Model- and Exemplar-based Robust Head Pose Tracking Under Occlusion and Varying Expression

Le Lu[†*], Zhengyou Zhang[‡ §], Heung-Yeung Shum[†], Zicheng Liu[‡], Hong Chen[†]

† Microsoft Research China, Beijing, 100080, P.R.China
‡ Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
§ Contact author. Email: zhang@microsoft.com

Abstract

We present a model- and exemplar-based technique for head pose tracking. Because of the dynamic nature, it is not possible to represent face appearance by a single texture image. Instead, we sample the complex face appearance space by a few reference images (exemplars). By taking advantage of the rich geometric information of a 3D face model and the flexible representation provided by exemplars, our system is able to track head pose robustly under occlusion and/or varying facial expression. The system starts with a simple learning stage. The user moves his/her head with a neutral expression in front of the camera within the working space. Our system automatically builds a personalized 3D face model by fitting a generic mesh model to a near frontal facial image, and acquires a few reference images at distinct poses to sparsely sample the facial appearance space. When tracking the head under occlusion and varying expression, we match the current view against the most appropriate reference image according to the predicted pose, which is much easier and more robust than if only a single texture image is used. A robust motion segmentation algorithm is used to separate point matches corresponding to rigid head motion from those corresponding to facial deformation. The head pose can then be reliably estimated from the rigid-motion points with the help of the 3D face mesh model, even when the number of points is small. Since we use reference images during tracking, the accumulative error inherent in frame-by-frame tracking is avoided and more accurate pose estimation is achieved. We demonstrate the validity of our approach with several video sequences acquired in a casual environment.

Keywords: Head pose determination, head tracking, model-based tracking, exemplar-based tracking, faces and gestures, facial expression.

1. Introduction

In the last few years, 3D head tracking in a video sequence or relative pose estimation from multiple images has been recognized as an essential prerequisite for robust facial expression/emotion analysis, synthesis and face recognition. 3D head pose information is also very important for user attention detection, vision-based interface and head gesture recognition. In multimedia applications, video coding also requires 2D or 3D motion information to reduce the redundant data [15].

When there is no expression change on the face, relative head pose can be solved as a rigid object tracking problem through traditional 3D vision algorithms for multiple-view analysis [19, 33, 18]. However, in practice, expressional deformation or even occlusion frequently occurs, together with head pose changes. Furthermore, facial expression analysis or face recognition also needs to deal with the alignment problem between different head orientations. Therefore, it is necessary to develop effective techniques for head tracking under the condition of expression changes.

The remaining part of the paper is organized as follows: Section 2 reviews several face tracking works reported in the literature. Section 3 provides an overview of our head pose tracking system, and Section 4 describes the system in details. Section 5 shows several pose tracking experiments with real video images. Section 6 concludes the paper.

2. Previous Work

The first category of research in 3D head tracking involves optical flow computation [5, 6, 12, 32]. This approach works well when the changes of illumination and head motion are small. It is in general difficult to handle facial deformations.

In the second category, the 3D head tracking problem is formulated as the registration of an input video image with the texture mapped 3D head model [21, 24, 13]. La Cascia et al. [21] proposed a simple 3D cylindrical model to warp texture from the input images, whereas Schodl et al. [24] employed a full texture-mapped polygonal head model to compensate for the registration errors. Dellaert et al. [13] followed a Kalman-filter-based approach for tracking of planar patches, which uses texture mapping as the measurement model in their framework. Our method improves on this model-based approach by introducing two novel techniques. First we use exemplar images to sample the facial appearance space thus avoiding matching between two images with very different illuminations which are caused by head pose changes. Second we use an expression mask model to help separate point matches corre-

*Current address: Computer Science Department, The Johns Hopkins University 3400 North Charles Street, Baltimore, MD 21218, USA

sponding to rigid head motion from those corresponding to facial deformations thus resulting in significantly more robust head pose estimation.

The techniques in the third category are based on tracking of some salient facial points, features, or patches in images [3, 20, 25, 27]. In [3], 3D structure, camera focal length and head pose can be obtained in an extended Kalman filter framework, which only utilizes tracked 2D points as input. In [20], the head orientation is estimated from tracking of five salient facial points: four eye corners and nose top. This approach is in general very fast. The drawback is that it is usually not very accurate nor very stable because some of the features may not be completely rigid during facial expressions and the number of these features is small.

Bascle and Blake [4] and Blake et al.[7] described useful techniques to separate head motion from facial expressions for a given set of tracked points and contours such as feature boundaries and expression wrinkles. Because most of these contours are in the expression area, their motions contain both pose changes and facial deformations. They separated the two motions by parameterizing head pose in terms of 2D affine transformations with parallax. The drawback is that they were not able to obtain 3D head rotations.

Attempt was made in [26] to determine the structure and motion simultaneously in the bundle adjustment framework while taking deformation into account. The techniques introduced in our paper including the exemplar images and facial expression masks can be potentially used in their system as well to improve their robustness and convergence rate.

3. Overview of Our Tracking System

In this paper, we propose a new model-based technique for robust head pose tracking under changing expression and/or occlusion. Here, the model information consists of a 3D face model, a set of exemplars sampling the face appearance space, a facial expression mask model and a Bayesian model. Our system starts with a simple initialization phase to acquire the 3D face model and exemplars, followed by a tracking phase.

During the initialization phase, we first capture a video sequence of the user's head under a neutral expression, which is considered to be a rigid object. We build a personalized 3D face model from only one or two pictures of the user at near frontal positions. No user interaction is needed during this stage. Several reference images at distinct poses are then automatically extracted by tracking the built mesh model in the video sequence, and the head poses are estimated in a similar way to that for texture blending of the face model in [22]. Note that the reference images actually sample the complex face appearance space which is difficult, if not impossible, to be represented by a single texture image.

After the warming-up, our system is able to track the head of the same person with occlusion or changing expression. We start tracking when the person is at a close-to-frontal position. We match the current view against the most appropriate reference image according to the predicted pose, based on the motion smoothness assumption. Since the matching is performed between the current view and the predicted reference image, our head tracking is much easier and more robust than if only a single texture image is used. For matching between a rigid reference image and the current image with facial deformation, we use a robust motion segmentation algorithm and the disparity continuity constraint to separate point matches corresponding to rigid head motion from those corresponding to facial deformation. Because of this, each reference image can cover a relatively large range (see experiments). Furthermore, we analyze the facial deformation regions under different expressions, and the rigid and non-rigid features grouping result is evaluated by a MAP-based coarse expression estimation. Since facial deformation tends to exist in some localized areas, the globally consistent rigid matches can be considered to be from the dominant head motion. The head pose is then estimated from the rigid motion points, and can be done reliable thanks to the 3D face mesh model, even when the number of points is small (see below). Since we use reference images during tracking, the accumulative error inherent in frame-by-frame tracking is avoided and more accurate pose estimation is achieved. Finally, a failure alert and tracking recovery mechanism is also implemented, to make the whole tracking system even more robust and flexible.

4. Inside Our Head Pose Tracking System

We now describe in details the major steps of our head pose tracking algorithm.

4.1. Acquiring Personalized Face Information

Our system starts with an initialization phase to acquire personalized face information, which includes a personalized 3D face model and a set of reference images sampling the face appearance space.

The system first captures a video sequence of the person to be tracked by having the person to turn his/her head from one side to the other with a neutral expression in front of the camera. A frontal face detector identifies one near frontal view, and an Active Shape Model (ASM) algorithm [11] is applied to that image to detect face feature points and silhouettes.

4.1.1. Constructing personalized 3D Face Model

The person's 3D face model is constructed from only the frontal view by using a model-based approach. We use the same face mesh model as the one used in the rapid modeling system developed by Liu et al [22]. It contains a generic face mesh and a number of deformation vectors called metrics. Each metric provides a way to deform the mesh such

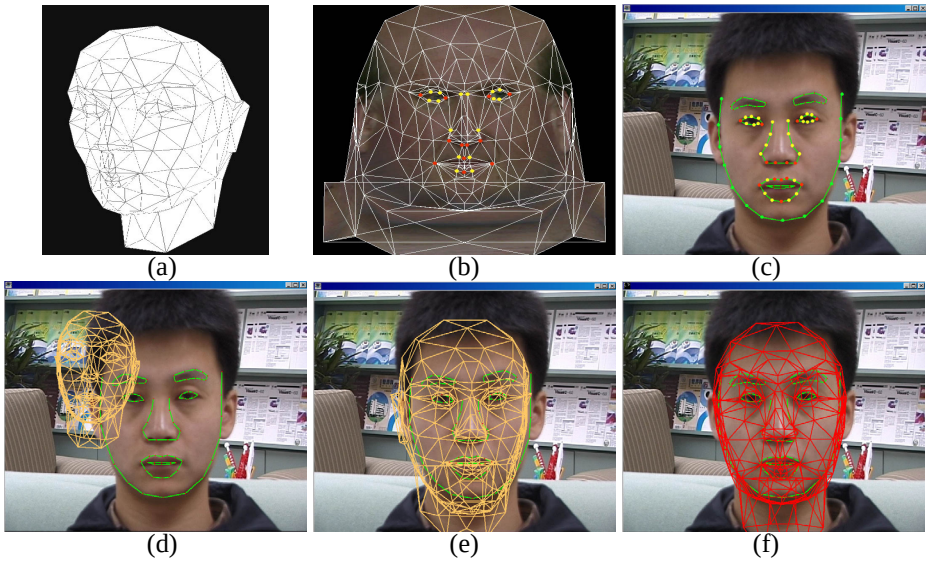


Figure 1: *Automatic Initialization Procedure: a) a general 3D face mesh model; b) mesh model with texture; c) a facial image with ASM converge; d) before pose initialization; e) after pose initialization; f) personalized face model adjustment.*

as to make the nose bigger, head wider, etc.

The problem of constructing the 3D face model is formulated as searching for the head pose and the metric coefficients to best fit the face feature points and silhouettes which are detected by ASM. As shown in Figure 1 (b) and (c),there are three different sets of point correspondences between the 3D face mesh model and the point distribution model of ASM. Each red point in (c) corresponds to a vertex on the face mesh in (b). These are point-to-point correspondences. Each yellow point corresponds to a point on the same facial curve in the face mesh. For example, the yellow points on the lips correspond to points on the lip boundary of the face mesh in (b). These are called point-to-curve correspondences. The green image points are those on the silhouettes which do not have simple association with the face mesh model but are useful in updating the face structure.

To obtain a reasonable initial guess of the head pose, we first treat the generic face mesh as rigid and solve for the pose to satisfy the point-to-point and point-to-curve correspondences. Notice that these correspondences are 3D to 2D correspondences since we treat the face mesh as rigid. Given a set of 3D to 2D correspondence, we can compute the head pose using a technique to be described in Section 4.3. The point-to-curve correspondences are converted to point-to-point correspondences using an iterative closest point approach. Figure 1 (d) and (e) shows the generic mesh before and after pose initialization respectively.

We then fix the head pose, and search for the metric coefficients by using the silhouette information as well as the other face features.

This pose estimation and structure update are alternated until no significant progress is made. We find that it usually terminates in 3 or 4 iterations.

Figure 1 (f) shows the final face mesh. We can see that its silhouettes match the image silhouettes a lot better than (e). The pose is also more accurate.

4.1.2. Reference Images

Once we get the 3D face model, we track the head in the whole image sequence, and determine the head pose recursively for each image as follows. Starting from the frontal views used in pose and model adjustment, successive images (I_i, I_{i+1}) are matched. The head pose for I_i is known, so we know the 3D points on the face corresponding to the matched points in I_i. The head pose for I_{i+1} is then determined based on 3D-2D point correspondences. Experiments have shown that this gives a much more accurate pose estimation than if only 2D points are used.

As our matching technique to be described in the next subsection can deal with relatively large motion, it is not necessary to use all images in this video sequence as reference images. We develop an automatic image selection algorithm during tracking. Let us call the amount of head rotation between two consecutive frames the *rotation speed*. If s is the current rotation speed and α is the desired angle between each pair of selected images (*reference images*), the next image is selected α/s frames away, assuming the head motion is pretty smooth. In our implementation, the initial guess of rotation speed is set to 1 degree/frame and the desired separation angle is equal to 5 degrees. We must point out that each reference image is a sparse sample in both the head pose space and the face appearance space. Figure 4 displays a few reference images selected during one head tracking session.

During tracking, we predict reference images, instead of traditional frame-by-frame prediction. Because each reference image can cover a relative large range thanks to our matching technique, we can search for head pose efficiently within a larger space. As will be evidenced by the experiments, this representation by a set of sparse samples contributes to both the flexibility and robustness for head pose

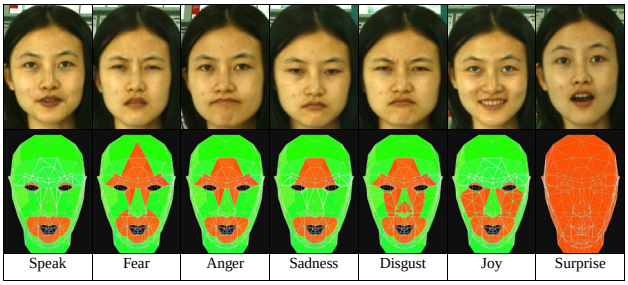


Figure 2: *3D facial mask model for rigid and non-rigid regions under different expressions.*

tracking. Furthermore, because we use reference images, the accumulative error inherent in frame-by-frame tracking is avoided and more accurate pose estimation is achieved.

We do not directly use the global textured image provided by the face modeling system for two reasons. The first is that the face mesh model is very coarse, resulting visible texture distortion. The second is that the global texture image is the blending of images under varying illumination condition (because the head is turning). Matching with such a textured image is hard. We use the original images at several distinct poses as reference images, making it more flexible and robust to handle expressional deformation, occlusion and illumination change.

4.2. Robust Matching and Clustering for Rigid-Motion Points

We use corner points for head tracking and pose estimation. We do not consider region appearance such as the texture map of the head model because appearance change is very difficult to model when expressional deformation and occlusion need to be dealt with.

The technique described in this section tries to find in a robust way point correspondences due to rigid motion between a reference image and the image to track. The reference images we acquired during the initialization phase contain the information of head pose and appearance under neutral expression. The current image contains facial deformation and/or occlusion, but we can usually expect that there exist some corner points encoding the rigid part of the head motion if that image was taken near where the reference image was taken. From the dense facial motion analysis under different expressions [10, 31] shown in Figure 2, we can find an important phenomenon that most facial expressional deformations due to speaking, smiling, frowning, etc., are partial, radial and symmetric, except for surprise. For occlusion, we can reasonably assume that occlusion only occupies a small portion of the face, while the remaining face parts undergo a rigid motion.

Given an image to track (called the target image) and a predicted reference image, we first extract points of interest with Harris corner detector [17], and then match them with a normalized cross-correlation technique [34]. Let us denote the set of obtained point matches by Ω_m. Notice that there are two subsets of point matches: one subset corresponding to rigid motion, denoted by Ω_g, and the other subset corresponding to non-rigid motion, denoted by Ω_n. We are seeking to separate Ω_m into Ω_g and Ω_n. Obviously, these two sets of points satisfy different types of constraints.

4.2.1 Rigid Motion Constraints

It is well known that point matches due to rigid motion satisfy the same epipolar constraint [16, 35], which can be described by the fundamental matrix $\mathbf{F}$. For each pair of points $\mathbf{m}_i^r$ and $\mathbf{m}_i^t$, we define a measure of distance

$$EC(\mathbf{m}_i^r, \mathbf{m}_i^t) = d(\mathbf{m}_i^r, \mathbf{l}_i^r) + d(\mathbf{m}_i^t, \mathbf{l}_i^t) \quad (1)$$

where $\mathbf{l}_i^r = \mathbf{F}\mathbf{m}_i^t$ and $\mathbf{l}_i^t = \mathbf{F}^T\mathbf{m}_i^r$ are epipolar lines, and $d(\mathbf{m}, \mathbf{l})$ is the distance from point $\mathbf{m}$ to line $\mathbf{l}$ on the image. If point match $(\mathbf{m}_i^r, \mathbf{m}_i^t) \in \Omega_g$, then $EC(\mathbf{m}_i^r, \mathbf{m}_i^t) = 0$.

Since the epipolar constraint cannot detect false matches along the epipolar line, we impose another constraint based on the assumption that the disparity of matched points should vary smoothly within a local region. Consider a point match $(\mathbf{m}_i^r, \mathbf{m}_i^t)$. Its disparity is defined as $\mathbf{d}_i = \mathbf{m}_i^r - \mathbf{m}_i^t$. In the neighborhoods around $\mathbf{m}_i^r$ and $\mathbf{m}_i^t$, we can find some other point matches, from which an average disparity, $\bar{\mathbf{d}}_i$, is computed. If $\mathbf{d}_i$ is very different from $\bar{\mathbf{d}}_i$, then this point match is likely to be invalid. So we define the disparity continuity measure as the following

$$DC(\mathbf{m}_i^r, \mathbf{m}_i^t) = \|\mathbf{d}_i - \bar{\mathbf{d}}_i\| \quad (2)$$

4.2.2. Approximate Expression Interpretation of Non-rigid Matches

The intuition behind the non-rigid matches constraint is that they should be interpretable with a meaningful facial expression. At the learning stage, we built several non-rigid motion models on the 3D facial mesh model. The corresponding motion regions (triangles) $\mathbf{U}_i$ are labeled for several expressions, as shown in Figure 2. Red regions are the meshes under expression deformations, and green regions are rigid for an emotion. Given their projection regions, Ψ_i, on the reference view and the set of non-rigid features Ω_n, the most likely interpretation of the facial expression is obtained by

$$P = \max_{\Psi_i} \Pr(\Psi_i|\Omega_n) \quad (3)$$

Using Bayes' Rule, this can be rewritten as

$$P = \max_{\Psi_i} \left(\frac{\Pr(\Psi_i)\Pr(\Omega_n|\Psi_i)}{\Pr(\Omega_n)}\right) \quad (4)$$

where $\Pr(\Omega_n|\Psi_i)$ is the probability of the non-rigid matching set given the expression interpretation, $\Pr(\Psi_i)$ is the prior probability of the expression, and $\Pr(\Omega_n)$ is the probability of the grouping result. $\Pr(\Omega_n)$ does not depend on Ψ_i, and can be considered as a constant.

Because the prior probability of the expression $\Pr(\Psi_i)$ is very difficult to obtain, we assign a uniform probability

distribution[1]. The likelihood $\Pr(\Omega_n|\Psi_i)$ is obtained from a simple observation model: the ratio between the number of non-rigid matches inside Ψ_i, $\mathcal{N}(\Omega_n^i)$, and the number of all non-rigid matches, $\mathcal{N}(\Omega_n)$, i.e.,

$$\Pr(\Omega_n|\Psi_i) = \frac{\mathcal{N}(\Omega_n^i)}{\mathcal{N}(\Omega_n)} \tag{5}$$

4.2.3. Clustering Algorithm

Now, we describe our algorithm to cluster the matches into rigid and non-rigid point matches.

Let us use $\mathbf{w} = \{w_i\}$ to record the correspondence status for all matches in Ω_m. If the ith point match is a rigid one, then $w_i = 1$; otherwise, $w_i = 0$. Basically $\mathbf{w}$ is what we are trying to solve for.

For a set of rigid point matches, we define three measures. The first is the ratio of the number of rigid point matches to the total number of matches, i.e.,

$$\phi = \frac{\mathcal{N}(\Omega_g)}{\mathcal{N}(\Omega_m)} \tag{6}$$

The other two measure the rigidity quality:

$$EC(\mathbf{w}) = \sum_{1 \le i \le \mathcal{N}(\Omega_m)} w_i EC(\mathbf{m}_i^r, \mathbf{m}_i^t), \tag{7}$$

and

$$DC(\mathbf{w}) = \sum_{1 \le i \le \mathcal{N}(\Omega_m)} w_i DC(\mathbf{m}_i^r, \mathbf{m}_i^t). \tag{8}$$

The objective of the clustering is to find the maximum number of rigid matches with desired quality and in the same time to be able to interpret the non-rigid matches with certain facial expression. Therefore, we formulate clustering as the following optimization problem:

$$\max_{\mathbf{w}} \quad \phi \max_{\Psi_i} \Pr(\Psi_i|\Omega_n) \tag{9}$$

$$\text{subject to} \quad EC(\mathbf{w}) \le \epsilon_e \tag{10}$$

$$DC(\mathbf{w}) \le \epsilon_d \tag{11}$$

where ϵ_e and ϵ_d are some pre-specified small constants.

This optimization problem is solved using a genetic algorithm as outlined below.

Initialization: Randomly generate a population.

Step 1: For each candidate in the population, check whether they satisfy (10) and (11). The ones that violate either of the two constraints are discarded.

Step 2: For the remaining candidates, evaluate (9) as their fitness levels. Perform evolutionary computation to generate next generation population based on their fitness levels. Then go to Step 1 and repeat until the maximum number of generations is reached.

In our implementation, we run through 300 generations. We choose the one with the maximum objective value in the last generation as the final result.

[1]We are planning to learn the expressional prior information from audio and video, and combine it with our tracking system in the future work.

4.3. Head Pose Estimation, Failure Alert and Auto-recovery

We now have a set of rigid matches $\{(\mathbf{m}_i^r, \mathbf{m}_i^t)\}$, where $\mathbf{m}_i^r$ and $\mathbf{m}_i^t$ are points in the reference and target image, respectively. For each point $\mathbf{m}_i^r$ in the reference image, we cast a 3D ray from the camera center through that point, and compute the intersection $\mathbf{x}_i$ of that ray with the face mesh model corresponding to the reference pose. Then the relative pose $\hat{\mathbf{T}} = \begin{pmatrix} \hat{\mathbf{R}} & \hat{\mathbf{t}} \\ \mathbf{0}^T & 1 \end{pmatrix}$ can be computed according to the following equation

$$\mathcal{A}\mathcal{P}\hat{\mathbf{T}}\tilde{\mathbf{x}}_i = \lambda \tilde{\mathbf{m}}_i^t \tag{12}$$

where $\tilde{\mathbf{x}}_i = (\mathbf{x}_i^T, 1)^T$, and $\tilde{\mathbf{m}}_i = (\mathbf{m}_i^T, 1)^T$. The intrinsic matrix $\mathcal{A}$, the standard projection matrix $\mathcal{P}$, $\mathbf{x}_i$ and $\mathbf{m}_i^t$ are known. Each of the above equation gives two constraints on $\hat{\mathbf{T}}$. We compute $\hat{\mathbf{T}}$ with a linear least-squares technique described in [16]. In order to get a higher accuracy, we refine the estimation by minimizing the sum of squared distances between the observed image coordinates $\mathbf{m}_i^t$ and the reprojected values on the target image, i.e.,

$$\min_{\hat{\mathbf{T}}} \sum_i \|\hat{\mathbf{m}}_i^t - \mathbf{m}_i^t\|^2 \tag{13}$$

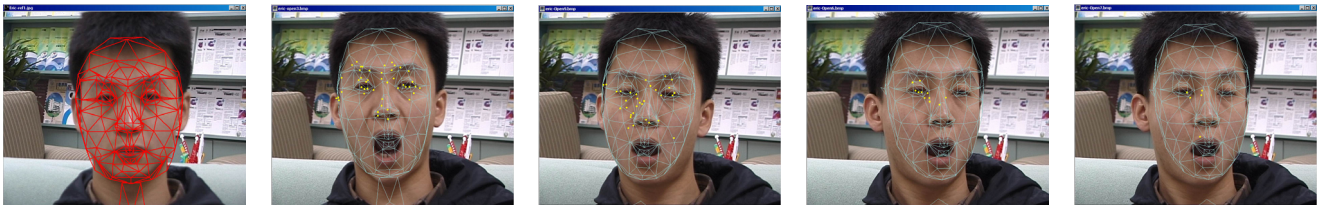
where $\hat{\mathbf{m}}$ is given by $\lambda \widetilde{\hat{\mathbf{m}}}_i^t = \mathcal{A}\mathcal{P}\hat{\mathbf{T}}\tilde{\mathbf{x}}_i$. After $\hat{\mathbf{T}}$ is determined, the head pose for the target image in the camera frame is given by

$$\mathbf{T}^t = \hat{\mathbf{T}}\mathbf{T}^r \tag{14}$$

where $\mathbf{T}^r$ is the head pose for the reference image. Note that since we use reference images during tracking, the accumulative error inherent in frame-by-frame tracking is avoided and more accurate pose estimation is achieved.

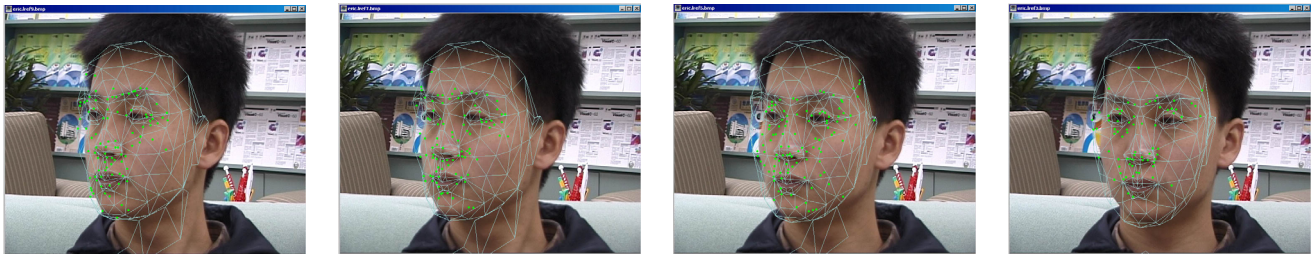
Failure alert and auto-recovery is critical for a robust tracking system. Tracking failure mainly occurs in two situations: the expression deformation is too violent (such as surprise), or the motion between facial views is too large. The first problem is very difficult to deal with. This is also true for other approaches such as the explicit 3D reconstruction technique [26]. Fortunately, this kind of special expression is infrequent and can be detected with our approximate expression interpretation. We can recover the head pose as soon as that expression finishes. The second problem is usually due to some agile head motion, leading to a wrong prediction of the reference image. In this case, we can easily solve it by relaxing the prediction to consider multiple neighboring reference images. The best reference image and the best pose estimate are selected according to the following criteria:

1. The number of resulting rigid matching pairs is above a certain value.
2. Their spatial distribution is not concentrated in a small region.
3. The average distance between a rigidly matched point

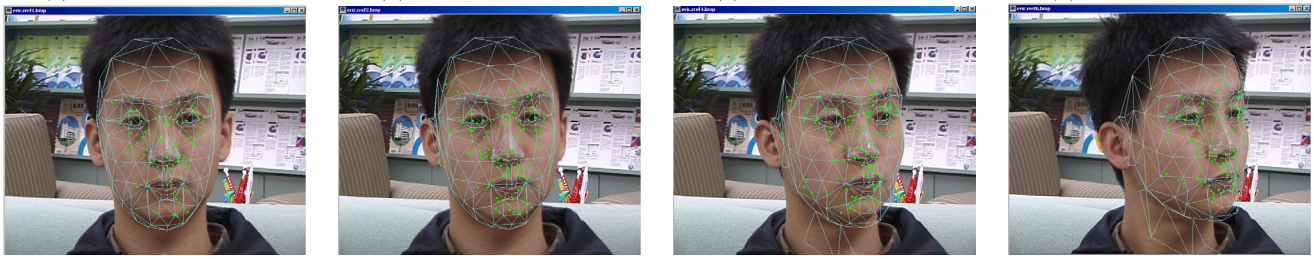


(a) reference view (b) a tracked image (c) a tracked image (d) a tracked image (e) a tracked image

Figure 3: *The range of head orientation that can be tracked with a near frontal reference image.*

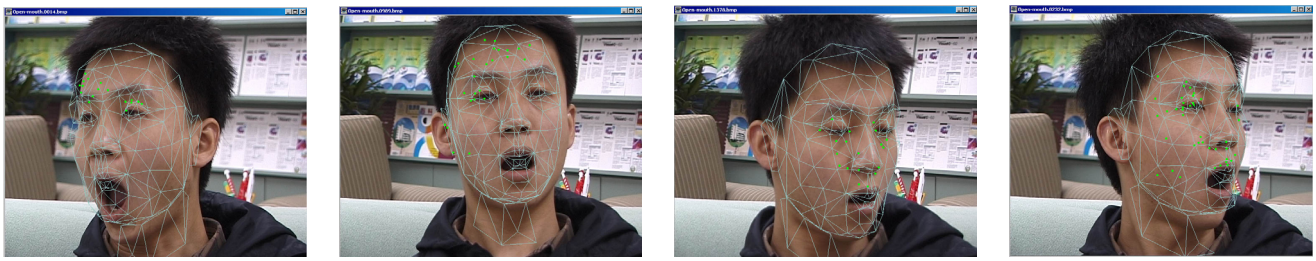


(a) reference view 1 (b) reference view 3 (c) reference view 5 (d) reference view 7

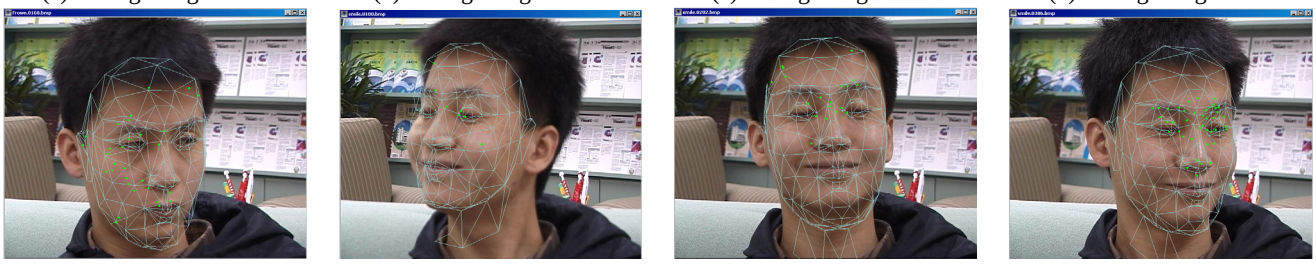


(e) reference view 9 (f) reference view 11 (g) reference view 13 (h) reference view 15

Figure 4: *A sample set of the reference images picked by computer for human head rolling.*



(a) talking using ref. 5 (b) talking using ref. 11 (c) talking using ref. 14 (d) talking using ref. 15



(e) frowning using ref. 12 (f) smiling using ref. 4 (g) smiling using ref. 10 (h) smiling using ref. 13

Figure 5: *Head pose tracking with changing facial expression.*

and the projection of its corresponding 3D point on the image is below a certain value.

4. The expression interpretation of non-rigid matching points is reasonable.

If one of the following four criteria is not satisfied, the system will issue a tracking failure alert.

Compared with multi-hypothesis tracking approaches [7], the robustness of our method is ensured by the alert-

Figure 6: *Sample results on head pose tracking of two other people*

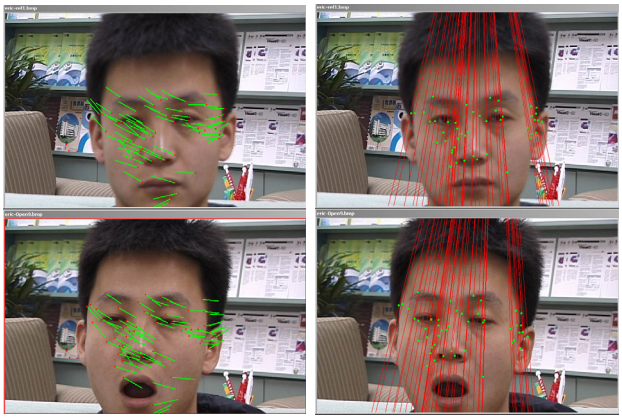


Figure 7: *The feature matches before and after motion segmentation.*

and-recovery scheme. Our method usually runs in a single pose hypothesis mode. It becomes a multiple hypothesis tracking only when the system issues a tracking failure alert and runs in a recovery mode. Our method can be easily extended to consider multiple pose hypotheses, but the computation cost will be increased by as many times.

5. Experiments

In this section, we first show the robustness of our pose estimation algorithm for two views. For example, the reference image in Figure 3(a) can be used for pose tracking for other images if the 3D orientation is within a range of about $(-10°, 10°)$ in all directions. In Figure 3, we show the tracking results for four images. Although it may be hard to see, the matched rigid points are indicated with yellow dots. The result is quite accurate for images shown in Figure 3 (b) and (c). For images shown in Figure 3 (d) and (e), since the orientations are quite different from the reference view, it is not surprising that only a few feature points are retained as belonging to the rigid head motion. The results are still reasonable.

Figure 8: *Tracking results under occlusion with certain illumination variation.*

An example of feature matches before and after GA segmentation are shown in Figure 7. We can see that some matches around the mouth are grouped as non-rigid features. Experiments show that our approach works well with disparity range up to 60 pixels.

Figure 4 shows the set of reference images selected during a demonstration when a person was turning his head. This type of out-of-plane rotation is usually difficult for head tracking, but we can see that our algorithm can determine accurately the head pose, thanks to the 3D mesh model. These reference images are then used to track a head with changing expression, and a result is shown in Figure 5, where the corresponding reference image number used for matching is indicated.

We have conducted many more experiments. Figure 6 shows some tracking results with two other people, while

Figure 8 shows that our tracking system works successfully even when occlusion and changing illumination occur.

6. Conclusions and Discussions

In this paper, we have proposed a new model- and exemplar-based technique for robust head pose tracking under changing expression and/or occlusion. Our system starts with a simple learning stage. It builds a personalized 3D face mesh model from two near frontal views and automatically acquires a few reference images at distinct poses which actually sample the complex face appearance space. During head tracking with deformation, we match the current view against the most appropriate reference image according to the predicted current pose. We also use a robust motion segmentation algorithm which separates point matches corresponding to rigid head motion from those corresponding to facial deformation. The head pose can then be reliably estimated from the rigid motion points thanks to the 3D face mesh model, even when the number of points is small. Since we use reference images during tracking, the accumulative error inherent in frame-by-frame tracking is avoided and more accurate pose estimation is achieved. We have demonstrated the validity of our approach with several video sequences acquired in a casual environment. We can indeed deal with large head motion, changing facial expression, talking, and limited occlusion and illumination change.

In our current work, the reference images only sample the appearance space of a face under a single expression. We are planning to extend our work to include exemplars with various expression, and formulate the head tracking problem in a probabilistic paradigm similar to that described in [29].

References

[1] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *IJCV*, 2:283–310, 1989.

[2] N. Ayache, *Vision Stereoscopique et Perception Multisensorielle*, InterEditions, Paris, 1989.

[3] A. Azarbayejani, et al., " Visually controlled graphics," *PAMI*, 15(6): 602–605, 1993.

[4] B. Bascle and A. Blake, "Separability of Pose and Expression in Facial Tracing and Animation," *6th ICCV*, pp. 323-328, 1998.

[5] S. Basu, et al., "Motion Regularization for Model-based Head Tracking," *13th ICPR*, 1996.

[6] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," *5th ICCV*, pp. 374-381, 1995.

[7] A. Blake, et al., "Statistical models of visual shape and motion," *Proc. Roy. Soc. Lond. A*, 356, pp. 1283–1302, 1998.

[8] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," *SIGGRAPH*, pp.187-194, 1999.

[9] J. Chai and S. Ma, "An Evolutionary Framework For Stereo Correspondence," *14th ICPR*, 1998.

[10] J.F. Cohn, A. Zlochower, J.J. Lien, W. Hua, T. Kanade, "Automated face analysis", C. Rovee-Collier and L. Lipsitt (Eds.),*Progress in infancy research*, (In press) 1. Hillsdale, NJ: Erlbaum.

[11] T.F. Cootes, D. Cooper, C.J. Taylor and J. Graham, Active Shape Models - Their Training and Application. Computer Vision and Image Understanding. Vol. 61, No. 1, Jan. 1995, pp. 38-59.

[12] D. DeCarlo and D. Metaxas, "The Integration of Optical Flow and Deformable Models with Applications to Human Face Shape and Motion Estimation," *CVPR*, 1996.

[13] F. Dellaert, et al., "Jacobian Images of Super-resolved Texture Maps for Model-based Motion Estimation and Tracking," *Workshop Applications of Computer Vision*, 1998.

[14] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*, Consulting Psychologists Press Inc., San Francisco, CA, 1978.

[15] Essa. I., and A. Pentland. "Coding, Analysis, Interpretation and Recognition of Facial Expressions", Vol. 19(7), *IEEE Computer Society Press*, July, 1997.

[16] O. Faugeras, *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, 1993.

[17] C. Harris and M. Stephens, "A combined corner and edge detector," *4th Alvey Vision Conf.*, pp. 189-192, 1998.

[18] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[19] H. C. Longuet-Higgins, "A Computer Algorithm for Reconstructing A Scene from Two Projections," *Nature*, Vol. 293, pp. 133-135, 1981.

[20] T. Horprasert, et al., "Computing 3-D Head Orientation from a Monocular Image," *Int'l Conf. Automatic Face and Gesture Recognition*, pp.242-247, 1996.

[21] M. La Cascia, et al., "Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Robust Registration of Texture-Mapped 3D Models," *PAMI*, 22(4):322-336, 2000.

[22] Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen, "Rapid Modeling of Animated Faces From Video," in *Proceedings of The Third International Conference on Visual Computing* (Visual 2000), pages 58-67, September 2000, Mexico City. Also *Technical Report MSR-TR-99-21*, Microsoft Research, Feb. 1999.

[23] H.S. Sawhney and S. Ayer, "Compact Representation of Videos through Dominant Multiple Motion Estimation", *PAMI*, 18(8): 814-830.

[24] A. Schodl, et al., "Head Tracking using a Textured Polygonal Model," *Workshop Perceptual User Interfaces*, 1998.

[25] H. Tao, et al., "Tracking of Face Features using Probabilistic Network," *Workshop Face and Gesture Recognition*, 1998.

[26] H. Tao and T. S. Huang, "Explanation-based Facial Motion Tracking using A Piecewise Bezier Volume Deformation Model," *CVPR*, 1999.

[27] Y.-L. Tian, T. Kanade, and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis", *PAMI*, 23(2):97–115, 2001.

[28] P. Torr, *Motion Segmentation and Outlier Detection*, Ph.D Thesis, Dept. of Engineering Science, University of Oxford, 1995.

[29] K. Toyama and A. Blake, "Probabilistic tracking in a metric space", in *Proc. International Conference on Computer Vision*, Vancouver, Canada, July, 2001.

[30] Y. Wu, K. Toyama, T.S. Huang, "Wide-Range, Person- and Illumination-Insensitive Head Orientation Estimation," *Proc. of Int'l Conf. on Face and Gesture Recognition* , pp. 183-188, Grenoble, France, March, 2000.

[31] Y.T. Wu, T. Kanade, C.C. Li, J.F. Cohn, "Image registration using wavelet-based motion model", *IJCV* (In press).

[32] Y. Zhang and C. Kambhamettu, "Robust 3D Head Tracking Under Partial Occlusion," *Int'l Conf. Automatic Face and Gesture Recognition*, 2000.

[33] Z. Zhang, "Motion and Structure From Two Perspective Views: From Essential Parameters to Euclidean Motion Via Fundamental Matrix," *JOSA-A*, 14(11):2938-2950, 1997.

[34] Z. Zhang, et al., "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry," *Artificial Intelligence J.*, Vol. 78, pp. 87-119, October 1995.

[35] Z. Zhang, "Determining the Epipolar Geometry and its Uncertainty: A Review," *IJCV*, 27(2):161-195, 1998.