# Integrating active face tracking with model based coding

Lijun Yin, Anup Basu

ELSEVIER

Pattern Recognition Letters 20 (1999) 651–657

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

# Integrating active face tracking with model based coding

Lijun Yin, Anup Basu *

*Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2H1*

Received 26 January 1999; received in revised form 18 February 1999

**Abstract**

In this paper, input from an active camera is used for MPEG4 model based coding. First, the background is compensated considering a moving camera (tilt or pan). Second, the talking face is segmented from the compensated background using frame differences fusion. A morphological filter is then applied to make the system less sensitive to noise. Third, Hough Transform and deformable template coupled with color information are exploited to detect the facial features, e.g., eyes, mouth. Fourth, a wireframe model is adapted to the extracted face. The feasibility of the proposed system is demonstrated using a real active video sequence. 

*Keywords:* Active tracking; MPEG-4; Model-based coding; Feature detection

## 1. Introduction

The emerging MPEG4 standard provides a fundamental framework for multimedia applications at low bit rates (MPEG Requirements, 1998; MPEG Video & SNHC, 1998). However, MPEG4 does not specify the techniques to be used for feature detection and tracking to realize an actual implementation. This allows researchers to investigate alternative techniques for different applications. Up to now, most work in MPEG4 (or related areas) has been limited to the still camera scenario (Choi et al., 1994; Essa et al., 1996; Zhang, 1998; Aizawa and Huang, 1995; Kompatsiaris et al., 1998; Meier and Ngan, 1998; Moscheni et al., 1998; Reinders et al., 1995; MPEG Requirements, 1998; MPEG Video & SNHC, 1998). In this paper, we address the problem considering a talking face in front of an active camera. The detection and tracking of the active talking face is a fundamental step towards realizing an application of MPEG4 in real situations.

Various approaches to the segmentation and feature detection of face images have been attempted, mainly in the field of face recognition. From a large amount of previous work (Chellappa et al., 1995), it has become obvious that face detection and recognition is still a very difficult subject. Segmenting and tracking a talking face from a background is a prerequisite when applying model-based coding schemes (MPEG4-SNHC (MPEG Video & SNHC, 1998)) to compress face-to-face video communication data (Moscheni et al., 1998; Gu and Lee, 1998; Meier and Ngan, 1998; Kompatsiaris et al., 1998; Zhang, 1998). For this purpose, the region of interest (i.e., face region) must be detected and tracked temporally. Detecting the moving part in an image reveals the silhouette region of the speaker, which can be relatively easily detected on the basis of frame differences if the speaker is the only moving object

* Corresponding author: E-mail: anup@cs.ualberta.ca

0167-8655/99/$ – see front matter 
PII: S0167-8655(99)00029-X

within the scene in the case of a still camera. However, in the situation of a movable camera, the difficulties increase since the background viewed is dynamically changeable.

In general, there are two approaches to tracking a moving object, which are recognition-based tracking and motion-based tracking (Murray and Basu, 1994; Aizawa and Huang, 1995; Darrel et al., 1996). Recognition-based tracking is really based on the object recognition technique, the performance of the tracking system is limited by the efficiency of the recognition method (Chellappa et al., 1995; Aizawa and Huang, 1995). Motion-based tracking relies on the motion detection technique, which can be divided into the optical-flow method and the motion-energy method. In optical-flow method, determining a complete optical flow field is ill-posed, the difficulties increase with the active camera for searching and matching feature points in successive images since the scene viewed is dynamically changeable. The complexity of this problem makes it unsuitable for real application. Motion-energy tracking method can segment an image into regions of motion and inactivity by calculating the temporal derivative of an image sequence and thresholding at a suitable level to filter out noise. This method is relatively simple and efficient. However, it is not suitable for application on an active camera system without modification. Since the active camera system can induce apparent motion on the scenes they view, compensation for camera motion must be made before motion-energy detection technique can be used.

In this paper, we present a system to track a talking face with an active camera. After the background compensation in successive frames (Murray and Basu, 1994), the motion-energy tracking approach is used coupled with a morphological filter to reduce the noise. Evidence about moving objects in the scene gathered from a single frame difference may not suffice to portray a speaker entirely. It is necessary to recover the speaker's silhouette by observing a number of successive frame differences (called motion masks). We believe that the motion masks are temporally correlated if the motion of the speaker is assumed to be slow with respect to the frame rate. We propose a progressive silhouette generation method, to detect the motion of a talking head, in which the consecutive motion masks are accumulated to complete the silhouette estimation. After the face region is detected, the deformable template technique coupled with color information and Hough Transform can be used to extract the facial features (e.g., eyes, mouth), then a model fitting procedure is applied to complete the facial model adaptation and animation. Taking the input of an active camera, our system can implement the face detection, tracking, adaptation and animation automatically.

In Section 2, a background compensation technique with an active camera is explained. In Section 3, a face motion detection algorithm is described. Sections 4 and 5 introduce the facial feature extraction and wireframe model adaptation. Section 6 shows some experimental results. Conclusion and final remarks are given in Section 7.

## 2. Background compensation

Before applying the motion detection technique, we must compensate for the apparent motion of the background of a scene caused by camera motion. Our camera is mounted on a pan/tilt device and hence is constrained to rotate only. The objective in background compensation is to find a relationship between pixels representing the same 3D point in images taken at different camera orientations. For camera rotation, the only components of the system that move are the camera coordinate system and the image plane. An example of this motion is shown in Fig. 1. The relationship between every pixel position in two images taken from different position of rotation about the lens center has been derived by Kanatani (1987) and Murray and Basu (1994). For an initial inclination ($\theta$) of the camera system and pan and tilt rotation of $\alpha$ and $\gamma$, respectively, this relationship is

$$x_{t-1} = f\frac{x_t + \alpha \sin\theta y_t + f\alpha\cos\theta}{-\alpha\cos\theta x_t + \gamma y_t + f}, \tag{1}$$

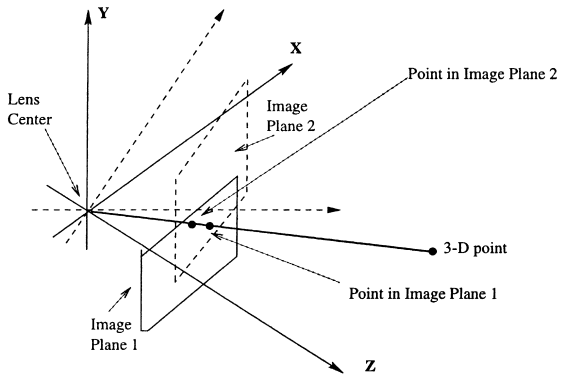


Fig. 1. 3D point projected on two image planes with the same lens center.

$$y_{t-1} = f\frac{-\alpha\sin\theta x_t + y_t - f\gamma}{-\alpha\cos\theta x_t + \gamma y_t + f}, \tag{2}$$

where $f$ is the focal length. With knowledge of $f, \theta, \gamma$ and $\alpha$, for every pixel position $(x_t, y_t)$ in the current image we can calculate the position $(x_{t-1}, y_{t-1})$ of the corresponding pixel in the previous image.

## 3. Moving head detection

The inaccuracies in the inputs to the compensation algorithm and small amount of camera translation may induce noise and corrupt the compensation method. A single frame difference may not suffice to portray a speaker entirely. It is necessary to recover the speaker's silhouette by observing a number of successive frame differences. Fig. 2 shows a block diagram describing the moving head detection algorithm, which consists of the following steps.

- After the background is compensated in the previous frame, the difference of the current frame and the previous frame is calculated, the absolute value is thresholded.
- To reduce the background noise, a morphological filtering (*opening* operator) is applied. The kernel size of erosion and dilation operations depends on the noise characteristics caused by compensation error (Murray and Basu, 1994), the filter must be at least as wide as the error. In our system, the kernel of filtering is set as $9 \times 9$.
- To generate the head silhouette, a frame motion fusion algorithm is developed in which the multiple frame differences are integrated to generate the continuous motion areas.
- *Progressive motion fusion.* The output image of the morphological filtering (*opening*) in Fig. 2 is called motion mask (MM), denoted by $m_t(x)$ ($t$ stands for time and $x$ for pixel position). The successive MMs are processed with a spatiotemporal filter. The function of the spatiotemporal filter is to *fuse* a number of consecutive masks. The temporal fusion in the consecutive $N$ MMs from time $t_0$ to time $T = t_0 + N$ is implemented as

$$\begin{aligned}&\text{for } t = t_0, t_0+1, \ldots, t_0+N\\ &\quad\{\\ &\quad I_t(x) = \mathrm{ADD}\{m_t(x), s_{t-1}(x)\}\\ &\quad s_t(x) = \mathrm{Median}\{I_t(x)\}\\ &\quad\}\end{aligned}$$

where $m_t(x) \in \{0, 1\}$ and $s_{t_0}(x) = 0$. To reduce noise effects, the combined images are further smoothed by a spatial median filter. The parameter $N$ controls how long frame information is integrated. For example, keeping $N$ large will tend to create connected and smooth foreground blobs at the expense of smearing the silhouette, especially if the object motion is excessive (in our experiment, $N$ is set to 8 as a

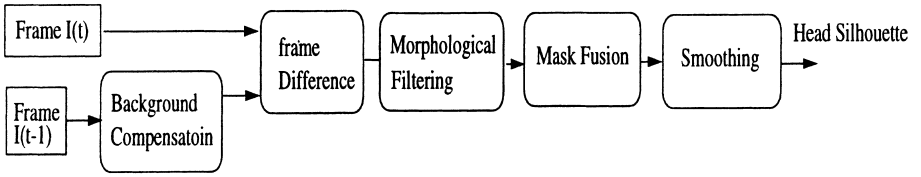


Fig. 2. Diagram of head detection.

tradeoff of the motion mask integration and the contour smearing). Finally, the temporally fused image $s_T(x)$ is thresholded to reveal the motion regions of the head. The threshold value is set to one so that all the motion area reflected in the fused image is taken into account.

- After obtaining the fused image $s_T(x)$, a morphological filter (*closing* operator) is applied to make the image more smooth. Then, the edges of motion areas are detected simply by a gradient operator. Although a few connected contours are detected, the head silhouette is the largest connected contour, and it can be extracted by a region growing method.

## 4. Facial feature detection

Since the head region is detected by the above work, we can continue to detect the facial features (such as eyes and mouth) restricted in this area. Our approaches to detecting eyes and mouth are similar, both use deformable template matching and exploit color information. In addition, eye detection uses Hough transform to search the iris position and size to determine the initial location of the eyes. The detailed algorithm is described in our previous work (Bernoegger et al., 1998).

The algorithm for eye detection can be outlined as follows:

- Determine two coarse regions of interest for the eyes.
- Search the iris of the eyes using a gradient based Hough transform.
- Determine a fine region of interest for extracting the boundaries of the eyes.
- Using color information (saturation) get an initial approximation for the eye lids.
- Localize the eye lids using deformable templates.

The algorithm for mouth detection can be outlined as follows:

- Determine a coarse region of interest for the mouth.
- Using color information (hue and saturation) get an initial approximation for the lip. This is done by minimizing the following energy ($E_{\text{HueSat}}$) which is similar to the valley energy in (Yuille et al., 1992):

$$E_{\text{HueSat}} = \frac{1}{|A_w|}\int_{A_w} \Phi_{\text{hue}}(\vec{x})\ \mathrm{d}A - \frac{1}{|A_w|}\int_{A_w} \Phi_{\text{sat}}(\vec{x})\ \mathrm{d}A, \tag{3}$$

$A_w$ is the total area inside the parabolas of upper lip and lower lip, $\Phi_{\text{hue}}(\vec{x})$ and $\Phi_{\text{sat}}(\vec{x})$ are the values of the hue-angle and the saturation of the color image.

- Localize the mouth contour using deformable templates.

## 5. Facial model adaptation

Based on the information extracted from head motion and facial features (i.e. eyes, mouth), a 3D wireframe model can be fitted to the moving face. Based on our previous work (Yin and Basu, 1997), an individual facial model of a person is created from two image views. Since the individual facial model has been generated off-line, so the shape of a person's head is known. The positions of the eyes and mouth can determine the position of the face, their shapes can determine the facial expressions. We developed a so called "coarse-to-fine" adaptation algorithm using the dynamic mesh to implement the model adaptation procedure, in which the features (eyes and mouth contours) are the mesh boundaries, the adaptation follows the energy minimization procedure to converge the mesh to fit the face image. Details are available in a technical report (Yin and Basu, 1998).

## 6. Experimental results

We use a camera mounted on an active platform (pan/tilt) to take an active video sequence, which shows a talking person with an unconstrained background. The camera rotation is less than $\pm 5^\circ$. Figs. 3 and 4 show the results of the active head detection and the model adaption, respectively.

Fig. 3. 1st row: Original active video sequence (frame 10, 24, 41); 2nd row: Compensated frame differences; 3rd row: Noise remove using morphological filtering ($kernel = 9 \times 9$); 4th row: Consecutive frames fusion; 5th row: Motion areas detected by smoothed fusion frames (closing); 6th row: Contours of motion areas; 7th row: Detected silhouette of the motion head.

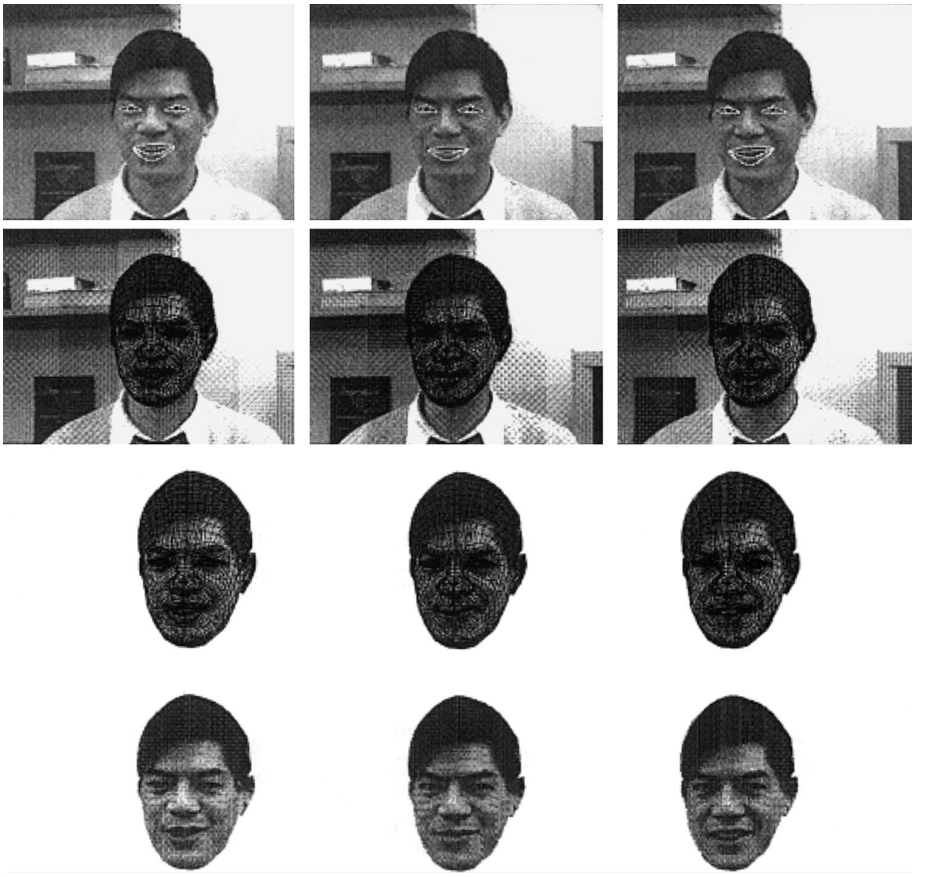

Fig. 4. 1st row: Detected facial features (iris, eye, mouth); 2nd row: Adapted models; 3rd row: Animated model with texture overlapped; 4th row: Texture-mapped model animation (frame 10, 24 and 41).

## 7. Conclusion

In this paper we proposed a system for tracking a face observed with an active camera for an arbitrary background. Initial results show that this approach is feasible in practical applications. More robust methods for face tracking and expression detection are needed in future work.

## Acknowledgements

This work is supported in part by the Canadian Natural Sciences and Engineering Research Council.

## References

Aizawa, K., Huang, T., 1995. Model-based image coding: Advanced video coding techniques for very low bit-rate application. Proceedings of the IEEE 2, 259–271.

Bernoegger, S., Yin, L., Basu, A., Pinz, A., 1981. Eye tracking and animation for MPEG-4 coding ICPR'98. In: Proceedings 14th IAPR International Conference on Pattern Recognition, August 1998, Vol. II, pp. 1281–1284.

Chellappa, R., Wilson, C., Sirohey, A., 1995. Human and machine recognition of faces: A survey. Proceedings of the IEEE 83, 705–740.

Choi, C., Aizawa, K., Harashima, H., Takebe, T., 1994. Analysis and synthesis of facial image sequences in model-based image coding. IEEE Transactions on Circuits and Systems for Video Technology 4 (3), 257–275.

Darrel, T., Moghaddam, B., Pentland, A., 1996. Active face tracking and pose estimation in an interactive room. In: Proceedings of IEEE CVPR'96.

Essa, I., Basu, S., Pentland, A., 1996. Motion regularization for model-based head tracking. In: ICPR'96: Proceedings 13th IAPR International Conference on Pattern Recognition, Vienna, Austria, August 1996, Vol. 3, pp. 611–616.

Gu, C., Lee, M.C., 1998. Semiautomatic segmentation and tracking of semantic video objects. IEEE Transactions on Circuits and Systems for Video Technology 8 (5), 572–584.

Kanatani, K., 1987. Camera rotation invariance of image characteristics. Computer Vision, Graphics, Image Processing 39 (3), 328–354.

Kompatsiaris, I., Tzovaras, D., Strintzis, M.G., 1998. 3-D model-based segmentation of videoconference image sequences. IEEE Transactions on Circuits and Systems for Video Technology 8 (5), 547–561.

Meier, T., Ngan, K.N., 1998. Automatic segmentation of moving objects for video object plane generation. IEEE Transactions on Circuits and Systems for Video Technology 8 (5), 525–538.

Moscheni, F., Bhattacharjee, S., Kunt, M., 1998. Spatiotemporal segmentation based on region merging. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (9), 897–915.

MPEG Requirements, 1998. Overview of MPEG-4 profiles and levels. ISO/JTC1/SC29/WG11 N2325, Dublin MPEG Meeting.

MPEG Video & SNHC, 1998. Final text for FCD 14496-2:visual. Doc. ISO/MPEG N2202, Tokyo MPEG Meeting.

Murray, D., Basu, A., 1994. Motion tracking with an active camera. IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (5), 449–459.

Reinders, M., Beek, P., Sankur, B., Lubbe, J., 1995. Facial feature localization and adaptation of a generic face model for model-based coding. Signal Processing: Image Communication 7, 57–74.

Yin, L., Basu, A., 1997. MPEG4 face modeling using fiducial points. In: Proceedings of IEEE International Conference on Image Processing, Santa Barbara, CA, October 1997, Vol. 1, pp. 109–112.

Yin, L., Basu, A., 1998. A robust method for realistic facial model adaptation using self-adaptive energy minimization. Technique Report, Department of Computing Science, University of Alberta.

Yuille, A.L., Hallinan, P.W., Cohen, D.S., 1992. Feature extraction from faces using deformable templates. International Journal of Computer Vision 8 (2), 99–111.

Zhang, L., 1998. Automatic adaptation of a face model using action units for semantic coding of videophone sequences. IEEE Transactions on Circuits and Systems for Video Technology 8 (6), 781–795.