Face Tracking

Amit K. Roy-Chowdhury and Yilei Xu

Department of Electrical Engineering, University of California, Riverside, CA 92521, USA
{amitrc,yxu}@ee.ucr.edu

Synonyms

Facial Motion Estimation

Definition

In many face recognition systems, the input is a video sequence consisting of one or more faces. It is necessary to track each face over this video sequence so as to extract the information that will be processed by the recognition system. Tracking is also necessary for 3D model-based recognition systems where the 3D model is estimated from the input video. Face tracking can be divided along different lines depending upon the method used, e.g., head tracking, feature tracking, image-based tracking, model-based tracking. The output of the face tracker can be the 2D position of the face in each image of the video (2D tracking), the 3D pose of the face (3D tracking), or the location of features on the face. Some trackers are able to output other parameters related to lighting or expression. The major challenges encountered by face tracking systems are robustness to pose changes, lighting variations, and facial deformations due to changes of expression, occlusions of the face to be tracked and clutter in the scene that makes it difficult to distinguish the face from the other objects.

Main Body Text

Introduction

Tracking, which is essentially motion estimation, is an integral part of most face processing systems. If the input to a face recognition system is a video sequence, as obtained from a surveillance camera, tracking is needed to obtain correspondence between the observed faces in the different frames and to align the faces. It is so integral to video-based face recognition systems that some existing methods integrate tracking and recognition [1]. It is also a necessary step for building 3D face models. In fact, tracking and 3D modeling are often treated as two parts of one single problem [2, 3, 4].

There are different ways to classify face tracking algorithms [5]. One such classification is based on whether the entire face is tracked as a single entity (sometimes referred to as head tracking) or whether individual facial features are tracked. Sometimes a combination of both is used. Another method of classification is based on whether the tracking is in the 2D image space or in 3D pose space. For the former, the output (overall head location or facial feature location) is a region in the 2D image and does not contain information about the change in the 3D orientation of the head. Such methods are usually not very robust to changes of pose, but are easier to handle computationally. Alternatively, 3D tracking methods, which work by fitting a 3D model to each image of the video, can provide estimates of the 3D pose of the face. However, they are usually more computationally intensive. Besides, many advanced face tracking methods are able to handle challenging situations like facial deformations, changes of lighting, and partial occlusions.

We will first give a broad overview of the basic mathematical framework of face tracking methods, followed by a review of the current state of the art and technical challenges. Next, we will consider a few application scenarios, like surveillance, face recognition and face modeling, and discuss the importance of face tracking in each of them. We will then show some examples of face tracking in challenging situations, before concluding this article.

Basic Mathematical Framework

We provide here an overview of the basic mathematical framework that explains the process in which most trackers work. Let $\mathbf{p} \in \Re^p$ denote a parameter vector that is the desired output of the tracker. It could a 2D location of the face in the image, the 3D pose of the face, or a more complex set of quantities that also include lighting and deformation parameters. We define a synthesis function $f : \Re^2 \times \Re^p \rightarrow \Re^2$ that can take an image pixel $\mathbf{v} \in \Re^2$ at time $(t-1)$ and transform it to $f(\mathbf{v}, \mathbf{p})$ at time t. For a 2D tracker, this function f could be a transformation between two images at two consecutive time instants. For a 3D model-based tracker, this can be considered as a rendering function of the object at pose $\mathbf{p}$ in the camera frame to the pixel coordinates $\mathbf{v}$ in the image plane. Given an input image $I(\mathbf{v})$, we want to align the synthesized image with it so as to obtain

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} g(f(\mathbf{v}, \mathbf{p}) - I(\mathbf{v})), \tag{1}$$

where $\hat{\mathbf{p}}$ denotes the estimated parameter vector for this input image $I(\mathbf{v})$.

The essence of this approach is the well-known Lucas-Kanade tracking, an efficient and accurate implementation of which has been proposed using the inverse compositional approach [6]. Depending on the choice of $\mathbf{v}$ and $\mathbf{p}$, the method is applicable to the overall face image, a collection of discrete features, or a 3D face model. The cost function g is often implemented as an L_2 norm, i.e., the sum of the squares of the errors over the entire region of interest. However, other distance metrics may be used. Thus a face tracker is often implemented as a least-squares optimization problem.

Let us consider the problem of estimating the change, $\triangle \mathbf{p}_t \triangleq \mathbf{m}_t$, in the parameter vector between two consecutive frames, $I_t(\mathbf{v})$ and $I_{t-1}(\mathbf{v})$ as

$$\hat{\mathbf{m}_t} = \arg\min_{\mathbf{m}} \sum_{\mathbf{v}} \big(f(\mathbf{v}, \hat{\mathbf{p}}_{t-1} + \mathbf{m}) - I_t(\mathbf{v})\big)^2, \tag{2}$$

and

$$\hat{\mathbf{p}}_t = \hat{\mathbf{p}}_{t-1} + \hat{\mathbf{m}_t}. \tag{3}$$

The optimization of the above equation can be achieved by assuming a current estimate of $\mathbf{m}$ is known and iteratively solve for increments $\triangle \mathbf{m}$ such that

$$\sum_{\mathbf{v}} \big(f(\mathbf{v}, \hat{\mathbf{p}}_{t-1} + \mathbf{m} + \triangle \mathbf{m}) - I_t(\mathbf{v})\big)^2 \tag{4}$$

is minimized.

Performance Analysis

While the basic idea of the face tracking algorithms is simple, the challenge comes in being able to perform the optimization efficiently and accurately. The function, f, will be non-linear, in general. This is because f will include camera projection, the 3D pose of the object, the effect of lighting, the surface reflectance, non-rigid deformations and other factors. For example, in [7] the authors derived a bilinear form for this function under the assumption of small motion. It could be significantly more complex in general. This complexity makes it difficult to obtain a global optimum for the optimization function unless a good starting point is available. This initialization is often obtained through a face detection module working on the first frame of the video sequence. For 3D model-based tracking algorithms, it also requires registration of the 3D model to the detected face in the first frame.

The need for a good initialization for stable face tracking is only one of the problems. All trackers suffer from the problem of drift of the estimates and face tracking is no exception. Besides, the synthesis function f may be difficult to define precisely in many instances. Examples include partial occlusion of the face, deformations due to expression changes and variations of lighting including cast shadows. Special care needs to be taken to handle these situations since direct optimization of the cost function (2) would give an incorrect result.

Computational speed is another important issue in the design of tracking algorithms. Local optimization methods like gradient descent, Gauss-Newton and Levenberg-Marquardt [8] can give a good result if the starting point is close to the desired solution. However, the process is often slow because it requires recomputation of derivatives at each iteration. Recently, an efficient and accurate method of performing the optimization has been proposed by using an inverse compositional approach that does not require recomputation of the gradients at each step [6]. In this approach, the transformation between two frames is represented by a warping function which is updated by first inverting the incremental warp and then composing it with the current estimate. Our independent experimental evaluation has shown that on real-life facial video sequences, the inverse compositional approach leads to a speed-up by at least one order of magnitude, and often more, leading to almost real-time performance in most practical situations.

Challenges in Face Tracking

As mentioned earlier, the main challenges that face tracking methods have to overcome are (i) variations of pose and lighting, (ii) facial deformations, (iii) occlusion and clutter, and (iv) facial resolution. These are the areas where future research in face tracking should concentrate. We will now briefly review some of the methods that have been proposed to address these problems.

- **Robustness to Pose and Illumination Variations:** Pose and illumination variations often lead to loss of track. One of the well-known methods for dealing with illumination variations was presented in [9], where the authors proposed using a parameterized function to describe the movement of the image points, taking into account illumination variation by modifying the brightness constancy constraint of optical flow. Illumination invariant 3D tracking was considered within the Active Appearance Model (AAM) framework in [10], but the method requires training images to build the model and the result depends on the quality and variety of such data. 3D model based motion estimation algorithms are the usually robust to pose variations, but often lack robustness to illumination. In [7], the authors proposed a model-based face tracking method that was robust to both pose and lighting changes. This was achieved through an analytically derived model for describing the appearance of a face in terms of its pose, the incident lighting, shape and surface reflectance. Figure 1 shows an example.
- **Tracking through Facial Deformations:** Tracking faces through changes of expressions, i.e., through facial deformations, is another challenging problem. An example of face tracking through changes of expression and pose is shown in Figure 2. A survey of work on facial expression analysis can be found in [12]. The problem is closely related to modeling of facial expressions, which has applications beyond tracking, notably in computer animation. A well-known work in this area is [13], which has been used by many researchers for tracking, recognition and reconstruction. In contrast to this model-based approach, the authors in [14] proposed a data-driven approach for tracking and recognition of non-rigid facial motion. More recently, the 3D morphable model [15] has been quite popular in synthesizing different facial expressions, which implies that it can also be used for tracking by posing the problem as estimation of the synthesis parameters (coefficients of a set of basis functions representing the morphable model).
- **Occlusion and Clutter:** As with most tracking problems, occlusion and clutter affect the performance of most face trackers. One of the robust tracking approaches in this scenario is the use of particle filters [16] which can recover from a loss of track given a high enough number of particles and observations. However, in practice, occlusion and clutter remain serious impediments in the design of highly robust face tracking systems.
- **Facial resolution:** Low resolution will hamper performance of any tracking algorithm, face tracking being no exception. In fact, [5] identified low resolution to be one of the main impediments in video-based face recognition. Figure 3 shows an example of tracking through scale changes and illumination. Super-resolution approaches can be used to overcome these problems to some extent. However, super-resolution of faces is a challenging problem by itself because of detailed facial features that need to be modeled accurately. Recently, [17] proposed a method for face super-resolution using AAMs. Super-resolution requires registration of multiple images, followed by interpolation. Usually, these two stages are treated separately, i.e., registration is obtained through a tracking procedure followed by super-resolution. In a recent paper [18], the authors proposed feeding back the super-resolved texture in the n^{th} frame for tracking the $(n+1)^{\text{th}}$ frame. This improves the tracking, which, in turn, improves the super-resolution output. This could be an interesting area of future work taking into consideration issues of stability and convergence.

Some Applications of Face Tracking

We highlight below some applications where face tracking is an important component.

- **Video Surveillance:** Since faces are often the most easily recognizable signature of identity and intent from a distance, video surveillance systems often focus on the face [5]. This requires tracking the face over multiple frames.
- **Biometrics:** Video-based face recognition systems require alignment of the faces before they can be compared. This alignment compensates for changes of pose. Face tracking, especially 3D pose estimation, is therefore an important component of such applications. Also, integration of identity over the entire video sequence requires tracking the face [1].
- **Face Modeling:** Reconstruction of the 3D model of a face from a video sequence using structure from motion requires tracking. This is because the depth estimates are related non-linearly to the 3D motion of the object. This is a difficult non-linear estimation problem and many papers can be found that focus primarily on this, some examples being [2, 3, 4].

- **Video Communications and Multimedia Systems:** Face tracking is also important for applications like video communications. Motion estimates remove the inter-frame redundancy in video compression schemes like MPEG and H.26x. In multimedia systems like sports videos, face tracking can be used in conjunction with recognition or reconstruction modules, or for focusing on a region of interest in the image.

Summary

Face tracking is an important problem for a number of applications, like video surveillance, biometrics, video communications, and so on. A number of methods have been proposed that work reasonably well under moderate changes of pose, lighting and scale. The output of these methods vary from head location in the image frame to tracked facial features to 3D pose estimation. The main challenge that future research should address is robustness to changing environmental conditions, facial expressions, occlusions, clutter and resolution.

Related Entries

Face Alignment, Face Recognition, Face Features

References

1. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding **91** (2003) 214–245
2. Fua, P.: Regularized Bundle-Adjustment to Model Heads from Image Sequences without Calibration Data. International Journal of Computer Vision **38** (2000) 153–171
3. Shan, Y., Liu, Z., Zhang, Z.: Model-Based Bundle Adjustment with Application to Face Modeling. In: Proc. of IEEE International Conference on Computer Vision. (2001) 644–651
4. Roy-Chowdhury, A., Chellappa, R., Gupta, R.: 3D Face Modeling From Monocular Video Sequences. In: Face Processing: Advanced Modeling and Methods. Academic Press (2005)
5. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Transactions (2003)
6. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision **56** (2004) 221–255
7. Xu, Y., Roy-Chowdhury, A.: Integrating Motion, Illumination and Structure in Video Sequences, With Applications in Illumination-Invariant Tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence (2007) 793–806
8. Luenburger, D.: Optimization by Vector Space Methods. John Wiley and Sons (1969)
9. Hager, G.D., Belhumeur, P.: Efficient Region Tracking With Parametric Models of Geometry and Illumination. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 1025–1039
10. Koterba, S., Baker, S., Matthews, I., Hu, C., Xiao, H., Cohn, J., Kanade, T.: Multi-view aam fitting and camera calibration. In: IEEE Intl. Conf. on Computer Vision. (2005)
11. Lepetit, V., Fua, P.: Monocular Model-Based 3D Tracking of Rigid Objects. Now Publishers Inc. (2005)
12. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition (2003)
13. Terzopoulos, D., Waters, K.: Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. IEEE Trans. on Pattern Analysis and Machine Intelligence **15** (1993) 569–579
14. Black, M., Yacoob, Y.: Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Models of Image Motion. In: International Conf. on Computer Vision. (1995) 374–381
15. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. on Pattern Analysis and Machine Intelligence **25** (2003) 1063–1074
16. Arulampalam, M., Maskell, A., Gordon, N., Clapp, T.: A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. IEEE Trans. on Signal Processing **50** (2002)
17. Dedeoglu, G., Baker, S., Kanade, T.: Resolution-aware fitting of active appearance models to low-resolution images. In: European Conference on Computer Vision. (2006)
18. Yu, J., Bhanu, B., Xu, Y., Roy-Chowdhury, A.: Super-resolved facial texture under changing pose and illumination. In: Intl. Conf. on Image Processing. (2007)

Definitional Entries

Cost Function

Tracking is an estimation process that computes the position of a target based on optimization of a criterion that relates the observations with the estimates. The criterion is represented mathematically using the cost function.

Optimization

Given a cost function, different strategies could be used to obtain the estimate. This is called the optimization strategy and the solution often depends upon the exact strategy that is used.

Motion Estimation

Face tracking can be looked upon as estimating the motion of the face over subsequent video frames. This can be the 2D motion on the image plane or the 3D pose of the face.

Illumination

The lighting on the face affects the quality of the image, and hence the tracking. Illumination is therefore one of the major challenges in face recognition. The effects of illumination could be locally constrained to a particular facial region, or can be global over the entire image. Some face trackers are able to estimate the illumination, in addition to motion.

Deformation

Since expressions are common in faces, robust face tracking methods should be able to perform well in spite of large facial expressions. Also, many face trackers are able to estimate the expression.

Fig. 1. Tracked points on a face through changes of pose and illumination. These points are projections of a 3D face mesh model.

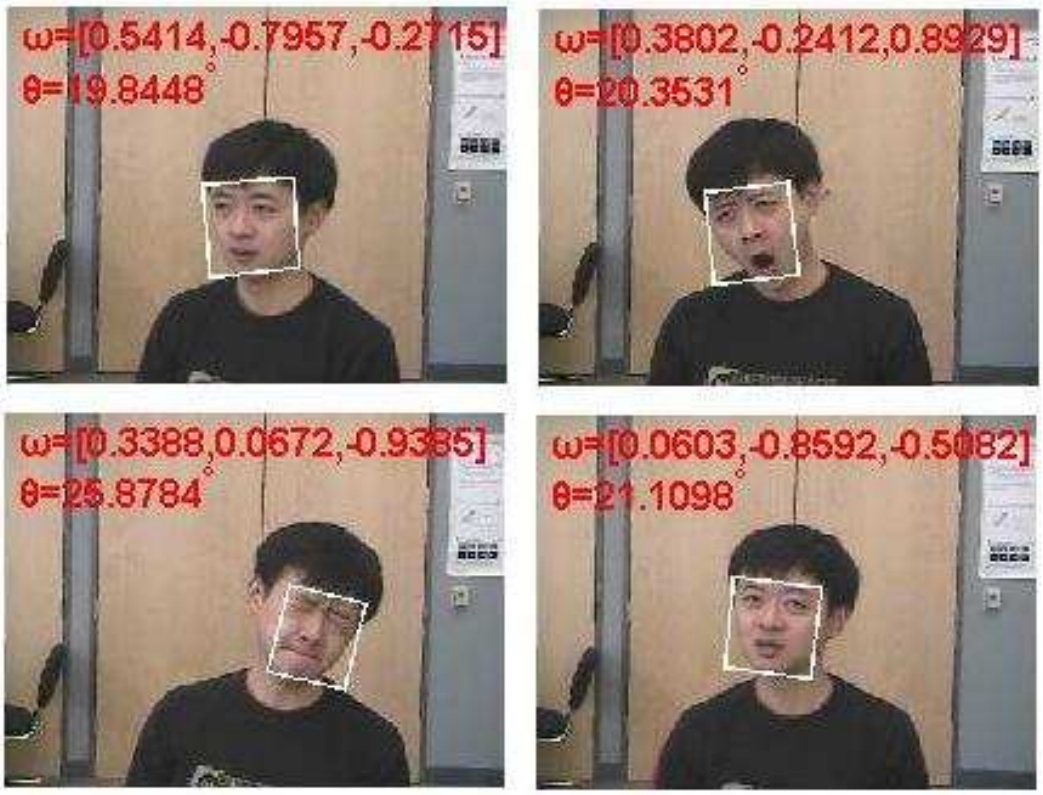

Fig. 2. An example of face tracking under changes of pose and expressions. The estimated pose is shown on the top of the frames. The pose is represented as a unit vector for the rotation axis, and the rotation angle in degrees, where the reference is taken to be the frontal face.

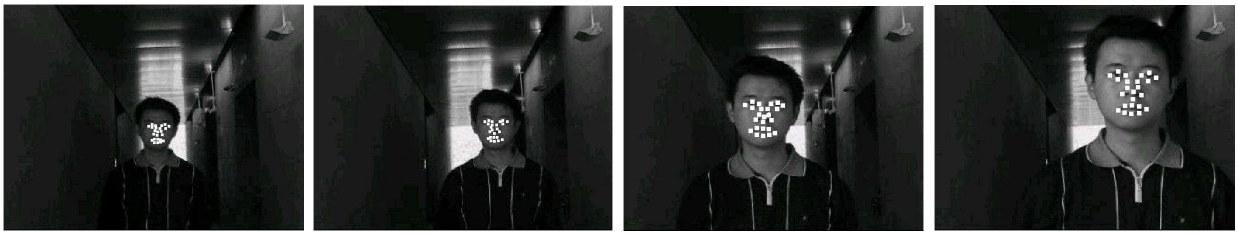

Fig. 3. Tracked points on a face through changes of scale and illumination.