社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Face recognition based on virtual frontal view generation using LVTM with local patches clustering

Xi LI†, Tomokazu TAKAHASHI†,††, Daisuke DEGUCHI†, Ichiro IDE†, and Hiroshi MURASE†

† Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
†† Faculty of Economics and Information, Gifu Shotoku Gakuen University,
Nakauzura 1-38, Gifu, Gifu, 500-8288, Japan
E-mail: †{li_xi,ddeguchi,ide,murase}@murase.m.is.nagoya-u.ac.jp, ††ttakahashi@gifu.shotoku.ac.jp

Abstract One of the major difficulties encountered by face recognition is the varying poses caused by in-depth rotations. The intra-person appearance differences caused by rotations are often larger than the inter-person differences, which makes the traditional face recognition methods such as eigen-face infeasible. This paper presents a framework for face recognition across pose based on virtual frontal view generation using Local View Transition Model(LVTM) with local patches clustering. Previous study on LVTM shows that more accurate appearance transition model can be achieved by first dividing the original face image plane into overlapping local patch regions and then the learned transition models for each patch are aggregated for the final transformation. In this paper we show that the accuracy the appearance transition model and the recognition rate can be further improved by better exploiting the inherent linear relationship between frontal-nonfrontal face image pairs. This is achieved based on the observation that variations in appearance caused by pose are closely related to the corresponding 3D face structure and intuitively frontal-nonfrontal pairs from more similar local 3D face structures should have a stronger linear relationship. For each specific location, instead of learning a common transformation as in LVTM, the corresponding local patches are first clustered based on appearance similarity distance metric and then the transition models are learned separately for each cluster. In the testing stage, each local patch for the input nonfrontal probe image is transformed using the learned local view transition model corresponding to the most visually similar cluster. The experimental results on real life face dataset demonstrate the effectiveness of the proposed method.

Key words face recognition, cross pose, view transition model, local patch, clustering

1. Introduction

Face recognition is one of the most active research topic in computer vision and pattern recognition communities due to its wide range of potential applications such as identity authentication, public security, surveillance, human-computer interface and so on. Unlike fingerprint recognition or iris recognition, face recognition has the advantage of being natural and passive and is inherently a non-intrusive technique that is able to identify an uncooperative face in arbitrary unconstrained condition. Within the past two decades many methods have been proposed for face recognition. Most of those traditional methods can only successfully recognize faces when face images are captured under constrained conditions. Usually the performance of the traditional methods will degrade greatly when face images are captured in unconstrained conditions caused by factors such as varying viewpoint, illumination, expression and poses. This work studies the problem of face recognition across poses, where each subject has a frontal gallery face image stored in the database and the probe image is not necessarily frontal. It is of great interest in many real life face recognition application scenarios such as surveillance systems, where the captured face images are usually low-resolution and non-frontal.

Pose variation caused by in-depth rotation was identified as one of the prominent unsolved problems in the research of face recognition. The intra-person appearance differences caused by rotations are often larger than the inter-person differences, which makes the traditional face recognition methods such as eigen-face [1] infeasible. [2] proposed the 3D Morphable Model method for pose invariant face recognition. The 3D Morphable Model is built using principal component analysis on 3D facial shapes and textures that obtained from laser scan device and then the 3D face is reconstructed by fitting the model to the input 2D image. Face recognition across pose with the assistance of 3D face models can deal with both the pose and illumination variations. Those 3D model

based methods can successfully handling pose variation. However it is difficult to apply this kind of strategy to low-resolution face images because they require a large number of accurate point correspondences between a face image and a face model to fit the image to the model. Instead of 3D model based method, 2D techniques such as the real view-based matching have also been proposed to tackle the pose invariant problem for face recognition [3]. For the purpose of tolerating pose variations, one can actively compensate pose variations by providing gallery views in rotation to recognize rotated probe views.The natural way to realize a face recognition system against pose variations in this direction is to prepare multiple real view templates for every known individual. The number of required real gallery images can be significantly reduced by quantization on the in-depth rotations. Usually it is generally impractical or unfavorable to collect multiple images in different poses for real view-based matching. 2D appearance based virtual view generation is another possible solution for pose invariant face recognition [4] [5] [6] [7] [8]. One of the noteworthy works is the View-Transition Model (VTM) [4], which transforms views of an object between different postures by linear transformation of pixel values in images. For each pair of postures, a transformation matrix is calculated from image pairs of the postures of a large number of training dataset. VTM was further extended to Local VTM(LVTM) in a patch-wise way and more satisfactory result was achieved [5]. Locally Linear Regression [7] is another work that takes a similar approach for pose-invariant face recognition,which generates a virtual frontal view from a single relatively high-resolution non-frontal face image by applying a patch-wise image transformation method. This paper further extends the LVTM and presents a framework for face recognition across pose based on virtual frontal view generation using LVTM with local patches clustering(c-LVTM). The experimental results on real life face dataset demonstrated the effectiveness of the proposed method.

The following of this paper is organized as follows: in section 2, the flowchart of face recognition across pose using virtual frontal face generation is illustrated and the original VTM and LVTM methods are introduced briefly; Section 3 describes the proposed clustering based local VTM method(c-LVTM) in detail. Section 4 is the experimental result and section 5 draws the conclusion.

図 1 The flowchart of cross pose face recognition based on virtual frontal face generation

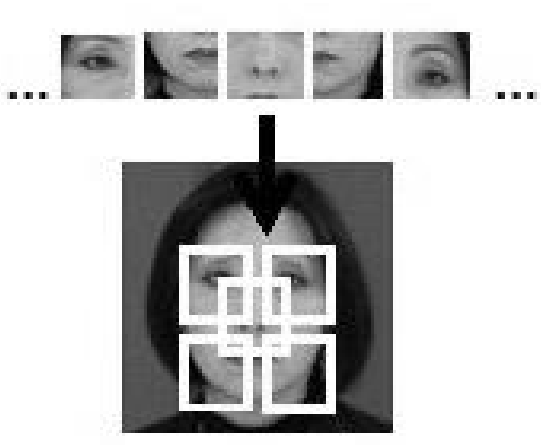

図 2 Synthesis by face patches aggregation.

2. Cross pose face recognition by virtual frontal view generation

Cross pose face recognition by virtual frontal face recognition is to synthesize virtual views to substitute the demand of real views from a limited number of known views, even from just a single view. Instead of directly classifying the probe nonfrontal face image, the nonfrontal face image is firstly transformed to its virtual frontal counterpart and then a general face recognition procedure is applied. The flowchart of cross pose face recognition based on virtual frontal face generation is illustrated in Fig 1. One of the typical method in this category is the View Transition Model(VTM). VTM method was proposed for virtual frontal face generation by transforming from multiple low-resolution non-frontal faces. To achieve this, VTM method uses a general training image dataset consisting of faces of a large number of individuals viewed from various angles other than the input individual. The linear transformations learned from the training dataset are applied to the probe nonfrontal face images to generate the counterpart virtual frontal face image that is fed into a general traditional face recognition engine.

More specifically, given a training dataset $\Theta : \{\mathbf{Q}_{\phi}^{1}, \mathbf{Q}_{\theta_1}^{1}, ..., \mathbf{Q}_{\theta_L}^{1}, ..., \mathbf{Q}_{\phi}^{N}, \mathbf{Q}_{\theta_1}^{N}, ..., \mathbf{Q}_{\theta_L}^{N}\}$, where N is the number of training subjects, $\mathbf{Q}_{\phi}^{n}, n = 1, ..., N$ represents the frontal face image for the ith subject in the vector form which is a column vector that has pixel values of the image as its elements and $\mathbf{Q}_{\theta_l}^{n}, l = 1, ..., L, n = 1, ..., N$ represents the nonfrontal face image for the ith subject with the rotation angle of θ_l. For an input probe nonfrontal face image $\mathbf{Q}_{\theta_l}^{probe}$, the purpose is to generate its virtual frontal image $\mathbf{Q}_{\phi}^{probe}$ using the linear transformation learned from the training dataset. VTM can be applied for frontal face recognition by one or any number of input images. However, in the interest of simplicity, we describe the frontal face generation algorithm for one non-frontal face input

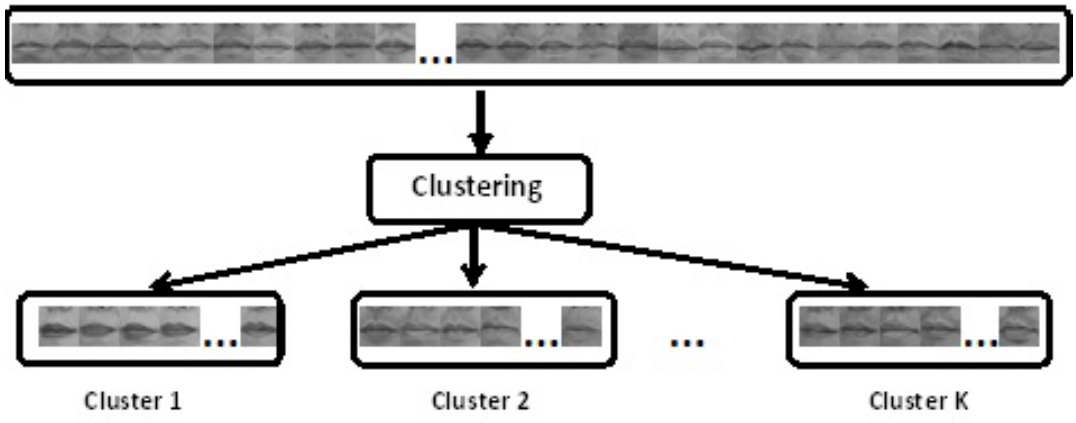

図 3 Clustering the local patches into several clusters according to the underlying 3D structure and corresponding 2D appearances and then learning linear mapping for each clusters.

image only and assume that the training dataset consists of frontal -nonfrontal face images pairs with only one rotation degree θ only. The VTM calculates the linear transformation $\mathbf{T}$ beforehand using the training dataset by solving the following equation [4]:

$$\left[\begin{array}{ccc} \mathbf{Q}_{\phi}^{1} & \cdots & \mathbf{Q}_{\phi}^{N} \end{array} \right] = \mathbf{T} \left[\begin{array}{ccc} \mathbf{Q}_{\theta}^{1} & \cdots & \mathbf{Q}_{\theta}^{N} \end{array} \right] \quad (1)$$

Then VTM method generates $\mathbf{Q}_{\phi}^{probe}$, which denotes the virtual frontal face image for the probe image, from the input nonfrontal probe face image $\mathbf{Q}_{\theta}^{probe}$as follows:

$$\mathbf{Q}_{\phi}^{probe} = \mathbf{T} \ \ \mathbf{Q}_{\theta}^{probe} \quad (2)$$

Faces of two persons might have similar parts although these faces are not totally similar. Thus transforming the input face image using the information of the entire face image of other individuals might degrade the characteristics of the input individual's face. In order to solve this problem, VTM was further extended in a local patch based way called Local View Transition Model(LVTM) [5], which achieves face pose transformation not by considering its three-dimensional structure, but by synthesizing a face image with a different pose from partial face image patches calculated from a large number of general individual's faces. That is to say, instead of transforming directly the entire face image, LVTM transform face patches that are partial images of a face image for each location in the face image.

Let $\mathbf{q}_{\phi(x,y)}, \mathbf{q}_{\theta(x,y)}$ represent face patches with patch center location at (x, y) of corresponding frontal and nonfrontal global face image plane $\mathbf{Q}_{\phi}, \mathbf{Q}_{\theta}$ respectively. LVTM learns the location specific linear transforms $\mathbf{T}_{(x,y)}$in a similar way with the VTM as follows,

$$\left[\begin{array}{ccc} \mathbf{q}_{\phi(x,y)}^{1} & \cdots & \mathbf{q}_{\phi(x,y)}^{N} \end{array} \right] \quad (3)$$
$$= \mathbf{T}_{(\mathbf{x},\mathbf{y})} \left[\begin{array}{ccc} \mathbf{q}_{\theta(x,y)}^{1} & \cdots & \mathbf{q}_{\theta(x,y)}^{N} \end{array} \right]$$

It should be noted that the LVTM method transforms each local area of an image while the VTM method transforms the entire area of an image. Then the virtual frontal appearances for each local patches can be generated as follows:

$$\mathbf{q}_{\phi(x,y)}^{probe} = \mathbf{T}_{(\mathbf{x},\mathbf{y})} \ \ \mathbf{q}_{\theta(x,y)}^{probe} \quad (4)$$

After this, the LVTM method synthesizes an output frontal face image $\mathbf{Q}_{\phi}^{probe}$ from all the transformed patches $\mathbf{q}_{\phi(x,y)}^{probe}$. The pixel values of regions where face patches are overlapped are calculated by averaging the pixel values of the overlapped patches, which is shown in Fig 2. Experimental results showed that LVTM can achieve smaller transformation residue error and higher recognition rate [5].

3. Frontal view generation using LVTM with local patches clustering(c-LVTM)

The key point of VTM-like methods is the underlying linear relationship in the frontal and nonfrontal face image pairs. LVTM improves the original VTM by learning linear transformations for each local patch, rather than a global one. Previous studies show that more accurate appearance transition model can be achieved by first dividing the original face image plane into overlapping local patch regions and then the learned transition models for each patch are aggregated for the final transformation. In this paper we show that the accuracy the appearance transition model and the recognition rate can be further improved by better exploiting the inherent linear relationship between frontal-nonfrontal face image patch pairs. This is achieved based on the observation that variations in appearance caused by pose are closely related to the corresponding 3D face structure and intuitively frontal-nonfrontal pairs from more similar local 3D face structures should have a stronger linear relationship. For each specific location, instead of learning a common transformation as in LVTM, the corresponding local patches

図 4 The difference of VTM, LVTM and the proposed c-LVTM. VTM learns a global linear mapping on the whole face image plane; LVTM learns location specific linear mapping for each local patch; c-LVTM learns linear mappings that are both location specific and local 3D structure specific.

are first clustered based on texture similarity distance metric and then the transition models are learned separately for each cluster. Fig 3 illustrates the observation using local patched for mouth location as an example. The original LVTM methods learns just a single common linear mapping using all the patch pairs for this specific location. Intuitively, those mouth patches with similar 3D shape (thus similar 2D appearance) should have more precise linear mapping relationship. In order to describe the relationship of frontal-nonfrontal pairs more precisely, it is better to first cluster patches first and then learn specific transformations for each cluster separately, just as Fig 3 illustrated.

More specifically, we first cluster the local patches $\mathbf{q}_{\theta(x,y)}$ for each location (x, y) into K clusters based on the appearance similarity where cluster k has c_k samples as $\{\mathbf{q}^1_{\theta(x,y)}, ..., \mathbf{q}^{c_k}_{\theta(x,y)}\}$. Then for each cluster, the corresponding linear transformation $\mathbf{T}^k_{(x,y)}$, which is both location specific and local 3D structure specific, is learned as follows,

$$\begin{bmatrix} \mathbf{q}^1_{\phi(x,y)} & \cdots & \mathbf{q}^{c_k}_{\phi(x,y)} \end{bmatrix} \quad (5)$$
$$= \mathbf{T}^k_{(x,y)} \begin{bmatrix} \mathbf{q}^1_{\theta(x,y)} & \cdots & \mathbf{q}^{c_k}_{\theta(x,y)} \end{bmatrix}, k = 1, ..., K$$

In the testing stage, each local patch for the input nonfrontal face image is transformed using the learned local view transition model corresponding to the most visually similar cluster, which is denoted as $k_{optimum}$.

$$\mathbf{q}^{probe}_{\phi(x,y)} = \mathbf{T}^{\mathbf{k_{optimum}}}_{(\mathbf{x},\mathbf{y})} \quad \mathbf{q}^{probe}_{\theta(x,y)} \quad (6)$$

And the final transformed global frontal face image is aggregated from $\mathbf{q}^{probe}_{\phi(x,y)}$ in a similar way as in LVTM. The difference in the linear mapping learning between VTM, LVTM and the proposed c-LVTM is illustrated as in Fig 4. VTM learns a global linear mapping on the whole face image plane. LVTM learns location specific linear mapping for each local patch. The proposed c-LVTM learns linear mappings that are both location specific and local 3D structure specific.

4. Experimental result

We used a subset of the face image dataset provided by SOFTPIA JAPAN [9] to demonstrate the effectiveness of the proposed method. The subset consists of 250 individuals take from frontal image and horizontal angles of 30 degree profile image. We compared the performance of direct eigen-face based recognition, View Transmission Model(VTM), local View Transmission Model(LVTM) and the proposed clustering based LVTM(c-LVTM) using 5-fold cross-validation. We transformed non-frontal face images to a frontal face image and then input the transformed images to a system that recognizes persons from the frontal face images using the traditional eigen-face algorithm. We first aligned all images by affine transformation using landmark correspondences. The image size was 32×32 pixels and the face patch size was set from 10×10 to 24×24 in pixels. The number of the cluster centers was set from 2 to 5. The clustering results at some example locations such as left eye, right eye and mouth are illustrated in Fig 5. The visual effects of transformed virtual frontal face images using different methods are illustrated in Fig 6.

It can be seen that the generated virtual frontal face image using the proposed c-LVTM method has higher fidelity than that of other methods. This trend is further demonstrated in the following face recognition rate comparison which is illustrated in Table 1. The recognition rate comparison results validate out assumption that learning both location specific and local 3D structure specific linear transforms can better capture the relationship between frontal and nonfrontal patch pairs than just learning a single common linear transformation.

5. Conclusion

In order to better exploit the underlying linear relationship between frontal and nonfrontal pairs, This paper presents a framework for face recognition across pose based on virtual frontal view generation using Local View Transition Model(LVTM) with local patches clustering. The proposed method further extends the LVTM by learning not only local patch specific transformations, but also local 3D structure specific linear transforms. For each local patched , rather than learning a single location specific linear mapping, we first clustering all the patches in this location into several clusters

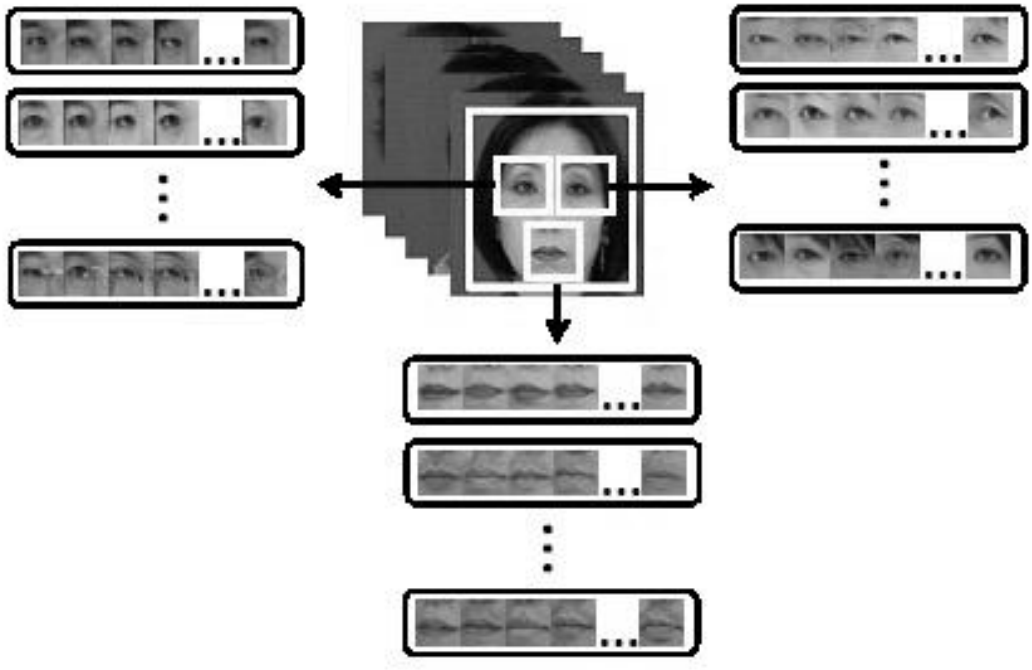

図 5　The clustering results at some example locations such as left eye, right eye and mouth

using the appearance similarity, then for each cluster we learn a local 3D structure specific linear transformation. For a probe patch at a specific location, we find its corresponding cluster based on appearance similarity and the use the corresponding linear transform to generate the frontal patch. Experimental results show the effectiveness of the proposed method.

Future works include 1) To investigate the effects of the number of clusters and the size of the local patches on the performance of virtual frontal face generation and the corresponding recognition rate; 2) To validate the performance of the proposed strategy on more face databases.

Recognition rate	**K=2**	**K=3**	**K=4**	**K=5**
Without transformation	37.6%			
VTM	73.6%			
LVTM(patch size = 10)	29.1%			
LVTM(patch size = 12)	43.6%			
LVTM(patch size = 14)	51.4%			
LVTM(patch size = 16)	60.0%			
LVTM(patch size = 20)	73.2%			
LVTM(patch size = 24)	76.3%			
c-LVTM(patch size = 10)	45.2%	39.5%	38.2%	35.1%
c-LVTM(patch size = 12)	56.4%	55.1%	49.7%	47.5%
c-LVTM(patch size = 14)	63.3%	59.8%	55.1%	51.4%
c-LVTM(patch size = 16)	74.8%	78.4%	**80.8%**	79.6%
c-LVTM(patch size = 20)	76.8%	78.0%	74.8%	73.6%
c-LVTM(patch size = 24)	79.2%	74.8%	70.8%	69.6%

表 1　The face recognition rate comparison.

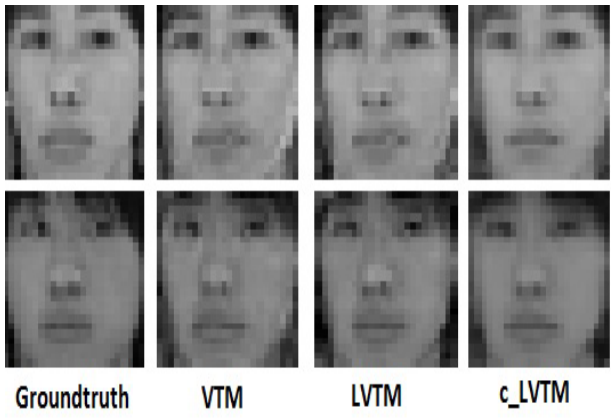

図 6　The visual effect of transformed virtual frontal face images using different methods

Acknowledgement

This work was supported by R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society, Special Coordination Fund for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

文　　献

[1] M. A. Turk and A. P. Pentland, Face recognition using eigenfaces, Proc. Computer Vision and Pattern Recognition, 1991

[2] V. G. Blanz and P. J. Phillips and T. Vetter, Face recognition based on frontal views generated from non-frontal images, Proc. Computer Vision and Pattern Recognition, 2005

[3] D. Beymer, Face recognition under varying pose, Proc. Computer Vision and Pattern Recognition, 1994

[4] A. Utsumi and N. Tetsutani, Adaptation of appearance model for human tracking using geometrical pixel value distribution, Proc. 6th Asian Conference on Computer Vision, 2004
[5] Y. Kono, T. Takahashi, D. Deguchi, I. Ide and H. Murase, Frontal face generation from multiple low-resolution non-frontal faces for face recognition, Proc. Asian Conference on Computer Vision, 2010
[6] S. Baker and T. Kanade, Hallucinating faces, Proc. Int. Conf. on Automatic Face and Gesture Recognition, 2000
[7] X. Chai and S. Shan and X. Chen and W. Gao, Locally linear regression for pose-invariant face recognition, IEEE Trans. Image Processing, pp.1716–1725, 2007
[8] D. Beymer and T. Poggio, Face recognition from one example view, Proc. 5th International Conference on Computer Vision, 1995
[9] SOFTPIA JAPAN face image databse, http://www.softpia.or.jp/rd/facedb.html