Fisher Vector Faces in the Wild

Karen Simonyan
karen@robots.ox.ac.uk
Omkar M. Parkhi
omkar@robots.ox.ac.uk
Andrea Vedaldi
vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

Figure 1: **Face Verification.** Given a pair of face images, the task is to verify if both depict the same person. This is a challenging task due to the variations in lighting, pose, face expression, etc.

Overview. The face verification and recognition domain has mostly been dominated by carefully designed representations based on features computed around numerous facial landmarks [1, 3]. On the other hand, for image classification, image representations are capable of capturing discriminative image information without any domain-specific knowledge by using densely computed features, coupled with a non-linear encoding. For instance, the high-dimensional Fisher vector encoding [6] of SIFT features [5] achieves state-of-the-art performance on several image classification benchmarks [2].

In this paper, we address this discrepancy and make two contributions. First, we show that Fisher vector encoding of dense SIFT, an off-the-shelf image representation, achieves state-of-the-art face verification performance on the challenging "Labeled Faces in the Wild" (LFW) benchmark. Second, we show that the high-dimensional face representation using Fisher vector encoding is amenable to discriminative dimensionality reduction. The resulting compact descriptor has equal or better recognition accuracy and is very well suited to large-scale face recognition tasks.

The overview of our face descriptor computation pipeline is shown in Fig. 2. The learnt dimensionality reduction model is visualised in Fig. 3. It is able to automatically extract the discriminative regions of the face.

Results. In the unrestricted setting of LFW, we achieve $93.03 \pm 1.05\%$ face verification accuracy, closely matching $93.18 \pm 1.07\%$ obtained by the state-of-the-art method [3], based on high-dimensional LBP descriptors sampled around 27 face landmarks. It should be noted that the best result of [3] using SIFT is 91.77%. In the restricted setting of LFW, we achieve the verification accuracy of $87.47 \pm 1.49\%$, setting the new state-of-the-art. This is better than the second best method of [4] by 3.4%.

The ROC curves of our method as well as the other methods are shown in Fig. 4.

[1] T. Berg and P. N. Belhumeur. Tom-vs-Pete classifiers and identity-preserving alignment for face verification. In *Proc. BMVC.*, 2012.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. BMVC.*, 2011.

[3] D. Chen, X. Cao, F. Wen, and Sun J. Blessing of dimisionality: High dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013.

[4] H. Li, G. Hua, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Proc. CVPR*, 2013.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[6] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

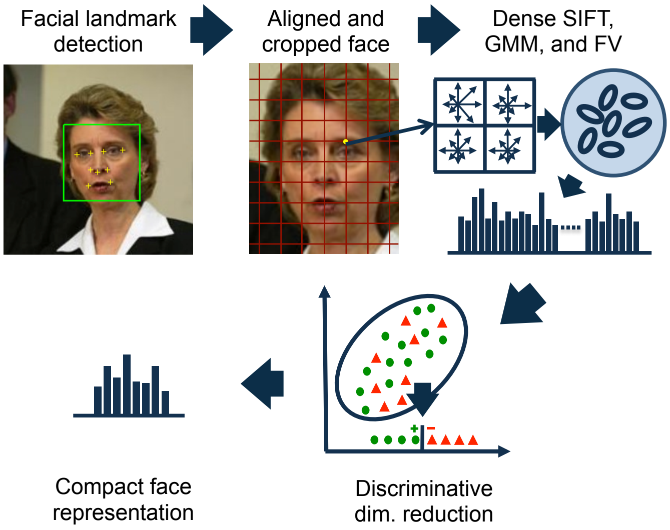
Facial landmark detection
Aligned and cropped face
Dense SIFT, GMM, and FV
Compact face representation
Discriminative dim. reduction

Figure 2: **Method overview.** Given an image, we first run an off-the-shelf face and facial landmark detectors to align the face to a canonical frame. Spatially-augmented SIFT features are then **densely** extracted from the face region and encoded into a high-dimensional **Fisher vector face representation**. Finally, discriminative dimensionality reduction is learnt on these features to compress them into a compact representation, while improving the discriminative ability.

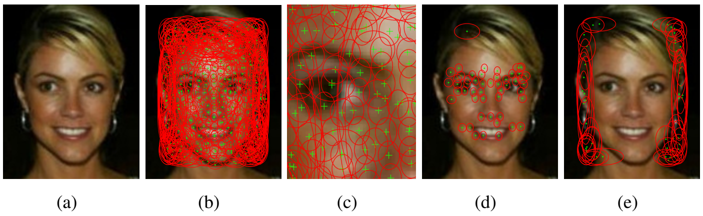

Figure 3: **Learnt model.** A Fisher vector, coupled with discriminative dimensionality reduction, can automatically capture the discriminative parts of the face. (a): an aligned face image; (b): unsupervised GMM clusters densely span the face; (c): a close-up of a face part covered by the Gaussians; (d): 50 Gaussians of the GMM, corresponding to the learnt projection matrix columns with the highest energy; (e): 50 Gaussians corresponding to the learnt projection matrix columns with the lowest energy.

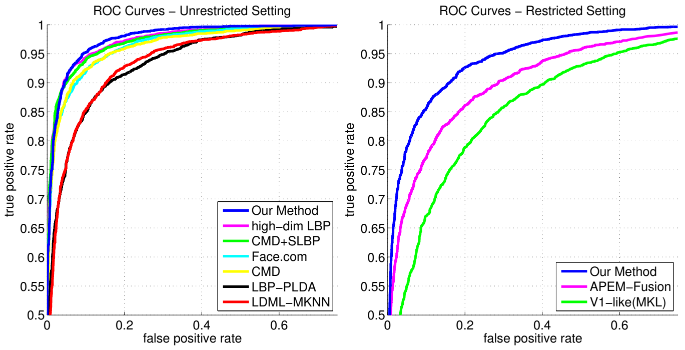

Figure 4: **Comparison with the state-of-the-art.** ROC curves of our method and the state-of-the-art techniques in LFW-unrestricted (left) and LFW-restricted (right) settings.