

On Edge Structure Based Adaptive Observation Model for Facial Feature Tracking

Xiaoyan Wang, Yangsheng Wang, Xuetao Feng and Mingcai Zhou
Institute of Automation, Chinese Academy of Sciences
{xiaoyan.wang, yangsheng.wang}@ia.ac.cn

Abstract

Facial feature tracking is a crucial and challenging task in computer vision. Recently online-learning methods have become increasingly popular on account of their strong ability to adapt to variations and have achieved good results in tracking. However, all previous work used only raw intensity to build the model, which is very sensitive to condition changes. In this work, we present a real time, fully automatic facial feature detection and tracking approach using adaptive observation models based on edge structure, which is more reliable especially when the lighting state alters during tracking. Experimental results demonstrate that using edge map measures in observation modeling can improve the accuracy and robustness of tracking.

1. Introduction

Facial feature tracking is one of the significant problems in fields of computer vision and graphics. It plays an important role in applications such as Human-Computer Interaction, surveillance, entertainment, and is highly relevant to the techniques of facial expression analysis, face recognition, realistic 3D face model generation, etc. The aims of facial feature tracking include tracking the rigid movement of head and the transformation caused by expression and actions, which make it quite challenging.

In literatures, many researches have been conducted to deal with the object tracking problem. The two main portions of tracking problems are target modeling and model fitting. Target model should have the ability to handle object changes caused by various reasons, and provide a proper reference for the fitting. Meanwhile fitting algorithms need to be promoted according to the four aspects presented in [10]: efficiency, robustness, accuracy and automation. Various algo-

rithms have been proposed and improved to meet these requirements[2][3].

When it comes to target modeling, fixed template models cannot adapt to appearance changes, while statistical models like ASM and AAM[4] are restricted by training sets and will fail if the imaging conditions are significantly changed. Recently online-learning methods become very popular. They combine the training stage with the searching part and have achieved very good results. Adaptive Gaussian mixture methods have been chosen for real-time video background subtraction and object tracking[7][12], because of its good features in both theory and implementation. Jepson et al.[9] introduced one kind of adaptive Gaussian mixture model named online appearance model for visual tracking. It is a generative model which combines both stable and motion constraints. Dornaika et al[6] use the 3D deformable wireframe model named Candide[1] to describe the pose variation and face deformation simultaneously, and apply a simplified online appearance model to track facial actions. However, only raw intensity is used in the model, and the intensity tends to be very sensitive to changes in conditions such as imaging conditions. Meanwhile non-linear representations of local image structure, especially edge strength, have been successfully used to improve the performance of model matching algorithms and object verification tasks[5][11].

In this paper, a real time, fully automatic facial feature detection and tracking approach using adaptive observation models based on edge structure is presented. The system runs automatically without any user interaction and any precalculation. From experiments, we demonstrate that using edge strength measures rather than intensity in observation modeling can improve accuracy and robustness of tracking.

The remaining part of the paper is organized as follows: Section 2 introduces edge map computing and normalizing. Section 3 presents the details of modeling facial features. The online observation model and

model fitting process are described in section 4 and section 5 respectively. The experimental results are given in Section 6. Finally a conclusion is addressed in Section 6.

2. Edge Strength Computation

2.1 Edge Map Structure

The corner and edge map of an image is introduced by Harris and Stephens[8], and then enhanced by Scott and Cootes[11]. A local descriptor is constructed by calculating the Euclidian distance between the image and the one shifted from itself in a small windows. The distance could be recognized as an energy E , which is descended with respect to the displacement.

$$E(x, y) = \sum_{u,v} [I(u+x, v+y) - I(u, v)]^2 \quad (1)$$

Considering the displacement (x, y) is very small, we could make first order approximation:

$$E(x, y) = \sum_{u,v} [x \frac{\partial I}{\partial u}(x, y) + y \frac{\partial I}{\partial v}(x, y) + o(x^2 + y^2)]^2 \quad (2)$$

Expanding the square-term gives

$$E(x, y) = Ax^2 + 2Bxy + Cy^2 = (x, y)M(x, y)^T \quad (3)$$

where $A = \sum_{u,v} [\frac{\partial I}{\partial u}]^2$, $B = \sum_{u,v} [\frac{\partial I}{\partial u}][\frac{\partial I}{\partial v}]$, $C = \sum_{u,v} [\frac{\partial I}{\partial v}]^2$, $M = \begin{pmatrix} A & C \\ C & B \end{pmatrix}$. At this point Harris and Stephens pointed out that for a point, if the value $\det(M) - k[\text{tr}M]^2$ is negative, it's identified as corner, if the value is positive, the point is edge, if the value is near zero, the point is in a flat area. Scott and Cootes deducted formulations to measure the cornerness and edgeness.

$$r = 2AB - 2C^2 \quad (4)$$

$$e = (A + B)\sqrt{(A + B)^2 - 2r} \quad (5)$$

where r , e indicate the cornerness and edgeness respectively. Only edge strength is used in this paper.

2.2 Non-linear Normalization

The edge strength detected needs to be normalized to obtain a representation. Sigmoid function is chosen for this transformation:

$$T(x) = \frac{|x|}{|x| + x_0} \quad (6)$$

where x_0 is the mean of the expected values of x . Figure 1 Shows a picture and its edge map smoothed twice with a sigmoid function.



Figure 1. An example of edge strength

3. Modeling Facial Features

A 3D wire-frame face model Candide built by Ahlberg [1] is used to depict the diversity between different human beings as well as different face expressions. The model is given as:

$$g = \bar{g} + \sum_{i=1}^{n_s} s_i \cdot S_i + \sum_{i=1}^{n_a} a_i \cdot A_i \quad (7)$$

where \bar{g} is the 3D standard shape, S_i denote the shape modes, A_i denote the animation modes, n_s and n_a are the numbers of shape and animation modes used respectively, s_i and a_i are the respective parameters.

In this study, thirteen parameters for the shape and six parameters for the action are used. Note that the shape modes represent inter-person variety, and are ascertained by some detection procedure at the beginning. Therefore, only the tracking of animation parameters are considered:

$$a = (a_1, a_2, \dots, a_{n_a})^T \quad (8)$$

After the face model is built, it is necessary to project the 3D wire-frame to 2D image coordinate system, so a weak perspective projection model is adopted:

$$g' = f \cdot R \cdot (\bar{g} + \sum_{i=1}^{n_s} s_i \cdot S_i + \sum_{i=1}^{n_a} a_i \cdot A_i) + t \quad (9)$$

where f is the camera focal length, $t = (t_x, t_y)$ is denoted as the translation vector and R is the rotation matrix. These compose the 3D head pose and camera parameters, and all the parameters to be updated by the tracking algorithm can be denoted as follows:

$$\rho = (\theta_x, \theta_y, \theta_z, t_x, t_y, a)^T \quad (10)$$

To obtain an normalized texture observation, we make the standard shape \bar{g} as the reference mesh by projecting it onto the image system using a centered frontal pose. Afterward each input image is warped so that the vertices of new mesh with parameter ρ match

the corresponding ones of the reference mesh. A piece-wise affine transform is used to obtain a shape-free image patch as the observation, and the shape-free texture mapping from input image I is denoted as:

$$x(\rho) = W(I, \rho) \quad (11)$$

If edge structure detector is taken, the shape-free patch is described as:

$$x(\rho) = f(W(I, \rho)) \quad (12)$$

where $f(\cdot)$ is an operator includes edge strength computation and non-linear normalization. Some examples are shown on the top right corner of pictures in figure 2.

When a rectangle on the 3D face model is nearly perpendicular to the image plane, the texture in the corresponding area is distorted. In order to prevent the tracking from being disturbed by the distortion, a reliability variable for each pixel in the shape-free texture is used:

$$r_i(b) = \begin{cases} h(\varphi_i), & 0 < \varphi_i < \pi/2 \\ 0, & \varphi_i \geq \pi/2 \end{cases} \quad (13)$$

where $i = 1, \dots, N$, N is the number of pixels in the shape-free texture. The function h is monotone decreasing with $h(0) = 1$ and $h(\pi/2) = 0$. φ_i is the angle between the normal of the image plane and the normal of the rectangle.

4. Adaptive Observation Model

By assuming that the pixels of shape-free observation are independent of each other, the observation can be defined as a single multivariate Gaussian with a diagonal covariance matrix and then an on-line learning method is taken to make the observation adaptive. In this model, each pixel is considered as a random variable having Gauss distribution $N(\mu_i, \sigma_i)$, and the model parameters μ_i and σ_i can be updated from time t to time $(t + 1)$, according to the update factors α_i and growing factor β_i :

$$\alpha_i = (1 - \beta_{i(t)}) + \frac{1}{t} \beta_{i(t)} r_i \quad (14)$$

$$\mu_{i(t+1)} = (1 - \alpha_i) \mu_{i(t)} + \alpha_i x_{i(t)} \quad (15)$$

$$\sigma_{i(t+1)}^2 = (1 - \alpha_i) \sigma_{i(t)}^2 + \alpha_i (x_{i(t)} - \mu_{i(t)})^2 \quad (16)$$

$$\beta_{i(t+1)} = \beta_{i(t)} + k \alpha_i r_i \quad (17)$$

where $k > 0$, $\beta_{i(0)}$, $\beta_{i(t)}$ is restricted to be less than one, r_i is the reliability of each pixel at this time, calculated by equation 13.

5. Model Fitting

The parameters ρ can be calculated by performing a image registration on each input image with respect to the adaptive observation model. In order to find the correct state ρ , the following Mahalanobis distance between the warped patch and the current observation model is minimized:

$$\min_{\rho_t} D(x(\rho_t), \mu_t) = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \quad (18)$$

Adopting a gradient-descent registration method, we can calculate ρ in an iterative manner:

$$\Delta b = -G_t^\# (x(\rho) - \mu_t) \quad (19)$$

$$\rho = \rho + \Delta \rho \quad (20)$$

where $G_t^\# = (G_t^T G_t)^{-1} G_t^T$ is the pseudo-inverse of the gradient matrix G_t , where $G_t = \partial W(y_t, \rho_t) / \rho_t$. In [6], each column of the gradient matrix G_t is estimated using several steps around the value of each element of the state parameter ρ . It is very time-consuming. In this work, before estimating the j th column of G_t , compare the texture $W(I_t, \rho_{t-1})$ with $W(I_{t-1}, \rho_{t-1})$ on the region which is relevant to the j th element of ρ . If the difference is smaller than a threshold, the j th column of G_{t-1} is used to approximate the values of G_t .

6. Experiments

The tracking approach proposed in this paper is tested on a series of videos. The size of video frames is 320×240 , and the resolution of observation is set to be 40×46 . The platform is a PC with an Intel C2D 3.0GHz CPU. In experiments, fourteen shape modes and six animation modes of the Candide model are used. The six animation units contain: (1) upper lip raiser A_1 ; (2) jaw drop A_2 ; (3) lip stretcher A_3 ; (4) eyebrow lowerer A_4 ; (5) lip corner depressor A_5 ; and (6) outer brow raiser A_6 . Most common facial expression can be represented by these actions. Model initialization is needed to detect the face and facial feature points, then the shape parameter is determined, and an initial value is given to the pose parameter. The model initialization is carried out using the first frame of the sequence, under the assumption that the target is facing the camera (out-plane rotations $\alpha = 0$ and $\beta = 0$) and has a neutral expression (all action parameters are 0) at this time. In our work, an Adaboost classifier is used to detect the face, and a weighted AAM algorithm is adopted to detect facial feature points.

Figure 2 displays the results of three frames from sequences with illumination changes during tracking. The

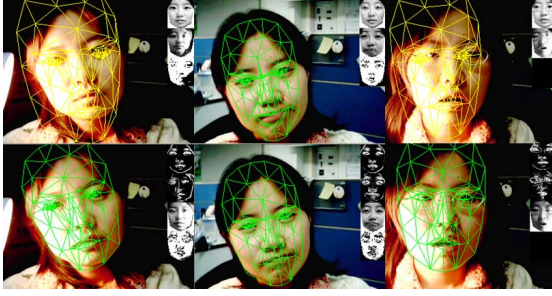


Figure 2. Examples of tracking results

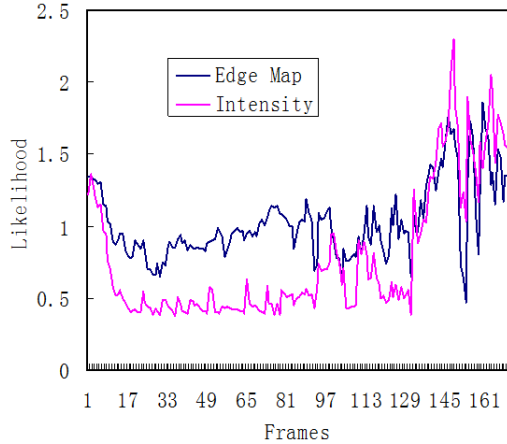


Figure 3. Likelihood measure

top right corner of each image shows the current observation template, edge or intensity based shape-free patch and growing factor respectively. The upper row shows the results using intensity, and the lower row was the results of the corresponding frames with edge. It can be seen that our method can still track smoothly when the intensity models fail in difficult lighting conditions. To be specific, we evaluate the tracking results by using a likelihood measure based on PCA[6]:

$$p(x(\rho)|\rho) = \exp\left(-\frac{1}{2} \sum_{i=1}^M \frac{\xi_i^2}{\lambda_i^2}\right) \exp\left(-\frac{e}{2\rho^*}\right) \quad (21)$$

The intensity shape-free textures with tracked parameters by both methods are scored by a PCA model trained using 799 aligned frontal face images. The result on a 204-frame-long sequence (only the tracking process is recorded) is shown in Figure 3. When the lighting condition start to change, a higher likelihood is achieved by our method. Note that the online intensity model had accommodated the new lighting condition at the end of sequence after learning for a period of time.

7. Conclusion

A real time, fully automatic facial feature detection and tracking approach using adaptive observation models is presented in this work. We made an improvement to the observation models by using edge strength. Experimental results performed on live video sequences demonstrate that our method is more accurate and robust than the original online appearance models, especially when the illumination changes during tracking.

References

- [1] J. Ahlberg. Candide-3 - an updated parameterized face, 2001. Tech. Report No.LiTH-ISY-R-2326. Image Coding Group, Dept.of EE, Linkping University, Sweden.
- [2] D. Chen and J. Yang. Robust object tracking via online dynamic spatial bias appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2157–2169, 2007.
- [3] Q. Chen, Q.-S. Sun, P.-A. Heng, and D.-S. Xia. Robust object tracking via online dynamic spatial bias appearance models. *Pattern Recognition Letters*, 29(2):126–141, 2008.
- [4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Proc. European Conf. Computer Vision*, 2:484–489, 1998.
- [5] T. Cootes and C. Taylor. On representing edge structure for model matching. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:1114–1119, 2001.
- [6] F. Dornaika and F. Davoine. On appearance based face and facial action tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(9):1107–1124, 2006.
- [7] C. Grimson and W. Grimson. Adaptive background mixture models for real-time tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2(6346911):246–252, 1999.
- [8] C. Harris and M. Stephens. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.
- [9] A. D. Jepson, D. J. Fleet, and T. R. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1(10):415–422, 2001.
- [10] S. Romdhani and T. Vetter. Efficient, robust and accurate fitting of a 3d morphable model. *IEEE International Conference on Computer Vision*, 1(8301595):59–66, 2003.
- [11] I. Scott, T. Cootes, and C. Taylor. Improving appearance model matching using local image structure. *Proc. Information Processing in Medical Imaging*, pages 258–269, 2003.
- [12] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Process*, 13(11):1491–1506, 2004.