Multi-cue Facial Feature Detection and Tracking

Jingying Chen and Bernard Tiddeman

School of Computer Science, University of St Andrews
Fife, Scotland, UK KY16 9SX
{jchen,bpt}@cs.st-and.ac.uk

Abstract. An efficient and robust facial feature detection and tracking system is presented in this paper. The system is capable of locating a human face automatically. Six facial feature points (pupils, nostrils and mouth corners) are detected and tracked using multiple cues including facial feature intensity and its probability distribution, geometric characteristics and motion information. In addition, in order to improve the robustness of the tracking system, a simple facial feature model is employed to estimate the relative face poses. This system has the advantage of automatically detecting the facial features and recovering the features lost during the tracking process. Encouraging results have been obtained using the proposed system.

1 Introduction

Facial feature detection and tracking is important in vision related applications such as human machine interaction [1], facial expression analysis [2], facial image transformation [3] and head pose tracking [4]. These applications need to track the facial features robustly and efficiently. A robust facial feature tracking system should incorporate automatic feature detection, tracking failure detection and feature recovery capability. However, the high variability of face morphology, head motion and complex background together with unknown and variable illumination conditions make the detection and tracking task difficult and complex. Hence, we propose a system capable of automatically locating a human face, analyzing its orientation, and then detecting and tracking six facial feature points, (pupils, nostrils and mouth corners), in a real time video. Integrating detection and tracking into a single system is important for recovering features after tracking failure e.g. due to temporary occlusion of the tracked features.

The facial feature detection and tracking literature includes image-based approaches [5, 6], template-based approaches [7, 8, 9], appearance-based approaches [10, 11] and motion-based approaches [12]. Each of these approaches has its own strengths and limitations. Image-based approaches use color information, properties of facial features and their geometric relationships to locate facial features. Yang and Stiefelhagen [5] presented a technique for tracking based on human skin color. This approach is in general very fast, however, color alone does not provide enough reliable information to track facial features. Stiefelhagen et al. [6] used

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 356 – 367, 2008.

color information and certain geometric constraints on the face to detect and track six facial feature points (pupils, nostrils and mouth corners) in real time for lip reading. This method works properly under good lighting conditions, however the mouth corners may drift away when the illumination changes. Template based approaches are usually applied to intensity images where a predefined template of facial feature is matched against image blocks. Tian et al. [7] used multiple state templates to track the facial features. Feature point tracking together with masked edge filtering is used to track the upper facial features. The system requires that templates be manually initialized in the first frame of the sequence, which prevents it from being automatic. Kapoor and Picard [8] used eyebrow and eye templates to locate upper facial features in a real time system. However, specialized hardware (an infrared sensitive camera equipped with infrared LEDs) is needed to produce the red eye effect in order to track the pupils. Matsumoto and Zelinsky [9] detected the facial features using an eye and mouth template matching method, which was implemented using the IP5000 image processing board.

Appearance-based approaches use facial models derived from a large amount of training data. These methods are designed to accommodate possible variations of human faces under difference conditions. Cootes et al. [10] proposed active appearance models (AAM) and Matthews and Baker [11] improved the performance of the original AAM. However, these methods need large amounts of delineated training data and involve relatively expensive computations. Also, the AAM fitting requires expensive computations which make the real time tracking difficult. Cristinacce and Cootes [12] proposed Constrained Local Model (CLM) for feature detection and tracking, they used a joint shape and texture appearance model to generate a set of region template detectors. The model is fitted to an unseen image in an iterative manner by generating templates using the joint model and the current parameter estimates, correlating the templates with the target image to generate response images and optimising the shape parameters so as to maximise the sum of response. In their method, Viola and Jones's [13] features are used to detect face. Within the detected face region they applied smaller Viola and Jones's feature detectors constrained using the Pictorial Structure Matching (PSM) method [14], to detect initial feature points. The PSM combines feature responses and shape constraints, which is very efficient due to the use of pairwise constraints and a tree structure. They claimed their proposed method is more robust and accurate than the original AAM. The method described here is similar to this method in terms of detection quality, but requires less processing time per frame. Bourel et al. [15] proposed a motion based facial feature point tracking system. In their method a Kanade-Lucas-Tomasi (KLT) tracker is employed and robust results have been obtained. However, manual initialization is required.

The facial feature detection and tracking approach using a single cue about the image sequence is insufficient for reliable performance. A robust tracking system should use as much knowledge about the image sequence as possible to handle all sources of variability in the environment. Hence, we propose to use the multi-cue of Haar-like features, intensity and its probability distribution, geometry constraints, motion and a

facial feature model to build a robust facial feature tracking system. In this paper, the proposed approach locates a human face without any artificial marks on it. This system is capable of detecting and tracking six facial features (i.e. two pupil centers, two nostril centers and two mouth corners) automatically when a human face appears in front of the camera. A simple facial feature model (locations of the feature points) is used to detect tracking failure and recover from it.

The outline of the paper is as follows. The proposed facial feature detection and tracking are presented in Section 2 and Section 3, respectively. Section 4 describes the experimental results while Section 5 presents the conclusions.

2 Facial Feature Detection

In the proposed approach, a face is first detected, which relies on a boosting algorithm and a set of Haar-like features. Then the pupils are searched for inside the face area based on their intensity characteristics and Haar-like features. Next, the mouth is found using the pupil positions and mouth intensity probability distribution. Finally the nostrils are located using their intensity and geometric constraints. The detail of the detection procedure is given below.

2.1 Face Detection

Viola and Jones's [13] face detection algorithm, based on Haar-like features is used to detect a face. Haar-like features encode the existence of oriented contrast between regions in the image. A set of these features can be used to encode the contrast exhibited by a human face and their special relationships. In Viola and Jones's method, a classifier (i.e. a cascade of boosted classifier working with Haar-like features) is trained with a few hundreds of sample views of face and non-face examples, they are scaled to the same size, i.e.24x24. After the classifier is trained, it can be applied to a region of interest in an input image. To search for the face, one can move the search window across the image and check every location using the classifier.

After the face is detected, Principal Component Analysis (PCA) [16] is employed within face region for estimation of the face direction. In 2D shape, PCA can be used to detect principal directions of the spatial shape. Since faces are approximately symmetric and there are many edge features around eyes, the first principal axis indicates the upright direction of the face while the second principal axis gives the direction of eyes (i.e. the line connecting the two pupils). Figure 1 shows the face edge map and the principal directions of the edge map.

Fig. 1. Principal axes of face skin region. The major axis Vy represents face direction while the minor axis Vx represents eye direction.

After obtaining the eye direction, one can rotate the face until the eye direction is parallel to the horizontal axis using the middle point of the line connecting the two pupils as the rotation centre, which facilitates the following pupil, nostril and mouth corner detection.

2.2 Eye Detection

2.2.1 Eye Region Localization

After the face is detected and aligned, the eye region can be located based on the face vertical edge map. Pixel intensities change in the vertical direction more near the eyes than in the other parts of the face (e.g. the eyebrows and the side of face boundary). Hence it is reasonable to use a vertical edge map to decrease false positive eye detection. First, a horizontal integral projection [17] is used on the upper face vertical edge map to estimate the vertical position of the eyes. The vertical position of the eyes is usually the global maximum of the horizontal integral projection in the upper face. According to the vertical position and the height of the face, we can estimate the vertical eye region (See Figure 2).

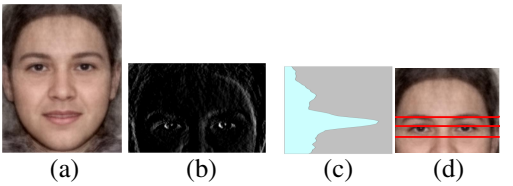


Fig. 2. Vertical eye region detection, (a) face image, (b) upper face vertical edge map, (c) horizontal projection performed on image(b), and (d) estimated vertical eye region

Second, a vertical projection is performed on the face vertical edge map to estimate the right and left boundary of face which correspond the two peaks of projection values (See Figure 3).

Finally, the eye region can be located based on the vertical eye region and right and left face boundary [18] (see Figure 4).

2.2.2 Pupil Detection

In order to improve the accuracy and efficiency of detection, eyes are searched for within the obtained eye region instead of the entire face, which decreases false positive and speeds up the detection process. Similar to face detection described above, eyes are found using a cascade of boosted tree classifiers with Haar-like features. A statistical model of the eyes is trained in this work. The model is made of a cascade of boosted tree classifiers. The cascade is trained on 1000 eye and 3000 non-eye samples of size 18x12. The training set contains different facial expression and head rotation. The 18x12 window moves across the eye region and each sub-region is classified as eye or non-eye. An example of the eye detection is shown in Figure 5.

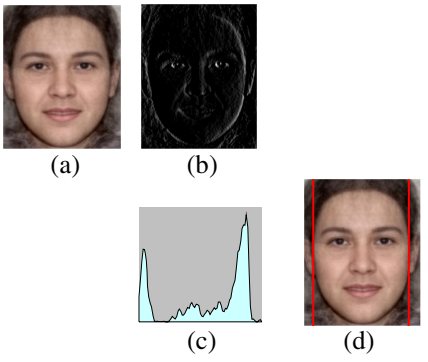


Fig. 3. Face boundary detection, (a) face image, (b) face vertical edge map, (c) vertical projection performed on image(b), and (d) estimated right and left face boundary

Fig. 4. Estimated eye region

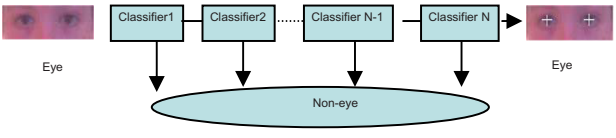


Fig. 5. Example of the eye detection using a cascade of boosted tree classifiers

2.3 Mouth Corners Detection

An estimated mouth region can be obtained using the pupil positions and face size. Then a novel approach based on entropy analysis of the partial histogram within the mouth region is proposed to segment the mouth. In this approach, the entropy E_j is iteratively calculated according to different parts of the histogram, until its value is greater than a threshold E_{th} which is found from the training data.

$$E_j = \sum_{i=0}^{j} H(i) \log H(i) \qquad \text{j=0,…,n. n}\in(1, 255)$$

where i and $H(i)$ are histogram index and value respectively. When $E_j > E_{th}$, j is used as a threshold to segment mouth. Mouths generally have lower intensities than neighbouring pixels and contain a relatively fixed proportion of the information within the mouth region, hence it is reasonable to segment them based on the threshold chosen from partial histogram entropy, which is insensitive to illumination variations. For example, in Figure 6 one can see that the two mouth images under different illumination conditions can be segmented correctly because the segmentation does not depend on the absolute intensity. Finally, the mouth corner positions are estimated based on the largest connected region of the segmented image (see Figure 6(b)). Extremities of the bright areas around the left and right parts are searched for as mouth corners.

Fig. 6. Mouth corner detection in the segmented image, (a) mouth intensity and (b) segmented images

2.4 Nostrils Detection

The nostrils appear dark relative to the surrounding area of the face under a wide range of lighting conditions. As long as the nostrils are visible, they can be found by searching for two dark regions, which satisfy certain geometric constraints. Here, the search region is restricted to an area below the pupils and above the mouth. Then the automatic thresholding, with a threshold corresponding to the lower 5 percent of the local histogram for the search window is applied. Then, the centers of the dark regions are computed as the centers of nostrils (see Figure 7).

Fig. 7. The illustration of nostrils detection

3 Facial Feature Tracking

Once the feature points have been detected, the Lucas-Kanade (LK) algorithm [19] is performed to track the pupils and nostrils. The algorithm detects the motion through the utilization of optical flow. Since the mouth has relatively higher variability (i.e. closed mouth or open mouth, with/without teeth appearance) compared to the eyes and nose, mouth corners are tracked based on the segmented mouth image. Extremities of the bright areas are searched for around the previously found left and right corners.

In order to build a robust tracking system, the system has to be able to detect tracking failure and recover from it. In this paper, the face is assumed to be planar and a simple facial feature model is used comprising five facial feature points: two centres of pupils, two nostrils and the centre of the mouth. The mouth centre is more reliable and less deformable than the mouth corners.

For each pixel (x_1, y_1) in a given image frame I_1 and the corresponding image point (x_2, y_2) in another image frame I_2, the relative orientation of two face poses is estimated using the basic projection equation of a weak perspective camera model for planar 3D object points[20].

$$\begin{pmatrix} x_2 - x_{c2} \\ y_2 - y_{c2} \end{pmatrix} = M_2 \times M_1^{-1} \begin{pmatrix} x_1 - x_{c1} \\ y_1 - y_{c2} \end{pmatrix}$$

where M_1 and M_2 are the projection matrices for image I_1 and I_2 respectively, and (x_{c1}, y_{c1}) and (x_{c2}, y_{c2}) are the projection points of the same reference point (X_c, Y_c, Z_c) in the image I_1 and image I_2. This equation is the fundamental weak perspective homographic projection equation that relates image projections of the same 3D points in two images with different face poses. The homographic matrix $P = M_2 \times M_1^{-1}$ gives the relative orientation between the two face poses [21]. Instead of using all the feature points and some sort of least-square pose fitting method [22], a minimal subset of feature points is used to estimate the pose. So long as one subset of good, accurate measurements exists, the rest of the feature points can be ignored and gross errors will have no effect on the tracking performance. The selection of a good subset can be done within the RANSAC regression paradigm [23]. The computed pose helps the tracking system to detect tracking failure and improves the robustness of the tracking system.

3.1 Tracking Failure Detection and Recovery

In this system, the failure detection includes two steps. First, all the found feature points are checked to see if they lie within the face region and satisfy certain constraints inherent in facial geometry. If not, the model points are projected back onto the image using the computed pose. In the case of mild occlusion, the lost feature points can be recovered (see Figure 9). Second, if the average distance between the back-projected model points and the actual found points exceeds a certain threshold, it is considered a failure.

Once the tracking failure has been detected, the feature points have to be searched for again. The failure recovery can be solved using the previously found pose just before the failure occurs.

If a pupil is lost during the tracking process, a search window is computed. Its center and size are chosen based on the previously found pose. The search window center is the previous position of the pupil and its size is proportional to the Euclidean distance between the two last known pupil positions. This can scale the search window automatically when the person gets closer or further from the camera. Then, the pupil detection described in the previous section is applied within the search window. The nostril recovery is based on the automatic detection of a dark region within a search window, the center of the dark region is computed as the recovered nostril center. The mouth corners are recovered based on entropy analysis of the partial histogram within the mouth region (see Figure 6, as described in the mouth corner detection).

4 Experimental Results

The proposed method has been implemented using C++ under MS windows environment with Pentium M 715 processor and tested with both static face databases and live video sequences.

4.1 Detection Results

Four face databases are used here. The first database includes 181 face images with standardized pose, lighting and camera setup. The second database includes 121 un-standardized mobile phone face images with different lighting, pose and face expression. The third database is an un-standardized set of face images from the internet containing 200 face images including different skin color, various illuminations and face poses and expressions. The final database is publicly available and known as the BIOID database (http://www.bioid.com) which consists of 1521 images of front faces under cluttered background and various illuminations. The detection rates (i.e. images with successful detected eyes, nostrils and mouth corners relative to the whole set of the database images) of the eyes, nostrils and mouth corners are 96%, 91% and 92%. Examples of the detected facial features using the proposed method are shown in Figure 8. The white crosses represent the detected features.

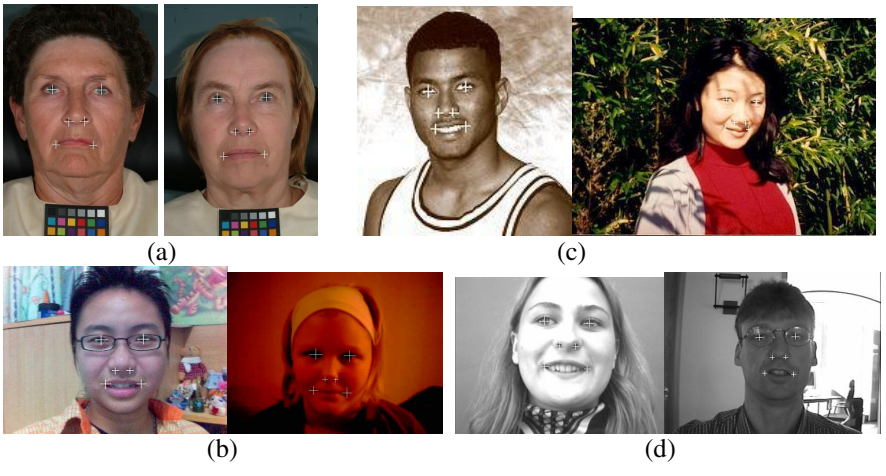
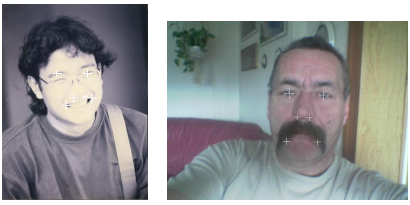


Fig. 8. The results of detected eyes, nostrils and mouth corners from four face image databases. (a) The database of highly standardised photos, (b) the database of un-standardized mobile phone image, (c) un-standardized images from the internet and (d) images from the BIOID database.

From these results, one can see that the proposed approach can detect the eyes and mouth corners accurately under various illumination conditions. However, false detection could exist under some circumstance (see Figure 9). One can notice that the eyebrows and eyes are quite similar and right nostril is invisible from Figure 9 (a) and the moustache occludes the mouth from Figure 9 (b).

To compare the proposed method with other methods, the publicly available BIOID database is used here which was specifically designed to capture faces in realistic authentication conditions and a number of methods have been evaluated on this data set [24, 25, 26]. An eye distance measure (i.e. D_e) introduced by Jersorsky et al. [24] is adopted here, which records the maximum displacement error (the displacement



(a) (b)

Fig. 9. False detection on eyes and nostrils(a) and mouth(b)

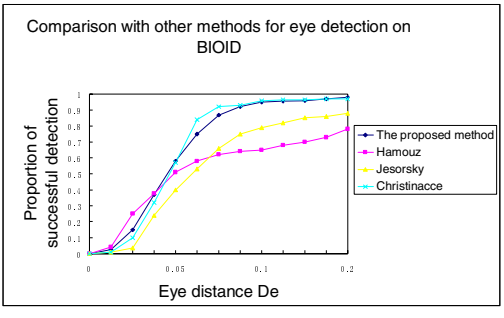


Fig. 10. Comparison with other methods for eye detection on BIOID

error is calculated as the distance between the manually determined positions of the feature points and the detected feature points) between both eyes, normalised by the true eye separation. The comparison results are given in Figure 10. From this figure, one can see that 58% and 94% success rates are obtained when D_e reaches 0.05 and 0.1 using our proposed method. In [24], Jersorsky et al. used a face matching method based on the Hausdoff distance followed by a Multi-Layer Perceptron eye finder. They achieve 40% and 79% success rate with D_e equal to 0.05 and 0.1. Hamouz et al. [25] used a method combining Gabor based feature detection to produce a list of face hypotheses, and which are then tested using a SVM face model. They also presented eye detection results on BIOID, they obtain 51% and 62% success rates when D_e approaches 0.05 and 0.1. Cristinacce et al. [26] presented a Pairwise reinforcement of feature responses approach combined with Active appearance model to detect facial feature points. They acquire 57% and 96% success rates with D_e equal to 0.05 and 0.1 on BIOID. Their method can provide high success rate, however, it needs ~1400ms to search a BIOID image using 500Mhz PII processor. The approach described here requires only ~220ms to search a BIOID image.

4.2 Tracking Results

The tracking experiments have been performed using both live data capture and pre-recorded video sequences. Tracking results from two test sequences are given below

(see Figure 10 and 11). Both sequences were captured using an inexpensive web camera with a resolution of 320 x 240 at 25 frames per second. The white crosses represent the tracked points.

From these results, one can see that the proposed approach can track the six feature points accurately when the person is moving or rotating his head, even in the case of temporary occlusion thanks to the use of the simple model which predicts occluded points during the tracking process. Figure 13 gives an estimate of the tracking accuracy of the proposed tracking system. The measurements were taken using 1027 frames of six subjects in this experiment. Manually corrected positions of the feature points were used as reference for measuring displacement error. The error was computed as the Euclidean distance between the reference points and the points obtained

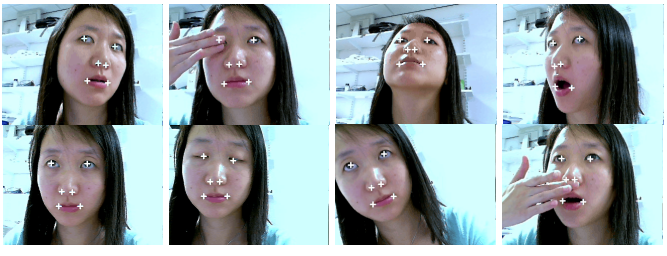


Fig. 11. Tracking results for sequence 1

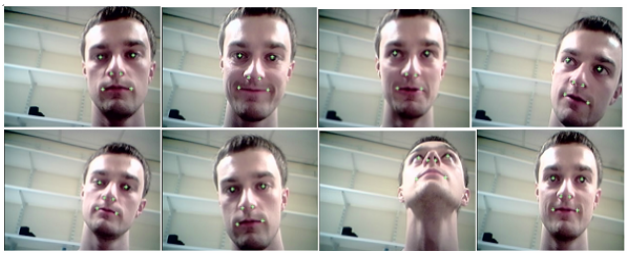


Fig. 12. Tracking results for sequence 2

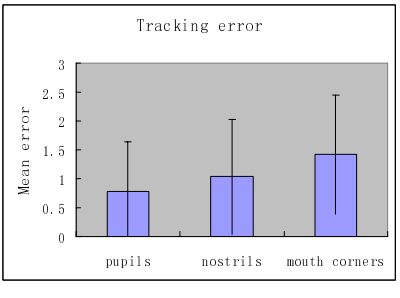


Fig. 13. The average and standard deviation of the distances between the tracked feature points and manually corrected feature points of the pupils, nostrils and mouth corners

by the proposed method. The average and standard deviations of the distances were computed across all pupils, nostrils and mouth corners during the tracking processing (see Figure 13).

The performance of the proposed approach is very good, which can cope with large angle head rotation, different facial expressions and mild occlusion of the face. The approach is fully automatic and Lucas-Kanade is a computationally efficient motion detector. They can be easily implemented in a real time system. On the other hand, a simple facial feature model (locations of the five feature points) is used to improve the system robustness, its simplicity needs inexpensive computation. Hence, the proposed tracking system is efficient and robust, suitable for putting into practice.

5 Conclusions

A multi-cue facial feature detection and tracking system is proposed in this paper, which detects a human face using a boosting algorithm and a set of Haar-like features, determines the face orientation using PCA, locates the pupils based on their intensity characteristics and Haar-like features, finds the mouth corners from the mouth intensity probability distribution, estimates the nostrils based on their intensity and geometric constraints and tracks the detected facial points using optical flow based tracking. The system is able to detect tracking failure using constraints derived from a facial feature model and recovery from it by searching for one or more features using the feature detection algorithms. The results obtained suggest the method has strong potential as alternative method for building a facial feature tracking system. In the future we hope to include additional features in the tracking.

References

1. Barreto, J., Menezes, P., Dias, J.: Human-robot interaction based on haar-like features and eigenfaces. In: Proceedings of the International Conference on Robotics and Automation, New Orleans, pp. 1888–1893 (2004)
2. Fasel, B., Luettin, J.: Automatic Facial Expression Analysis: A Survey. Pattern Recognition 36(1), 259–275 (2003)
3. Tiddeman, B., Perrett, D.: Moving Facial Image Transformations using Static 2D Prototypes. In: Proceedings of the 9-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2001 (WSCG 2001), Plzen, Czech Republic, February 5-9 (2001)
4. Chen, J., Tiddeman, B.: A stereo head pose tracking system. In: Proceedings of the 5th IEEE International Symposium on Signal Processing and Information Technology, Greece, December 18-21, 2005, pp. 258–263 (2005)
5. Yang, J., Stiefelhagen, R., Meier, U., Waibel, A.: Real time face and facial feature tracking and applications. In: Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP 1998, pp. 207–212 (1998)
6. Stiefelhagen, R., Meier, U., Yang, J.: Real-time lip-tracking for lip reading. In: Proceedings of the Eurospeech 1997, 5th European Conference on Speech Communication and Technology, Rhodos, Greece (1997)
7. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing upper face action unit for facial expression analysis. In: Proceedings of the International Conference on Computer Vision and Pattern recognition, South Caroline, USA, June 2000, pp. 294–301 (2000)

8. Kapoor, A., Picard, R.W.: Real-Time, Fully Automatic Upper Facial Feature Tracking. In: Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition, Washinton DC, USA, May 2002, pp. 10–15 (2002)
9. Matsumoto, Y., Zelinsky, A.: An algorithm for real time stereo vision implementation of head pose and gaze direction measurement. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, France, pp. 499–505 (2000)
10. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
11. Matthews, I., Baker, S.: Active Appearance Models Revisited, Technical report: CMU-RI-TR-03-02, the Robotics Institute Carnegie Mellon University (2002)
12. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: Proceedings of British Machine Vision Conference, UK, pp. 929–938 (2006)
13. Viola, P., Jones, M.: Robust real time object detection. In: Proceedings of the 2nd International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling, Vancouver, Canada (July 2001)
14. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision 61, 55–79 (2005)
15. Bourel, F., Chibelushi, C.C., Low, A.A.: Robust Facial Feature Tracking. In: Proceedings of the Eleventh British Machine Vision Conference, Bristol, UK (September 2000)
16. Feng, G., Yuen, P.: Multi-cue eye detection on grey intensity image. Pattern Recognition 34, 1033–1046 (2001)
17. Kanade, T.: Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University (1973)
18. Peng, K., Chen, L., Ruan, S., Kukharev, G.: A Robust Algorithm for Eye Detection on Gray Intensity Face without Spectacles. Journal of Computer Science & Technology 5(3), 127–132 (2005)
19. Lucas, B., Kanade, T.: An interactive image registration technique with an application in stereovision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
20. Trucco, E., Verri, A.: Introductory techniques for 3-d computer vision. Prentice-Hall, New Jersey (1998)
21. Zhu, Z., Ji, Q.: 3D Face Pose Tracking From an Uncalibrated Monocular Camera. In: Proceedings of the 17th International Conference on Pattern Recognition, UK, pp. 400–403 (2004)
22. Yao, P., Evans, G., Calway, A.: Using affine correspondence to estimate 3-d facial pose. In: Proceedings of the International Conference on Image Processing, Greece, pp. 919–922 (2001)
23. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm. of the ACM 24, 381–395 (1981)
24. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using Hausdorff distance. In: Proceedings of the 3rd International Conference on Audio and Video-based Biometric Person Authentication, Halmstad, Sweden, pp. 90–95 (2001)
25. Hamouz, M., Kittler, J., Kamarainen, J.K., Kalvioinen, H.: Affine invariant face detection and localization using GMM-based feature detectos and enhanced appearance model. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 67–72 (2004)
26. Cristinacce, D., Cootes, T., Scott, I.: A multi-stage approach to facial feature detection. In: Proceedings of British Machine Vision Conference, UK, pp. 277–286 (2004)