LANDMARK-BASED FISHER VECTOR REPRESENTATION FOR VIDEO-BASED FACE VERIFICATION

Jun-Cheng Chen, Vishal M. Patel and Rama Chellappa

Center for Automation Research, University of Maryland, College Park, MD 20742

{pullpull, pvishalm, rama}@umiacs.umd.edu

ABSTRACT

Unconstrained video-based face verification is a challenging problem because of dramatic variations in pose, illumination, and image quality of each face in a video. In this paper, we propose a landmark-based Fisher vector representation for video-to-video face verification. The proposed representation encodes dense multi-scale SIFT features extracted from patches centered at detected facial landmarks, and face similarity is computed with the distance measure learned from joint Bayesian metric learning. Experimental results demonstrate that our approach achieves significantly better performance than other competitive video-based face verification algorithms on two challenging unconstrained video face datasets, Multiple Biometric Grand Challenge (MBGC) and Face and Ocular Challenge Series (FOCS).

Index Terms— face verification, facial landmarks, Fisher vector

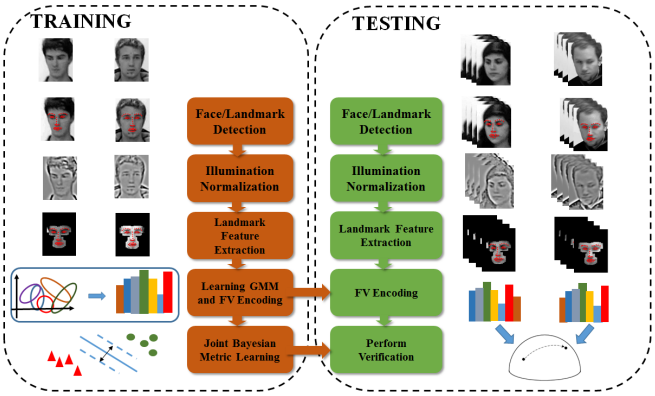


Fig. 1: An overview for our landmark-based Fisher vector video-based face verification algorithm.

1. INTRODUCTION

Face recognition is one of the active research areas in computer vision and has a wide range of practical applications including surveillance, social networks, and mobile platform authentication [1]. However, unconstrained face recognition is still a challenging open problem due to large variations in pose, lighting, blur, expression, and occlusion.

In general, face recognition can be broadly classified into three tasks: identification, verification, and watchlist. In this work, we mainly focus on unconstrained video-to-video face verification in which the goal is to determine whether two face videos belong to the same person or not. To handle large variations in pose, expression and illumination, extracting invariant and discriminative representation from face images/videos is an important issue. Chen *et al.* [2] have shown that the high-dimensional multi-scale Local Binary Pattern (LBP) descriptors extracted from local patches centered at each facial landmarks have strong discriminative power for the still-face recognition problem. However, directly applying this idea to videos is infeasible because of the high dimensionality of videos. On the other hand, the Fisher Vector (FV) representation is one of many bag-of-visual-word encoding methods, originally proposed for object recognition problem and subsequently shown to work well for face verification problems [3][4]. Even though FV descriptors are compact for videos, their dimension is still high and increases linearly with the number of components in the Gaussian Mixture Model (GMM). More components in GMM representation usually allow FVs to encode more discriminative information from image and video data. However, having many mixture components may be impractical for large face databases. Motivated by the successes of these two approaches, we propose a landmark-based FV representation for video-based face verification. Instead of learning the mixture model from the dense features of the whole face, we fit a Gaussian model for each landmark with multi-scale dense features extracted from patches centered at each landmark. In this way, we can greatly reduce the number of mixture components and the dimensionality of the FVs while preserving sufficient discriminative power.

The rest of the paper is organized as follows: We briefly review some related works in Section II. In Section III, we present the details of the training and testing algorithms of our landmark-based FV representation and present experimental results on two challenging video datasets in Section IV. We concludes the paper in Section V with a brief summary and discussion.

2. RELATED WORK

In this section, we briefly review several related works on video-based face verification as follows. Generally, there are two major components of a face verification system: (1) robust feature representation and (2) designing a similarity measure.

Learning invariant and discriminative representation is the first

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

step towards realizing a successful face verification system. Ahonen *et al.* [5] showed that LBP is effective for face recognition. In addition, Gabor wavelets [6] also have been widely used to encode multi-scale and multi-orientation information for face images. On the other hand, Coates *et al.* [7] showed that an over-complete representation is critical to achieving high recognition rates regardless of the encoding methods used. For still-face recognition, Chen *et al.* [2] demonstrated excellent results using the high-dimensional multi-scale LBP features extracted from facial landmarks. These works showed that over-complete and high-dimensional features are important for face recognition. Li *et al.* [8] proposed a probabilistic elastic model for face recognition by learning a GMM using dense local spatial-appearance features and selecting sparse representative features for each Gaussian component. On the other hand, Simonyan *et al.* [3] and Parkhi *et al.* [4] showed that FV, a feature encoding method widely used for object and image classification, can be successfully applied to face recognition. Their experiments showed that FV can effectively encode over-complete and dense features yielding a robust representation.

Designing the similarity measure is the other key component in a face verification system. Guillaumin *et al.* [9] proposed two robust distance measures: Logistic Discriminant-based Metric Learning (LDML) and Marginalized kNN (MkNN). The LDML method learns a distance by performing logistic discriminant analysis on a set of labeled image pairs, and the MkNN method marginalizes a k-nearest-neighbor classifier to both images of the given test pair using a set of labeled training images. Taigman *et al.* [10] learned the Mahalanobis distance for face verification using the Information Theoretic Metric Learning (ITML) method proposed in [11]. Wolf *et al.* [12] proposed the one-shot similarity (OSS) kernel based on a set of pre-selected reference images mutually exclusive to the pair of images being compared. Chen *et al.* [13] proposed a joint Bayesian approach which models the joint distribution of a pair of face images instead of modeling the difference vector from them.

Our approach mainly takes advantage of the discriminative power of landmark features and effectively encodes the FVs to perform face verification. Furthermore, we apply the joint Bayesian metric learning to learn the projection matrices to reduce the feature dimensionality for efficiency and improve discriminative performance.

3. PROPOSED APPROACH

Our method can be divided into two stages: training and testing stages. For training, we use the well-known "Label Face in the Wild" (LFW) dataset [14]. First, we apply preprocessing steps to detect faces, facial landmarks and to normalize the face images/videos. Then, we extract multi-scale dense SIFT features around each landmark and learn a Gaussian model for each landmark using the mean and diagonal sample covariance of the features. After feature extraction, we perform the FV encoding and train a similarity measure using the augmented face pairs (i.e. we generate positive and negative pairs using the identity information available in the unrestricted setting of LFW). For testing, we use the learned metric on our proposed feature representation to compute the similarity of each test pair of the face images/videos. Fig. 1 presents an overview of our method. In the following subsections, we describe in detail each step used in training and testing stages.

Fig. 2: The first row shows the original image before preprocessing. The second row is the image after illumination normalization. The final row demonstrates the facial landmarks and patches used in this work.

3.1. Preprocessing

Before feature extraction and metric learning, we perform the following preprocessing steps to normalize the face data:

Landmark detection: We perform landmark detection for face alignment and for landmark-based feature representation. The approaches proposed in [15] and subsequent work [16] are adopted because of their computational efficiency and excellent performance on low-resolution and lower-quality face images/videos. We use the detected landmarks to align each face into the canonical coordinates using similarity transform. After alignment, the face image resolution is 63×80 pixels, and the distance between centers of two eyes is about 10 pixels.

Illumination normalization: Local block-wise illumination normalization approaches, such as self-quotient image (SQI) [17] which divides each pixel value by the weighted average of its neighborhood, have shown better illumination normalization performance for face recognition than histogram equalization which enhances the dynamic range by adjusting the intensity distribution of the entire image. Therefore, we adopt the SQI approach proposed by Tan *et al.* [18] which takes the Gamma correction, difference of Gaussian filtering, masking, and contrast equalization into consideration for image normalization. The normalization results are presented in Fig. 2.

3.2. Landmark-based Fisher vector face representation

In this subsection, we show how to extract the proposed landmark-based FV face representation (LFRV) and to apply metric learning on the extracted representation to compute the face similarity of a pair of face images/videos.

Fisher vector encoding: The FV is one of bag-of-visual-word encoding methods which aggregates a large set of feature vectors (e.g. dense SIFT features) into a high-dimensional vector. In general, the whole process is done by fitting a parametric generative model (e.g. GMM) and encoding the features using the derivatives of the log-likelihood of the learned model with respect to the model parameters. As in [19], a GMM model with diagonal covariances is used in this work, and the FV encoding can be thus computed by the

first- and second-order statistics of the dense features as follows

$$\mathbf{\Phi}_{ik}^{(1)} = \frac{1}{N\sqrt{w_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{\mathbf{v}_{ip} - \boldsymbol{\mu}_{ik}}{\boldsymbol{\sigma}_{ik}} \right) \quad (1)$$

$$\mathbf{\Phi}_{ik}^{(2)} = \frac{1}{N\sqrt{2w_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{(\mathbf{v}_{ip} - \boldsymbol{\mu}_{ik})^2}{\boldsymbol{\sigma}_{ik}^2} - 1 \right), \quad (2)$$

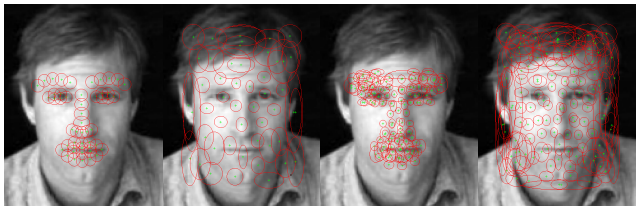
where w_k, $\boldsymbol{\mu}_k$, $\mathbf{\Sigma}_k = diag(\boldsymbol{\sigma}_{1k}, ..., \boldsymbol{\sigma}_{dk})$ are the weight, mean, and diagonal covariance of the kth Gaussian component of the GMM. Here, $\mathbf{v}_p \in \mathbb{R}^{d \times 1}$ is the pth feature vector and N is the number of feature vectors. Parameters of the GMM can be estimated from the training data using the EM algorithm. $\alpha_p(k)$ is the weight of $\mathbf{v}_p$ belonging to the kth mixture component. In addition, the final FV, $\mathbf{\Phi}(\mathbf{I})$, of an image $\mathbf{I}$ is obtained by concatenating all the $\mathbf{\Phi}_k^{(1)}$ and $\mathbf{\Phi}_k^{(2)}$s into a high-dimensional vector $\mathbf{\Phi}(\mathbf{I}) = [\mathbf{\Phi}_1^{(1)}, \mathbf{\Phi}_1^{(2)}, ..., \mathbf{\Phi}_K^{(1)}, \mathbf{\Phi}_K^{(2)}]$, whose dimensionality is $2Kd$ where K is the number of Gaussians in the GMM and d is the dimensionality of the extracted features.

To incorporate spatial information, we augment each extracted SIFT feature with the normalized x and y coordinates [8][3] as $[\mathbf{a}_{xy}, \frac{x}{w} - \frac{1}{2}, \frac{y}{h} - \frac{1}{2}]^T$ where $\mathbf{a}_{xy}$ is the SIFT descriptor at (x, y), and w and h are the width and height of the image, respectively. (i.e. For K, we use 49 and 128 in this work. For d, it is 130 after augmentation.) In addition, FV is further processed with signed square-rooting and L_2 normalization as suggested in [19] for improved performance.

Dense landmark features extraction: We extract dense root-SIFT features at three scales from the 16 $\times$ 16-pixel patches centered at each facial landmark of inner faces with a scaling factor of $\sqrt{2}$ (i.e., 49 landmarks are used here). For training, we aggregate the extracted features around each landmark and take the mean and diagonal sample covariance, $\mathbf{\Sigma}_k = diag(\boldsymbol{\sigma}_{1k}, ..., \boldsymbol{\sigma}_{dk})$, to fit a Gaussian for each landmark as follows:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{p=1}^{N_k} \mathbf{v}_p, w_k = \frac{1}{K}, \quad \boldsymbol{\sigma}_{ik} = \frac{1}{N_k - 1} \sum_{p=1}^{N_k} (\mathbf{v}_{ip} - \boldsymbol{\mu}_{ik})^2,$$

where N_k and $\mathbf{v}_p$ are respectively the number of features and SIFT features extracted from the patch centered at kth landmark. The fitted Gaussians are illustrated in Fig. 3.



(a) (b) (c) (d)

Fig. 3: (a) and (b) illustrate the GMM with 49 components learned from 49 facial landmarks and from the whole image, respectively. (c) and (d) show the GMM with 128 components learned from the neighborhood regions of 49 facial landmarks using EM algorithm and learned from the entire image respectively.

For testing, we aggregate the extracted features with augmented spatial information into a feature matrix, $\mathbf{F} \in \mathbb{R}^{130 \times N_F}$ for each frame, where N_F is the total number of aggregated features. Because some patches overlaps, we take the union of them to remove the duplicate features. Detected landmarks and patches for feature extraction are shown in Fig. 2. Then, we perform FV encoding for each frame within a video and average all the FVs into one for each video. (i.e. the other choice is to use pooling.)

3.3. Joint Bayesian Metric Learning

Recently, the joint Bayesian method to face metric learning has shown good performance for face verification [13][20]. Instead of modeling the difference vector between two faces, the approach directly models the joint distribution of feature vectors of both ith and jth images, $\{\mathbf{x}_i, \mathbf{x}_j\}$, as a Gaussian. Let $P(\mathbf{x}_i, \mathbf{x}_j | H_I) \sim N(0, \mathbf{\Sigma}_I)$ when $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class, and $P(\mathbf{x}_i, \mathbf{x}_j | H_E) \sim N(0, \mathbf{\Sigma}_E)$ when they are from different classes. In addition, each face vector can be modeled as, $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu}$ stands for the identity and $\boldsymbol{\epsilon}$ for pose, illumination, and other variations. Both $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$ are assumed to be independent zero-mean Gaussian distributions, $N(0, \mathbf{S}_\mu)$ and $N(0, \mathbf{S}_\epsilon)$, respectively. It was shown in [13] that the log likelihood ratio of intra- and inter-classes, $r(\mathbf{x}_i, \mathbf{x}_j)$, which has a closed-form solution can be computed as follows:

$$r(\mathbf{x}_i, \mathbf{x}_j) = log \frac{P(\mathbf{x}_i, \mathbf{x}_j | H_I)}{P(\mathbf{x}_i, \mathbf{x}_j | H_E)} = \mathbf{x}_i^T \mathbf{M} \mathbf{x}_i + \mathbf{x}_j^T \mathbf{M} \mathbf{x}_j - 2\mathbf{x}_i^T \mathbf{R} \mathbf{x}_j \quad (3)$$

where $\mathbf{M}$ and $\mathbf{R}$ are negatively semi-definite matrices of $\mathbf{S}_\mu$ and $\mathbf{S}_\epsilon$. The equation can be written as $(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) - 2\mathbf{x}_i^T (\mathbf{R} - \mathbf{M}) \mathbf{x}_j$. Instead of using the EM algorithm to estimate $\mathbf{S}_\mu$ and $\mathbf{S}_\epsilon$, we optimize the closed-form distance in a large-margin framework with hinge loss. However, directly learning $\mathbf{M} \in \mathbb{R}^{D \times D}$ and $\mathbf{R} \in \mathbb{R}^{D \times D}$ are intractable because of the high dimensionality of FVs where $D = 2Kd$. Thus, we let $\mathbf{M} = \mathbf{H}^T \mathbf{H}$ and $\mathbf{B} = (\mathbf{R} - \mathbf{M}) = \mathbf{V}^T \mathbf{V}$ where $\mathbf{H} \in \mathbb{R}^{r \times D}$ and $\mathbf{V} \in \mathbb{R}^{r \times D}$ and choose $r = 128 \ll D$ in our work. We solve the following optimization problem

$$\underset{\mathbf{H}, \mathbf{V}, b}{\operatorname{argmin}} \sum_{i,j} \quad max[1 - y_{ij}(b - (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^T \mathbf{H} (\mathbf{x}_i - \mathbf{x}_j) + 2\mathbf{x}_i^T \mathbf{V}^T \mathbf{V} \mathbf{x}_j), 0] \quad (4)$$

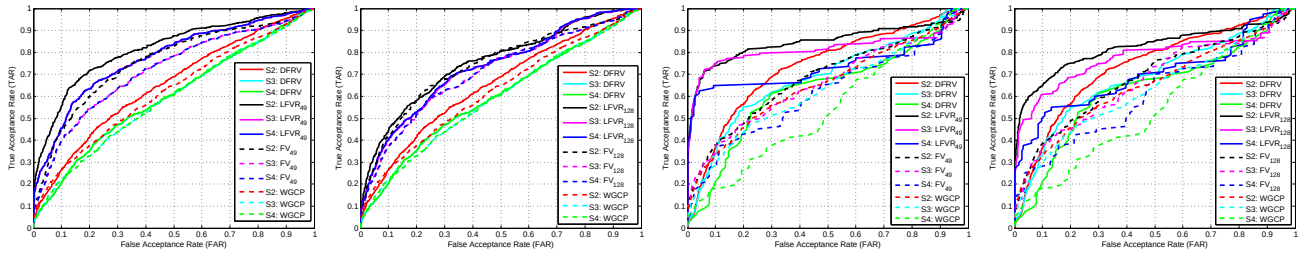
where $b \in \mathbb{R}$ is a threshold, and y_{ij} is the label of a pair: $y_{ij} = 1$ if person i and j are the same and $y_{ij} = -1$, otherwise. For simplification, we denote $(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^T \mathbf{H} (\mathbf{x}_i - \mathbf{x}_j) - 2\mathbf{x}_i^T \mathbf{V}^T \mathbf{V} \mathbf{x}_j$ as $d_{\mathbf{H},\mathbf{V}}(\mathbf{x}_i, \mathbf{x}_j)$. In addition, $\mathbf{H}$, $\mathbf{V}$, and b can be updated using a stochastic gradient descent algorithm as follows and are equally trained on positive and negative pairs in turn:

$$\mathbf{H}_{t+1} = \begin{cases} \mathbf{H}_t, & \text{if } y_{ij}(b_t - d_{\mathbf{H},\mathbf{V}}(\mathbf{x}_i, \mathbf{x}_j)) > 1 \\ \mathbf{H}_t - \gamma y_{ij} \mathbf{H}_t \mathbf{\Psi}_{ij}, & \text{otherwise,} \end{cases}$$

$$\mathbf{V}_{t+1} = \begin{cases} \mathbf{V}_t, & \text{if } y_{ij}(b_t - d_{\mathbf{H},\mathbf{V}}(\mathbf{x}_i, \mathbf{x}_j)) > 1 \\ \mathbf{V}_t + \gamma y_{ij} \mathbf{V}_t \mathbf{\Gamma}_{ij}, & \text{otherwise,} \end{cases}$$

$$b_{t+1} = \begin{cases} b_t, & \text{if } y_{ij}(b_t - d_{\mathbf{H},\mathbf{V}}(\mathbf{x}_i, \mathbf{x}_j)) > 1 \\ b_t + \gamma_b y_{ij}, & \text{otherwise,} \end{cases} \quad (5)$$

where $\mathbf{\Psi}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$, $\mathbf{\Gamma}_{ij} = \mathbf{x}_i \mathbf{x}_j^T + \mathbf{x}_j \mathbf{x}_i^T$, and γ is the learning rate for $\mathbf{H}$ and $\mathbf{V}$, and γ_b for the bias b. We perform whitening PCA to the extracted features and initialize both $\mathbf{H}$ and $\mathbf{V}$ with r largest eigenvectors. Note that $\mathbf{H}$ and $\mathbf{V}$ are updated only when the constraints are violated.



(a) (b) (c) (d)

Fig. 4: (a) and (b) show the ROC curves of face verification for subsets of S2, S3, and S4 for MBGC dataset where target and query videos are from the same set. (c) and (d) for the FOCS dataset. For these figures, we compare the results of LFVR of 49 (i.e. in (a)(c)) and 128 (i.e. in (b)(d)) components with DFRV and their FV counterparts using the same number of components respectively.

4. EXPERIMENTAL RESULTS

We present face verification results using the receiver operating characteristic (ROC) curves on two well-known public datasets for unconstrained video-based face recognition: (1) Multiple Biometric Grand Challenge (MBGC)[21], and (2) Face and Ocular Challenge Series (FOCS)[22].

4.1. Multiple Biometric Grand Challenge

In the MBGC dataset, there are 146 subjects in total, and videos are available in two resolutions: standard definition (SD, 720 $\times$ 480 pixels) and high definition (HD, 1440 $\times$ 1080 pixels). It consists of 399 walking sequences where 201 of them are in SD and 198 in HD, and 371 activity sequences where 185 in SD and 186 in HD. Fig. 5 shows the sample frames For the walking sequences, subjects usually walk toward and keep their faces facing the camera for most of the time and turn their faces sideways at the end. The main challenge of the dataset comes from blur caused by motion, frontal and non-frontal faces with shadows which also lead to difficulty in tracking the faces in the video. To the best of our ability, we implemented the dictionary-based method for video-based face recognition, DFRV, proposed in [23] and the manifold-based method, WGCP, proposed in [24]. These methods produced favorable results compared to several manifold and image set-based methods. As a result, we use them as the baseline algorithms. We perform the verification experiments on the subsets of S2, S3, and S4 from the walking sequences where S2 is the set of subjects who have at least two face videos available, S3 at least three available, and S4 at least four available (S2: 144 subjects, 397 videos in total, S3: 55 subjects, 219 videos in total, and S4: 54 subjects, 216 videos). The verification results are shown in Fig. 4 and Table 1. It can be seen from this figure that the proposed approach achieves better results than the one based on FV with the same number of components as our LFRV method, the DFRV and WGCP methods. The results essentially demonstrate the effectiveness of dense multi-scale facial landmark features.

4.2. Face and Ocular Challenge Series

In addition to the MBGC dataset, we tested our approach on another challenging dataset, FOCS. The FOCS UT-Dallas dataset contains 510 walking and 506 activity video sequences for 295 subjects with the resolution, 720 $\times$ 480 pixels. The sample frames are shown in Fig. 5. The sequences were acquired on different days. For the walking sequences, subjects initially stand far away from the camera, and then walk toward and keep their faces facing the camera keeping their face and then turn away at the end. We conducted the same verification tests as we did for MBGC subsets: S2 (189 subjects, 404 videos), S3 (19 subjects, 64 videos), and S4 (6 subjects, 25 videos) for UT-Dallas walking videos. The verification results are shown in Fig. 4 and Table 1. As in the MBGC case, the FOCS results also show that our proposed LFRV works more effectively than FV whose GMM is trained over the entire face. However, we can find from results of both MBGC and FOCS that the performance of LFVR_{128} is worse than LFVR_{49}. One possible reason is that the resolution of detected faces of these two datasets is smaller than the PaSC (i.e. about the half on average.) After alignment, the face images become blurred with fewer textural details. Thus, the performance saturated earlier when increasing the number of GMM components.

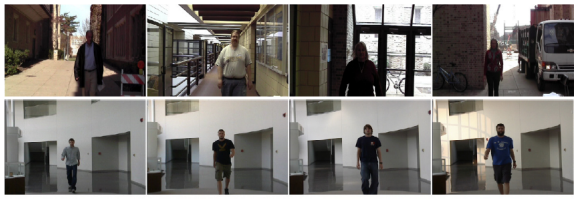


Fig. 5: The upper row is the sample frames of MBGC walking sequences in four different scenarios, and the bottom row shows the sample frames from FOCS UT-Dallas walking videos.

MBGC	WGCP [24]	DFRV [23]	FV_{49} [3]	FV_{128} [3]	LFVR_{128} Ours	LFVR_{49} Ours
S2	0.27	0.26	0.45	0.42	0.45	**0.58**
S3	0.22	0.22	0.40	0.38	0.40	**0.45**
S4	0.22	0.22	0.40	0.38	0.40	**0.45**
FOCS	WGCP	DFRV	FV_{49}	FV_{128}	LFVR_{128}	LFVR_{49}
S2	0.33	0.36	0.38	0.39	0.65	**0.74**
S3	0.29	0.34	0.38	0.37	0.61	**0.75**
S4	0.18	0.21	0.29	0.28	0.51	**0.65**

Table 1: it shows the verification rates of each algorithm at FAR=0.1. Our LFVR_{49} achieves the best results.

5. CONCLUSIONS

In this paper, we proposed a landmark-based Fisher vector representation for video-based face verification problems. Our experimental results demonstrate that if the landmarks are available, we should always utilize them. In addition, our approach greatly reduces the training time to learn a GMM and the dimensionality for the final feature representation while achieving better performance than the original Fisher vector counterpart. For future work, we plan to use dictionaries to perform feature encoding instead of GMM and FV. Inspired by deep learning, we also plan to develop hierarchical feature learning and encoding methods for video-based face verification.

6. REFERENCES

[1] W. Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] D. Chen, X. D. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[3] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *British Machine Vision Conference*, 2013, vol. 1, p. 7.

[4] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman, "A compact and discriminative face track descriptor," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[5] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[6] B. C. Zhang, S. G. Shan, X. L. Chen, and W. Gao, "Histogram of gabor phase patterns (hgpp): a novel object representation approach for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 57–68, 2007.

[7] A. Coates, A. Y. Ng, and H. L. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 215–223.

[8] H. X. Li, G. Hua, Z. Lin, J. Brandt, and J. C. Yang, "Probabilistic elastic matching for pose variant face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3499–3506.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *IEEE International Conference on Computer Vision*, 2009, pp. 498–505.

[10] Y. Taigman, L. Wolf, and T. Hassner, "Multiple one-shots for utilizing class label information.," in *British Machine Vision Conference*, 2009, pp. 1–12.

[11] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *International Conference on Machine learning*, 2007, pp. 209–216.

[12] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *Asian Conference on Computer Vision*, pp. 88–97. 2010.

[13] D. Chen, X. D. Cao, L. W. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*, pp. 566–579. Springer, 2012.

[14] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition*, 2008.

[15] A. Asthana, S. Zafeiriou, S. Y. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3444–3451.

[16] A. Asthana, S. Zafeiriou, S. Y. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[17] H. T. Wang, S. Z. Li, and Y. S. Wang, "Face recognition under varying lighting conditions using self quotient image," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2004, pp. 819–824.

[18] X. Y. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[19] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*, pp. 143–156. 2010.

[20] X. D. Cao, D. Wipf, F. Wen, G. Q. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *IEEE International Conference on Computer Vision*. IEEE, 2013, pp. 3208–3215.

[21] "National institute of standards and technology: Multiple biometric grand challenge, http://www.nist.gov/itl/iad/ig/mbgc.cfm," .

[22] "National institute of standards and technology: Face and ocular challenge series, http://www.nist.gov/itl/iad/ig/focs.cfm," .

[23] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *European Conference on Computer Vision*, pp. 766–779. 2012.

[24] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2273–2286, 2011.