# Virtual View Generation using Clustering based Local View Transition Model

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

†Graduate School of Information Science,
Nagoya University, Japan
††Department of Economics and Information,
Gifu Shotoku Gakuen University, Japan.
††† Information and Communications Headquarters,
Nagoya University, Japan

**Abstract.** This paper presents an approach for realistic virtual view generation using appearance clustering based local view transition model, with its target application on cross-pose face recognition. Previously, the traditional global pattern based view transition model (VTM) method was extended to its local version called LVTM, which learns the linear transformation of pixel values between frontal and non-frontal image pairs using partial image in a small region for each location, rather than transforming the entire image pattern. In this paper, we show that the accuracy of 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. For each specific location, instead of learning a common transformation as in the 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 non-frontal probe image is transformed using the learned local view transition model corresponding to the most visually similar cluster. The experimental results on a real-world face dataset demonstrated the superiority of the proposed method in terms of recognition rate.

## 1 Introduction

Due to its wide range of potential real-life applications such as identity authentication, intelligent surveillance, human-computer interface and so on, face recognition has been one of the most active research topics in the biometric field within the computer vision and the pattern recognition communities [1]. Unlike other biometric techniques such as fingerprint recognition, palm print recognition or iris recognition, face recognition is inherently a passive and non-intrusive technique that has the advantage of not requiring cooperative subjects. That is to say, a practical face recognition system is supposed to have the ability to recognize the face of an uncooperative subject in an arbitrary situation and uncontrolled environment setting, even without the target subject noticing. This advantage of environment setting generality also poses great challenges to the problem of face recognition because as the viewing condition changes, the captured face appearances might vary too drastically to be easily identified. Within the past several decades, many methods have been proposed for face recognition. However, most of those traditional methods can successfully recognize faces only when face images are captured under constrained condition and controlled environment, for example recognize frontal faces with normal expressions and typical indoor illuminations, which are usually unrealistic in many real-life application scenarios. Usually the performance of these traditional methods will degrade greatly when face images are captured in unconstrained conditions caused by factors such as varying viewpoints, illumination changes, occlusions, aging, expressions 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-world face recognition application scenarios such as surveillance systems, where the subjects are either indifferent or uncooperative, so the captured face images are usually low-resolution and non-frontal. Pose variation has been identified as one of the prominent difficult problems in the research of face recognition [1]. The major difficulty of the cross-pose face recognition is that the intra-person appearance differences caused by rotation are often larger than the inter-person differences. That is to say, the distance between appearance vectors of two faces of different persons under similar viewpoints is much smaller than that of the same person under different viewpoints. This phenomenon makes the traditional face recognition methods such as eigen-face [2] or fisher-face [3] infeasible. Obviously one straightforward method for cross-pose face recognition is to actively compensate pose variations by providing gallery views in each rotation angles to recognize rotated non-frontal probe views. This can be achieved by first collecting and preparing multiple real-view templates beforehand for every known individual in each specific pose condition. Although the number of required real gallery images can be reduced by proper quantization on the rotation angles due to the fact that general face recognition algorithms are able to tolerate small pose variations to some extent, the tedious process of collecting multiple face images in different poses for real-view based matching is still unfavorable and even impractical in some cases. For example, in the application of airport security surveillance systems, there is only one frontal passport photo per person that could be collected and stored in the database.

Previously, both 3D model based methods [4][5] and 2D appearance based methods [6][7][8][9][10] have been proposed for pose invariant face recognition. 3D Morphable Model [4] is a typical 3D model based method for pose invariant face recognition. The 3D morphable model is built using the principal component analysis of the 3D facial shapes and textures obtained from laser scanner devices where inter-person pixel correspondences are established using the optical flow on the 3D surfaces. The 3D Morphable Model can then realize recognition either by transforming non-frontal face images to frontal view or by directly performing

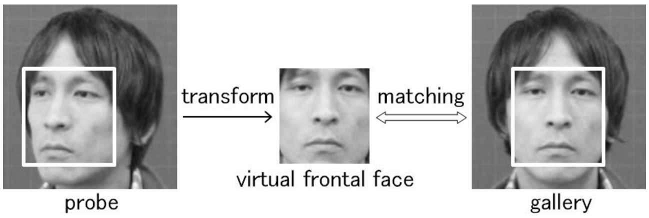


**Fig. 1.** Cross-pose face recognition based on virtual view generation.

the recognition by using the coefficients of the morphable model. But usually it is difficult to detect dense facial feature points that are accurate enough for the model fitting from low-resolution surveillance camera images.

Among the 2D appearance based methods, one of the successful approaches is to first generate a virtual frontal view by applying pose transformation on any given non-frontal face view. The View-Transition Model (VTM) [6] is a noteworthy work for pose transformation that can construct human appearance models for different poses which have proper texture information from a limited number of input images. The VTM method transforms views of an object between different poses by linear transformation of pixel values in images. For each pair of poses, a transformation matrix is calculated from image pairs of the poses of a large number of training data. The VTM was further extended to Local VTM (LVTM) in a patch-wise way [7] and it was shown that a more satisfactory face recognition result can be achieved using the virtual frontal face view generated by the local patch based LVTM than the original global patch based LVTM. This paper further extends the LVTM and presents a framework for face recognition across poses based on virtual frontal view generation using the LVTM with local patches clustering, which is denoted as c-LVTM hereafter. The proposed c-LVTM can describe the inherent transforming relationship between pixel values of patch pairs in a more precise way, thus more realistic virtual frontal face images can be generated and a higher recognition rate can be obtained. The experimental results on a real-world face dataset demonstrated the superiority of the proposed method.

The rest of this paper is organized as follows: in section 2, the underlying principle of the original VTM for pose transformation and the LVTM based face recognition methods are introduced briefly. Section 3 describes the proposed clustering based local VTM method (c-LVTM) in detail. Section 4 introduces the experimental result and section 5 is the summary.

## 2 Cross-pose face recognition by virtual frontal view generation

Instead of directly classifying the probe non-frontal face image, VTM or LVTM based cross-pose face recognition methods firstly synthesize a virtual frontal face view before a general face recognition procedure is applied, as shown in Fig. 1. Both the VTM and the LVTM methods use a general training image dataset consisting of faces of a large number of individuals viewed from both frontal and various profile angles. The linear transformations learned from the training dataset are applied to the probe non-frontal face images, either in a global way as in the VTM or in a local patch based way as in the LVTM, to generate the counterpart virtual frontal face image that is then fed into a general traditional face recognition engine.

More specifically, given a training multi-pose face image dataset $\varTheta$:$\{\mathbf{Q}_{\phi}^{1}, ..., \mathbf{Q}_{\phi}^{N}$ $, \mathbf{Q}_{\theta_1}^{1}, ..., \mathbf{Q}_{\theta_L}^{1}, ..., \mathbf{Q}_{\theta_1}^{N}, ..., \mathbf{Q}_{\theta_L}^{N}\}$, where $N$ is the number of training subjects, $\mathbf{Q}_{\phi}^{n}$, $(n = 1, ..., N)$ represents the frontal face image for the $n$-th subject as a vector 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 non-frontal face image for the $n$-th subject with the pose rotation angle $\theta_l$. For an input probe non-frontal face image $\mathbf{P}_{\theta_l}$, the purpose is to generate its virtual frontal image $\mathbf{P}_{\phi}$ using the linear transformation learned from the training dataset. The VTM can be applied for virtual frontal face generation 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 image only and assume that the training dataset consists of frontal-nonfrontal face image pairs with a single rotation degree $\theta$. The VTM calculates the linear transformation $\mathbf{T}$ beforehand using the training dataset by solving the following equation [6]:

$$\left[\mathbf{Q}_{\phi}^{1} \cdots \mathbf{Q}_{\phi}^{N}\right] = \mathbf{T}\left[\mathbf{Q}_{\theta}^{1} \cdots \mathbf{Q}_{\theta}^{N}\right] \tag{1}$$

Then the VTM generates $\mathbf{P}_{\phi}$, which denotes the virtual frontal face image for the probe image, from the input non-frontal probe face image $\mathbf{P}_{\theta}$ as follows:

$$\mathbf{P}_{\phi} = \mathbf{T}\mathbf{P}_{\theta} \tag{2}$$

Faces of two persons might have similar parts although these faces are not in total 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, the VTM was further extended in a local patch based way called Local View Transition Model (LVTM) [7], which achieves face pose transformation by synthesizing a face image from partial face image patches. That is to say, instead of transforming directly the entire global face image, the LVTM transforms face patches that are partial images of a face image for each location in the face image.

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

$$\left[\mathbf{q}_{\phi(x,y)}^{1} \cdots \mathbf{q}_{\phi(x,y)}^{N}\right] = \mathbf{T}_{(x,y)}\left[\mathbf{q}_{\theta(x,y)}^{1} \cdots \mathbf{q}_{\theta(x,y)}^{N}\right] \tag{3}$$

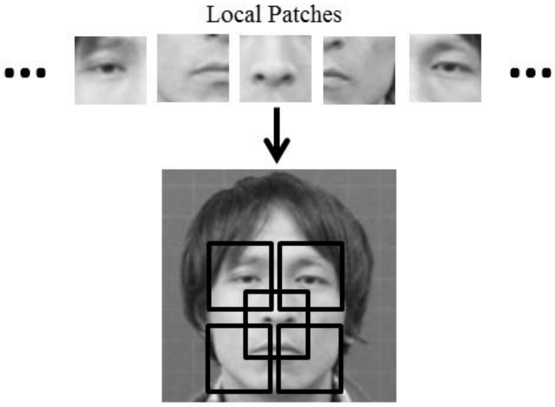
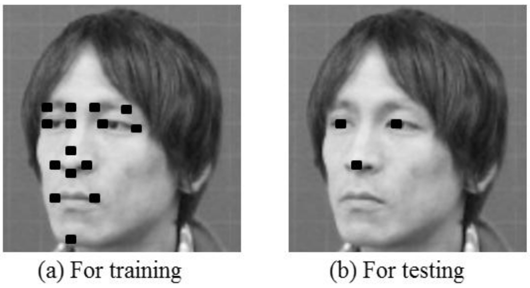


**Fig. 2.** Face image synthesis by local patches aggregation.

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

$$\mathbf{p}_{\phi(x,y)} = \mathbf{T}_{(x,y)}\, \mathbf{p}_{\theta(x,y)} \tag{4}$$

After this, the LVTM synthesizes an output frontal face image $\mathbf{P}_{\phi}$ from all the transformed local patches $\mathbf{p}_{\phi(x,y)}$. The pixel values of regions where face patches are overlapped are calculated by averaging the pixel values of the overlapped patches as illustrated in Fig. 2. Experimental results showed that the LVTM can


(a) For training

(b) For testing

**Fig. 3.** Affine alignment using landmarks. Different strategies are used for training and testing stages. (a) In the training stage, in order to learn the linear transformations more accurately, the face images are finely affine aligned using multiple (15) landmarks labeled manually. (b) While in the testing stage, the input probe face image is only roughly affine aligned using three landmarks (left eye, right eye, and nose tip), which can be easily detected by any off-the-shelf facial point detectors.

achieve a higher recognition rate than that of using VTM for pose transformation [7].

## 3 Virtual view generation using clustering based LVTM (c-LVTM)

The key point of VTM-like methods is the underlying linear relationship in the frontal and non-frontal face image pairs. Next we will show that the accuracy of 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 structure, and intuitively, frontal-nonfrontal patch pairs from more similar local 3D face structures should have a stronger linear relationship. Thus for each specific location, instead of learning a common transformation as in the LVTM, in the proposed c-LVTM, the corresponding local patches are first clustered based on the appearance similarity distance metric and then the transition models are learned separately for each cluster. We assume that those patches with similar 3D shapes and thus similar 2D appearances should have a more precise linear mapping relationship. For the purpose of describing the relationship of frontal-nonfrontal pairs more precisely, it is better to learn the transformations specific for each cluster separately, rather than learning just a single common linear mapping using all the patch pairs for a specific location. As Fig. 3(a) shows, in order to learn the linear transformations in a precise way, the training face image pairs are finely affine aligned using multiple landmarks. However in the testing stage, as Fig. 3(b) shows, the input probe face image is only roughly affine aligned using three landmarks (left eye, right eye, and nose tip), which can be easily detected by any off-the-shelf facial point detectors.

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 using the Normalized Cross-Correlation score[11], 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,

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

In the testing stage, the probe non-frontal face image is first roughly affine aligned, for example using only three landmark points at left eye, right eye and mouth, which can be easily obtained using any standard facial feature point detector. Then for each local patch of the input non-frontal face image $\mathbf{p}_{\theta(x,y)}$, the most visually similar cluster in the training set is searched in the neighborhood regions $([x-\epsilon, x+\epsilon], [y-\epsilon, y+\epsilon])$ space of a specific location $(x, y)$. If we denote

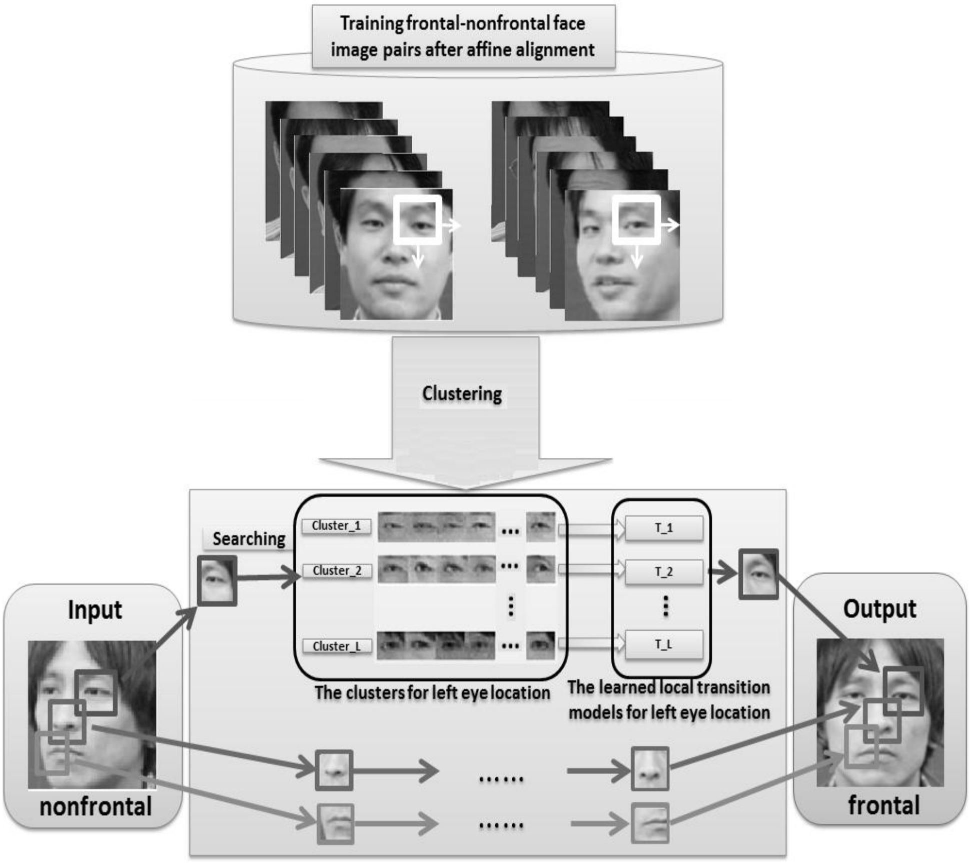

Training frontal-nonfrontal face image pairs after affine alignment
Clustering
Searching
Cluster_1
Cluster_2
Cluster_L
T_1
T_2
T_L
The clusters for left eye location
The learned local transition models for left eye location
Input
nonfrontal
Output
frontal
......
......

**Fig. 4.** The illustration of main steps of the proposed c-LVTM method. The steps of the appearance clustering based local transition models computation and the optimum transition model searching are depicted by taking the local patches located on the left eye as an example. First, the local patches location on the left eye are clustered into clusters of **cluster_1**, **cluster_2**,..., **cluster_L** based on appearance similarity. Then for each cluster, the local transition models **T_1**, **T_2**,..., **T_L** are computed using the corresponding local patches. Then for left eye local patch of the input nonfrontal face image, the most visually similar clusters in the training set is searched in the neighborhood regions and local transition model corresponding to the most visually similar patch found is used to perform the transformation. The final transformed global frontal face image is the aggregation of all transformed local patches where the pixel values of the overlapped patches are averaged.

the most visually similar patch found resides in the $k_{\mathrm{opt}}$-th cluster of location $(x_{\mathrm{opt}}, y_{\mathrm{opt}})$, then

$$\mathbf{p}_{\phi(x,y)} = \mathbf{T}^{k_{\mathrm{opt}}}_{(x_{\mathrm{opt}}, y_{\mathrm{opt}})}\, \mathbf{p}_{\theta(x,y)} \tag{6}$$

The final transformed global frontal face image is aggregated from $\mathbf{p}_{\phi(x,y)}$ in a similar way as in the LVTM. The main idea of the appearance clustering based local transition models computation and the optimum transition model searching is illustrated in detail in Fig. 4 and the flowchart of the proposed c-LVTM is described in Fig. 5.

The differences between the VTM, the LVTM and the proposed c-LVTM are illustrated in Fig. 6. The VTM learns a global linear mapping on the holistic face image plane. The 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.

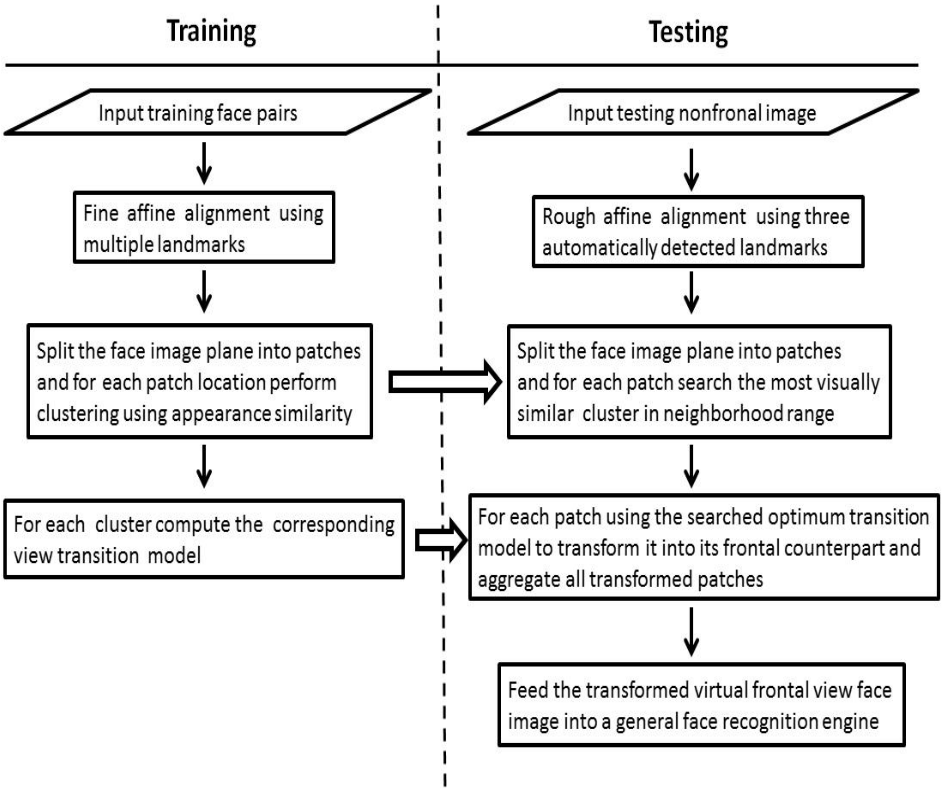

Training
Testing
Input training face pairs
Fine affine alignment using multiple landmarks
Split the face image plane into patches and for each patch location perform clustering using appearance similarity
For each cluster compute the corresponding view transition model
Input testing nonfronal image
Rough affine alignment using three automatically detected landmarks
Split the face image plane into patches and for each patch search the most visually similar cluster in neighborhood range
For each patch using the searched optimum transition model to transform it into its frontal counterpart and aggregate all transformed patches
Feed the transformed virtual frontal view face image into a general face recognition engine

**Fig. 5.** The flowchart of the proposed c-LVTM method.

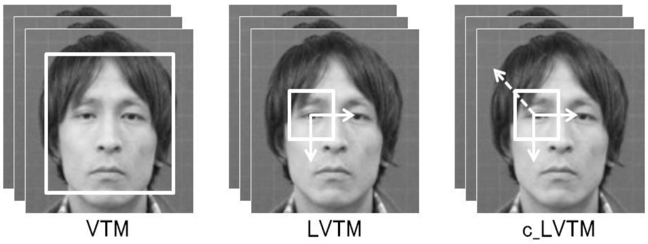


**Fig. 6.** The difference in how the image patterns are selected for the transition model computation between VTM, LVTM and the proposed c-LVTM method.

## 4 Experiment

We used a subset of the face image dataset provided by SOFTPIA JAPAN to demonstrate the effectiveness of the proposed method. The subset consists of 250 individuals' images. They were taken with horizontal angles varying from $-30$ degrees to 30 degrees at 10 degrees interval as shown in Fig. 7. We compared the performance of using input images directly, the VTM, the LVTM and the proposed c-LVTM by 5-fold cross-validation. We transformed non-frontal face images to virtual frontal face images and then input the transformed images to a system that recognizes persons from the virtual frontal face images using a common subspace based face recognition algorithm, where the subspace for each face image was spanned by the slide window shifting extended sample set. The training images were precisely affine aligned using 15 landmark points and the testing images were roughly aligned using only 3 landmark points at left eye, right eye and mouth.

The image size was $32 \times 32$ in pixels and the face patch size was set to be $16 \times 16$ in pixels. The number of the cluster centers $K$ was set to 4. The region of neighborhood searching $\epsilon$ was set to 5. The visual effects of the transformed virtual frontal face images using different methods are illustrated in Fig. 8. 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 Fig. 9. The recognition rate of the straightforward baseline method that using the non-frontal face images directly as input is much lower than that

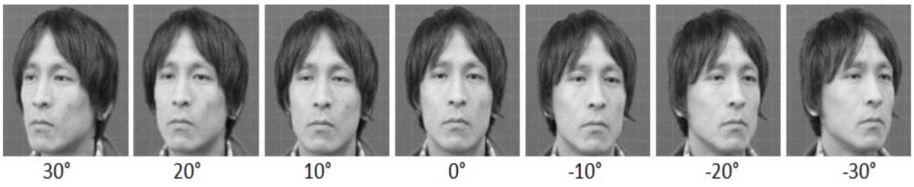


**Fig. 7.** The sample images of the multiple pose faces.

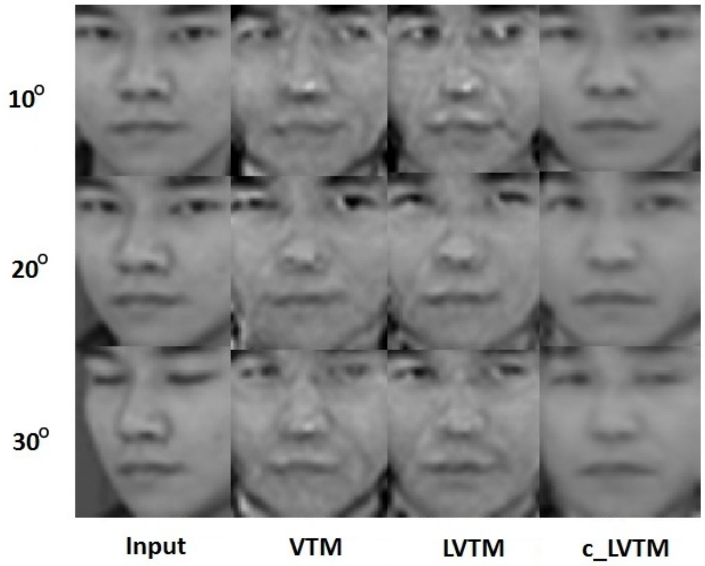


**Fig. 8.** The comparison of the visual effect of the transformed virtual frontal face image using different methods. It can be clearly seen that the virtual frontal face images generated using the proposed c-LVTM method have the highest visual fidelity.

of using the virtually generated frontal face images as input, either using VTM, LVTM or the proposed c-LVTM. Furthermore, the recognition performance of the proposed c-LVTM outperforms the VTM and LVTM in two ways: 1) c-LVTM has a higher recognition rate than VTM and LVTM; 2) Though all methods have a rate decreasing trend as the pose angle increases, the proposed c-LVTM has a more robust property against pose angle degree. That is to say, as pose angle increases, the curve of rate-vs-angle for c-LVTM drops less drastically than that of VTM and LVTM. The recognition rate comparison results validate our assumption that learning both location specific and local 3D structure specific linear transforms can better capture the relationship between frontal and non-frontal patch pairs than just learning a single common linear transformation.

## 5 Summary

In order to better exploit the underlying linear relationship between frontal and non-frontal pairs, this paper presented a framework for face recognition across pose based on virtual frontal view generation using the Local View Transition Model (LVTM) with local patches clustering. The proposed method further extended the LVTM by learning not only the local patch position specific transformations, but also the local 3D structure specific linear transforms. Experimental

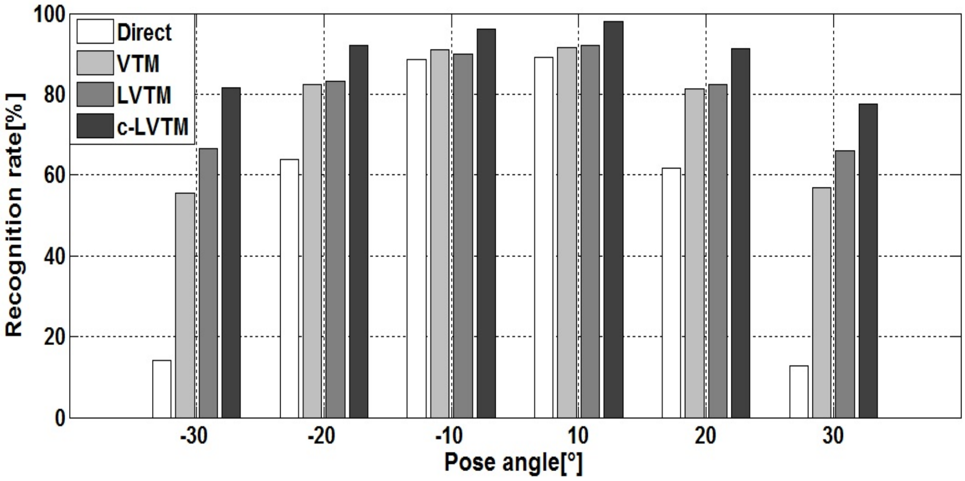


**Fig. 9.** Comparison of recognition rates across different angles. The input non-frontal face images are transformed using the VTM, LVTM and the proposed c-LVTM, respectively. The rate for the straightforward method of using the input non-frontal face images directly is also included for comparison.

results showed the effectiveness of the proposed method. Although the main focus of this paper is on the problem of face recognition, the proposed framework for realistic virtual view generation is quite general. In the future, we would like to further investigate its performance evaluation not only on other more facial datasets, but also on databases in other domains such as multi-view object recognition or view invariant person identification using body images.

## 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. The face image dataset used in this work was provided by SOFTPIA JAPAN.

## References

1. Zhao, W., Chellappa, R., Philips, P.J., Rosenfeld, A.: Face recognition: A literature survey (2003) ACM Computer Survey, vol. 35, no. 4, pp. 399–459.
2. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces (1991) Proc. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–591.

3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection (1997) IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no.7, pp. 711–720.
4. Blanz, V.G., Phillips, P.J., Vetter, T.: Face recognition based on frontal views generated from non-frontal images (2005) Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 454–461.
5. Beymer, D.: Face recognition under varying pose (1994) Proc. 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 756–761.
6. Utsumi, A., Tetsutani, N.: Adaptation of appearance model for human tracking using geometrical pixel value distribution (2004) Proc. 6th Asian Conference on Computer Vision, pp. 794–799.
7. Kono, Y., Takahashi, T., Deguchi, D., Ide, I., Murase, H.: Frontal face generation from multiple low-resolution non-frontal faces for face recognition (2010) Proc. 10th Asian Conference on Computer Vision, pp.175–183.
8. Baker, S., Kanade, T.: Hallucinating faces (2000) Proc. 2000 IEEE Conference on Automatic Face and Gesture Recognition, pp.83–88.
9. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition (2007) IEEE Transactions on Image Processing, vol. 16, no. 7, pp.1716–1725.
10. Beymer, D., Poggio, T.: Face recognition from one example view (1995) Proc. 5th International Conference on Computer Vision, pp.500–507.
11. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited (2006) Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.2402–2409.