Selective Encoding for Recognizing Unreliably Localized Faces

Ang Li, Vlad I. Morariu, Larry S. Davis
University of Maryland, College Park
{angli,morariu,lsd}@umiacs.umd.edu

Abstract

*Most existing face verification systems rely on precise face detection and registration. However, these two components are fallible under unconstrained scenarios (*e.g.*, mobile face authentication) due to partial occlusions, pose variations, lighting conditions and limited view-angle coverage of mobile cameras. We address the unconstrained face verification problem by encoding face images directly without any explicit models of detection or registration. We propose a selective encoding framework which injects relevance information (*e.g.*, foreground/background probabilities) into each cluster of a descriptor codebook. An additional selector component also discards distractive image patches and improves spatial robustness. We evaluate our framework using Gaussian mixture models and Fisher vectors on challenging face verification datasets. We apply selective encoding to Fisher vector features, which in our experiments degrade quickly with inaccurate face localization; our framework improves robustness with no extra test time computation. We also apply our approach to mobile based active face authentication task, demonstrating its utility in real scenarios.*

1. Introduction

As face recognition techniques have gradually matured over the past few decades, the research focus has shifted from recognizing faces with controlled variations to unconstrained real-world scenarios [3]. Modern approaches based on high dimensional feature encoding [4, 13, 16, 19] and deep neural networks [20, 21] have recently emerged and achieved promising results on unconstrained face databases [6, 25]. However, most existing face recognition systems depend on accurate face detection and registration. Unfortunately, these two components are a significant source of error in real-world environments or real-time applications.

In the application of mobile face authentication, for example, faces recorded from a front-facing smartphone camera often exhibit rare non-horizontal poses (*i.e.*, neither frontal nor profile) and are often partly outside the camera's viewpoint. This problem is exacerbated when users are performing other tasks (as opposed to actively ensuring that their face is within the camera view) in which case the facial video quality becomes even worse, further challenging existing face detection and registration systems. For example, one of our experiments shows that the popular Viola-Jones face detector [23] fails on a significant portion of a smartphone-recorded face dataset [5] (Fig. 1).

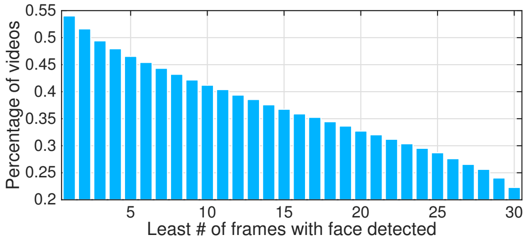


Figure 1. Performance of Viola-Jones (OPENCV) multi-scale face detector on a mobile based video face authentication dataset [5] with a total of 19,158 sampled video clips each 30 frames long. The x-axis is the number of frames in each video and the y-axis shows the percentage of video clips with at least the corresponding number of frames having faces detected. While all of the video clips contain faces, only 54% of the videos have at least one face detected and 22% have faces detected across all the 30 frames.

Most current face recognition datasets use images viewed from a distance for benchmarking. This type of data involves other challenges, compared to those from mobile applications: low image resolution and background distractions, because of which we can still expect some degree of errors in the detection step, *i.e.*, improper estimation of face centers and bounding box sizes. A statistical illustration of the face detection errors using FDDB benchmark data [7] is shown in Fig. 2.

Motivated by these observations, we explore the possibility of addressing unconstrained face verification problems without explicit face detection or registration. The central idea of our approach is that the codebook can be optimized to encode additional information for discriminating relevant image patches from irrelevant background distractions. We propose a unified codebook-based framework, named "selective encoding", the core of which is a compo-

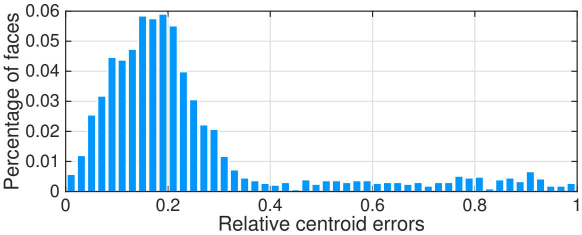


Figure 2. Viola-Jones (OPENCV) multi scale face detection results on Face Detection Dataset and Benchmark (FDDB) [7]: the relative centroid errors are computed as the centroid distance between detected faces and their closest ground truths, divided by the averaged axis length of ground truth ellipses. The chart shows 68% of faces are detected faces while the other 32% are false alarms with no overlap with any of the ground truth faces. Notably, 50% of the detected faces produce some levels of offsets from 0% to 25% where the peak is around 20% of the face size (*e.g.*, for 150 × 150 faces, the peak of errors is 30 pixels).

nent named "selector" which injects trained relevance information into codewords via a set of "relevance weights" and utilizes these weights to select semantically relevant patch descriptors and codewords at test time. Patch descriptors and codewords that successfully pass the selector will be used for encoding images. The selector essentially finds a good relevant sub-matrix of the posterior probability (assignment) matrix for feature encoding.

For recognizing unreliably localized faces, we define the descriptor relevance as foreground probabilities, so image patches belonging to the facial region are selected over those that do not. The relevance distribution training involves counting for each codeword the foreground/background distribution of its assigned patch descriptors. These distributions are used for computing the foreground probability of each newly observed patch in testing. Background distractions are thereafter removed from the descriptor set so that the encoded representation can achieve spatial robustness.

Fisher vector encoding [18] is one of the most powerful codebook based feature encoding techniques. However, its most recent applications in face verification require face detection and registration. One of our experiments shows that this method degrades quickly with inaccurate estimation of face centers and bounding box sizes due to the inclusion of more distractive patches. We validate our framework using the Fisher vector encoding on public datasets and show that our method is capable of robustifying such encoding technique with respect to uncertain face localization. We further apply our framework to a mobile based active face authentication task to show its applicability in real-world scenarios.

Contribution. The main contributions of our work include (1) a generic and unified framework for selecting and encoding relevant features which does not require accurate detection or registration, (2) its application to Fisher vector encoding for spatially robust face verification, and (3) its application to mobile based active face authentication.

2. Related work

Feature encoding. The bag of visual words model [10] is the most popular feature encoding framework for many computer vision tasks. In this model, a codebook is built using K-means clustering and each feature is assigned a weight for each cluster center (aka. codeword) according to their distances. An image is thereafter represented by the distribution (histogram) of those assignments. Most modern feature encoding techniques are extensions of this codebook model such as Fisher vectors [14] and the vector of locally aggregated descriptors [8]. The central idea is that, instead of using only an assignment distribution, an image can also be represented using the first order (mean of difference) and the second order (standard deviation) statistics of all the (soft or hard) assigned features for each codeword. Fisher vector encoding is now among the state-of-the-art on various computer vision applications such as image classification [14, 16, 18], image retrieval [15] and face verification [13]. Our work is built upon Fisher vectors and integrates additional supervised information into the codebook for encoding semantically relevant patches.

Unconstrained face recognition. The upsurge of research on unconstrained face recognition gave rise to the creation of Labeled faces in the wild (LFW) dataset [6]. Besides the Fisher vector faces [19], many works have been developed on this topic, such as high dimensional local binary patterns [4], deep learning based approaches [20, 21] and sparse coding based approaches [24, 26]. Considering that face recognition problems are often challenged by pose variations, many works try to improve recognition accuracy by means of robust facial alignment and correction using sophisticated 3D models or shape matching [2, 3, 21, 24]. However, the vulnerablility of face detectors under real-world scenarios is usually overlooked and most existing face verification methods generally assume that detected and well aligned faces are given [13, 19]. The goal of our work is to remove the strong dependency on face detection by improving the encoding scheme to be significantly more robust to spatial misalignment.

Joint localization and classification. The general image object classification task is also affected by the performance of object localization. Most approaches try to find good localization and segmentation of the objects to relieve the subsequent recognition task [1, 22]. However, detection is even harder than classification in some sense (*e.g.*, robust bounding box estimation). A few recent approaches are motivated by the idea of jointly detecting and classifying objects in images in the hope that the two tasks help each other. Nguyen *et al.* [12] proposed to jointly localize discriminative regions and train a region-based SVM for im-

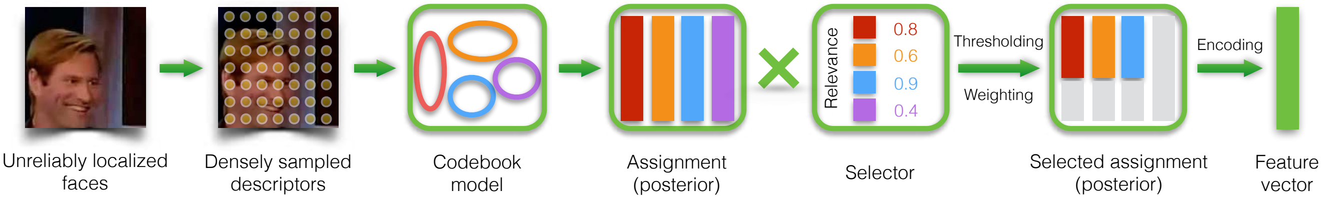


Figure 3. The proposed selective encoding framework: Images or videos with unreliably localized faces are the direct input to our model. Posterior probabilities (assignment) for densely sampled local descriptors are computed according to the trained codebook model. The relevance weight of each descriptor is calculated according to its posterior probability distribution and the relevance of corresponding codewords. The selector component is trained offline using weakly supervised features. A subset of the assignment matrix (or a new assignment matrix) is generated by thresholding (or re-weighting using) the descriptor relevance, and used for image feature encoding.

age categorization. Lan *et al.* [9] proposed a figure-centric model learned by latent SVM for joint action localization and recognition. The most similar work to ours is object-centric pooling [17]. Its main idea is to infer, jointly with classification, tight object bounding boxes and pool features within detected regions. They developed an MIL-like SVM formulation for joint object localization and classification. However, our work differs in that (1) instead of finding perfect detections, we explore the implicit feature selection power of the codebook, and (2) our framework is designed for feature encoding and does not depend on any subsequent classification.

3. Preliminary – Fisher vector encoding

The Fisher vector (FV) encoding was first proposed in [14] and applied to face verification problems in [19] and [13]. The central idea of Fisher vector encoding is to aggregate higher order statistics of each codebook into a high dimensional feature vector. More specifically, a Gaussian mixture model (GMM) is trained as the visual codebook. First-order and second-order distance statistics w.r.t. each of the Gaussian mixture components are concatenated into the final feature representation. Let $\mathbf{x}_p$ be the p-th descriptor and (μ_k, σ_k^2) be the k-th Gaussian component. The assignment coefficient (posterior probabilities) of $\mathbf{x}_p$ with respect to the k-th Gaussian is represented using $\alpha_k(\mathbf{x}_p)$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ be the descriptor set, the Fisher vector representation is computed as $\phi(\mathbf{X}) = [\Phi_1^{(1)}, \Phi_1^{(2)}, \ldots, \Phi_K^{(1)}, \Phi_K^{(2)}]$ where

$$\Phi_{ik}^{(1)} = \frac{1}{N\sqrt{\pi_k}} \sum_{p=1}^{N} \alpha_k(\mathbf{x}_p) \left(\frac{x_{ip} - \mu_{ik}}{\sigma_{ik}} \right) , \tag{1}$$

$$\Phi_{ik}^{(2)} = \frac{1}{N\sqrt{2\pi_k}} \sum_{p=1}^{N} \alpha_k(\mathbf{x}_p) \left[\left(\frac{x_{ip} - \mu_{ik}}{\sigma_{ik}} \right)^2 - 1 \right] . \tag{2}$$

Most algorithms using Fisher vectors apply signed square root and ℓ^2 normalization to the feature vectors which tend to further improve the representation capability of Fisher vectors [16, 19].

4. Our approach – Selective encoding

4.1. Framework overview

The proposed selective encoding framework is illustrated in Fig. 3. Existing codebook based face recognition approaches require detection and registration beforehand, while our framework reduces the need for such prerequisites. Generally speaking, our framework is composed of three main stages: (1) building a vocabulary (2) descriptor and codeword selection (selector) and (3) feature encoding. The key component for achieving spatial robustness is the selector, which selects relevant descriptors and codewords for the feature encoding stage. The selector is trained with weakly supervised prior knowledge on the descriptor relevance (*i.e.*, rough detection bounding boxes). An advantage of our framework is that we do not require any extra computational cost during testing because the selector is essentially performed on the matrix of posterior probabilities (assignment) for the codebook, which is necessarily computed in the conventional codebook framework.

4.2. Vocabulary

Descriptor extraction. Following [19], we extract densely sampled SIFT descriptors [11] at 5 different scales. The 128-D descriptors are further reduced to 64-D by principal component analysis. Fisher vectors are often learned using an augmented descriptor which adds two additional dimensions for the spatial coordinates of each SIFT descriptor. A normalization is utilized for the augmented dimension, *i.e.*, $[x_{\text{aug}}, y_{\text{aug}}] = [\frac{x}{w} - 0.5, \frac{y}{h} - 0.5]$ where w, h are the width and height of the window.

Codebook construction. The Fisher vector encoding uses Gaussian mixture models to provide softer structures and capture smoother feature distributions in the encoding than the K-means clustering based codebook. As [19], we use 512 Gaussian components for our experiments.

4.3. Selector

The selector consists of two parts: (1) descriptor selection and (2) codeword selection. Both stages are executed based on the trained relevance weights of each codeword and their corresponding posterior probabilities w.r.t. newly observed image patches.

Codeword relevance. Given a trained codebook (Gaussian mixture model), the selector is trained to associate additional foreground/background information with each codeword (Gaussian component). The training involves calculation of the relevance weights for each codeword.

Let $\mathbf{x}_i$ be the i-th patch descriptor, $\boldsymbol{\theta}_k$ be the k-th Gaussian mixture component and their corresponding posterior probability be $p(\boldsymbol{\theta}_k|\mathbf{x}_i)$. The selector is trained using n-dimensional patch descriptors $\mathbf{x}_i \in \mathbb{R}^n$ with their binary labels $y_i \in \{0, 1\}$ which represent whether they should be selected for feature encoding, by counting for each codeword the expected descriptor relevance, *i.e.*,

$$p_s^c(\boldsymbol{\theta}_k) = \frac{\sum_{i=1}^N p(\boldsymbol{\theta}_k|\mathbf{x}_i) y_i}{\sum_{i=1}^N p(\boldsymbol{\theta}_k|\mathbf{x}_i)} \ . \tag{3}$$

The codeword relevance value ranges between 0 and 1. Codewords with higher relevance weights (larger than 0.5) are more likely to aggregate foreground descriptors while those with lower relevance weights (lower than 0.5) have higher chance of being background. Although keeping unnecessary codewords will not damage the encoding space, discarding those background codewords naturally reduces the feature dimension and in some cases improves the recognition accuracy (Fig. 11(b)).

For recognizing unregistered faces, the training patches and their semantic labels are obtained by using images with valid detection outputs. Those features located within detected face bounding boxes are labeled as 1 and those outside labeled as 0. In our experiments we are using loose detection bounding boxes which contain background areas; however, the learned relevance distributions is sufficient for improving the encoding robustness.

Descriptor relevance. At test time, the posterior probabilities for each patch descriptor are given from the codebook model. The descriptor relevance weight is then computed by counting the relevance contribution from each codeword with respect to their posterior probabilities, *i.e.*,

$$p_s^d(\mathbf{x}_i) = \sum_{k=1}^K p(\boldsymbol{\theta}_k|\mathbf{x}_i) p_s^c(\boldsymbol{\theta}_k) \ . \tag{4}$$

The posterior probability can be computed via either soft or hard assignment (in hard assignment settings, the highest posterior probability for each descriptor is lifted to 1 and all the others reduced to 0). The descriptor relevance also ranges between 0 and 1, similar to codeword relevance. Intuitively, the descriptor selection plays a key role in achieving spatial robustness of feature encoding by removing background patches. In our experiment, we remove all descriptors with relevance lower than 0.5 (a threshold for separating foreground from background) for patch selection.

4.4. Encoding

The encoding stage receives from the selector a subset (or a modified version) of the posterior probability matrices and encodes them as Fisher vectors (as described in Section 3). The encoded Fisher vectors can be further reweighed or reduced to lower dimensions by multiple metric learning approaches; however, with restricted training samples, learning a low rank metric is difficult [19]. The mobile face authentication problem comes with a limited training set – users are not likely to spend much time actively training the smartphones. So in our experiments, we employ the ℓ^2 metric and diagonal metric learning (*i.e.*, training a diagonal metric using support vector machines) proposed in [19] for evaluating encoding performance.

4.5. Learning with spatial-sensitive features

Intuitively, the location features help when the face images are properly registered. However, when the registration is poor, augmented location information may instead hurt the performance. The GMM model can smooth out the Gaussian component on the location dimensions (Fig. 4) and may also learn the location distribution of patches when the training images have some underlying mis-registration patterns. However, the robustness to localization errors is not sufficient for unconstrained spatial patterns, in which case performance drops quickly and becomes worse than ignoring location information altogether. The main reason is because patches belong to the same facial part are assigned to different codewords due to the influence of the augmented location dimension. However, our framework can adapt to such location sensitive augmented features. The central idea is that we can identify relevant patches in the codebook and renormalize the augmented dimensions of their corresponding descriptors so that patches belonging to close facial parts can be aggregated into the same codewords.

Since the augmented dimensions are spatially sensitive, they should not be involved in learning the descriptor and codeword relevance distributions. As a result, we use the appearance-based dimensions (first 64D) of each Gaussian mixture component when computing the relevance weights of codewords and descriptors. Once patches are selected, the last two augmented dimensions of corresponding descriptors are reduced by their mean values, *i.e.*, $[x'_{\text{aug}}, y'_{\text{aug}}] = [x_{\text{aug}} - \bar{x}_{\text{aug}}, y_{\text{aug}} - \bar{y}_{\text{aug}}]$, and the updated descriptors are used in feature aggregation and encoding.

5. Experiments

We validate our approach on three face datasets with different foci: (a) image based face verification (b) video based

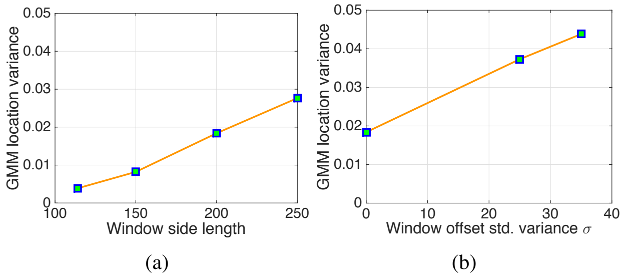


Figure 4. LFW: averaged variance of Gaussian components on augmented location dimensions vs. (a) window side length with zero offset and vs. (b) standard deviation of window offsets (window side length 200). As the window spatial uncertainty increases, the learned GMM increases the variance of Gaussian distributions on location dimensions, which essentially reduces the influence of location information on codeword assignment.

face verification and (c) mobile based face authentication. In the first two datasets, we perform random shifts to the detected face bounding box to compare the spatial robustness of the original Fisher vector encoding and the proposed selective Fisher vector encoding.

5.1. Image based face verification

Labeled faces in the wild (LFW) [6] is an image based face verification dataset. The dataset contains 13,233 images of 5,749 celebrities. The evaluation set is divided into 10 disjoint splits each of which contains 600 image pairs. Of these 300 are positive pairs describing the same person and the other 300 are negatives representing different identities. Two protocols are used for the benchmark: restricted and unrestricted. The restricted protocol prohibits using any outside data for training the models while the unrestricted version allows that. We validate our framework on the restricted protocol to show its performance with limited access to training data.

Perturbation generation. To study the sensitivity of localization, we randomly shift the annotated face centers (which are detected by Viola-Jones detector) using a Gaussian distribution $N(0, \sigma^2)$ where σ is chosen from 0, 25, 35 and 50 pixels. We set the window side length to 200 pixels, around 1.7 times the size of the tight facial bounding box.

Evaluation. Performance is evaluated using true positive rates at equal error rate (TPR@EER) averaged over the 10 splits. The codebook is trained using perturbed images with 512 Gaussian mixture components. For selective encoding, codeword relevance distributions are learned using 150×150 windows at the face center detected by Viola-Jones detector in the training set. It is worth noting that these windows do not tightly bound the faces.

Comparison with original Fisher vectors. Comparison with the original Fisher vectors is shown in Fig. 5 using both appearance and augmented descriptors. The proposed

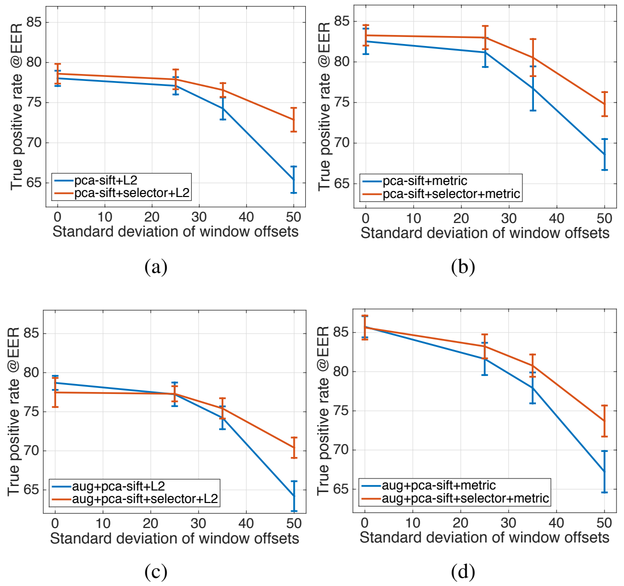


Figure 5. LFW: Original FV vs. hard selective FV encoding with PCA-SIFT descriptors with (a) ℓ^2 and (b) diagonal metric learning; original FV vs. soft selective FV encoding with *augmented* descriptors with (c) ℓ^2 and (d) diagonal metric learning.

selective encoding outperforms conventional Fisher vectors using both ℓ^2 metric and diagonal metric learning with 64-D PCA-SIFT descriptors. Interestingly, our method performs better even when there is no centroid perturbation. This might be because even the true facial bounding box includes a small number of distractive patches from the background. With augmented descriptors, a 1% performance drop of our framework is observed with no center offset using ℓ^2 metric. However, this performance gap vanishes using diagonal metric learning. Our approach also produces more stable performance across multiple levels of window offsets.

Comparison with perfect face localization. Since our goal is to make the original encoding technique more robust to localization, we compare our framework with the ideal case, where the ground truth face bounding box is known (this will serve as an upper bound on performance, since localization will be perfect). The results with both PCA-SIFT and augmented descriptors are shown in Fig. 6, where under ℓ^2 metric there is less than 0.5% difference between our approach and the ideal one. A larger gap is seen with diagonal metric learning. The ideal case is about 2% better with offset $\sigma = 0, 25, 35$; our approach performs better when more severe face occlusions occur with offset $\sigma = 50$.

Appearance-only vs. augmented descriptors. Fisher vectors are usually computed over descriptors augmented with their spatial coordinates, encoding spatial structures into the feature representation. These coordinate features are spatially sensitive and not suitable for learning foreground/background distributions. However, our frame-

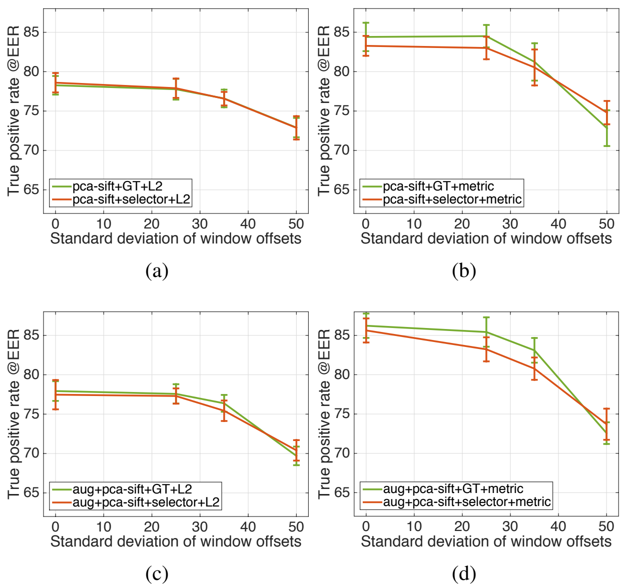


Figure 6. LFW: Hard selective FV encoding on perturbed images vs. Original FV encoding on ground truth facial windows with PCA-SIFT descriptors with (a) ℓ^2 and (b) diagonal metric learning; and Soft selective FV encoding on perturbed images vs. FV encoding on ground truth facial windows with *augmented* descriptors with (c) ℓ^2 and (d) diagonal metric learning.

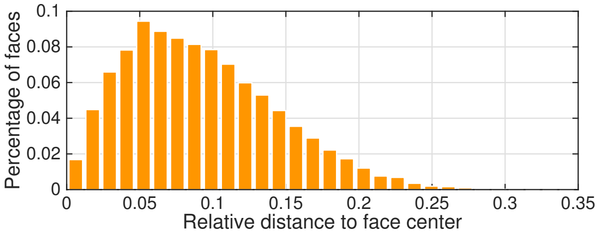


Figure 7. LFW: Histogram of relative distances (window side length equals 200 and standard deviation of offset 25) between mean of selected patch locations to true face centers.

work can adapt to such spatially sensitive features by "re-centering" selected patches. Fig. 7 shows the relative distance between the true face center and the mean coordinates of those selected patches when the window side length is 200 and offset standard deviation is 25. The peak error is around 5% (*i.e.*, 10 pixels). Our experiments suggest that, compared to appearance-only descriptors, the spatially augmented descriptors perform better with low spatial uncertainty (85.63 ± 1.53 vs. 83.27 ± 1.26 with zero offset and 200 window side length) and gradually degrades with similar performance when the spatial uncertainty increases (80.77 ± 1.42 vs. 80.53 ± 2.28 with 35 offset standard deviation and 200 window side length).

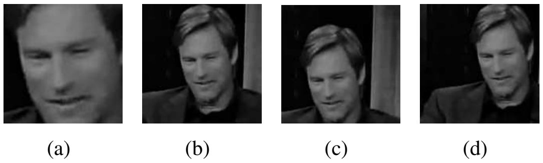


Figure 8. Sample perturbed face images in Youtube Faces dataset: $(\mu_{\text{scale}}, \sigma_{\text{scale}}, s_{\text{offset}})$ = (a) (1, 0, 0), *i.e.*, labeled face bounding box, (b) (2, 0, 0), (c) (2, 0, 0.5) and (d) (2, 0.5, 0.5).

5.2. Video based face verification

Youtube Faces (YTF) [25] is a benchmark for video based face verification. The dataset contains 3,425 videos for 1,595 celebrities collected from YouTube movies. All of the faces are localized by the Viola-Jones face detector. The evaluation set is composed of 5,000 pairs of tracks which are also divided into 10 splits. In each split, 250 pairs are positive and the other 250 are negative. For each of the 10 runs, 9 splits are used for training and the remaining split is used for testing. Similar to LFW, the dataset has restricted and unrestricted protocols. Our experiment adopts the restricted protocol in which only 4,500 pairs of videos are available for training the model and the similarity metric.

Data preparation. Youtube Faces contains a set of original video frames (faces and background) and a set of cropped and registered face videos. We randomly shift the annotated centers of the faces on each of original videos obeying a uniform distribution $U[-s_{\text{offset}}W, s_{\text{offset}}W]$ in both x and y directions to guarantee that perturbed images have intersections with detector bounding boxes, where s_{offset} is a scale factor and W is the side length of the detected facial bounding box, which differs from person to person. We choose the scale factor s_{offset} among values 0, 0.25, 0.5 and 0.75. For the scale of the windows, we enlarge the side length with another scale factor chosen from a Gaussian distribution $N(\mu_{\text{scale}}, \sigma^2_{\text{scale}})$. The mean μ_{scale} is chosen between 1 (original size) and 2 (double size). The σ_{scale} values are chosen from 0, 0.25 and 0.5. We resize all of the perturbed windows to 150×150 for feature encoding. Sampled perturbed images are shown in Fig. 8.

Evaluation. Verification accuracy is also evaluated using TPR@EER, averaged over 10 splits. We downsample each video to 5 frames long. It is worth noting that increasing the sample rate to 20 frames per video produces only 0.04% higher TPR@EER (80.88%) on tightly bounded detected faces than 80.84% obtained from sampling 5 frames per video. Following [13], we apply the incremental "video pooling" for encoding each video, *i.e.*, patch descriptors across frames from the same video are pooled together before being encoded into one Fisher vector. We train PCA and GMM using perturbed training images and learn codeword relevance distributions using detection bounding boxes in sampled training frames for each split.

Table 1. Youtube Faces: TPR@EER averaged over 10 folds for different perturbation settings using augmented PCA-SIFT descriptors and diagonal metric learning, comparing the proposed selective encoding with original Fisher vectors. Each row represents a setting of face window scaling and relative centroid offset distributions. The better result for each setting is annotated in red.

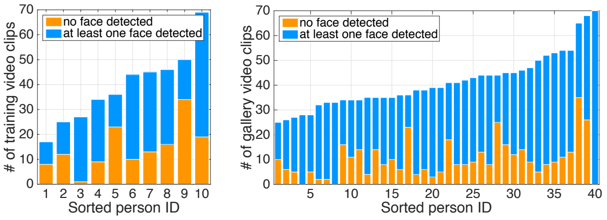
μ_{scale}	σ_{scale}	s_{offset}	Original FV	Selective FV
1	0	0	80.84 ± 1.91	81.00 ± 2.32
2	0	0	76.72 ± 3.33	77.24 ± 2.02
2	0	0.5	74.52 ± 1.81	76.96 ± 1.73
2	0.25	0	76.84 ± 2.27	77.40 ± 1.53
2	0.25	0.25	75.04 ± 1.92	77.72 ± 2.40
2	0.25	0.5	74.44 ± 1.26	75.76 ± 2.08
2	0.25	0.75	69.64 ± 1.87	72.88 ± 1.60
2	0.5	0	74.52 ± 1.90	75.32 ± 1.60
2	0.5	0.5	70.92 ± 1.35	72.72 ± 2.07

Result. The results comparing the proposed selective encoding and the original Fisher vectors are shown in Tab. 1, with different configurations of window scale and offset uncertainty. Both methods use the augmented descriptors and the selector in our approach is trained with soft assignment and tested with no codewords discarded. The results show that our approach outperforms the original Fisher vectors in all settings. Even for the true detected face windows ($\mu_{\text{scale}} = 1, \sigma_{\text{scale}} = s_{\text{offset}} = 0$), our approach obtains slightly improved accuracy. Both approaches experience a 3% performance drop when μ_{scale} is increased from 1 to 2, which is due to the decrease in face resolution, and a 2% drop when σ_{scale} increases from 0.25 to 0.5 with no window offset. Fortunately such high scale uncertainty is typically rare for face detectors and mobile applications. When the scale uncertainty ranges between 0 and 0.25, the encoding quality is relatively stable. The performance gap between the two approaches becomes larger when offset uncertainty increases (over 3% gain when $\mu_{\text{scale}} = 2, \sigma_{\text{scale}} = 0.25, s_{\text{offset}} = 0.75$).

5.3. Active face authentication on mobile devices

The use of mobile devices has increased dramatically over the last decades. The privacy protection of mobile phone users has always been an important problem. Verifying the faces recorded by the smartphone camera plays a central role in identifying the users. However, authentication is passively performed in the background, and users may not be actively trying to ensure that their face is viewed clearly by the camera. This results in face videos with unconstrained poses, some of which are raised faces because users are likely to read while their smartphones are below their faces instead of looking directly at the phone.

Dataset. We validate our approach on a dataset that contains 750 long videos recorded from the viewpoint of mobilephone cameras when user activities are present [5].



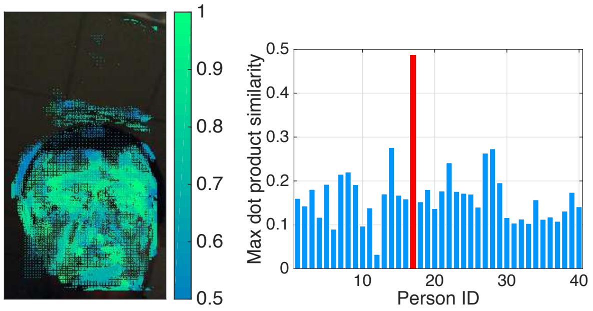
(a) (b)

Figure 9. Distribution of the video numbers in (a) training set and (b) gallery set. Identities are sorted in ascending order of their video numbers. Orange bars show the number of videos with no face detected at any of their frames and the blue bars show the number of those with at least one face detected. The training set contains 393 videos in total and the gallery set contains in average 43 videos clips per person.

More specifically, there are 50 persons (subjects) participated in the video recording. Each subject is asked to use the same smartphone to perform 5 different tasks, *i.e.*, Enrollment, Scrolling, Popup, Picture and Document, under three different lighting conditions, *i.e.*, well-lit, dim-lit and natural. The Enrollment task is to ask the user to record their faces in different poses and this data will be the gallery in the face verification protocol. All the other four tasks involve the users performing some activities on the cellphone (refer [5] for details); these videos make up the probe set.

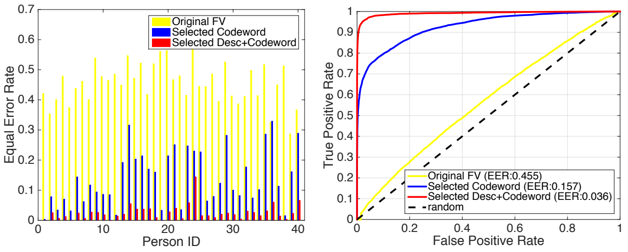
In practice, it is sufficient to identify users every few seconds. So we sample 30 short clips, each 30 frames long (approximately one second) for each test video. For the gallery set, each enrollment video is segmented into consecutive clips of 30 frames uniformly instead of random sampling. We use the Enrollment data of 10 persons for training and use those of the remaining 40 persons for constructing the gallery set. The lengths of enrollment videos vary for different persons. Fig. 9 shows the distribution of the training videos and the gallery. Eventually, we have a training set of 393 video clips and a gallery set that contains on average 43 video clips per person. The probe set contains 4 tasks for each person out of 40 for each of the 3 illumination conditions, *i.e.*, 360 video clips per person and 14,400 in total.

Evaluation. The evaluation protocol is different from LFW and YTF datasets because, for face authentication, each device has access to only the videos of the owner. So during test time, only the gallery of the corresponding identity is accessible. More specifically, each test clip is compared to all the gallery clips of the corresponding person and a maximum similarity score is calculated. Thereafter, an ROC curve can be generated either by averaging over identities with independent similarity score thresholding or by using a global similarity threshold for all persons. According to our experiments, there is no significant difference between using person-specific thresholds and using a global



(a) (b)

Figure 10. Probe identity #17: (a) patch centers with relevance (color annotated) larger than 0.5 are shown on top of the origin image and (b) max dot product similarity scores between the Fisher vector of selected patches and that of each gallery video clip. Red color shows the similarity for the ground truth identity.



(a) (b)

Figure 11. Results on active authentication dataset: (a) Equal error rate (EER) for each person and (b) ROC curves. Three approaches are compared: the original Fisher vector (Original FV), selective encoding with only codeword selection (Selected Codeword), selective encoding with both descriptor selection and codeword selection (Selected Desc+Codeword).

threshold. So, in all of our experiments, we use global thresholding for ROC curves. Equal error rates (EER) are also used for performance evaluation and comparison.

Result. We use the training clips which cover only 10 identities (Fig. 9) for training PCA and GMM of SIFT descriptors. Also we use all of the images with detected faces in the training set for learning the relevance distribution for selective encoding. Sometimes, real applications may not have large amount of data available for training. So we use such limited training data to evaluate the generalization ability of our trained selector. This experiment is based on appearance descriptors without location features.

We first run an example experiment on a sampled video frame from identity #17. The frame is taken under dark lighting condition and the chin of the identity is slightly out of sight. We apply the selector to dense multi-scale descriptors extracted over the image and obtain for each descriptor a relevance weight. The centroids of patches with higher than 0.5 relevance are plotted on top of the original image in Fig. 10(a). Most patches within the facial area are selected, although we still see a few background patches selected above the face on the ceiling. These incorrectly selected patches have an insignificant influence on the descriptor distribution when pooled with a large number of facial patches. We use these selected patch descriptors and the selected codewords (with 0.5 relevance or higher) for encoding the image and compare the feature representation with those from the 40 gallery sets using dot product similarity (equivalent to ℓ^2 since features are normalized). Similarity scores are shown in Fig. 10(b). The top scored identity is the ground truth and its score is over 0.2 larger than that of the second most similar identity which shows that even using such a dark and low quality image, we are still able to distinguish the identity from all other 39 identities.

The face authentication results are shown in Fig. 11. We compare our selective encoding framework (based on hard assignment selector) with the original Fisher vectors and a variant of our framework which discards only the codewords with relevance weights lower than 0.5. While the original Fisher vectors achieve 0.455 equal error rate, our approach improves significantly and achieves 0.036 equal error rate. Using only codeword selection achieves 0.157 equal error rate. That means the codeword selection is useful; however the selection of visual descriptors plays a more central role in robustifying feature encoding.

It is worth noting that the detector used for learning the relevance distribution is not specifically tuned in this experiment, so it might still produce errors. However, the experimental results suggest that our selection strategy is robust and does not require accurate registration.

6. Conclusion

We have proposed a generic selective encoding framework for representing objects of interest that are unreliably localized in images. Our framework introduces the selector component into the codebook model so that it does not require test time detection or registration and becomes robust to localization errors in real scenarios. Our method is also computationally efficient which can benefit real-time applications. We have applied selective encoding to general face verification and mobile phone face authentication. Experimental results suggest that our approach is able to improve the spatial robustness of feature encoding when face detectors produce errors or even fail to localize faces. We expect that our framework could be applied to general image classification and object recognition in the future.

Acknowledgement. This work was supported by cooperative agreement FA8750-13-2-0279 from DARPA. The authors acknowledge the University of Maryland supercomputing resources (`http://www.it.umd.edu/hpcc`) made available for conducting the research reported in this paper.

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 811–818. IEEE, 2013.

[2] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–11, 2012.

[3] L. Best, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. Technical Report MSU-CSE-14-1, Department of Computer Science, Michigan State University, East Lansing, Michigan, March 2014.

[4] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimisionality: High dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] M. Fathy, V. Patel, and R. Chellappa. Face-based active authentication on mobile devices. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1687–1691, April 2015.

[6] 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, University of Massachusetts, Amherst, October 2007.

[7] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[8] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 34(9):1704–1716, Sept. 2012.

[9] T. Lan, Y. W. 0003, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, editors, *Proceedings of International Conference on Computer Vision (ICCV)*, pages 2003–2010. IEEE, 2011.

[10] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '05, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, Nov. 2004.

[12] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.

[13] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2007.

[15] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3384–3391, June 2010.

[16] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

[17] O. Russakovsky, Y. Lin, K. Yu, and F.-F. Li. Object-centric spatial pooling for image classification. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Proceedings of European Conference on Computer Vision (ECCV)*, volume 7573 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2012.

[18] J. Sanchez, F. Perronnin, T. E. J. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision (IJCV)*, 2013.

[19] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *British Machine Vision Conference (BMVC)*, 2013.

[20] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, June 2014.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[22] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 2013.

[23] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, May 2004.

[24] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 34(2):372–386, 2012.

[25] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534. IEEE, 2011.

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 31(2):210–227, 2009.