Face Recognition in the wild

Tiecheng Su

University of Rochester
Dept. of Electrical Engineering

Abstract

Using Fisher vectors to encode densely sampled SIFT features. Since Fisher vectors are very high dimensional, we show that a compact descriptor can be learned by using discriminative metric learning. This compact descriptor has a better recognition accuracy and is well suited to large scale identification tasks.

I. Introduction

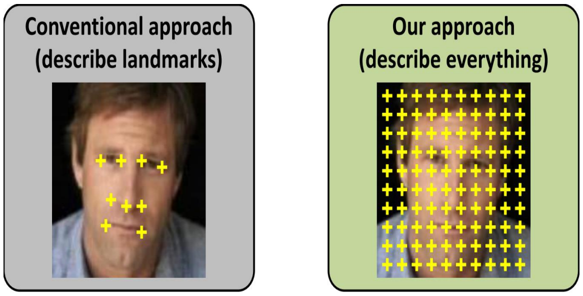
Over the last ten years, face recognition has become a popular area of research in computer vision and one of the most successful applications of image analysis and understanding. The database "Labeled Faces in the Wild" [1] contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the person pictures.

II. Dataset Information

In order to gain a higher recognition accuracy, we need more training data. Therefore, we first extract the data with at least 12 faces per person. That ends up with 127 persons. Then we split the dataset into training set and test set with a ratio of 8:2.

III. Fisher vector faces representation

Dense SIFT. We first need to extract SIFT feature from the image. The idea is to compute SIFT densely on an image, rather than on a sparse and potentially unreliable set of points obtained from an interest point detector. After cropping and rescaling the face to a 150 × 150 image, we run the algorithm and ends up with 20K 128-dimensional descriptors per face.



Fisher vectors. The FV encoding aggregates a large set of dense SIFT features into a Gaussian Mixture Model(GMM)features. We train a GMM with diagonal covariances, using KNN as initial value, and only consider the derivatives with respect to the Gaussian mean and variances. This leads to the representation which captures the average first and second order differences between the dense features and each of the GMM centers:

$$\Phi_k^1 = \frac{1}{N\sqrt{w_k}} \sum_{p=1}^{N} \alpha_p(k)(\frac{x_p - \mu_k}{\delta_k}), \Phi_k^2 = \frac{1}{N\sqrt{w_k}} \sum_{p=1}^{N} \alpha_p(k)(\frac{x_p - \mu_k}{\delta_k^2} - 1) \quad (1)$$

Here,$\{w_k, \mu_k, \delta_k\}_k$ are the mixture weights, means, and diagonal covariances of the GMM, which is computed on the training set and used for the description of all face images; $\alpha_p^k$is the soft assignment weight of the p-th feature x_p to the k-th Gaussian. An FV ϕ is obtained by stacking the differences: $\phi = [\Phi_1^{(1)}, \Phi_1^{(2)}, ..., \Phi_K^{(1)}, \Phi_K^{(2)}]$. The encoding describes how the distribution of features of a particular image differs from the the distribution fitted to the features of all training images.

But before we using the FV encoding, we first need to applied PCA to SIFT, reducing its dimension from 128 to 64. The FV dimensionality is 2Kd, where K is the number of Gaussians in the GMM, and d is the dimensionality of the patch feature vector(d=64).The performance of an FV is further improved by passing it through signed square-rooting and L_2 normalization.

Spatial information. The Fisher vector is an effective encoding of the feature space structure. However, it does not capture the distribution of features in the spatial domain. Therefore, we need to augment the visual features with their spatial coordinates, in this case $[S_{xy}; \frac{x}{w} - \frac{1}{2}; \frac{y}{h} - \frac{1}{2}]$, where S_{xy} is the (PCA-SIFT) descriptor of a patch centered at(x,y), and w and h are the width and height of the face image, and then using the FV encoding of the augmented features as the image descriptor.

Discriminative dimensionality reduction. The aim of discriminative dimensionality reduction is to obtain smaller image descriptors, while preserving or even improving their ability to discriminate images based on their content. To realize that, we need to find a low-rank linear projection W of the descriptors that minimizes the distances between same faces and maximizes it otherwise.

IV. Algorithm

- **Face alignment and extraction.**

  Given an image, we fist need to extract the face by face detection algorithm. We can simply call the matlab function vision.CasadeObjectDetector. And then we need to resize the image to the same size. In the aligned image, we extract a 150 × 150 face region. As we can see, some of the image may contains several faces, in case of that, the best solution is to just select the largest face, since we can not recognize the face yet.
- **Face descriptor computation.**

  For dense SIFT computation and Fisher vector encoding, we utilized publicly available packages [2]
- **Dimensionality reduction.**

  In order to compress a high-dimensional FV encoding into a small discriminative representation, we introduce a linear projection function, which serve two purposes: (i)it dramatically reduces the dimensionality of the face descriptors. (ii)it improves the recognition performance.

  The linear projection $W \in \mathbb{R}^{p \times d}$,$p \ll d$, which projects high-dimensional Fisher vectors $\Phi \in \mathbb{R}^d$ to low dimensional vectors $W\Phi \in \mathbb{R}^p$, such that the squared Euclidean distance $d_W^2(\Phi_i, \Phi_j) = \|W\Phi_i - W\Phi_j\|_2^2$between images i and j is smaller than a learnt threshold $b \in \mathbb{R}$ if i and j are the same person, and larger otherwise. We further impose that these conditions are satisfied with a margin of at least one, resulting in the constraints:

$$y_{ij}(b - d_W^2(\Phi_i, \Phi_j)) > 1 \tag{2}$$

where $y_{ij} = 1$ if images i and j contain the faces of the same person, and $y_{ij} = -1$ otherwise.

Note that the Euclidean distance in the p-dimensional projected space can be seen as a low-rank Mahalanobis metric in the original d-dimensional space:

$$d_W^2(\Phi_i, \Phi_j) = \|W\Phi_i - W\Phi_j\|_2^2 = (\Phi_i - \Phi_j)^T W^T W(\Phi_i - \Phi_j) \tag{3}$$

Learning W optimizes the following objective function, incorporating the constraints(2) in a hinge loss formulation:

$$argmin_{W,b} \sum_{i,j} max[1 - y_{ij}(b - \Phi_i - \Phi_j)^T W^T W(\Phi_i - \Phi_j), 0] \tag{4}$$

Using stochastic gradient descent. At each iteration t, the algorithm samples a single pair of face images(i,j)(sampling with equal frequency positive and negative labelsy_{ij})and performs the following update of the projection matrix:

$$W_t + 1 = \begin{cases} W_t, & \text{if } y_{ij}(b - \Phi_i - \Phi_j) > 1 \\ W_t - \gamma y_{ij} W_{ij} \Psi_{ij}, & \text{otherwise} \end{cases} \tag{5}$$

where $\Psi_{ij} = \Phi_i - \Phi_j)(\Phi_i - \Phi_j)^T$ is the outer product of the difference vectors, and γ is a constant learning rate.

Finally, note that the object(4) is not convex in W, so initialization is important. In practice, we initialize W to extract the p largest PCA dimensions. Furthermore, differently from standard PCA, we equalize the magnitude of the dominant eigenvalues(whitening) as the less frequent modes of variation tend to be among the most discriminative.

- **Learning.**

  In order to fit multi-class models with support vector machine(SVM), we can just call the matlab function fitcecoc

V. Results and analysis

SIFT density	GMM Size	Desc. Dim.	Distance Function	accuracy,%
2 pix	256	33792	diag. metric	79.1
2 pix	256	128	low-rank Mah. metric	83.7

From the table above, we can find out that the by doing dimensionality reduction, the accuracy was improved and since decision dimensionality was largely reduced, both the training and testing process become much faster than before.

In this project, we use dense features to represent faces which avoid applying a large number of sophisticated face landmark detectors. Also, we present a large-margin dimensionality reduction framework, well suited for high-dimensional Fisher vector representations. As a result, we obtain an effective and efficient face descriptor computation pipeline, which can be easily applied to large-scale face image repositories.

In this project, we only use a single feature type. In our future work, we are planning to investigate multi-feature image representations, which can be incorporated into our framework.

References

[1] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, 20:317–330. University of Massachusetts, Amherst, 2007.

[2] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. *ACM Multimedia*, 2010

[3] Face Recognition with MATLAB, Technology

[4] Relja ArandjeloviÂť c, Andrew Zisserman. Three things everyone should know to improve object retrieval.

[5] Florent Perronnin and Christopher Dance. Fisher Kernels on Visual Vocabularies for Image Categorization.

[6] Yan Ke1, Rahul Sukthankar2,1 PCA-SIFT: A More Distinctive Representation for Local Image Descriptors.

[7] Florent Perronnin, Christopher Dance, Gabriela Csurka, and Marco Bressan Adapted Vocabularies for Generic Visual Categorization

[8] Ken Chatfield,Victor Lempitsky,Andrea Vedaldi,Andrew Zisserman The devil is in the details: an evaluation of recent feature encoding methods

[9] Karen Simonyan,Omkar M. Parkhi,Andrea Vedaldi,Andrew Zisserman Fisher Vector Faces in the Wild