複数の視野により最適化した頭部姿勢推定方式

Optimization-based Multi-view Head Pose Estimation for Driver Behavior Analysis

熊 怀欣*
Huaixin XIONG

要　旨

頭部姿勢推定という研究課題はドライバの安全運転行動分析として必須である．単一のカメラでは視野が狭く人の頭部の動きを正確に推定することができないため，複数のカメラを配置することによって，より広い視野からドライバの頭部の動きを取得する方式が一般的に採用されている．本稿では最適化したマルチ視角による頭部姿勢推定の新方式を提案する．本方式では，運転席でのカメラとドライバの相対的な位置関係における姿勢制約条件を利用し，複数のカメラによる頭部姿勢の推定結果を融合させる．本方式の特徴は以下の3点である．1) 理想的な姿勢制約条件により，それぞれの視角での姿勢推定値を調整する．2) 2次元上に投影したエラーの平均最小化を最適化目標として姿勢推定値の調整を行う．3) 各視角における姿勢推定値の調整量を計算する．シミュレーションと実際の測定により，本方式がシステムの推定精度と信頼性を効率よく高めることを示すことができた．この頭部姿勢推定方式に基づき，我々は簡単なプロトタイプシステムを開発し，これを利用しドライバの頭部の動きについて分類を行った．

ABSTRACT

Head pose estimation plays an essential role in driver behavior analysis for safe driving. Since a single camera cannot accurately estimate head pose if the head moves to a large degree, we need to deploy multiple cameras to obtain a larger view space to cover the head movements during driving. In this paper, we present an optimization-based multi-view head pose estimation method that takes advantage of the constraint relationship formed by the relative positions of the cameras and the driver's head to fuse multiple estimation results and generate an optimized solution. The proposed method is novel in the following ways: 1. it introduces the ideal pose constraint conditions for each view pose self-adjustment, 2. it sets up the optimization goal of minimizing the average of the 3D projection error in the 2D plane to guide pose estimated value adjustment, and 3. it determines the adjustment through the iteration process for each view pose. The proposed method can improve the accuracy and confidence of the system estimation through self-adjustment of each view's pose estimation result by using the proposed optimization rules. We verified the method by simulation and experiment and developed a prototype to classify the driver's head movements to study the driver's behavior.

* リコーソフトウェア研究所（北京）有限公司
Ricoh Software Research Center (Beijing) Co., Ltd.

This paper is based on a paper for the 12th International Symposium on Visual Computing. The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-50832-0_45.

1. Introduction

The behavior of the driver is the most important thing for safe driving, and head pose estimation, as a basis for fatigue and visual distraction detection, has attracted the attention of more and more researchers to reduce the number of traffic accidents. Among them, the vision-based method is widely used because it is non-intrusive, easy, and less expensive. Usually the head pose has three degrees of freedom (DOF) and can be characterized by *pitch, roll,* and *yaw* angles (Fig. 1), which correspond to a single rotation matrix. Vision-based 3D pose estimation is used to find the proper rotation and translation of a head from a 2D image.

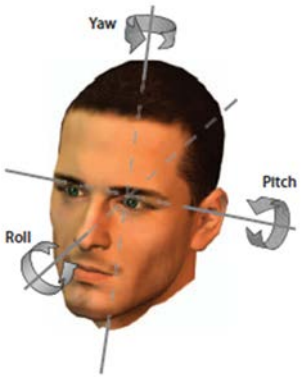

Fig. 1 Three degrees of freedom of head pose.

Currently most head pose estimation methods can be categorized into two types: model-based methods[1-8] and image appearance-based methods[9-13]. The former is usually based on the correspondence between 2D facial landmark points and the 3D head model[1-6] or on the geometric features of those facial landmark points[7,8]. The image appearance-based method is used to estimate pose directly through classification[9-12], regression[13], or Manifold embedding[13], which in most cases can obtain only coarse estimation.

The human head can be regarded as a sphere or cylinder. When the head turns to a large degree, no matter which method is applied, the estimation result is likely to be less accurate because of the limited view plane of a single camera. To obtain a larger viewing space, multiple camera fusion solutions have emerged. Ren et al.[14] pick up one head image with minimum *Yaw* angle from multiple cameras and calculate pose based on this image. Ruddarraju et al.[15] use a decision metric to switch from one camera set to another. Jiménez et al.[16] apply a weighted sum as a fusion operation based on conventional single-view head pose estimation. Michael Voit[17] fuses each result through scoring a pose hypothesis and finding the best one. Most existing fusion methods do not consider the camera's position relationship as well as the accuracy and confidence of the estimation values with different pose angles.

In fact, the multiple cameras are relevant to each other in a pose estimation system. Their stable position relationship results in the head pose estimation under a different view also having a certain correlation, which provides the possibility of improving the overall pose estimation.

In this paper, we present an optimization method to fuse head pose estimation from multiple views. Firstly, the ideal pose constraints between multiple views are derived from the position relationship of the cameras. Then, with the rotation matrix and projection as the bridge, each view estimation result is adjusted by the ideal pose constraint and constraint of the pose estimation from the image. Through setting the optimized target, it is possible to make the final head pose converge to the actual pose with the desired accuracy as good as possible. Thus, both the confidence and accuracy of the pose estimation are improved.

The rest of the paper is organized as follows. In the next section, single camera pose estimation method is given. Section 3 introduces the proposed optimization-based multi-view head pose estimation method. Section 4 presents the experimental result and finally conclusion is given in Section 5.

2. Single camera head pose estimation

2-1 Overview

In model-based pose estimation methods, the model refers to the geometric relationship in which different mathematical techniques can be used, such as analytical perspective solutions (PnP), affine solutions (POSIT), numerical perspective solutions, etc. These methods also need to obtain facial landmark points from the 2D image to describe face orientation. The known geometric correspondence between the landmark points and the model is the core feature of the model-based methods for single view pose estimation.

2-2 Landmark point location and face alignment

For a human face, the effective facial landmark points are usually on the areas near face contours, eyebrows, eyes, the nose, and the mouth. The locating of landmark points is called face alignment. The ASM (active shape model) is one of the most representative face alignment methods[18], which provides a framework to use prior knowledge from training samples to aid face alignment. In the ASM, each face shape is interpreted by model shape parameters, and it uses an iterative procedure to deform the model example to find the best fit to the image of the object by locally finding the best nearby match for each landmark point. Many researchers[19-21] have made improvements to the ASM, such as the AAM adding texture constraint to enhance shape matching[19], using SIFT descriptors to improve the local feature[20], or reorganizing the ASM face model as a hierarchical component model tree to solve the difficulties in shape optimization in high dimensional parameter space[21]. Besides, more and more new methods[22-24] have been presented in recent years to obtain a better face alignment result. Figure 2 shows an example of an ASM face alignment result with 76 points.

Fig. 2 Face alignment result with 76 landmarks.

2-3 POSIT and head 3D pose estimation

POSIT (pose from orthography and scaling with iterations) can find the pose of an object from a single image and does not require an initial guess. It is a widely used model-based object pose estimation method[1-3,14,16]. It requires four or more point correspondences between the 3D model and 2D image to calculate the object pose.

For single camera head pose estimation using POSIT, an anthropometric 3D rigid model of a human head is required. This can be acquired by a frontal laser 3D scan of a physical model, but the sparse density is enough. In this paper, five non-coplanar points on a human face are picked up from the face alignment result for pose estimation (Fig. 3), and the focal length in POSIT is obtained through camera calibration. In our paper, the tip of the nose is selected as the first 3D model point to establish the object coordinate system to describe the head 3D model.

The POSIT pose estimation accuracy depends on not only how well the 3D head model describes the current human face but also the positioning accuracy of the landmark points, which is also affected by the pose angle. With the increase in the rotation angle of the head in three directions, the corresponding estimation accuracy and confidence gradually weakens.

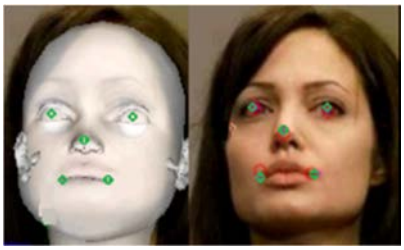

Fig. 3 Five non-coplanar points corresponding to 3D head model are used for POSIT head pose estimation.

3. Multi-camera head pose estimation

3-1 Overview

Multi-cameras can obtain a larger viewing space than covered by a single camera, which helps to track the head when there are continuous pose changes. In a multi-camera environment, each camera works independently, and the pose estimated value has stochastic properties associated with its pose. When all cameras face the same person jointly, they become relevant to each other. Unlike most multi-view pose fusion methods, our proposed method not only takes into account the confidence and accuracy of each estimated value under different pose angles but also pays more attention to the relevance among those cameras to improve both the confidence and accuracy of the pose estimation.

3-2 Multi-camera deployment and ideal pose constraint

To better capture the driver's head during regular activity, two cameras are deployed on each side of the driver's head, and they both face the driver. Once the position of the cameras and the driver are fixed, a stable pose difference is formed between the pose observed from each camera. The difference between the estimated values of the same object from a different view is called the ideal pose difference, which represents a constant constraint between the views. For example, Fig. 4 shows two cameras that are symmetrically deployed, and the angle between the camera and the head is 90°. When the face is facing forward, the *yaw* direction angle obtained by the left and right camera will be −45° and 45°. If the head turns 10° to the left, the new angle for the two cameras will change to 35° and 55°. They always keep a difference of 90° between them.

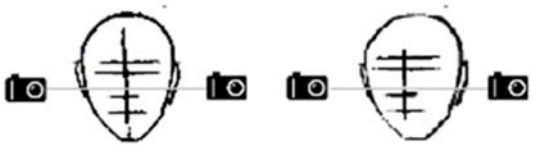

Fig. 4 Ideal pose constraint under different view.

Generally speaking, pose determination is equivalent to the exterior orientation of a camera, that is, determining the rotation and translation between the object coordinate system and the camera coordinate system[25]. Thus, we can calculate the ideal pose difference in advance through camera calibration. Each camera external parameter obtained by calibration is the rotation matrix of the camera coordinate respect to the world coordinate. Thus, firstly the world coordinate system centered at the head is established, and the left and right camera rotation matrices R_L and R_R can be obtained through the independent calibration operation. Then, the rotation matrix R_{LoR} from the left camera to the right camera can be calculated based on R_L and R_R. Finally, R_{LoR} can be decomposed into *pitch, roll,* and *yaw* angles through Euler angle calculation[26] to form the pose constraint *CONST* (*pitch, roll, yaw*).

Although each camera works independently, as part of the multi-camera system, they are relevant to each other. Thus, one can try to adjust each estimation value in a certain range from the perspective of relevance to make it meet a certain constraint. The adjustment is under the control of the optimization objective. The rotation matrix and projection are used as a bridge that connects the optimization and the adjustment.

3-3 Rotation matrix, projection, and optimization objective

Each camera estimates pose independently and the pose estimated value has stochastic properties associated with its pose. For this reason, the ideal constraints between multiple cameras are not guaranteed to be well every time. Thus, we adjust each estimated value in a controllable range to approximate the ideal pose constraint, while the controllable range is constrained by images. The estimated values, the image and pose constraint can be associated by rotation matrix, 3D projection error, and optimization goal.

A sequence of rotations around the *pitch, roll,* and *yaw* direction can be represented as a 3 x 3 rotation matrix *R*, and the pose of the 3D head is strictly a combination of its orientation *R* (a 3D rotation matrix) and its position *T* (a 3D translation vector) relative to the camera. So, the pose $P = [\, R \mid T \,]$ is a 3 x 4 matrix. Given a 3D point *(X, Y, Z)* of the head in the object coordinate system, their corresponding projection point *(x, y)* in the image is defined as

$$(x, y)^T = (x0 + fx * Xc / Zc, y0 + fy * Yc/Zc)^T \quad (1)$$

Where $(Xc, Yc, Zc)^T = [R|T]\ (X, Y, Z)^T$

Here, *(x0, y0)* is the center point of the image, and *(fx, fy)* is the focal length of the camera in the "x" and "y" direction.

It can be seen that the adjustment of the pose rotation angle can be reflected in the rotation matrix and is further associated with image landmark points through 3D projection. The distance between the 2D point re-obtained from the 3D projection and the corresponding landmark point detected in the image is called projection error. It is usually used to verify the correctness of pose estimation, and in this sense it is also considered as an image pose estimation constraint. Figure 5 shows five projection points with a 5° offset in the *yaw* direction.

Fig. 5 Green points are projected points with 5° offset in yaw direction, and white points are original landmark points.

Correction of each pose value should be carried out between the ideal pose constraint and the image pose estimation constraint. This forms our optimization goal: to minimize the average projection error while keeping each image projection error less than a threshold value. In fact, the accuracy of pose estimation has special probability distribution. The estimated value can be considered as a sampling of the distribution; therefore, the adjustment should be in a certain error range to ensure its credibility.

It can be expected that the adjustment for different views is different. Figure 6 shows that even with the same rotation angle, the projection error is different for different poses. It is a similar situation with the head model, and the pose adjustment for each view will be non-linear in our optimization.

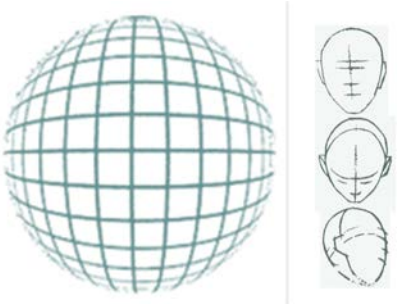

Fig. 6 Projection of equal interval grid sphere on 2D plane. Each intersection can be considered as 3D projection point.

3-4 Optimization-based method for multi-view pose estimation

To sum up, our multi-view pose estimation method is actually an optimization calculation, which can be described as below:

$$\Delta A^*, \Delta B^* = \arg \underset{\Delta A, \Delta B}{Min} \frac{1}{n} \sum_{i}^{n} (F_L(A + \Delta A, i) + F_R(B + \Delta B, i))$$

Meet condition

$$A + \Delta A + CONST \approx B + \Delta B \quad and$$

$$\frac{1}{n} \sum_{i} F_L(A + \Delta A, i) < V, \frac{1}{n} \sum_{i} F_R(B + \Delta B, i) < V$$

Here, $F_L(.)/F_R(.)$ is projection error calculation function for left / right camera.

$$F_L(A + \Delta A, i) = \left\| f_L(A + \Delta A, MP_i) - Pi(L) \right\|$$

$$F_R(B + \Delta B, i) = \left\| f_R(B + \Delta B, MP_i) - Pi(R) \right\|$$

Here, MPi is i-th head 3D point, and Pi(L/R) is 2D point in left / right view image corresponding to MPi. $f_L(.)$ / $f_R(.)$ is projection function for left /right camera

Though conventional single-view head pose estimation, we obtain Pose *A (yaw, pitch, roll)* and *B (yaw, pitch, roll)* from the left and right camera, the ideal pose constraint *CONST (yaw, pitch, roll)* is known when the positions of cameras and drivers are fixed. Our problem is to calculate $\triangle A$ and $\triangle B$, meet the condition of the ideal pose constraint . The objective of optimization is to minimize the average of the projection error "$\min_{\Delta A, \Delta B} \sum_i^n (F_L(A + \Delta A, i) + F_R(B + \Delta B, i))/n$" for all landmark points (including the left and right view image) in the 2D plane while keeping each view image projection error less than a threshold value *V*.

The simplest way to solve the optimization calculation is to apply discretization and enumerate all the possible combinations for $\triangle A$ and $\triangle B$. Since the optimization process involves matrix computation instead of image processing, it is not a time consuming operation.

The new value for pose A and B meet the ideal pose constraint. They can be used to derive the same result for the head pose relative to the front direction independently; thus, two independent events become concurrent events. According to the theory of probability, the system confidence will increase accordingly. Figure 7 shows the new flow chart for this optimization-based multi-view head pose estimation method.

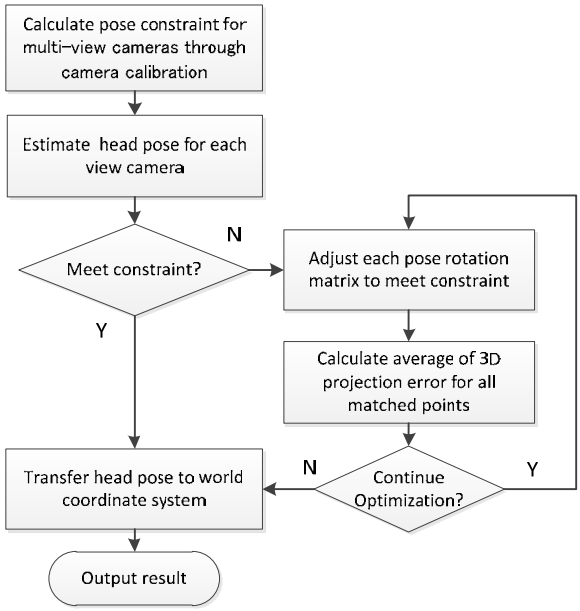

Fig. 7 Flow chart of optimization-based multi-view head pose estimation.

3-5 Calculation of each view pose self-adjustment in optimization

An iterative searching method to solve the optimization problem, which was inspired by the idea of half-interval search, is given below. For example, with the calculation in the *yaw* direction between the two cameras, since the *yaw* direction in the projection only affects the X coordinates, we can redefine the projection error based on the difference in the X coordinates as a measure of optimization.

Firstly, we calculate $\Delta D = B - A - CONST$.
Let $n = 0$, $S = \Delta D$, $A_0 = A$
Then, enter into iterative processing.
while ($S >=$ threshold V) {
n++;

Select best solution A_n based on optimization objective and conditions from 3 candidates $\{A_{n-1}, A_{n-1}+S, A_{n-1}-S\}$, each candidate C_i should be valid, meet condition, $Ci >= \min\{A, A+\Delta D\}$ and $Ci < \max\{A, A+\Delta D\}$;

$S = S/2$;
}
S is step length of each round of iteration.
If $S <$ threshold V, stop iteration
So finally, $\Delta A = An - A$,
correspondingly $\Delta B = \Delta A - \Delta D$

In the above process, A and B are still the pose values obtained from the left and right cameras, each candidate in iteration represents the new value after adjustment for pose A. Accordingly, the new value for pose B is determined under the ideal pose constraint; thus, the average projection error for each candidate can be calculated, and the candidate corresponding to the minimum projection error can be selected as a better new value for the current iteration. Since the process is iterated until S is less than a threshold value V, V indicates the accuracy for this method.

The adjustment of the other two directions can be carried out in a similar manner.

4. Experiment analysis

To evaluate the proposed method, we set up an experiment with three cameras placed at −30°, 0°, and 30° in the yaw direction. Among them, the camera at 0° represented the single camera system, and the other two cameras formed the multi-view system. All cameras were calibrated and individually estimated pose by the same method using the same 3D model. The driver's head movement was limited to ±45° in the yaw direction, and we took 10 pictures every 15° for every camera. In total, we collected 70 x 3 pieces, and all facial landmark points were manually marked to eliminate error caused by inaccurate location and to better show each algorithm performance. We compared our method with the single camera method and the Samsung patent method[14]. The results show that our method is better than the other two methods in improving accuracy. Fig. 8 shows each average estimation error at different poses in the yaw direction with ground truth facial feature landmarks.

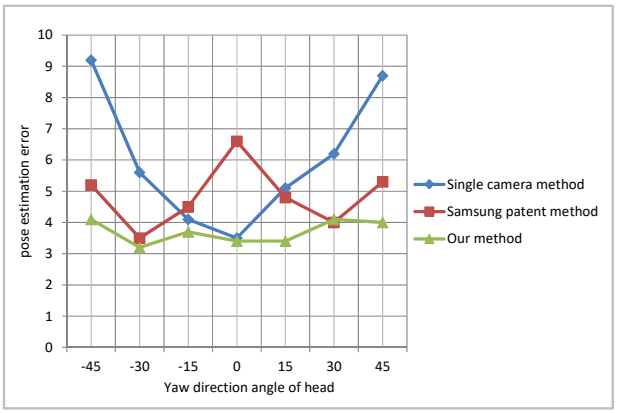

Fig. 8 Comparison of accuracy among three methods.

The confidence is also an important performance indicator that shows how much the probability can maintain the estimated value around the true value in a given range. Considering the capacity of the sample was not enough, the confidence was analyzed by sampling simulation method. We assume each camera pose estimation value obeys Gauss distribution and has different deviation under different pose angles. Although the simulation result is affected by assumed parameters, we can still find the trend change in confidence improving compared with the other two methods through trying different parameter combinations. Figure 9 is one of the simulation results with a failure number of 1000 samples.

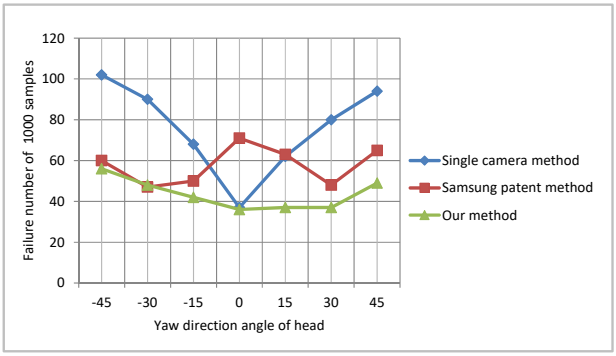

Fig. 9 Confidence simulation result with 1000 samples.

In addition to the experimental verification, we also developed a simple prototype to show the concept of vision-based driver behavior analysis. We captured image and estimated head pose with POSIT, through voting in short time. We divided the driver's head status into five kinds: front, left, right, etc. This prototype integrates five models, the camera calibration, face and eyes detection, face alignment, pose estimation, and behavior analysis. Figure 10 shows some snapshots outputted by this prototype.

Fig. 10 Snapshots of head pose estimation prototype.

5. Conclusion and future work

We have proposed an optimization-based method for multi-view head pose estimation in the driving environment. The core idea for the proposed method is to use the relevance between different views and promote the estimated values to adjust themselves in a certain range to improve the accuracy and confidence. And the experiment result confirms this idea from simulation and real measurement.

During the optimization adjustment, if the image estimation constraint cannot be effectively guaranteed, we can relax the ideal pose constraint within a certain range of accuracy and try again. If this attempt fails again, there is at least one mistake in those independent estimators, so the optimization method cannot be applied. In fact, the setting of the multi-view system is not only for single frame pose calculation but also more importantly for tracking the head pose changes continuously. Thus, when the confidence for a single estimation is not known, the priori probabilities and the historical information at that time should be considered together to decide which view result is more reasonable. Therefore, to strengthen collaborative tracking in the multi-view environment is the main task for future work.

References

1) D. F. DeMenthon, L. S. Davis: Model Based Object Pose in 25 Lines of Code, *International Journal of Computer Vision*, Vol. 15, Issue 1-2, pp. 123-141 (1995).

2) B. H. P. Prasad, R. Aravind: A Robust Head Pose Estimation System for Uncalibrated Monocular Videos, *Proc. 7th Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 162-169, ACM (2010).

3) P. Martins, J. Batista: Monocular Head Pose Estimation, *Proc. 5th ICIAR*, pp. 357-368, Springer (2008).
4) T. Petersen: A Comparison of 2D-3D Pose Estimation Methods, Aalborg University (2008).
5) 胡步发, 邱丽梅: 基于多点模型的3D人脸姿态估计方法, 《中国图像图形学报》, Vol 13, No. 7 (2008).
6) 邱丽梅, 胡步发: 基于仿射变换和线性回归的3D人脸姿态估计方法, 《计算机应用》, Vol. 26, No. 12 (2006).
7) 胡元奎, 汪增福: 快速的人脸轮廓检测及姿态估计算法, 《模式识别与人工智能》, Vol. 19, No. 5 (2006).
8) 王言群: 基于边缘统计和特征定位的人脸姿态估计方, 《计算机系统应用》, Vol. 20, No. 4 (2011).
9) T. Vatahska, M. Bennewitz, S. Behnke: Feature-based Head Pose Estimation from Images, *Proc. 7th IEEE-RAS International Conference on Humanoid Robots* (2007).
10) Z. G. Yang et al.: Multi-view face pose classification by tree-structured classifier, *IEEE International Conference on Image Processing*, Vol. 2 (2005).
11) 刘坤, 罗予频, 杨士元: 光照变化情况下的静态头部姿态估计, 计算机工程, Vol. 34, No. 10 (2008).
12) 张毅, 廖巧珍, 罗元: 融合二阶HOG与CS-LBP的头部姿态估计, 智能系统学报, Vol. 10, No. 5 (2015).
13) 范进富, 陈锻生: 流形学习与非线性回归结合的头部姿态估计, 中国图像图形学报, pp. 1002-1010 (2012).
14) 任海兵, 王西颖, 金智渊: 一种头部姿态检测设备及方法, CN 102156537 A (2010).
15) R. Ruddarraju et al.: Fast Multiple Camera Head Pose Tracking, Vision Interface (2003).
16) P. Jiménez et al.: Face tracking and pose estimation with automatic three-dimensional model construction, *Computer Vision, IET*, Vol. 3, Iss. 2, pp. 93‐102 (2009).
17) M. Voit: Multi-view Head Pose Estimation using Neural Networks, *Proc. 2nd Computer and Robot Vision*, IEEE (2005).
18) T. F. Cootes et al.: Active Shape Models-Their Training and Application, *Computer Vision and Image understanding*, Vol. 61, No. 1, pp. 38-59 (1995).
19) T. F. Cootes et al.: Active Appearance Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence archive*, Vol. 23, Issue 6, pp. 681-685 (2001).
20) S. Milborrow, F. Nicolls: Active Shape Models with SIFT Descriptors and MARS, *Computer Vision Theory and Applications (VISAPP)* (2014).
21) 熊怀欣: 一种人脸对齐方法装置及电子设备 CN 201610963243.4 (2016).
22) L. Liang, F. Wen, J. Sun: Face Alignment Via Component-Based Discriminative Search, US patent 8,200.017 B2 (2012).
23) S. Ren: Face Alignment at 3000 FPS via Regressing Local Binary Features, *Computer Vision and Pattern Recognition (CVPR)* (2014).
24) S. Z. Zhu et al.: Face Alignment by Coarse-to-Fine Shape Searching, *Computer Vision and Pattern Recognition (CVPR)* (2015).
25) I. Lopez et al.: Pose estimation from 2D to 3D for computer vision in an assembly node, CTB500-02-0000 (2002).
26) G. G. Slabaugh: Computing Euler angles from a rotation matrix, http://www.staff.city.ac.uk/~sbbh653/publications/euler.pdf (accessed 2016-10-05).