Learning One-Shot Exemplar SVM from the Web for Face Verification

Fengyi Song; Xiaoyang Tan*

Department of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract. We investigate the problem of learning from a single instance consisting of a pair of images, often encountered in unconstrained face verification where the pair of images to be verified contain large variations and are captured from never seen subjects. Instead of constructing a separate discriminative model for each image in the couple and performing cross-checking, we learn a single Exemplar-SVM model for the pair by augmenting it with a negative couple set, and then predict whether the pair are from the same subject or not by asking an oracle whether this Exemplar-SVM is for a client or imposter in nature. The oracle by itself is learnt from the behaviors of a large number of Exemplar-SVMs based on the labeled background set. For face representation we use a number of unlabeled face sets collected from the Web to train a series of decision stumps that jointly map a given face to a discriminative and distributional representation. Experiments on the challenging Labeled Faces in the Wild (LFW) verify the effectiveness and feasibility of the proposed method.

1 Introduction

In many computer vision applications, we often encounter the problem of comparing the similarity between two images which are captured from never seen objects. For example, in the unconstrained face verification, the task is to decide whether a pair of images are from the same person or not, in which not only the images given are never seen, but the subjects behind are usually never seen as well (see Fig. 1). This problem is challenging mainly due to the following two reasons: 1) the images by themselves contain large variations which have to be dealt with. 2) since the information source (subjects which generate the images) is hidden, the known knowledge to infer them is extremely scarce (actually only one shot of sample per subject). Moreover, these two issues seem to be closely related - the task of learning good representation that well supports the one-shot similarity evaluation is much more difficult than doing that when the training samples are abundant.

To address these issues, many different "pairwise" approaches have been developed in recent years - either directly learn same person/different person decision rules, or that learn pairwise similarity metrics that can be used to produce such rules [2–5]. The key idea behind these approaches roots from the attempt

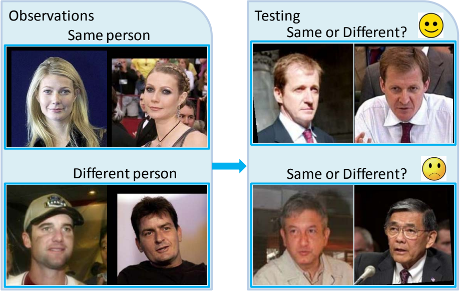

Fig. 1. Illustration of the two key problems in face verification, i.e., the pair of images to be verified contain *large variations* and are captured from *never seen subjects*. Images from the LFW face database [1].

to learn a suitable distance measure that compares pair of examples in spite of large appearance variations existed. For example, in [2] a logistic discriminant approach (LDML) was introduced to learn the metric from a set of labelled face pairs, while [4] propose to use an ensemble of extremely randomized binary trees to quantize the differences between pairs of "same" and "different" images. However, due to the complicated distribution of facial appearance, formulating an appropriate distance function is still very difficult.

Many recent works hence turn to mine domain-specific knowledge or to disentangle various explanatory factors of variation from a large amount of background data [3, 5]. One successful approach of this type is the attribute-based method [3], where the domain-specific knowledge is encoded as a bunch of attribute extractors (gender, race, hair color, etc.) which effectively facilitate similarity measuring. This method is extended in [5] by focusing more on automatically detecting such attribute-like features without human labeling. Their idea is based on the observation that the evidence about how two faces are different is much easier to be identified than how they are similar, and therefore, a large number of so-called "Tom-vs-Peter" classifiers are learned from images of two persons, while these training data are collected from the Web according to their identities. This method achieves state-of-the-art results on the challenging LFW face database [5].

However, there are some limitations in [5]. First, although the "Tom-vs-Peter" classifier significantly alleviates the burden of manual labeling of attributes, one does need to know his/her name before collecting his/her face images. Such a supervised way for data collecting would be less practical than an unsupervised one. Second, the collected data tends to be biased towards celebrated people since they are the people who are most likely to have lots of images per person while images from less familiar people might be ignored. Finally, the collected data may become too complex to be properly handled, since each subject may have a large number of images with large variations. Actually, to deal with such variations, a complicated face alignment algorithm has to be adopted to ensure very good correspondence for the "Tom-vs-Peter" classifier.

To address these issues, we propose a new method for face representation that does not require any manual labeling efforts during or after data collection and is less sensitive to the “celebrated people” problem. The key idea of our method is to collect the background data from the Web using the appearance of a query face instead of his/her name, due to the availability of numerous internet image search engines. In other words, we may utilize such “meta-learners” to model the local variability of the face space near each centre. For this we adopt the same method as [5] by constructing decision stumps between two face groups, and each provides us a ‘view’ about how two faces are different to each other in terms of their local variability, and jointly these decision stumps give a mapping from the image space to a discriminative and distributional representation space.

Another problem in face verification is that the subjects to be verified may be never seen by the system, which means that a fixed pre-trained same person/different person decision rule may be inappropriate since in this case samples from the test subjects may not be regarded as i.i.d ones with those in the training set any more. To address this issue, in this paper we propose a new strategy that essentially allows the verifier adaptive to the specific test sample. Actually the idea of training test-sample-specific model at the testing phase is not new. One of a successful early attempt in this regard is the SVM-KNN classifier [6], which built a new model for each test sample by finding the training data in the region around it. Another idea is to treat each image in a test couple as a single training example and the other one as a test example (hereafter the SVM trained with only one single positive sample and a large number of negative background samples is called Exemplar-SVM following [7]), and this is adopted in the so called One-Shot Similarity (OSS) method [8].

However, the above strategies somehow cut off the connection between the two images in a couple, and this could prevent the utilization of domain specific knowledge which characterizes how two images can be jointly similar (or dissimilar) to each other. Inspired by this, we treat the test couple as a whole as the single training example, and encode the similarity/dissimilarity relationship contained in the couple with an Exemplar-SVM. As a result, the problem of face verification boils down to decide whether this Exemplar-SVM is in nature client-biased or imposter-biased[1].

To this end, we construct an oracle which gains knowledge from client and imposter pairs of the same generic category (i.e., human face) (see Fig. 4). In implementation, it is just a classifier of classier which learns information from the behaviors of a large number of background Exemplar-SVMs, while the latter are trained respectively using a single client instance or a single imposter instance as the ‘Exemplar’. We call our method One-Shot Exemplar-SVM to emphasize the fact that it is used for one shot face verification with never seen persons. Hence our “one shot” is very different from many works which transfer knowledge from

[1] Here the terminologies ‘client’ and ‘imposter’ are used slightly different from those in the usual face verification context: a client instance means that both images in a couple are from the same person while an imposter instance means they are from different subjects.

related domains to *improve* generalization capability of the model learnt with very few training samples, e.g., Bayesian one shot learning [9].

The paper is organized as follows. Section 2, 3 introduces our face representation and the One-shot Exemplar-SVM scheme, respectively. Section 4 shows experimental results. The conclusion is in section 5.

2 Learning Face Representation from the Web

2.1 Approach Overview

As mentioned before our face representation method is partially inspired by the attribute-based representation [3, 5], which can be justified in two aspects: on one hand, there is little relationship between each individual attribute and a particular face image, and in this sense an attribute is highly invariant to various appearance variations from pose, illuminations and so on; on the other hand, combining a number of attributes can rapidly shrink the range of possible face images simultaneously satisfying these conditions. Consequently, different attributes provide us an invariant and abstract representation space, whose effectiveness has been witnessed by many recent successful applications in face verification [3] and object recognition [10].

However, among others, there are several challenges in attribute learning. First, attributes are difficult to be defined, in particular for non-experts it is very hard to define a suitable set of attributes for object representation. In addition, even the attribute set are properly defined, the task of checking the presence/absence of each attribute for a particular face is non-trivial, not to mention the additional labeling about the locations where the attribute appears. This labeling procedure involves exhaustive human labors, and can hardly guarantee high labeling quality for accurate attribute learning. Hence it is not surprising that in literatures there are many works which try to bypass these difficulties, for example, using various automatic attribute naming and discovery mechanisms [11], but unfortunately most of them are not designed for face recognition.

In our method, images in each reference set are essentially concrete instances which in all define a high level attribute template. This is possible since each of our reference set is constructed based on the criterion of visual similarity, meaning that the variations contained in the reference set can be somewhat controlled. In this sense the explored similarities can be thought of as special kind of attributes which are not necessary describable with human thesaurus [2], but might cover various types of attributes like the style of the contour of faces, the facial texture, skin color, affection states, and more complicated attributes that can be understood by human visual cognition. It is worthy mentioning that our method enjoys the same advantages as the previous attribute-based methods [3, 10] such as high-level visual semantic, sharable for images with different

[2] Of course if needed we can still name such template with some complex attribute sentence such as "white middle-aged man with beard", although we do not have to do this considering that our ultimate goal is for classification in this work.

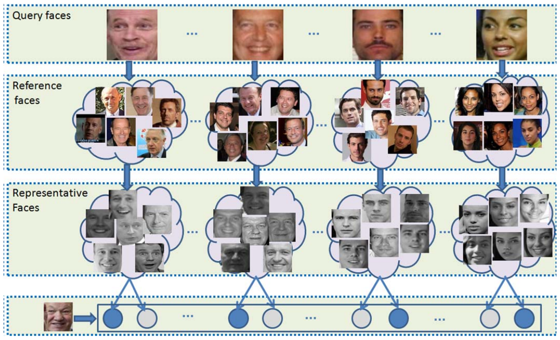

Fig. 2. An overview of the proposed method. A set of reference sets of images are collected from the Web according to their similarity to a query face in appearance. Then some representative images are selected from each of the reference set, serving as the attribute templates for a decision stump. Finally the ensemble of decision stumps maps a given face to its high-level representation.

identities and also description efficiency, and we obtain these cheaply in a pure unsupervised way without any attribute naming and labeling supervision except a query face.

The architecture of the method is shown in Fig. 2. Note that there is a key difference between our method with [5], i.e., we use appearance of the query face to collect the data while [5] uses his/her name. This results in very different reference sets between the two methods.

2.2 The Reference Sets

The reference sets are indexed by its query face and hence it is important to carefully choose diverse and representative face images as queries. To choose these queries, we run K-means (with cosine similarity measure) over the training set of the LFW face database [1] and in each cluster we select the face image that is the most similar to the cluster center as the query. Following this procedure we select 116 query face images among 7701 training images. Note that although these selected query images may appear in the test set of certain training/test set partition, the corresponding labels (information about their identification) are kept unknown to us. This is similar to the strategy adopted in the One-Shot Similarity (OSS) method [8]. All the images in the reference sets undergo the same geometric normalization (detailed in the next section) and are represented using Local Ternary Patterns (LTP) [12] prior to clustering.

Next, to model the local variability of the face space near each centre, we collected 1,000 images for each of the selected 116 faces by searching each face with an image search engine [3]. Note that the search engine we used is appearance based rather than name based. Fig. 2 gives some illustration of the resulting

[3] We used the Baidu image search engine (http://shitu.baidu.com/) in this work.

images and their corresponding query face. One can see that these images are quite similar to the query face in appearance (such as face contour style, skin color, expression and pose) and if the person to be searched is an ordinary people, the search engine just outputs similar face images from other subjects.

One possible criticism to our method is that since the search engine uses its own notion of locality, our results are likely to be influenced by the quality of the search engine output. To reduce such an effect, in our experiment we use only first 1000 most similar faces and check the results using a face detector and a facial points detector [13]. All those images that either don't contain a face or have large appearance deformations are discarded. Finally we obtained 77,408 images in total, belonging to 116 reference sets with each about 667 images.

2.3 Face Alignment

In [5], the "Tom-vs-Peter" classifier is a local classifier in the sense that each of them works on the corresponding small patches of two faces to detect the evidence that the two faces are different. Hence it is essential to use a carefully designed face alignment mechanism to ensure its performance [5], see also [14]. Considering that the reference sets themselves may not be very coherent at the level of a single local region even after alignment and that there are so many possible combinations of patches with different sizes and locations, we take an alternative method which relies on the global representation of appearance of faces to construct the decision stumps. This means that a relatively simple alignment would be enough (actually the low-level features we adopted are tolerant to the misalignment to some extent, see below).

Specifically, for each image a Viloa and Jones detector is first used to detect the boundary of its face region, then a face region as large as 1.5 times the radius of the detected face is cropped from the original image. Then we run the congealing alignment algorithm [15] over these cropped images for coarse alignment, which helps to reduce the variations in pose and scale. The congealing method is an unsupervised alignment approach which learns a particular affine transformation for each facial image such that the entropy of a group of them is minimized.

To get better global representation of a face, instead of directly cropping the face from the resulting image of congealing, we further fine-tune it with the positions of 9 key facial landmarks (see Fig. 3) estimated using the method of [13]. Based on these landmarks, we first rotate the facial image such that the straight line connecting the centers of two eyes coincides with the horizontal line. Next we should estimate a bounding box for face cropping, which has to be robust enough to tolerate the slight errors in landmark localization. Figure 3 illustrates the measurements we use to calculate the shape parameters which define the bounding box (i.e., its center, width and length).

In particular, to estimate the face width, we first estimate the width of eyes w_{eye} by averaging the width of both eyes (w_1,w_2) and the distance between the two eyes (w_3)(from the left eye corner to the right eye corner) based on

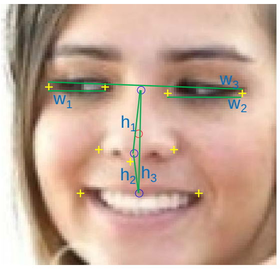

Fig. 3. Illustration of the measurements for face cropping in the final alignment, where the yellow crosses denote the nine facial landmarks estimated by [13].

the positions of four eye landmarks, i.e., $w_{eye} = (w_1 + w_2 + w_3)/5$, and then empirically estimate the face width as $w_{eye} \times 4.2$.

Similarly, we estimate the face height with the eye-mouth height h_{avg}, which is estimated by averaging the following three measurements: (1) the height between the center of two eyes and nose center (h_1); (2) the height between the nose center and the mouth center (h_2); (3) the height between the center of two eyes and the mouth center (h_3). Then the final face height is empirically estimated as $h_{avg} \times \frac{5}{3}$.

Finally, we assume that the face region is centered at the mean position (indicated by the red circle in Fig. 3) of the nine landmarks.

2.4 Mapping to the Representation

As the final step for our face descriptor, we should construct a series of discriminative decision stumps which jointly map a given face to a sequence of visual bits. The 'visual bits'-type descriptor is very popular in computer vision due to its capability to capture different aspects of the image information in a distributed and compact way.

This can be simply implemented as a binary classifier (decision stump) trained using a low-level feature on images of two different reference set. Not that we use the whole face instead of a local region as the input to the decision stump, as mentioned in Section 2.3. For n reference set and k low-level features, we will have $D = k \cdot n(n-1)/2$ binary classifiers. Here we use linear support vector machines which selects some most representative samples from the reference set and compares them to the test face to make a binary judgement. Denote the i-th decision stump as h_i, one can see that these decision stumps define for a test face x a map which transforms the face to a high-dimensional representation, i.e., $h(x) = (h_1(x), h_2(x), ..., h_D(x)) \in R^D$.

As for low-level features, we use a texture descriptor and a region descriptor which capture the local texture details and local shape information of the face respectively. In particular, for texture information extraction we use the Local Ternary Pattern (LTP) [12], which is a simple generalization of Local Binary Patterns (LBP) with 3-valued codes in discretization of the difference between

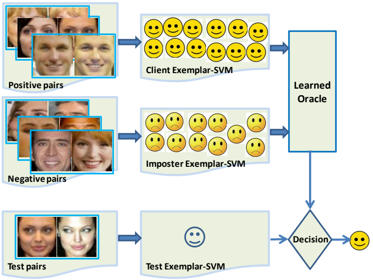

Fig. 4. Overview of the proposed One-Shot Exemplar-SVM method for face verification, where a smiling face icon denotes a client Exemplar-SVM model while an unhappy face icon denotes an imposter Exemplar-SVM. For a test Exemplar-SVM from a pair of test faces with never seen subjects, we want to know which kind of Exemplar-SVM it is, by asking the oracle learnt from the training data.

the central pixel and its surrounding pixels, tackling challenging conditions such as uneven illumination and image noise. While for local shape information we use a variant of Histogram of Oriented Gradients (HOG) [16] called Histogram of Principal Oriented Gradients (HPOG) proposed in [17]. In HPOG the gradient information at each pixel is computed using the eigenvector with the largest eigenvalue of a 2×2 covariance matrix which models the gradient distribution in a neighboring region of that pixel. We expect that such a gradient smoothing operation could help to alleviate the influence of small appearance changes due to image blur, noise, low resolution, etc.

3 One Shot Exemplar-SVM

In this section we give a detailed account on our One-Shot Exemplar-SVM method for face verification with never seen subjects. Fig. 4 gives an overview of the proposed method. Here, instead of pre-training one fixed verification model, we allow our model adaptive to the test pair. For this we train an Exemplar-SVM for the test face pair with a held-out set of imposter pairs as negative instances. Note that although the images of the test pair may be either from the same person or from different persons, we always treat it as the single positive instance. As a result, the problem of face verification is transformed into a problem of deciding whether this model is a client model (i.e., trained with a matched pair) or an imposter model (i.e., trained with a unmatched pair). In other words, a problem of comparing two faces becomes now a problem of model comparison. To make the final decision, we use an oracle which is learnt from behaviors of those client/imposter Exemplar-SVMs. The motivation to adopt this strategy is based on the assumption that face verification at the model level could be easier than that at the feature level, since a model can be thought of as the generalization of samples which essentially compensates for the insufficient information of a single instance.

3.1 Learning Exemplar-SVMs

Given a face pair, we first extract their high level representation from each face using the method described above, then concatenate the absolute difference and element-wise product of the two face descriptors as the representation for the pair, following [3]. To train an Exemplar-SVM for the pair, we use it as a single positive instance and a fixed set of 2700 pairs with imposter pairs as negative instances. Then we train the model by optimizing the following convex objective,

$$\begin{aligned} \min_{w,b} \tfrac{1}{2}||w||^2 &+ C_1 \max(1-(w^T x_E + b), 0) \\ &+ C_2 \textstyle\sum_{x \in \mathcal{N}_E} \max(1-(-w^T x - b), 0) \end{aligned} \tag{1}$$

Where the x_E is the exemplar of face pair and the $\mathcal{N}_E$ is the negative pairs set. C_1 and C_2 are the loss penalty coefficients for positive and negative samples respectively. In our setting, we set $C_1 = 0.5$ and $C_2 = 0.01$ (as did in [7]) for balancing the loss penalty between the two classes. Then we calibrate the output of the exemplar-SVM by fitting them to a sigmoid function on a validation set, which contains 5400 instances with both positive and negative pairs. Then the prediction is calibrated as follows.

$$f(x|w_E, \alpha_E, \beta_E) = 1/(1 + e^{-\alpha_E(w_E^T x - \beta_E)}) \tag{2}$$

where, w_E is the parameter of the learned Exemplar-SVM, α_E and β_E are the Sigmoid parameters.

3.2 Training the Oracle

To train the oracle, we use the Exemplar-SVMs as examples, including 2700 client Exemplar-SVM and 2700 imposter ones. These are trained respectively with 2700 positive pairs and 2700 negative pairs in the training set of the LFW face dataset, with each positive pair as the single positive instance for the client Exemplar-SVM and each negative pair for the imposter Exemplar-SVM. In training both kinds of Exemplar-SVMs we share the same 2700 negative pairs as negative instances. After this we can observe the behavior of a learned Exemplar-SVM on all the 5400 training face pairs by sending each pair into this Exemplar-SVM and concatenating the responses as a 5400-D vector. These response vectors are in turn used for training the oracle, which is a linear SVM in our case. See the top of Fig. 4 for illustration.

3.3 Face Verification Using the Oracle

For a test pair of images, we first train an Exemplar-SVM for it using the same 2700 negative pairs mentioned before, then attach a 5400-D response vector for the obtained model by running it on the 5400 training pairs, which is very efficient (about 0.027 ms per pair on average). We pass this response vector to the oracle to make the same-or-different decision.

4 Experimental results

In this section, we conduct a series experiments to validate the effectiveness and feasibility of the proposed method on the LFW face database [1], which contains 13233 face images of 5749 people collected from the Web, with large variations in pose, expression and illumination etc. These images are divided into ten folds with each containing 300 matched pairs and 300 unmatched pairs, and the identities between folds are mutually exclusive, which means that the subjects in the test fold will be never seen in the training folds. We follow the "image-restricted" evaluation protocol, in which only a same or different label is assigned to each face pair without any identity information about each face.

In what follows we first evaluate the effectiveness of the proposed method in two aspects, i.e., the face representation and the One-Shot Exemplar-SVM verification method, then we compare our method with other related state-of-the-art methods.

4.1 Effectiveness of Unsupervised Representation Learning

To demonstrate the advantage of the proposed unsupervised representation learning method, we compare its verification performance with that of the several other face descriptors. In particular, following the method of [3], we first construct those descriptors from each single face image, then represent a pair of images by concatenating the absolute difference and element-wise product of the two face descriptors. We use the standard linear SVM trained with the LIBLINEAR package [18] as the verifier.

- Low-level features: Use the LTP [12] and the HPOG [17] for face description as described in Section 2.4
- Attribute-based representation: Directly use the attribute data provided by [3].
- Random Splitting: Replace the query face indexed reference sets with randomly partitioned groups of face images.

Table 1 lists the results. Although directly using our low-level feature for verification looks effective (77.6%), its performance is much inferior to that of more high-level representation, such as the attribute based representation (84.75%). However, if we replace the original attribute representation [3] with our representation, the verification improves to 86.70%. In contrast to [3], our representation is constructed in a completely unsupervised way without any human supervision in terms of attribute labeling.

To further understanding our method, instead of grouping collected images according to query faces, we randomly split those data and use them for decision stumps training. One can see from Table 1 that such a random splitting degrades the performance significantly by over 20.0%, which indicates the importance of using relative consistent faces for constructing high-level features. As shown in Fig. 2, images in each reference set are similar to each other in many aspects,

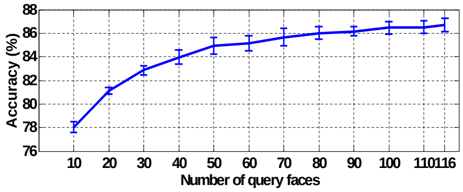

Fig. 5. The influence of the number of query faces on the verification performance in the restricted setting of LFW, using linear SVM as the verifier.

including the face shape, gender, age and so on, and that's the major reason why we regard them as sampling from a common (complex and unnamable) attribute template.

Method	Accuracy
Low-level features (LTP+HPOG)	0.7764±0.0056
Random splitting	0.6083±0.0083
Attribute-based representation [3]	0.8475±0.0051
Our method	**0.8670**±0.0057

Table 1. Comparison of verification performance of our method with other face representation methods in the restricted setting of LFW [1], using the linear SVM as the verifier.

We then investigate the influence of the different number of query faces to the performance. In particular, we conduct a series of experiments by varying the number of reference sets from 10 to 116 with the step as 10 and Fig. 5 gives the results. One can see a general tendency of increasing verification performance with the increasing of the number of reference sets. Specifically, our system reaches the performance of 78.03% with only 10 query faces, and the speed of performance improvement begins to slow down when 80 reference sets are used, with an accuracy of 86.02%, which is nearly comparable to the best performance of 86.7% achieved using 30 more query faces.

4.2 Effectiveness of the One-Shot Exemplar-SVM

To investigate the effectiveness of the proposed One-Shot Exemplar-SVM scheme, we compare it with several closely related verification methods. All these classifiers are based on our face representation method.

- KNN (feature level): Since KNN can also be regarded as a method which adjusts its model according to the test instance, we include it for comparison as well.

- KNN (model level): The same as above, but at the level where each pair is represented as an Exemplar-SVM model and the similarity between two models is measured in the same way as described in Section 3.2.
- SVM-KNN (feature level) [6]: Each time pair of images is verified by a new model trained using K nearest neighbors of the test couple.
- SVM-KNN (model level): A variant of [6] in which the K nearest neighbors are the Exemplar-SVMs on the training set, instead of a pair of images.
- One Shot Similarity (OSS) [8]: This implementation is based on the codes provided by the authors.

Method	Accuracy
KNN (feature level)	0.8083±0.0047
KNN (model level)	0.8442±0.0061
SVM-KNN (feature level)[6]	0.8332±0.0059
SVM-KNN (model level)	0.8392±0.0061
One-Shot Similarity (OSS) [8]	0.8650±0.0042
One-Shot Exemplar-SVM (ours)	**0.8805**±0.0054

Table 2. Comparison of the verification performance of the proposed method with closely related methods in the restricted setting of LFW, using our face representation method.

Table 2 gives the results. One can see from the table that our One-Shot Exemplar SVM scheme yields the best performance among the compared ones. This is mainly due to its capability to generalize beyond the test samples. Actually for both KNN and SVM-KNN [6], their model level version consistently gives better performance than the corresponding feature level version. The One Shot Similarity Kernel method [8] works better than the SVM-KNN on this dataset, but it models each image in a couple independently, while our method effectively exploits the correlation between the two images in a couple by incorporating them in a single model.

One obvious concern about our Exemplar-SVM is its performance since it is assumed to be a weak classifier trained with only one single positive instance. Figure 6 gives the histogram of verification accuracy of the trained Exemplar-SVMs over the training data. One can see that the accuracy of client Exemplar-SVMs tends to be distributed in a single mode at 73.3% with a narrow support ranged from 60.0% to 80.0% while that of the imposter Exemplar-SVMs is distributed in a relatively flat way. Such a difference is exploited by the oracle to make a prediction for a new Exemplar-SVM trained on the test pair.

Furthermore, Fig. 7 details the behavior of two typical Exemplar-SVMs (one client and one imposter Exemplar-SVM) over the training face pairs. The figure reveals that the behavior of a client Exemplar-SVM is very different to that of an imposter one. This is reasonable since the imposter Exemplar-SVM trained with an imposter pair serves only as a background model which is almost not sure of anything.

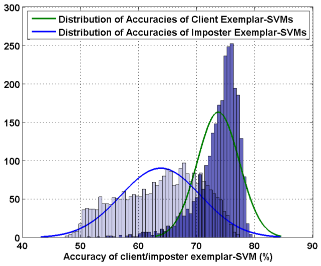

Fig. 6. Histogram of verification accuracy of Exemplar-SVMs over the training data.

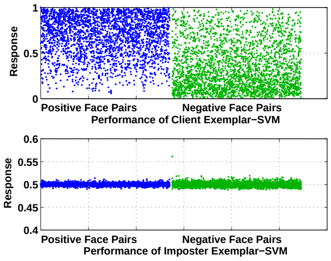

Fig. 7. Illustration of the typical behavior of the client/imposter Exemplar-SVM on labeled face pairs.

4.3 Comparison with state-of-the-art methods

We now compare our method with other related state-of-the-art methods on the LFW data set [1]. Table 3 lists the results. Note that each of these published results varies in its feature extraction (the first group in the table) or similarity measuring techniques (the second group), while our methods are most related to those methods in the third group in terms of methodology. It can be seen that our method outperforms most of these methods, and is comparable to [19], but loses about 2.0% in performance compared to the Tom-Peter Classifier with affine aligned faces [5]. However in the "Tom-vs-Peter" method [5] the identity of each face collected from the Web is known to the model (note that this is a rather strong assumption and thus making it not comparable to ours directly).

Method	Accuracy
LARK supervised, aligned [20]*	0.8510±0.0059
CSML + SVM, aligned [19]	0.8800±0.0037
Fisher vector faces [21]	0.8747±0.0149
SIFT Sub-SML, funneled [22]*	0.8642±0.0046
Nowak, funneled [4]	0.7393±0.0049
LDML-MkNN, funneled [2]*	0.8750±0.0040
LBP multishot, aligned [8]*	0.8517±0.0061
Attribute and Simile classifiers [3]	0.8529±0.0123
Tom-vs-Peter Classifier, **affine aligned** [5]	0.9047±N/A
Tom-vs-Peter Classifier, **full** [5]	0.9310±N/A
One-Shot Exemplar-SVM (ours)	**0.8805**±0.0054

Table 3. Comparison of the proposed method with other related state-of-the-art methods on the LFW [1], where the methods marked with '*' means they are evaluated in the image unrestricted setting otherwise in the restricted setting.

5 Conclusions

In this paper we propose a new method to deal with two key problems in unconstrained face verification. First, to address the large variation problem, we propose an unsupervised face representation learning method from the Web, with the major advantages of effectiveness and convenience in practice since no supervision is required in terms of the attribute or the identity labeling. Another problem we addressed concerns how to verify pair of images from never seen subjects and we propose the One-shot Exemplar-SVM scheme which is characterized by making the prediction at the model level rather than that at the feature level. Experiments on the challenging LFW database show that our method achieves encouraging verification performance comparable to other related state-of-the-art algorithms. Last but not least, it is worthy mentioning that the best performer on the LFW database has achieved an accuracy as high as 99.15% [23] based on the deep learning technique. Nevertheless, in our opinion exploring alternative methods like ours to learn feature representation from unsupervised data in a more efficient and more interpretable way is still useful.

Acknowledgement. The authors want to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (61073112, 61035003, 61373060), Jiangsu Science Foundation (BK2012793), Qing Lan Project, Research Fund for the Doctoral Program (RFDP) (20123218110033).

References

1. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)
2. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: ICCV. (2009) 498–505
3. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Describable visual attributes for face verification and image search. PAMI **33** (2011) 1962–1977
4. Nowak, E., Jurie, F.: Learning visual similarity measures for comparing never seen objects. In: CVPR. (2007)
5. Berg, T., Belhumeur, P.N.: Tom-vs-Pete classifiers and identity-preserving alignment for face verification. In: BMVC. Volume 1. (2012) 5
6. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: discriminative nearest neighbor classification for visual category recognition. In: CVPR. Volume 2. (2006) 2126–2136
7. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of Exemplar-SVMs for object detection and beyond. In: ICCV. (2011) 89–96
8. Wolf, L., Hassner, T., Taigman, Y.: The one-shot similarity kernel. In: CVPR. (2009) 897–902
9. Fe-Fei, L., Fergus, R., Perona, P.: A Bayesian approach to unsupervised one-shot learning of object categories. In: ICCV. (2003) 1134–1141
10. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)

11. Berg, T., Berg, A., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: ECCV. (2010) 663–676
12. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. TIP **19** (2010) 1635–1650
13. Everingham, M., Sivic, J., Zisserman, A.: “Hello! My name is... Buffy” – automatic naming of characters in TV video. In: BMVC. (2006)
14. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2233–2246
15. Huang, G.B., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: ICCV. (2007)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893
17. Song, F., Tan, X., Liu, X., Chen, S.: Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. Pattern Recognition **47** (2014) 2825–2838
18. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. JMLR **9** (2008) 1871–1874
19. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: ACCV. (2011) 709–720
20. Seo, H.J., Milanfar, P.: Face verification using the lark representation. TIFS **6** (2011) 1275–1286
21. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: BMVC. (2013)
22. Cao, Q., Ying, Y., Li, P.: Similarity metric learning for face recognition. In: ICCV. (2013)
23. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. CoRR **abs/1406.4773** (2014)