Dynamically Adaptive Tracking of Gestures and Facial Expressions*

D. Metaxas, G. Tsechpenakis, Z. Li, Y. Huang, and A. Kanaujia

Center for Computational Biomedicine, Imaging and Modeling (CBIM),
Computer Science Department, Rutgers University,
110 Frelinghuysen Rd, Piscataway, NJ 08854
{dnm, gabrielt, kanaujia}@cs.rutgers.edu, zhli@paul.rutgers.edu,
yuchi.huang@gmail.com

Abstract. We present a dynamic data-driven framework for tracking gestures and facial expressions from monocular sequences. Our system uses two cameras, one for the face and one for the body view for processing in different scales. Specifically, and for the gesture tracking module, we track the hands and the head, obtaining as output the blobs (ellipses) of the ROIs, and we detect the shoulder positions with straight lines. For the facial expressions, we first extract the $2D$ facial features, using a fusion between KLT tracker and a modified Active Shape Model, and then we obtain the 3D face mask with fitting a generic model to the extracted $2D$ features. The main advantages of our system are (i) the adaptivity, i.e., it is robust to external conditions, e.g., lighting, and independent from the examined individual, and (ii) its computational efficiency, providing us results off- and online with a rates higher than $20fps$.

1 Introduction

Behavioral indicators of deception and behavioral states are extremely difficult for humans to analyze. Our framework aims at analyzing nonverbal behavior on video, by tracking the gestures and facial expressions of an individual that is being interviewed.

Our system uses two cameras (one for the face and one for the whole body view), for analysis in two different scales, and consists of the following modules: (a) head and hands tracking, using Kalman filtering [9] and an data-driven adaptive (to each specific individual) skin regions detection method, (b) shoulders tracking, based on a novel texture-based edge localization method, (c) $2D$ facial features tracking, using a fusion between the KLT tracker [12, 15] and different Active Shape Models [2], and (d) $3D$ face and facial features tracking, using the $2D$ tracking results and our novel $3D$ face tracking method. The main advantages of our framework is that we can track both gestures and facial expressions with great accuracy and robustness, in rates higher than $20fps$.

* This research has been funded by an NSF-ITR/NGS-0313134 and an NSF-ITR-[ASE+ECS]-0428231 Collaborative Project to the first author.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part III, LNCS 3993, pp. 554–561, 2006.

This paper is organized as follows. In the next subsection we give a brief overview of the previous work on gestures and face tracking and in section 2 we describe our approach. In subsections 2.1-2.4, we describe the individual parts of our system, namely the head and hands tracking, the shoulders localization, the $2D$ facial features tracking, and the $3D$ face tracking, respectively. In section 3 we present the results of our system and in section 4 we present our conclusions.

1.1 Previous Work

Research efforts have investigated gesture analysis [8], but accurate tracking of gestures is still an open topic. According to our work presented in [11], using color analysis, eigenspace-based shape segmentation, and Kalman filters [9], we have been able to track the position, size, and angle of the face and hands regions. In this work we use the color distribution from an image sequence. A Look-Up-Table (LUT) with three color components (red, green, blue) is created based on the color distribution of the face and hands. This three-color LUT, called a 3D-LUT, is built in advance of any analysis and was formed using skin color samples. After extracting the hand and face regions from an image sequence, this method computes elliptical *blobs* identifying candidates for the face and hands. Thus, the most face-like and hand-like regions in a video sequence are identified. Although we have been able to track successfully a wide variety of individuals, we could only use this method under controlled lighting conditions and for a limited range of skin colors.

On the other hand, accurate facial feature localization and tracking under different head poses and facial expressions is a very challenging task. The general problem of feature tracking was presented in the early work of Lucas et.al [12], was developed fully by Tomasi et.al. [15], and was explained explicitly in the paper of Shi et.al. [13]. Later, Tomasi proposed a slight modification which makes the computation symmetric with respect to the two images; the resulting equation is derived in the unpublished note of Birchfield [1].

The appearance of a face can change dramatically as the head poses and the facial expressions change. Sometimes these changes constitute a very difficult problem. Several methods have addressed the issue, such as the multi-view Active Shape Models (ASMs) [2], Active Appearance Models (AAMs) [3], and their extensions [4]. In [10] a summary to multi-view face detection, alignment and recognition is presented, but facial features are not accurately tracked under large head rotations and exaggerated expressions.

Finally, for the 3D face tracking, several algorithms [5, 14] use low-level image feature extraction methods to recover the model state. Most parameterized three dimensional deformable models rely on the correct estimation of image features that are then used to obtain good estimates for the model's state. While the 3D deformable model can reduce some ambiguities to some degree, the unreliable low-level image cues can still cause error accumulation of the 3D model.

2 Our Framework

In this section we describe the different parts of our integrated system. For gesture tracking, we use a camera capturing the entire body of the interviewed individual, whereas a second camera is also used to capture the individual's face in a closer view.

2.1 Head and Hands Tracking

The drawback of our previous approach of [11] is that color is not a robust feature for skin detection and tracking, even if we collect and use in the 3D-LUT a wide variety of skin colors. This method may perform with great accuracy in a controlled environment, i.e., fixed light sources and smooth background (with a color much different from all skin samples of the 3D-LUT), but it is not robust under varying conditions.

To overcome the above problem, we use a data-driven adaptive alternative to the 3D-LUT, i.e., a method that is automatically adapted to the specific individual that is being interviewed. More specifically, our approach consists of the following steps. (i) The first step of our approach is to construct a skin color database to estimate a generic color distribution. Instead of using a 3D-LUT as previously, we estimate a gaussian distribution that fits to the collected color samples. This generic distribution will be modified and adapted to the specific individual's skin color. (ii) In the first frames of the video sequence (optimally 20 frames), we detect the face region of the interviewed individual, using the method proposed by Viola et.al. [16]. This method is experimentally proved to be efficient in terms of computational time, and robust under varying conditions; also it does not require any skin color information and avoids tracking error accumulation over time, since it detects the desired region in each frame separately. (iii) The next step of our approach is to track the face and hands, given the face detection results in the first frames of the sequence. Based on the skin color of the detected facial region, we modify the generic skin distribution (gaussian fitting to the skin color samples, including the specific individual's facial color), and use this new distribution for the detection of the skin regions. In this way, our system automatically adapts to different lighting conditions and all possible skin color variations. (iv) For fast and accurate tracking, free of error drift, we use primarily Kalman tracking [9], and every 10 frames we re-detect the skin regions (face and hands) for the tracker re-initialization. (v) The final result of our method, similarly to our previous work of [11], is the head (face) and hands blobs, extracted with ellipse fitting in the extracted skin regions. To refine our tracking results, we also extract the edges inside the detected skin regions, using the Canny method. For the hand regions, we estimate the edge densities; in cases of individuals with short-sleeve clothes, we segment the hands from the arms based on the hands' increased edge densities (compared to the arms' edge densities). In this way, we extract the blobs of the hand regions and not the entire arms.

2.2 Shoulder Tracking

Tracking the shoulders enables us to detect events such as shrugging, but also estimate relative positions of the hands to the shoulders. In the regions left and right of the estimated head region, we apply the Canny edge detection method. We extract all the edges in each one of these two regions, i.e., the actual (unknown) shoulder edges, background edges, and edges inside the individual's torso region. Our aim is to detect the shoulders, excluding all the undesired edges. To achieve that, we estimate the texture inside these regions, in a block-based manner: we estimate the texture of all blocks centered at the pixels in the normal direction of each edge. In this way, in the normal direction of each edge, at each edge pixel, we obtain a function of texture values. In our system we used the method of Zhang et.al. [17] for the texture estimation. If a detected edge corresponds to the desired shoulder, for all edge points the corresponding texture functions must have a *change-point* at these points respectively. The term *change-point* is obtained from the statistical change-point detection theory [6]; in our framework, in order to detect the texture change points, we use the CUSUM procedure [6]. On the other hand, if an edge does not correspond to a shoulder, the estimated texture functions have either random change-points (not corresponding to the same edge) or no change-points at all. The final result of the shoulder detection is a straight line of fixed length.

2.3 2D Facial Features Tracking

Our framework tracks robustly and accurately facial feature points under multiple views and different expressions. We use an ASM [2] to localize the facial feature points accurately in the first video frame, and to supervise the tracking results in the following frames. For accurate face localization, some local ASMs for the mouth and eyes are used. To track the facial feature points obtained by the multi-pose ASM, we use the KLT tracker [12, 15, 13]. In our framework, this coupling between the KLT tracker and the ASM provides us fast and accurate results, ideal for real-time tracking of facial expressions.

We assume that the face of the interviewed individual is in the frontal view in the first frame of the sequence. We also assume that there is only one person in front of the camera. For the first frame, we use the face detection method of Viola et.al. [16] to obtain a bounding box of the face. Then a mean ASM shape (initialization) is inserted into the bounding box, and the frontal-view ASM is used to localize the facial features accurately. After the facial features initialization, we use the KLT tracker to track the features over time. In parallel to the KLT tracker, we use different ASMs to ensure that the facial features are estimated accurately. This fusion between the KLT tracker and different ASMs makes our framework robust to different head positions and facial expressions.

The frontal-view ASM used in the first frame is trained off-line, using 100 frontal-view face images and their corresponding feature points. For every face we use 87 feature points, which determine different facial features. All these points are manually marked before training. In the ASM, PCA models are trained

both for shape variation and local profile variation. Note that four-level multi-resolution strategy is used here: we train profile variations and search for every point in every level.

Apart from the frontal-view ASM, we also train four ASMs for the left- and right-view, upward and downward head pose, using 70 individuals. These models are used in real-time face tracking to ensure that there is no error accumulation over time that may lead to the loss of track. After the feature points are localized by the frontal-view ASM, the KLT tracker is used to track those feature points over time. In our framework we also integrated Birchfield's code [1]: for every input frame we use the KLT tracker to track the feature points obtained from the previous frame; then the current points are introduced into one of the ASM shape subspaces. To decide which model will be used for each input frame, we assume that transitions between ASMs are possible only between the frontal-view and one of other four ASMs. We compute the distances between (a) the nose and the left and right cheek, and (b) the nose and the chin. The model is then chosen based on the changes of the ratios of these distances.

2.4 3D Face Tracking

The main advantage of deformable face models is the reduced dimensionality. The smaller number of degree of freedom makes the system more robust and efficient. However, the accuracy and reliability of a deformable model tracking application is strongly dependent on how well the object under tracking fits the family of shapes described by the parameters of the model.

A 3D deformable model is parameterized by a vector of parameters $\mathbf{q}$. Changes in $\mathbf{q}$ causes geometric deformations of the model. A particular point on the surface is denoted by $\mathbf{x}(\mathbf{q};\mathbf{u})$ with $\mathbf{u} \in \Omega$. The goal of a shape and motion estimation process is to recover parameter $\mathbf{q}$ from face image sequence. The parameters $\mathbf{q}$ can be divided into two parts: static parameter $\mathbf{q}_b$, which describes the unchanging features of the face, and dynamic parameter $\mathbf{q}_m$, which describes the global (rotation and translation of the head) and local deformation (facial expressions) of an observed face during tracking.

The deformations can also be divided into two parts: $\mathbf{T}_b$ for shape and $\mathbf{T}_m$ for motion (expression), so that:

$$\mathbf{x}(\mathbf{q};\mathbf{u}) = \mathbf{T}_m(\mathbf{q}_m;\mathbf{T}_b(\mathbf{q}_b;s(\mathbf{u}))) \tag{1}$$

The kinematics of the model is $\dot{\mathbf{x}}(\mathbf{u}) = \mathbf{L}(\mathbf{q};\mathbf{u})\dot{\mathbf{q}}$, Where $\mathbf{L} = \frac{\partial \mathbf{x}}{\partial \mathbf{q}}$ is the model Jacobian. Considering the face images under a perspective camera with focal length f, the point $\mathbf{x}(\mathbf{u}) = (x,y,z)^T$ projects to the image point $\mathbf{x}_p(\mathbf{u}) = \frac{f}{z}(x,y)^T$. The kinematics of the new model is given by:

$$\dot{\mathbf{x}_p}(\mathbf{u}) = \frac{\partial \mathbf{x}_p}{\partial \mathbf{x}}\dot{\mathbf{x}}(\mathbf{u}) = (\frac{\partial \mathbf{x}_p}{\partial \mathbf{x}}\mathbf{L}(\mathbf{q};\mathbf{u}))\dot{\mathbf{q}} = \mathbf{L}_p(\mathbf{q};\mathbf{u})\dot{\mathbf{q}} \tag{2}$$

Where

$$\frac{\partial \mathbf{x}_p}{\partial \mathbf{x}} = \begin{bmatrix} f/z & 0 & -fx/z^2 \\ 0 & f/z & -fy/z^2 \end{bmatrix} \tag{3}$$

In a physics based deformable model framework, optimization of the parameters is carried out by integrating differential equations derived form the Euler-Lagrange equations of motion:

$$\dot{\mathbf{q}} = \mathbf{f}_q \tag{4}$$

Where the generalized forces $\mathbf{f}_q$ are identified by the displacements between the actual projected model points and the identified corresponding 2D image features. They are computed as:

$$\mathbf{f}_q = \sum_j (\mathbf{L}_p(\mathbf{u}_j)^T \mathbf{f}_{image}(\mathbf{u}_j)) \tag{5}$$

Given an adequate model initialization, these forces will align features on the model with image features, thereby determining the object parameters. The dynamic system in equation 4 is solved by integrating over time, using standard differential equation integration techniques:

$$\mathbf{q}(t+1) = \mathbf{q}(t) + \dot{\mathbf{q}}(t)\Delta t \tag{6}$$

Goldenstein et.al. showed in [7] that the image forces $\mathbf{f}_{image}$ and generalized forces $\mathbf{f}_q$ in these equations can be replaced with affine forms that represent probability distributions, and furthermore that with sufficiently many image forces, the generalized force converges to a Gaussian distribution. In our framework, we take advantage of this property by integrating the contributions of ASMs with other cues, so as to achieve robust tracking even when ASM methods and standard 3D deformable model tracking methods provide unreliable results by themselves.

3 Experimental Results

Fig. 1 illustrates an example of tracking the head, hands and shoulders of an individual during an interview, using our framework as described in subsections 2.1 and 2.2. The detected shoulders are shown in red straight lines. The red rectangles illustrate the Kalman tracking results, and the blue rectangles show our tracking results before the blobs estimation, using both Kalman filtering and skin region detection. The head and hands final tracking results (blobs) are shown in white ellipses (along with their major and minor axes).

Fig. 2 illustrates an example of face tracking in six key-frames of a sequence. The the upper images show the results of our 2D face tracking method, whereas the lower images show our results for the 3D face tracking, using the extracted 2D features.

4 Summary and Conclusions

We presented dynamic data-driven framework for tracking gestures and facial expressions from monocular sequences. From the gesture tracking module, we

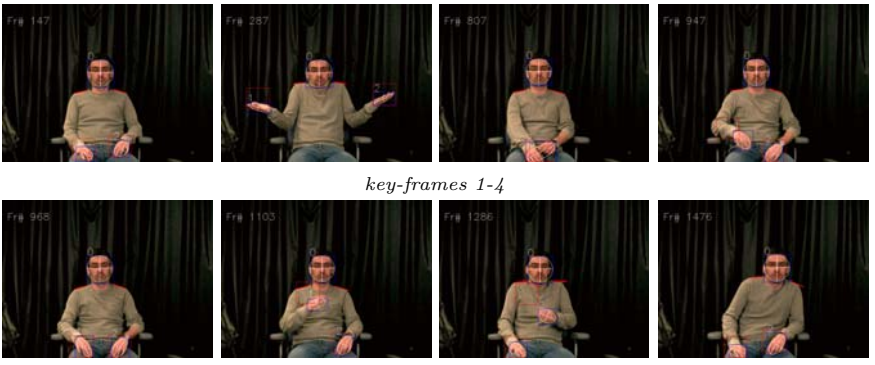

Fig. 1. Head, hands and shoulders tracking results

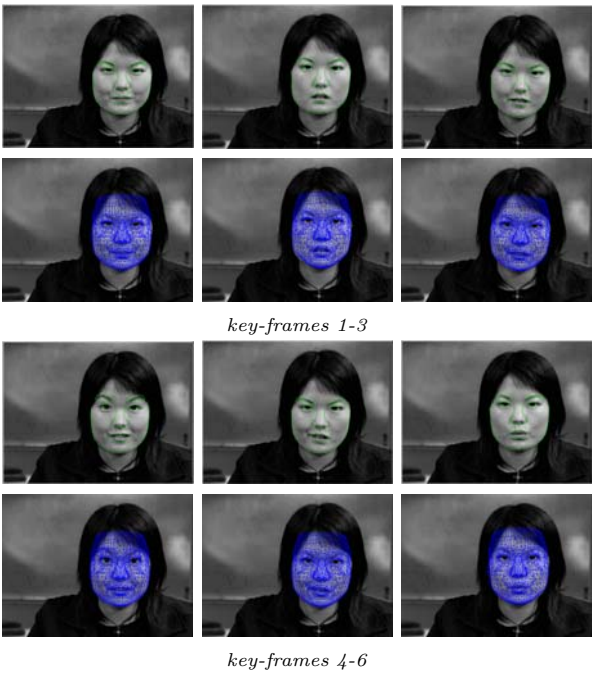

Fig. 2. 2D and 3D face tracking results for six key-frames of a sequence

obtain the blobs (ellipses) of the head and hands, and we detect the shoulder positions with straight lines. For the facial expressions, we first extract the $2D$ facial features, and then we obtain the 3D face information, using the extracted 2D features. The main advantages of our system are its robustness to lighting changes, its adaptivity to every examined individual, and its computational efficiency, with rates higher than $20fps$.

References

1. S. Birchfield, "Derivation of Kanade-Lucas-Tomasi Tracking Equation," *web-published at http://www.ces.clemson.edu/ stb/klt/*, May 1996.
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*, vol. 61(1), p. 389, January 1995.
3. T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Model," *5th Europen Conference on Computer Vision*, Freiburg, Germany, 1998.
4. T.F.Cootes, and P. Kittipanya, "Comparing Variations on the Active Appearance Model Algorithm," *Brittish Machine Vision Conderence*, University of Cardiff, September 2002.
5. D. de Carlo, and D. Metaxas, "Optical Flow Constraints on Deformable Models with Applications to Face Tracking," *International Journal of Computer Vision*, vol. 38(2), pp. 99-127, July 2000.
6. J.L. Devore, Probability and Statistics for Engineering and the Sciences, Pacific Grove, Calif.: Brooks/Cole Pub. Co, 2004.
7. S. Goldenstein, C. Vogler, and D. Metaxas, "Statistical Cue Integration in Deformable Models," *Pattern Analysis and Machine Intelligence*, vol. 25(7), pp. 801-813, 2003.
8. D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," Computer Vision and Image Understanding, Vol. 73(1), pp.82-98, 1999.
9. R.E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," ASMEJournal of Basic Engineering, pp. 35-45, March 1960.
10. S.Z. Li, X.L. Zou, Y.X. Hu, Z.Q. Zhang, S.C. Yan, X.H. Peng, L. Huang, and H.J. Zhang, "Real-Time Multi-View Face Detection, Tracking, Pose Estimation, Alignment, and Recognition," *IEEE Conference on Computer Vision and Pattern Recognition, Demo Summary*, Hawaii, December 2001.
11. S. Lu, G. Tsechpenakis, D. Metaxas, M.L. Jensen, and J. Kruse, "Blob Analysis of the Head and Hands: A Method for Deception Detection and Emotional State Identification," *Hawaii International Conference on System Sciences*, Big Island, Hawaii, January 2005.
12. B.D. Lucas, and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *International Joint Conference on Artificial Intelligence*, pp. 674-679, 1981.
13. J. Shi, and C. Tomasi, "Good Features to Track," *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994.
14. H. Tao and T. Huang, "Visual Estimation and Compression of Facial Motion Parameters: Elements of a 3D Model-based Video Coding System," *International Journal of Computer Vision*, vol. 50(2), pp. 111-125, 2002.
15. C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," *Carnegie Mellon University Technical Report CMU-CS-91-132*, April 1991.
16. P.A. Viola, and M.J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57(2), pp. 137-154, 2004.
17. H. Zhang, J.E. Fritts, and S.A. Goldman, "A Fast Texture Feature Extraction Method for Region-based Image Segmentation," *16th Annual Symposium on Image and Video Communication and Processing*, SPIE Vol. 5685, January 2005.