MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

An Efficient 3D Geometrical Consistency Criterion for Detection of a Set of Facial Feature Points

Mayumi Yuasa, Tatsuo Koazkaya and Osamu Yamaguchi
Corporate Research & Development Center, Toshiba Corporation
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 212–8582, Japan
{mayumi.yuasa,tatsuo.kozakaya,osamu1.yamaguchi}@toshiba.co.jp

Abstract

We propose a novel efficient 3-dimensional geometrical consistency criterion for detection of a set of facial feature points. Many face recognition methods employing a single image require localization of particular facial feature points and their performance is highly dependent on localization accuracy in detecting these feature points. The proposed method is able to calculate alignment error of a point set rapidly because calculation is not iterative. Also the method does not depend on the type of point detection method used and no learning is needed. Independently detected point sets are evaluated through matching to a 3-dimensional generic face model. Correspondence error is defined by the distance between the feature points defined in the model and those detected. The proposed criterion is evaluated through experiment using various facial feature point sets on face images.

1 Introduction

Most face recognition methods require localization and normalization of facial images, and their performance is highly dependent on localization accuracy in detecting these feature points[1]. Normalization based on relatively few facial points, typically the eyes, is standard. Recently, normalization methods based on increasing numbers, more than ten points, have been proposed[2]. These methods improve face recognition based on a single image, because the normalization of pose becomes resistant to localization error in a few points. Therefore, it is important to detect a large number of facial points automatically with sufficient precision. Figure 1 shows examples of pictorial facial feature points.

Facial feature point detection methods include those based on fitting to models of geometry, appearance or both[3, 4]. Other methods detect each feature point independently or only use a heuristic or two-dimensional relationship between points[5, 6].

In the case of detecting many facial feature points, the following problem arises. To detect many facial points at one time, it is necessary to ensure consistency among them. To coordinate the consistency among many points, there are two main types of methods: methods using models and heuristic methods. The method using models optimizes the degree of fitting to the models. As optimization often includes an iterative process, computational cost is relatively high. If the model supports three-dimensional variation, the cost increases significantly. There is also a problem in initialization. In the case of heuristic methods, it is difficult to define the relationships between points, and also to support three-dimensional variation.

We consider it to be desirable to develop a criterion to check the 3D geometrical consistency among facial feature points with ease. Such a criterion would be able to verify the facial feature points detected by some methods or to select the most appropriate one from them, when the number of detected point sets is greater than one.

We propose a three-dimensional consistency criterion that has this property. The proposed method has the following advantages: a) the criterion does not depend on the type of detection method of individual points, b) no learning process is necessary, c) the calculation is simple and does not need an iterative process. Note, however, that a sufficient number (greater than three) of non-planar labeled feature points should be detected with sufficient confidence.

2 Geometrical consistency criterion

A key question is how can we describe the facial geometric distribution (i.e. facial likelihood) of a set of feature points? Although the alignment of facial feature points varies from person to person, the distribution is similar to some extent. We consider the point set is similar to a face if its alignment is similar to that on a 3D generic face model. We found that a distance between the detected point set and the projected one by the motion matrix calculated by matching the detected points to those on the generic 3D face model has the property of a 3D geometrical consistency criterion of the detected point set; that is, the criterion describes a facial geometric likelihood for the detected feature points.

There are several approaches to find an individual shape model from images or to estimate head pose. However, the proposed approach uses the error in

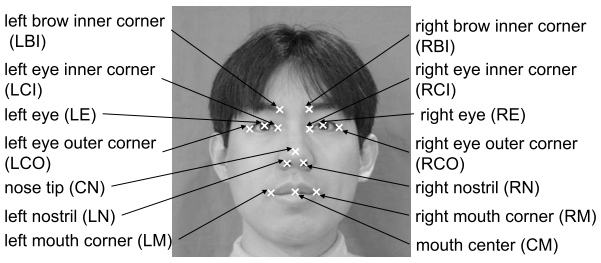


Figure 1: Examples of facial feature points: 14 points on a face image.

matching to the shape model as the criterion to eliminate incorrect point set candidates and as a measure to select the most appropriate ones.

2.1 Matching to generic 3D model

We use a part of the factorization framework[7] to obtain correspondences of facial feature points to those on the generic model. To aquire an individual shape model with a single image is difficult. It is, however, easy to find the correspondence between detected points and those on the model if the model is generic. We also assume that the feature points on the model are known.

If N feature points are detected, a $2 \times N$ measurement matrix W is defined as

$$W = \begin{bmatrix} u'_1 & u'_2 & \cdots & u'_N \\ v'_1 & v'_2 & \cdots & v'_N \end{bmatrix}, \quad (1)$$

where $(u_i, v_i)^T$ is the i-th feature point, $(\overline{u}, \overline{v})^T$ is the average of them, and $(u'_i, v'_i)^T = (u_i - \overline{u}, v_i - \overline{v})^T$. A $3 \times N$ shape matrix S is defined as

$$S = \begin{bmatrix} x'_1 & x'_2 & \cdots & x'_N \\ y'_1 & y'_2 & \cdots & y'_N \\ z'_1 & z'_2 & \cdots & z'_N \end{bmatrix}, \quad (2)$$

where $(x'_i, y'_i, z'_i)^T$ is the i-th feature point on the generic 3D model, $(\overline{x}, \overline{y}, \overline{z})^T$ is their mean and $(x'_i, y'_i, z'_i) = (x_i - \overline{x}, y_i - \overline{y}, z_i - \overline{z})^T$. Then a 2×3 motion matrix M is represented by the expression[7]:

$$W = MS \quad (3)$$

If the feature points are non-planar, the pseudo inverse matrix $S^\dagger$ is able to be calculated and motion matrix M is obtained by multiplying equation (3) with $S^\dagger$.

$$M = WS^\dagger \quad (4)$$
$$= WS^T(SS^T)^{-1} \quad (5)$$

The obtained motion matrix describes correspondence between the detected facial feature points and those on the generic model. The pseudo inverse matrix $S^\dagger$ is able to be calculated in advance as the model is constant. Therefore, the calculation is simply a matrix multiplication and is non-iterative.

2.2 Matching error measured in the 2D image plane

First we describe the method to project the feature points on the model to the original image plane through motion matrix M calculated by the above-mentioned matching process. The left side of Figure 2 illustrates the concept of this projection process. The obtained motion matrix M can be considered a projection matrix from the model to the image, which minimizes the projection error of the feature points on the model. By applying the projection matrix to the feature points $(X, Y, Z)^T$ on the model, the corresponding point $(u, v)^T$ on the original image can be calculated as follows:

$$\begin{bmatrix} u \\ v \end{bmatrix} = M \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (6)$$

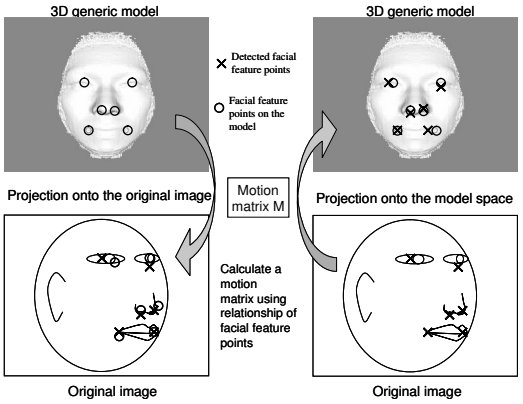


Figure 2: Conceptual diagram of the projections (an example of six feature points).

Thus, the projected feature points are obtained.

Then, the distance between the projected point set and the detected one is obtained. The distance has to be normalized by some measure of scale. Eye-to-eye distance is often used for this purpose.

2.3 Matching error measured in the 3D model space

Although the above-mentioned projection can be used for the proposed criterion, the normalization of the distance by the detected eye-to-eye distance will fluctuate owing to detection error or facial pose.

Instead of projecting the feature points on the model onto the original image plane, the detected points can be projected onto the model space; that is the point set is projected onto the error measurement plane in the model space (the right side of Figure 2). To do so, scale normalization is reasonable compared to the method described in the preceding section. This is because the scale is naturally normalized through this projection process and eye-to-eye distance is therefore less disturbed by pose or detection error.

If the detected facial feature point in the image is $(u, v)^T$, the corresponding point on the model $(X, Y, Z)^T$ is defined by equation (6). In equation (6), the number of unknown variables exceeds the number of equations to find $(X, Y, Z)^T$ from $(u, v)^T$; the solution is ill-defined. To solve this problem, we assume that the Z coordinate, depth, of the detected point is the same as the point on the model. Based on this hypothesis, $(X, Y)^T$ is obtained by the following equation.

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}^{-1} \begin{bmatrix} u - m_{13}Z \\ v - m_{23}Z \end{bmatrix}, \quad (7)$$

where

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{bmatrix} \quad (8)$$

2.4 Definition of distances

There are several variations of distances to measure the two point sets. In this paper, we use the maximum distance among feature points (similar to L_∞ distance).

We defined this as "paired L_∞ distance." Compared to the case where small errors exist for many points, the case where a few points have large errors is more important. The advantage of using the proposed distance is that it is able to distinguish the latter case.

For comparison, we use other types of distances: Euclidean distance(L_2) and L_∞. The validity of the proposed distances is evaluated later.

If the detected point set projected onto the model space is expressed by X and the point set on the model by X^0, where

$$\mathrm{X} = (X_1, Y_1, X_2, Y_2, \cdots, X_N, Y_N)^T \tag{9}$$

$$\mathrm{X}^0 = (X_1^0, Y_1^0, X_2^0, Y_2^0, \cdots, X_N^0, Y_N^0)^T \tag{10}$$

Paired L_∞ distance $D_{p\infty}(\mathrm{X}, \mathrm{X}^0)$ is defined as

$$D_{p\infty} = \max_{1 \le i \le N} \left(\sqrt{(X_i - X_i^0)^2 + (Y_i - Y_i^0)^2} / D_{\mathrm{eyes}} \right) \tag{11}$$

The distance is normalized by scale parameter D_{eyes}, which means the eye-to-eye distance on the model.

L_∞ distance $D_\infty(\mathrm{X}, \mathrm{X}^0)$ is

$$D_\infty = \begin{cases} D_\infty^x & \text{if } D_\infty^x > D_\infty^y \\ D_\infty^y & \text{otherwise} \end{cases}, \tag{12}$$

where

$$D_\infty^x = \max_{1 \le i \le N} \left(|X_i - X_i^0| / D_{\mathrm{eyes}} \right) \tag{13}$$

$$D_\infty^y = \max_{1 \le i \le N} \left(|Y_i - Y_i^0| / D_{\mathrm{eyes}} \right) \tag{14}$$

and L_2 distance $D_2(\mathrm{X}, \mathrm{X}^0)$ is

$$D_2 = \sqrt{\sum_{i=1}^{N} \{(X_i - X_i^0)^2 + (Y_i - Y_i^0)^2\}} / D_{\mathrm{eyes}} \tag{15}$$

3 Evaluation of the proposed criterion

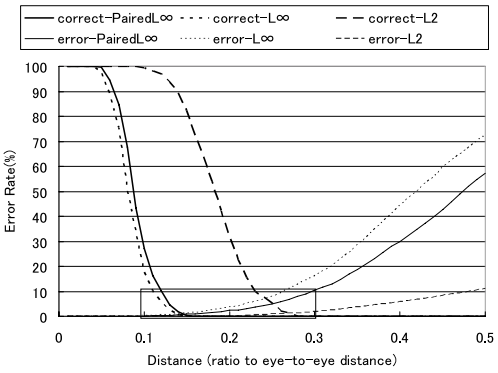
In this section, the effectiveness of the proposed criterion is evaluated through experiment using real face image data. All facial feature points used in this section are produced by manual input. Fourteen points are used for evaluation. They are eyes, nostrils, mouth corners, eye corners, brow inner corners, tip of the nose and mouth center (see Figure 1): the points that have distinct shape and therefore are easy to detect.

The 3D shape models used in this paper are acquired by a 3D digitizer. We captured several individuals and averaged them to make a generic 3D model.

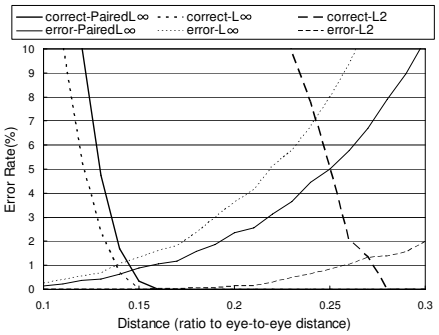
3.1 Evaluation of effectiveness for error measurement

Availability of the proposed criterion as a measure to detect inconsistent alignment of the points is validated. We use the XM2VTS database[8] for evaluation in this section.

The dataset used is the XM2VTS session1(295 individuals). The distances defined in section 2.4 are calculated. Figure 3 shows the distributions of these distances for both normal and artificially generated error data. For correct data, the rate over the threshold distance (false rejection rate), and for error data, the rate under the threshold distance (false acceptance rate) is shown. The error data are generated by scattering the feature point coordinates with random error. The displacement is generated randomly between 0.2 and 0.5 of the eye-to-eye distance. The rage of displacement is decided so that the error data are not too close to the correct ones.



(a)



(b)

Figure 3: Distributions of the defined distances for correct and error data ((b) is a magnified view of (a)).

The equal error rate (i.e. the cross point) obtained by this experiment is 0.8% for paired L_∞ distance, 1.0% for L_∞, and 1.3% for L_2. The results show that the proposed "paired L_∞ distance" is better than others and has an ability to eliminate incorrect point sets. Note that there are in principle incorrect point sets that cannot be eliminated by this criterion. For example, in the situation that all the points can be displaced in parallel one another.

3.2 Evaluation of pose robustness

Robustness of the proposed criterion for variation of facial poses is evaluated. The FERET database[9] with various poses is used for evaluation in this section.

Figure 4 shows the ROC curve for the FERET database with known facial poses. Note, however, that only paired L_∞ distance is shown.

The results show that the equal error rate increases with the face angle from frontal being larger. However, over 85% of the error data can be eliminated when FRR is under 0.1% even for the data of 60 degrees.

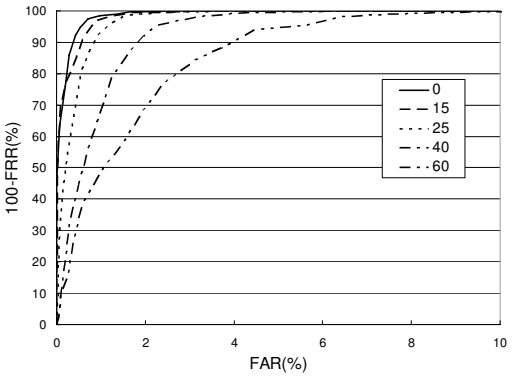


Figure 4: ROC curve of paired L_∞ distance for the FERET database for various poses (0, 15, 25, 40, and 60 degrees).

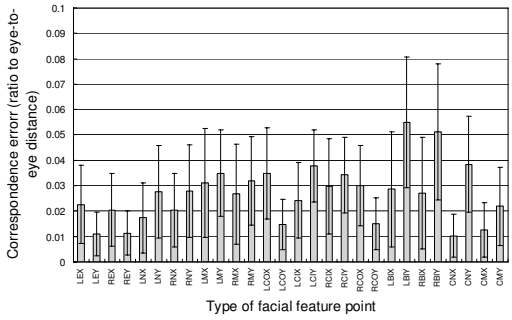


Figure 5: Average errors for each feature point with their standard deviation.

3.3 Discussion about error variation between feature points

In this section we discuss the effect of error variation between feature points. Figure 5 shows average errors for each feature point for XM2VTS database. The degree of error varies considerably according to the feature point. To eliminate this effect, we have to introduce an extended distance, considering an effect of variation by feature points. For this purpose, the Mahalanobis distance should be appropriate. For simplicity, we normalize the displacement for each point using an average and standard deviation of displacement on the assumption that the point errors are independent to some extent. Displacement of each point is normalized before calculating distances.

Figure 6 shows the normalized paired L_∞ distance distribution for the XM2VTS data. From this result, the equal error rate is 0.6%. It is lower than that of the unnormalized distance shown in Figure 3.

4 Conclusions

We have proposed a novel efficient 3D geometrical consistency criterion for detection of a set of facial feature points. If sets of facial feature points are detected, the detected point sets are evaluated through matching to a 3D general face model. The matching calculation is efficient because it is only necessary to project the coordinates in the original image plane onto the model using the motion matrix created from the model and to calculate the distance between them. The proposed criterion was evaluated through experiment using various image data. The suitability of the proposed criterion for large error of a small number of points was demonstrated.

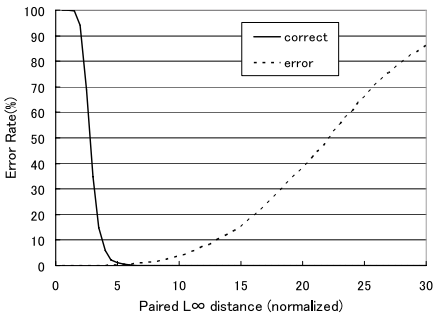


Figure 6: Paired L_∞ distance (normalized) distribution for the XM2VTS data.

In future work, we intend to develop a method of detecting each feature point more accurately and efficiently. We will also consider a method to estimate missing points and to deal with local deformation such as facial expressions.

References

[1] P. Wang, M.B. Green, Q. Ji and J. Wayman, "Automatic eye detection and its validation," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol.3, p. 164, 2005.

[2] T. Kozakaya and O. Yamaguchi, "Face recognition by projection-based 3D normalization and shading subspace orthogonalization," Proc. the 7th International Conference on Automatic Face and Gesture Recognition, pp.163–168, 2006.

[3] I. Matthews and S. Baker, "Active appearance models revisited," International Journal of Computer Vision, vol. 60, no. 2, pp. 135–164, 2004.

[4] L. Chen, L. Zhang and M. Abdel-Mottaleb, "3D shape constraint for facial feature localization using probabilistic-like output," Proc. the 6th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 302–307, 2004.

[5] D. Vukadiovic and M Pantic, "Fully automatic facial feature point detection using gabor feature based boosted classifiers," Proc. IEEE International Conference on Systems, Man and Cybernetics, pp. 1692–1698, 2005.

[6] D. Cristinacce and T. Cootes. "Facial feature detection using AdaBoost with shape constraints." Proc. British Machine Vision Conference, pp. 231–240, 2003.

[7] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," International Journal of Computer Vision, vol. 9, no. 2, pp. 137-154, 1992.

[8] K. Messer, J. Matas, J. Kitter, J. Luettin and G Maitre, "XM2VTSDB: The extended M2VTS database," Proc. Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.

[9] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, 2000.