FACIAL FEATURE TRACKING COMBINING MODEL-BASED AND MODEL-FREE METHOD

Jianyu Wang[1], Wen Gao[1,2], Shiguang Shan[2], XiaoPeng Hu[3]

[1]Department of Computer Science, Harbin Institute of Technology, Harbin, China, 150001
[2]FRJDL, Institute of Computing Technology, CAS, P.O.Box 2704, Beijing, China, 100080
[3]Department of Computer Science, Wuhan University, Wuhan, China, 430000
{jywang, wgao, sgshan, xphu}@jdl.ac.cn

ABSTRACT

In this paper we propose a novel facial feature tracker, which integrates model-free and model-based method to reliably track dense facial features with complex non-rigid motions in low quality video sequence. The tracker consists the mouth motion model part and the enhanced KLT tracker part. There are three key elements in our algorithms: dense facial feature tracking, noise removal with global rank constraints and characteristic non-rigid motion description. Experiments show good results on tracking dense facial features under various expressions, even some facial features have degenerate features.

1. INTRODUCTION

Feature tracking is one of the most fundamental tasks in computer vision, as it is probably the most popular way of extracting motion information. Many vision applications, such as motion interpretation, object tracking, object recognition, etc. rely heavily on accurate correspondence of the feature points.

Existing feature tracking algorithms can be divided into two categories: model-based tracking algorithms and model-free tracking algorithms. Model-based tracking algorithms generally try to select the same set of fiducial features from each frame, then attempt to establish correspondences between both sets of features using a prior established model. Model-free Tracking algorithms generally select a set of features from the reference frame only. The positions of these features in subsequent frames is found by doing a local search inside a suitable sized window for the position which correlates best with the texture around the feature in the reference frame, no prior knowledge is required.

Visual feature tracking has been extensively studied with the development of computer vision. McKenna et al. [7] proposed an approach to track rigid and non-rigid face motion based on a point distribution model (PDM) and Gabor wavelets. Many lip trackers were variants of the snake method of Kass and Terzopoulos [8] or of the deformable template technique of Yuille [9]. Luettin [10] used an Active Shape Model (ASM). Lucas and Kanade [11] have proposed a method for registering two images for stereo matching based on a translation model between images. From the initial work of Lucas and Kanade, Tomasi and Kanade [4] developed a feature tracker based on the 'sum of squared intensity differences (SSD)' matching measure, using a translation model. In [3], L. Torresani proposed a model-free tracking method that makes use of global rank constraints. In [6], Brand combined probabilistic error estimation and global rank constraints to track facial features.

In this paper we describe a novel facial feature tracker, which integrates model-free and model-based method to reliably tracking dense facial features.

The rest of the paper is organized as follows: We describe the tracker briefly in section 2. The mouth motion model is described in Section 3. In section 4, we extend the Kanade-Lucas-Tomasi (KLT) tracker to enable it to track feature points with degenerate textures. And we report our experiment results in section 5. In section 6, we conclude our work and give some future research direction.

2. DESCRIPTION OF THE TRACKER

The proposed facial feature tracker integrates model-based and model-free tracking method for dense facial feature tracking. The model-free tracking method is general in use and often more computational efficient. But it cannot predict and track the object with complex characteristic motions, such as face, in some situations. The model-based tracking method can compensate this shortcoming and provides characteristic motion tracking when model-free method fails.

Our tracker combines these two reciprocal methods to tracking complex non-rigid motions. The tracker consists a mouth model part and enhanced KLT (EKLT) tracker part. The mouth model is trained from example

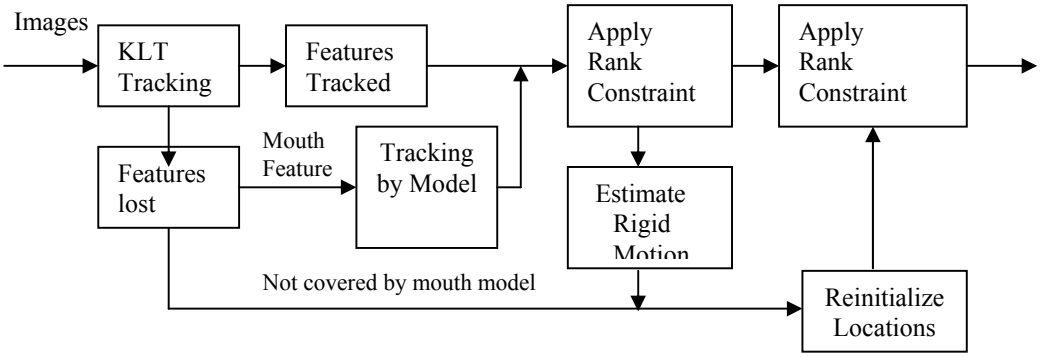

Figure. 1 Facial feature tracker diagram

mouth images. It can track characteristic mouth motion and ensure that the subspace of the non-rigid face motion does not fall into rank degeneration. The EKLT tracker is a basic KLT tracker with global rank constraints. Other prior knowledge such as topography consistency is incorporated into the tracker for robust facial feature tracking. The general tracking process is show in Figure 1.

3. ACTIVE MOUTH MOTION MODEL

Human face has delicate motion. We believe that facial motion has its own characteristic and should be learned from observations for accurate and robust tracking. Though most facial feature points' motion can be assumed as near-rigid, for face parts such as mouth, it is not practical to track its intrinsically non-linear motion due to large variance using model-free method such as optical flow. Here we employ example-based learning method to establish a mouth motion model for mouth tracking.

3.1. Database of the mouth images

The database used in this work includes 2D mouth images from 15 subjects, 200 images for each subject. 16 landmarks on the mouth are labeled on each mouth image manually.

3.2 Mouth Motion Model Training

In this subsection, a morphable mouth model is trained from example images.

First, we concatenate landmarks in each example image in dataset $X=\{x_i \mid i=1{:}n\}$ into a vector.

$$x_i = \{u_1, v_1 \ldots u_j, v_j \ldots u_m, v_m\},$$

where (u_i, v_i) is the location of landmark i. Normalize x_i to unit and align all mouth shapes into a common co-ordinate frame.

Secondly, for each landmark on one mouth image, we define a rectangle window centered at the location of that landmark and sample pixel intensities and do normalization to make the sum of all pixel value are unit.

Thirdly, we perform Principle Component Analysis (PCA) on the normalized shape examples and normalized texture examples, we get vector $\bar{x}$, the mean shape and vector $\bar{g}$, the mean texture, E_s *and* E_g are corresponding shape eigen matrix and texture eigen matrix of the first several significant eigen vectors

$$E_s = \{e_{s1}, e_{s2} \ldots e_{sp}\}$$

$$E_g = \{e_{g1,} e_{g2} \ldots e_{gq}\}$$

Then one mouth pattern can be represented by a vector b_s and a vector b_g in the learned subspace

$$b_s = E^{\mathrm{T}}(x_i - \bar{x})$$

$$b_g = E_g{}^{\mathrm{T}}(g_i - \bar{g})$$

and during tracking, one observed pattern x_0 can be matched by tuning parameter vector b in a reasonable range

$$x_o = Eb + \bar{x}$$

Note that considering the representation method, our method is different from Active Appearance Model (AAM) in that our method need not to warp textures from a specific shape to the mean shape, which can reduce additional noise. Simple linear subspace based methods, such as PCA, are more suitable to model the simple motion like mouth rather than the whole face motion.

Furthermore, our method is different from AAM and ASM due to following observation:

1) AAM method makes use of the global textures, that means it treat every pixel equally important. We argue that textures near fiducial feature points are more important than those far from them.
2) ASM has too simple local model, it only samples texture on the profiles.

3.3. Model-Based Mouth Motion Tracking

The mouth motion tracking procedure is treated as an optimization problem. The difference vector δI between the real mouth image I_r and the synthesized mouth image I_s is evaluated to tune the parameter vector b. To best match the real mouth image and the synthesized mouth image, we want to minimize the magnitude of δI . We adopt Coots's[1] stochastic searching algorithm to find parameters to minimize δI .

4. ENHANCED KLT FEATURE TRACKER

In this section we extend KLT Feature Tracker to make it fit for dense facial feature tracking due to the following three reasons:

1) The model-free tracking method is generally more computationally efficient and accurate than model-based tracking method for those features that do not change their appearance violently.
2) The purpose of global constraints we employ here is to remove tracking noise from the initial result. It can enforce more effective constraints as the number of feature points increase.
3) Face feature topography consistency is adopted to aid face feature tracking.

4.1 Global Rank Constraint

Bregler[5] has shown that a non-rigid shape state can be appropriate as a combination of several basic rigid shapes:

$$S = \sum_{i=1}^{K} l_i \cdot S_i \qquad S, S_i \in IR^{3xP}, l_i \in IR$$

Where S_i is the i-th basic shape that can represent some characteristic motion of the non-rigid target.

Let $\{(u_{fp}, v_{fp}) \mid f = 1...F, p = 1...P\}$ denote the pixel location of the feature p in frame f. We stack all feature points' trajectory into one observation matrix W as in [5]. Under a weak perspective projection, W can be rewritten as:

$$W = \begin{bmatrix} l_1^{\,1}R^1 & \dots & l_K^{\,1}R^1 \\ l_1^{\,2}R^2 & \dots & l_K^{\,2}R^2 \\ & \dots & \\ l_1^{\,F}R^F & \dots & l_K^{\,F}R^F \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_K \end{bmatrix}$$

The observation matrix W lies in a subspace whose dimension is no larger than $3K$[5]. This strong constraint is different from the heuristic constraints such as appearance constancy.

4.2 Noise Removal by Strong Constraints

We divide observation matrix W into two sub-matrices W_{rel} and W_{unrel}, where W_{rel} contains the feature points' trajectory that can be reliable tracked by the basic KLT Tracker and mouth model matching. We notice that though some feature points are declared reliably tracked, they are often contaminated by some kind of noise. Global rank constraints are first applied to W_{re} and force these features within a reasonable subspace.

$$W_{rel} \xrightarrow{SVD} L_{rel} M_{rel} R_{rel} .$$

We project the W_{rel} into a $3K$ lower subspace by retaining the largest $3K$ singular values and reconstruct a more accurate W_{rel}. Then, we estimate the translation T for every frame, i.e., for frame f, using all reliably tracked features.

For a human face, though its movement is non-rigid, it seldom changes its topography during motion. By modeling the feature points and its neighbor features points as a graph $G = (V, E)$, where V is a set of feature points, and E is a edge set of adjacent feature points, we can take advantage of the topography consistency to get a good estimation for those lost features from their neighbor feature points and treat the residual error as noise. Global rank constraints are applied to W through a thin SVD. $W \xrightarrow{SVD} LMR$. A new observation matrix W^n is obtained by retaining the largest $3K$ singular values

5. EXPERIMENTAL RESULTS

To test our tracker, we randomly select a piece of 139x177 10 Hz video of a subject who had been captured using home digital camera two years ago. The face region covers approximate 90x129 pixels in the image. There are some exaggerate expressions in the sequence. Note that this test sequence is of low frequency, which means the inter-frame displacement of the feature points maybe large. The person in the sequence has fairy smooth skins and there are little texture around some feature points.

In [2], Shi gives a criterion for feature selection that is optimal for tracking algorithms. Here we use Shi's criterion to select 63 features from the reference face image. Then, 16 mouth features are manually selected corresponding to our training model (see fig.2 (a)).

Basic KLT feature tracker is employed to track all features through 100 frames, which performs very well for rigid object tracking for those well-selected features based on Shi's criteriion. Only 17 features can be reliably tracked through sequence (see fig.2 (b)) due to the characteristic of the face and its motion. The feature points around mouth are all lost due to large appearance variance.

We use our algorithms to track all 79 features and only 7 features are lost during tracking (see fig.2(c)). The Enhanced KLT feature tracker with global rank constraint

employed demonstrates good performance to track face features even with little texture. The EKLT module provides good guidance for model-based tracking and reduces the searching space to a more reasonable degree. The residual error is approximate 0.0361-pixel error in horizontal direction and 0.0294-pixel error in vertical direction. The algorithm can run smoothly on a PIII 766 machine and process each frame in 3-4 seconds. Figure 3 shows some of our tracking results.

We also test our algorithms on news broadcaster sequence and it can track the dense facial features well for later processing such as motion analysis and speech recognition.

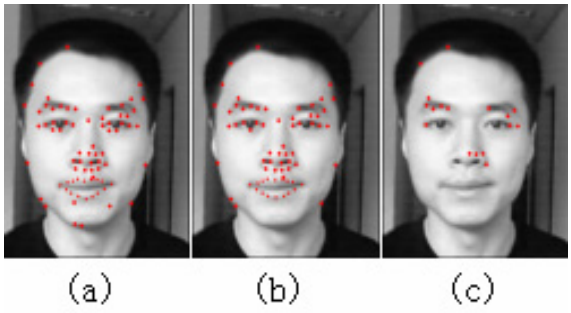

Fig.2 (a) Facial features selected for tracking; (b) Tracked facial features by our proposed facial feature tracker; (c) Tracked facial feature by general KLT feature tracker.

6. CONCLUSIONS AND FUTURE WORK

We propose a novel facial feature tracker that can track dense facial features through video sequence. Our tracker can deal with exaggerate face expressions and track those features with degenerate textures. Experimental results show that our facial tracker is robust and accurate enough. This framework can also be used to track other types of non-rigid motions. Future works include dealing with feature occultation and reappearance, long video sequence feature tracking with online feature window updating according to variance and wide view facial feature tracking.

ACKNOWLEDGEMENTS

This research is sponsored partly by National Hi-Tech Program of China (No.2001AA114160), SiChuan Chengdu YinChen Net. Co. (YCNC) and 100 Talents Foundation of Chinese Academy of Sciences.

REFERENCES

[1] T. Cootes, G. Edwards, and C.J.Taylor. Active appearance models. In *Proc. ECCV*, volume 2, pages 484–498, 1998.

[2] J. Shi and C. Tomasi. Good Features to track. IEEE Conference on Computer Vision and Pattern Recognition, pp 593-600, June 1994.

[3] L.Torresani and C. Bregler. Space-Time Tracking. Europe Conference on computer Vision(ECCV), 2002

[4] C. Tomasi and T. Kanade. Detection and tracking of feature points. Carnegie Mellon University Technical Report CMU-CS-91-132, Pittsburgh, PA, 1991.

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering nonrigid 3D shape from image streams. IEEE Proc. Computer Vision and Pattern Recognition, 2000.

[6] Brand, M.E., "Morphable 3D Models from Video", IEEE Proc. Computer Vision and Pattern Recognition, December 2001.

[7] S. McKenna, S. Gong, R. P. Wurtz, J. Tanner, and D. Banin. Tracking facial feature points with Gabor wavelets and shape models. Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication, Crans-Montana, Switzerland, pages 35–42, 1997.

[8] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. International Journal of Computer Vision, 1(4):321–331, 1987.

[9] A. Yuille. Feature extraction from faces using deformable templates. International Journal of Computer Vision, 8(2):99–111, 1992.

[10] J. Luettin, A. N. Thacker, and S. W. Beet. Locating and tracking facial speech features. Proc. ICPR'96, Vienna, Austria, 1:652–656, 1996.

[11] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. Proc. International Joint Conference on Artificial Intelligence, pages 674–679, 1981.

Fig. 3 Some of our tracking result.