilon_n \stackrel{\text{def}}{=} \frac{\|\mathbf{X}_n - \hat{\mathbf{X}}_n\|^2}{\|\mathbf{X}_n\|^2} + \frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_n\|^2}{\|\mathbf{Y}_n\|^2} \qquad (14)$$

where $\|\cdot\|$ represents the Frobenius norm. Now, Fig. 4 shows a sample of instantaneous relative landmark position estimation error ε_n versus n where the order of the AR channel model is found to be $p = 5$. As can be seen obviously the proposed tracking procedure has better performance than that of LMS. Fig. 5 shows a_1 in the proposed algorithm with different initial conditions, as it is obvious the method does not depend on the initial conditions for AR parameters. Finally Fig. 6 shows the final contour detection via ASM method using the tracked landmarks in some different frames with special positions. It is quite clear that the face shape and head rotations do not affect the detection process.

VI. Conclusion

In this paper, we have proposed a solution to lip contour detection that minimizes the user interaction. In this scenario it is required to mark a minimal number points on the mouth image manually in the few initial frames. The method is based on tracking of these points i.e, landmarks. Trying to optimize the tracking procedure one could have proper data about each landmark position in the corresponding frame. Profiting these spots we then utilize a well known contour detection algorithm called Active Shape Model(ASM) to identify the desired contour in each frame. The obtained results show better performance of our proposed approach compared with that of the well known LMS approach using a defined Frobenius norm index.

References

[1] R. Caucic et al., "Real time lip tarcking for audio-visual speech recognition applications,"*Proc. European Conf. Computer Vision*, Cambridge, UK, pp. 376-387, April 1996.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition,"*IEEE Transactions on Multimedia*, vol. 2, no. 3, pp.141-151, Sept 2000.

[3] S. L. Wang, W. H. Lau, and S. H. Leung. "A new real-time lip contour extraction algorithm" *Proc. IEEE international conference on Acoustics, Speech and Signal Processing, ICASSP' 03*, Hong Kong, Vol. 3, pp. 578-582, April 2003.

[4] I. Mattews, T.F Cootes, J.A Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, *IEEE Tran. on PAMI*, vol.24, pp.198-213, Feb. 2002.

[5] S. Mirsaidi and G. A. Fleury, and J. Oskman, "LMS like AR modeling in the case of missing observations," *IEEE Trans. on signal processing*, vol. 45, no. 6, pp.1574-1583 , June 1997.