Computer Vision and Image Understanding xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Exploiting relationship between attributes for improved face verification ☆

Fengyi Song, Xiaoyang Tan *, Songcan Chen

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

ARTICLE INFO

Article history:
Received 27 March 2013
Accepted 19 February 2014
Available online xxxx

Keywords:
Attribute relationship graph
Attribute-graph regularized SVM
Face verification

ABSTRACT

Recent work has shown the advantages of using high level representation such as attribute-based descriptors over low-level feature sets in face verification. However, in most work each attribute is coded with extremely short information length (*e.g.*, "is Male", "has Beard") and all the attributes belonging to the same object are assumed to be independent of each other when using them for prediction. To address the above two problems, we propose a discriminative distributed-representation for attribute description; on the basis of this description, we present a novel method to model the relationship between attributes and exploit such relationship to improve the performance of face verification, in the meantime taking uncertainty in attribute responses into account. Specifically, inspired by the vector representation of words in the literature of text categorization, we first represent the meaning of each attribute as a high-dimensional vector in the subject space, then construct an attribute-relationship graph based on the distribution of attributes in that space. With this graph, we are able to explicitly constrain the searching space of parameter values of a discriminative classifier to avoid over-fitting. The effectiveness of the proposed method is verified on two challenging face databases (*i.e.*, LFW and PubFig) and the a-Pascal object dataset. Furthermore, we extend the proposed method to the case with continuous attributes with promising results.

1. Introduction

Recently, there has been growing interest in using middle-to-high level feature descriptors for face representation. One typical example is the attribute descriptors [1–9]. N.Kumar et al. [3,4] have recently shown that using the outputs of a series of component classifiers with each tailored to some particular aspects of the human face images, called visual attributes, they are able to achieve close to state-of-the-art performance of face verification on the challenging Labeled Faces in the Wild (LFW) [10]. This result is interesting in several aspects. Firstly, the number of features used in their work is very small (*i.e.*, only 73 attributes), which means that it provides a very economical but powerful way to describe faces. This is in sharp contrast with the commonly used low-level features in image description, such as pixel values, gradient directions, scale-invariant feature transform (SIFT) [11], where usually thousands of features are needed. Secondly, the attribute descriptor is user-friendly in that its meaning is understandable to human beings (everyone knows what "white male" means) while the meaning of most previously mentioned low-level features is less intuitive to us. Last but not least, such a descriptor is generalizable and sharable, which makes it particularly suitable for such problems as zero-shot learning [12,13] or between-class transfer learning [2,14].

However, in most work each attribute is coded with extremely short information length (*e.g.*, using binary code such as "is Male", "has Beard") and all the attributes belonging to the same object are assumed to be independent of each other when using them for prediction. The one-bit information length of attribute coding makes the representation less stable, and could bring trouble to many interesting subsequent processing tasks, such as modeling the similarity between attributes. Actually, research in the field of cognitive discovery has shown the usefulness of the relationship between feature sets. For example, Bhatt and Rovee-Collier [15] experimentally showed that infants as young as three months of age gain the capability to encode the relations among object features, and use such a feature configuration for general object recognition. However, traditionally one of the major challenges in modeling the feature configurations lies in the huge number of low-level features (*e.g.*, the dimension of a 100×100 face image is as high as $10,000$ using the gray-value features). In addition, it

☆ This paper has been recommended for acceptance by Aleix M Martinez.

* Corresponding author. Fax: +86 25 8489 2452.
E-mail addresses: f.song@nuaa.edu.cn (F. Song), x.tan@nuaa.edu.cn (X. Tan), s.chen@nuaa.edu.cn (S. Chen).

http://dx.doi.org/10.1016/j.cviu.2014.02.010
1077-3142/

is very difficult for a human being to understand what exactly such a big feature configuration mean. Fortunately, both aforementioned problems can be addressed by the attribute descriptors due to its high level and compactness in object description. Indeed, despite the partial success of using attribute descriptors by treating them statistically independent of each other [1,3,4,16] or conditionally independent given the class label [2], recent work has shown that it is beneficial to exploit the relationship between attributes under various contexts [5,17–19]. Some of them will be discussed in the next section.

In this work, we propose a discriminative distributed-representation for attribute description; on the basis of this description, we investigate how to model the similarity relationship between attributes and how such relationship could be exploited to improve the performance of face verification. The idea of distributed representation was first introduced by Hinton [20], and successfully applied in statistical language modeling [21]. In this work, we develop a new distributed representation for each individual attribute by taking the information of subject identification into account. The method is inspired by the vector representation of words in the literature of text categorization, and the meaning of each attribute is embedded into a high-dimensional vector in the subject space (*cf.* Fig. 1). Such a representation allows us to model the similarity between attributes in a much stable and reliable way. In particular, we construct an attribute-relationship graph based on the distribution of attributes in the subject space, which effectively encodes the pairwise closeness relationship between any two attributes. For example, a "male" attribute is highly related to such attributes as "wearing necktie", "bushy eyebrows", "beard", and so on (*cf.* Fig. 9). To exploit such information for prediction, we integrate the attribute-relationship graph into a linear classifier to constrain the searching space of its parameters, based on the assumption that similar attributes should have similar weights. This is helpful to avoid over-fitting and improve the generalization capability of the learned classifier. The uncertainty in attributes responses is also taken into account in the final model.

This journal paper builds on the earlier conference work [22]. In this extended version, we extend above ideas and merge them into a single framework, which works for both discrete and continuous attributes. The effectiveness of the proposed methodology is empirically verified with encouraging results on two large-scale face databases, one object classification dataset and several UCI data sets. In what follows, we first review the related work in Section 2 and then present the proposed method in Section 3. Extensive experimental results are given in Section 4. Finally, we conclude this work in Section 5.

2. Related work

Recently, attribute-based representation has been extensively researched in and beyond the field of face recognition [3,4], including object recognition [5,9,23], scene understanding [18,24], image retrieval [25,26], activity analysis [27–29], and shows special advantages in active learning [30–32], transfer learning [2,12,14] and zero-shot learning [13]. Since this work is mainly about attribute representation and modeling their relationship, in what follows, we will not go into the details on how to extract attributes and apply them in various applications, but first give a brief discussion on how to define attributes and then focus on the related work on building the relationship between attributes.

2.1. Attribute definition

To use the attributes, we have to define them firstly. Attribute definition is the process of deciding which visual qualities should be used for depicting the objects or events. Most attributes are manually specified with respect to different application scenarios. These attributes are usually semantically understandable and can be seen as concepts in natural language. The specified attributes are then extracted from the images based on some low-level features. In this way, attribute can be thought of as a high-level representation which incorporates human understandable concepts into the machine learning process in a reasonable way.

Although attribute description is mostly intuitive, building a suitable taxonomy of attributes for a particular task is not easy. In [3], the authors proposed to describe each face with 73 attributes (*cf.* Fig. 2), which can be roughly categorized into four types: (1) appearance description of key facial parts, such as the shape, size and style of the nose, mouth, eyes, eyebrow, jaw, and hair; (2) high-level semantic features like gender, age, and ethnicity; (3) specification about imaging conditions, *e.g.*, lighting, expression, posture, accessory, and the environment; and (4) personal specific traits like bald, goatee, and attractiveness. In [9], by surveying multiple online cataloges, the authors produced 26 common attributes to describe clothing, covering 6 patterns, 11 colors, and 6 miscellaneous characteristics such as wearing the necktie or the scarf, and the collar or the placket presence. Patterson and Hays [33] gave a comprehensive discussion on attribute definition, discrimination and predictive power of attribute in the context of scenes description.

2.2. Attribute relationship exploitation

We now review how to model and exploit the relationship between attributes. As mentioned before, this is not trivial because an attribute is usually simply represented as a binary bit to denote its presence/absence. Despite this, there is some work which exploit various types of attribute relationship in different contexts to improve the performance of prediction.

In [19], the concept of binary attribute was introduced to describe the spatial relationship between a pair of attributes corresponding to two image segments respectively. Such relationship was shown to be very effective in describing simple geometric

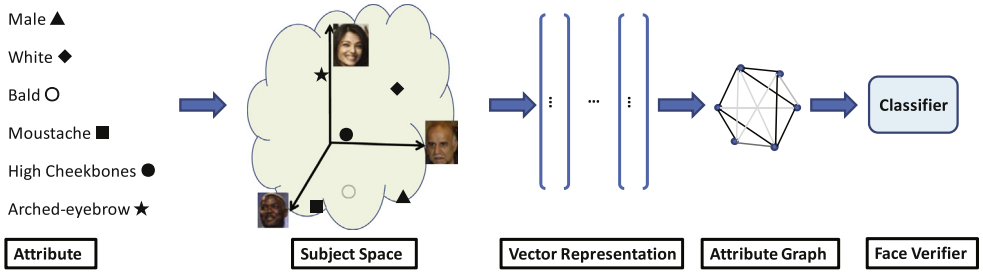

Fig. 1. The overall pipeline of the proposed algorithm. Each attribute descriptor is first projected into a common subject space to obtain a high-dimensional vector representation, which are then used to construct an attribute graph. The graph is finally exploited to regularize the objective of a linear SVM-based face verifier.

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

Facial Appearance Description: { Non-Black Hair, Non-Blond Hair, Non-Brown Hair, Curly- Hair, Wavy Hair, Non-Straight Hair, Receding Hairline, Non-Bangs, Non-Sideburns, Fully Visible Forehead, Non-Partially Visible Forehead, Non-Obstructed Forehead, Non-Bushy Eyebrows, Non-Arched Eyebrows, Narrow Eyes, Eyes Open, Big Nose, Non-Pointy Nose, Big Lips, Non-Mouth Closed, Mouth Slightly Open, Non-Mouth-Wide-Open, Non-Teeth Not Visible, Non-No Beard, Non-Goatee, Round Jaw, Double Chin, Non-Oval-Face, Non-Square Face, Non-Round Face, Non-Gray Hair, Bags Under Eyes, Non-Heavy Makeup, Non-Rosy Cheeks, 5 o'Clock Shadow, Strong Nose-Mouth-Lines, High-Cheekbones, Brown Eyes}

High-level Semantic Features: { Male, Non-Asian, White, Non-Black, Non-Baby, Non-Child, Non-Youth, Non-Middle Aged, Non-Senior, Non-Indian }

Imaging Conditions: { No Eyewear, Non-Eyeglasses, Non-Sunglasses, Non-Mustache, Smiling, Non-Frowning, Non-Chubby, Non-Blurry, Non-Harsh-Lighting, Non-Flash, Soft Lighting, Non-Outdoor, Non-Wearing Hat, Color Photo, Posed Photo, Non-Shiny Skin, Non-Pale-Skin, Non-Wearing Lipstick, Non-Flushed Face, Non-Wearing Earrings, Wearing Necktie, Non-Wearing Necklace }

Specific Traits:{ Bald, Non-Attractive Man, Non-Attractive Woman}

Fig. 2. Illustration of the 73 attributes descriptors used for face verification [4] (The figure is best viewed in electronic form).

patterns like stripes. In [5], Wang and Mori exploited more general relationships between attributes to improve the performance of object recognition. Particularly, they treated the correlations among attributes as augmented feature sets in the latent SVM framework [34]. However, one unwelcome consequence of this is that the number of possible combinations between attributes grows quadratically with the number of attributes. To address this issue, they had to simplify the undirected graph that encodes the attribute correlations to a tree by keeping only highly related attributes while pruning others. Recently, Parikh and Grauman [18] proposed the relative attribute descriptors to model the relative strength of disagreement among instances for each attribute, which resulted in a user-friendly way for object description. Using this method, for example, you do not have to describe explicitly whether a man is smiling when it is difficult to make such judgment, but only need to say that his expression is roughly between smiling and not smiling. In [17], even higher-order relationship between attributes was explored. They built for each attribute a regressor from all the other attributes responses and used the output of each regressor as the corresponding attribute value. In this way, the attribute response is "denoised".

Our method is different from the aforementioned ones in several ways. Firstly, all the above methods have shown its effectiveness in their particular contexts, *e.g.*, object category recognition [19,5] or scenario analysis [18], but little work addressed the question of whether this is true in face verification as well, which is exactly what we do in this work. Secondly, our way to model the attribute relationship is different from all the above methods, though it is closer to [5]. In particular, instead of learning a pairwise relationship between attributes independently as in [19,18], we try to model an attribute-relationship graph based on the understanding of the meaning of the attributes in a more general context of subjects to whom each attribute belongs (see Section 3.1 for more details). Finally, in contrast with previous work [18,5] where relationships among attributes are used as feature sets to augment the *input* of classifiers, we exploit attribute relationship to *improve* the generalization capability of the classifier in a more straightforward way, *i.e.*, by using it as a prior constraint on the searching space of model parameters.

In the field of machine learning, the graph-based prior is commonly adopted to control the model complexity of the learner. A typical example is the Laplacian SVM method proposed by Belkin et al. [35]. In their method, an instance-graph is organized to constrain the label value of neighboring instance, based on the manifold assumption that similar instances should have similar labels. Our method is similar to this, but instead of constructing an instance-graph, we build an attribute-relationship graph. One advantage of the attribute-graph is that its complexity is controllable since its size does not grow with the number of instances as in [35] but only with the number of attributes, and the latter is usually not too large in practice as mentioned before. Furthermore, our graph is not meant to constrain the output space of instances but the searching space of model parameters, based on a simple idea that similar attributes should play similar roles in the learned classifier.

In this sense, our method can also be thought of as a mechanism to automatically regularize the coefficients of the linear classifier using graph-based prior knowledge, and hence is related to many norm-based (*e.g.*, L_2 or L_1 norm) regularization methods in machine learning. Among them, our method is most related to those group-lasso-like methods commonly seen in the multi-task learning literature [36–38], where some groups of coefficients survived while other groups are forced to be quiet during optimization. However, there is no any within-group regularization except sparsity is imposed in those methods, while, in our method, we do not intend to cancel the contribution of any single coefficient but emphasize that the consistency between coefficients is of importance.

3. The approach

In this section, we give a detailed description of the proposed approach, whose overall pipeline is presented in Fig. 1.

3.1. Modeling the attribute relationship

Assuming that we are given a set of M attribute descriptors $A = \{A_i \in \{0,1\}\}_{i=1}^{M}$ for each face image. Although the meaning of each attribute is clear to human beings (see Fig. 2), the way to represent each attribute as a binary code might be too simple from the respect of subsequent processing. Therefore, we still need to find a method to properly represent each attribute in a richer manner so that they are computationally convenient to support the advanced inference.

One commonly used trick in computer vision for this purpose is to think of each face as a document which is described by words (attributes) [24,39]. Although this analogy between word and attribute is not so perfect, it makes it possible to borrow a great amount of ideas from textual analysis to represent the meaning of the attributes. One particular way we choose is the so-called featural representation [40], which is proven to have explanatory value by representing the word meaning as featural primitives.

To construct such featural primitives, we use the subjects available in the training set and call the space spanned by these subjects space (see Fig. 1). Hence for K subjects, we have a subject space

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

with K dimensions and the meaning of each attribute is represented as a high-dimensional vector in the subject space, with each entry representing whether the corresponding subject owns such an attribute. For several images of the same subject, the value of the corresponding entry is accumulated and then normalized by the total number of images of the subject.

After projecting all the attributes into the subject space, we may model their relationship based on the distribution of each attribute in an information theoretic framework. In particular, we first compute the point-wise mutual information $I(A_i, y_j)$ of each attribute A_i with each subject y_j, which are then collected as another vector AK_i,

$$AK_i = (I(A_i, y_1), I(A_i, y_2), \ldots, I(A_i, y_K)) \tag{1}$$

where $I(A_i, y_j)$ is defined to be,

$$I(A_i, y_j) = \log_2 \frac{p(A_i, y_j)}{p(A_i)p(y_j)} \tag{2}$$

After this, correlated information encoded by M attributes and K subjects is organized as the following matrix (Eq. (3)), based on which, the attribute graph can be constructed by treating each row as a node.

$$\begin{pmatrix} & y_1 & \cdots & y_j & \cdots & y_K \\ A_1 & I(A_1,y_1) & \cdots & I(A_1,y_j) & \cdots & I(A_1,y_K) \\ A_2 & I(A_2,y_1) & \cdots & I(A_2,y_j) & \cdots & I(A_2,y_K) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_i & I(A_i,y_1) & \cdots & I(A_i,y_j) & \cdots & I(A_i,y_K) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_M & I(A_M,y_1) & \cdots & I(A_M,y_j) & \cdots & I(A_M,y_K) \end{pmatrix} \tag{3}$$

Before proceeding, we briefly discuss how to calculate the mutual information in Eq. (2). There are three statistics involved, *i.e.*, $p(A_i, y_j)$, representing the probability of co-occurrence of the attribute A_i and the person y_j; $p(A_i)$ and $p(y_j)$, representing the probability of occurrence of the attribute A_i and the person y_j respectively. They are empirically evaluated using the Maximum Likelihood Estimation (MLE) method through the training set, as follows,

$$\begin{aligned} p(A_i, y_j) &= \frac{\text{number of images of person } y_j \text{ with attribute } A_i}{\text{number of images of person } y_j} \\ p(A_i) &= \frac{\text{number of images with attribute } A_i}{\text{total number of images}} \\ p(y_j) &= \frac{\text{total number of images of person } y_j}{\text{total number of images}} \end{aligned} \tag{4}$$

To improve the reliability of the MLE estimation for subjects with only a few face images, we use the Laplace smoothing strategy [41], *i.e.*, merely adding a constant to each count.

Finally, the attribute graph is built by computing the similarity between two attributes nodes through commonly used similarity measures such as Cosine similarity or Heat Kernel,

$$s_{ij} = \frac{AK_i^T AK_j}{\|AK_i\| \cdot \|AK_j\|} \text{ or } s_{ij} = e^{-\frac{1}{\sigma}\|AK_i - AK_j\|_2^2}, \quad i, j = 1, 2, \ldots, M \tag{5}$$

In our implementation, the Heat Kernel is adopted, and a brief discussion on the choice over the two similarity measures is given in Section 4.3.2. Also note that the size of our attributes graph depends only on the number of attributes but is independent with the number of subjects or the number of images.

3.2. Exploiting the attribute-graph model

Given a set of training data $D = \{x_i, y_i\}_{i=1}^N$, our goal is to estimate the posterior of the model parameter w. With the criterion of maximum a posteriori probability (MAP), we have $p(w \mid D) \propto p(D \mid w)p(w)$, where $p(D \mid w)$ is the likelihood while $P(w)$ is the prior on the distribution of w. This formulation has an equivalent form,

$$log(p(w \mid D)) = log(p(D \mid w)) + log(p(w)) + c \tag{6}$$

where c is a constant. According to this, MAP criterion is equivalent to minimize the total energy of the likelihood model and the prior model. In this work, we use the linear SVM as our base classifier. With the hinge loss, the objective energy function of linear SVM is,

$$\min_w \sum_{i=1}^N \max\{0, 1 - y_i(w^T x_i + b)\} + \frac{\lambda_1}{2} w^T w \tag{7}$$

Note that although it is a linear model, it may still face the risk of over-fitting since it works in a high-dimensional space and the number of training samples is small. To further control the complexity, we use the attribute-graph as one of the prior constraints,

$$\min_w \sum s_{ij}(w_i - w_j)^2 \tag{8}$$

where $w = (w_1, w_2, \ldots, w_M)$ are the model parameters and s_{ij} is defined in Eq. (5). Using the standard spectrum technique, we construct the Laplacian matrix L of the attribute-graph as $L = D - S$, where D is a diagonal matrix with $D_{ii} = \sum_j s_{ij}$. With these notations, it is well-known that Eq. (8) can be reformulated as $w^T L w$, and we add this to the standard SVM objective function,

$$\min_w \sum_{i=1}^N \max\{0, 1 - y_i(w^T x_i + b)\} + \frac{\lambda_1}{2} w^T w + \frac{\lambda_2}{2} w^T L w \tag{9}$$

Sometimes the uncertainty in the attribute response is available to us (*e.g.*, [3]), and we may take this into account. Suppose that we are given the accuracy π_i of each attribute classifier. We organize them as a diagonal matrix P with $P_{ii} = e^{-\pi_i}$, based on the intuition that the less accurate the attribute classifier the more punishment it should receive. By adding this to Eq. (9), we have,

$$\min_w \sum_{i=1}^N \max\{0, 1 - y_i(w^T x_i + b)\} + \frac{\lambda_1}{2} w^T P w + \frac{\lambda_2}{2} w^T L w \tag{10}$$

To the best of our knowledge, this modification to the linear SVM is novel, with advantages of flexibility and scalability as stated in Section 2. This objective is a usual quadratic programming problem with linear inequality constraints. The corresponding dual form is given by,

$$\begin{aligned} \min_\alpha \quad & \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^T (\lambda_1 P + \lambda_2 L)^{-1} x_j - \sum_{i=1}^N \alpha_i \\ s.t. \quad & 0 \leqslant \alpha_i \leqslant 1, i = 1, 2, \ldots, N \\ & \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned} \tag{11}$$

Such kind of optimization problem can be solved with many off-the-shelf methods either in primal form or in dual form. In our implementation, we use the Mosek Optimization Toolbox [42] as the solver for the primal problem. To set appropriate values for λ_1 and λ_2, we have to consider (1) trade-off between regularization terms and the loss term and (2) trade-off between the two regularization terms. For these purposes, we set $\lambda_1 + \lambda_2 = c$ and $\lambda_1/\lambda_2 = r$, and then do the grid search on c and r through cross validation. Typical parameter values selected on the validation

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

data set are $\lambda_1 = 0.16$ and $\lambda_2 = 0.8$, where the larger value of λ_2 emphasizes more importance of the attribute correlation constraint.

We also consider the kernel version of the formulation in Eq. (11). In particular, we first perform a Cholesky factorization to obtain $(\lambda_1 P + \lambda_2 L)^{-1} = R^T R$. Then, the dual form in Eq. (11) is rearranged as follows.

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (Rx_i)^T (Rx_j) - \sum_{i=1}^{N}\alpha_i \\ s.t. \quad & 0 \leqslant \alpha_i \leqslant 1, \quad i = 1, 2, \ldots, N \\ & \sum_{i=1}^{N}\alpha_i y_i = 0 \end{aligned} \tag{12}$$

Finally, we replace the inner product of Rx_i and Rx_j in Eq. (12) with any qualified kernel function $K(Rx_i, Rx_j)$.

It should be noted that this method for kernelization is a compromise because the embedded kernel functions may not completely depend on inner products in feature space but the term $(\lambda_1 P + \lambda_2 L)^{-1}$ as well. Actually, due to the existence of this term, we have $w = \sum_{i=1}^{n}\alpha_i y_i (\lambda_1 I + \lambda_2 L)^{-1} x_i$, which means that the representation theory may not be satisfied in the original input space. But by transforming the input space with R and letting $\dot{x}_i = Rx_i$, we can show that $\dot{w} = \sum_{i=1}^{n}\alpha_i y_i \dot{x}_i$, which could be served as a good starting point of kernelizing (from the transformed space).

3.3. Extensions to continuous attributes

We also investigate the possibility to extend our method to more general scenarios where the values of attributes are continuous (in this case the attributes can also be called features). For example, in the famous iris data set [43], each instance is described by four continuous attributes, describing the sepal length, the sepal width, the petal length, and the petal width in centimeter respectively. In this case, it will be pointless to project each attribute value onto a subject space (or class space) but it is still meaningful to exploit the relationship between attributes to control the notorious overfitting. As mentioned before, the benefit of exploiting the relationship between features in the general machine learning tasks is less studied, which is exactly the purpose of this section.

In particular, we propose several strategies (as follows) to model the relationship between attributes (features) when they are continuous, then build the attribute (feature)-graph as in Section 3.1, which is finally used as a penalty term in the linear SVM (Eq. (9)).

3.3.1. Mutual Information (MI)

The first method we explored is the mutual information method since it measures the information content of each individual feature x with regard to the output class y. We approximate the MI value $I(x, y)$ for each feature based on the Parzen window estimation [44], and then use it to evaluate the similarity between features (*cf.* Eq. (5)). The obtained feature relationship graph is finally plugged into the classifier (Eq. (9)) for regularization.

3.3.2. Conditional Class Variance (Var)

Estimating mutual information for continuous data is known to be complicated, and it is desirable to find some efficient way to approximate it. In some regard, the point-wise mutual information summarizes the information from the normalized class-conditional distribution,[1] which inspires us to evaluate the relationship between an individual feature and a class label through estimating the distribution $p(x \mid y)$ by projecting the feature x onto the class y. We model the distribution $p(x \mid y)$ as a 1D Gaussian $N(u, \sigma^2)$, then the variance σ^2 can be used as a good indicator of the uncertainty. In particular, the large variance value may indicate an unstable relationship between the feature and the class label; in contrary, the lower value may indicate short description length for the feature to encode the class. Therefore, we simply use the variance value as the representation of the corresponding feature, and in this way, each feature is represented as a K-dimensional vector with each entry being the σ_k^2 for class k. After this, we calculate the desired feature relationship graph using Eq. (5).

3.3.3. Relief (Relief)

Another way to construct the feature graph is based on the importance of each feature. This can be evaluated in different ways (with or without taking label information into account, see [45]), but to our knowledge, they have not been used for feature graph construction. Relief [46] is one of the most successful representatives of them, which assigns a weight to a particular feature based on the large margin criterion. We use the feature weights obtained from the Relief algorithm to compute the similarity between features and impose the constraint that features with similar weights should play a similar role in the classifier.

3.3.4. Principal Component Analysis (PCA)

If we arrange the training data with an $N \times d$ matrix X, where N is the number of samples and d is the number of features for each sample, then, in principle, we can evaluate the similarity between two features by taking each column of X as some kind of representation for the corresponding feature and then invoking Eq. (5). Alternatively, we can do this by removing the noise first. In particular, we decompose the covariance matrix Σ as $\Sigma \approx A\Lambda A'$, where A is a $d \times q$ matrix whose columns are the first q orthonormal eigenvectors with the largest eigenvalues of the matrix Σ. Then each row of A gives us a more compact and more robust feature representation, which is then used for similarity evaluation.

3.3.5. Linear Discriminant Analysis (LDA)

This is similar to the PCA method in spirit, except that supervised label information is used to guide the evaluation of the importance of each feature for classification. In particular, we use the Fisher criterion to find out the most discriminative projection directions from the training data, and the importance of each feature is then evaluated according to the magnitude of the components of the discriminant vector, which is finally fed into Eq. (5) to construct the feature graph.

4. Experiments

We illustrate the effectiveness of our methods by presenting experiments on two large-scale face databases with attributes annotated: LFW [10] and PubFig [4]. To verify the performance of the proposed method on the tasks beyond face recognition, we also present experimental results on the a-Pascal object recognition dataset [1] and UCI data sets. We divide the results into two sections, the first focusing on face verification and object recognition, and the second on more detailed issues about the proposed attribute-graph regularized SVM classifier itself, *e.g.*, parameter settings and its applications in general machine learning tasks.

4.1. Data sets

To verify the effectiveness of the proposed method, we conduct a series of experiments on two typical real-world face databases with attributes annotation. The first is the Labeled Face in the Wild (LFW) database [10], which is a de facto standard database to test

[1] $I(A_i, y_j) \propto P(A_i, y_j)/P(A_i)P(y_j) = P(A_i \mid y_j)/P(A_i)$.

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

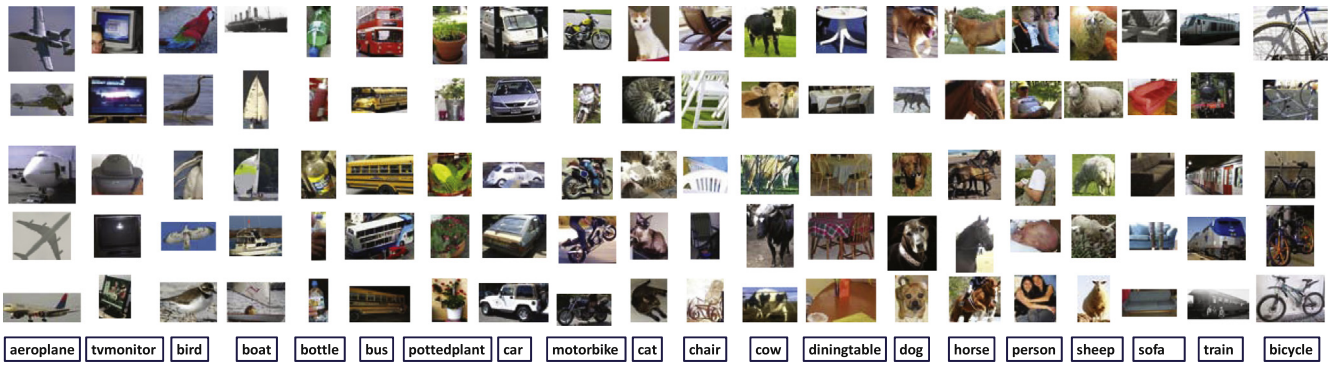

Fig. 3. Illustration of sampled images of the twenty object classes in the a-Pascal dataset [1].

the performance of face verification system under the unconstrained conditions. The collected face images are full of typical features of unconstrained conditions including great variations in pose, expression, lighting, occlusion and image resolution. The second is the Public Figures (PubFig) Face Database [3], which is similar in spirit to the LFW database, but is much deeper (more images per person) than LFW. Thanks to Kumar et al. [3], the attribute descriptors of face images in both databases are publicly available through the internet, and there are totally 73 facial attributes for each face, and we will use these directly as the high-level representation for each face throughout the experiments, see Fig. 2 for more details.

To test the performance of the proposed method in the general attribute-based object recognition task, we also perform experiments on an object recognition dataset, *i.e.*, a-Pascal [1]. This dataset contains 20 object classes in all (see Fig. 3) with the number of objects from each category ranging from 150 to 1000. A list of 64 attributes is designed to describe a-Pascal objects, along with 9751 dimensional basic features on each image to facilitate attribute extraction.

To evaluate the effectiveness of the extended feature-graph regularized SVM algorithm in the case of continuous attributes, we test them on eight popular UCI data sets, retrieved from the UCI Machine Learning Repository.[2] These data sets include the sonar (sonar), the Johns Hopkins university ionosphere (iono), optical recognition of handwritten digits (digits), SPECTF heart (heart), Spambase (spam), Hill-Valley (vally), and Breast Cancer Wisconsin Prognostic (wpbc). For the digits data set, we choose to distinguish the number '8' from '9' due to its difficulty.

4.2. Experimental settings

For face verification experiments, in Kumar et al.'s original paper [3,4], an SVM with RBF kernel is used as the classifier. The input for this classifier, however, involves two parts, one is the absolute difference between the attribute features of two face images to be verified, *i.e.*, $|A_i - A_j|$, while the other part is the bitwise product of these two attributes, *i.e.*, A_iA_j. Although adding the second part increases the performance by about 2%, in our experiment, we do not use this since it is not so natural for us – commonly we do not take the product of two feature vectors as new features since this will double the dimension of the input vector. Indeed, the focus of this paper is not to find a new way for feature extraction but to see whether exploiting the relationship between attributes could improve the performance of face verification. For the above reasons, we use the scheme of '$|A_i - A_j|$ + Linear SVM' as our baseline classifier and replacing it with RBF–SVM only slightly improves the performance (about 0.4%).

We also compared our method with the strategy of [5], where the relationship between attributes is encoded by a max-spanning-tree and is used as augmented feature sets for the training data. For better performance, in our implementation we augment the original feature sets with the product of correlated attributes and named this approach 'Augm.Fea.'.

For object recognition experiments, we follow the protocol proposed in [1]. Specifically, we first use the provided base features to train a set of across category attributes predictors, based on which, 20 object classifiers are trained using the one-vs-all multi-classification framework. The predicted object label is estimated according to the maximum response of the trained 20 object classifiers. Considering the skewed distribution of the number of samples over object classes (for example, the "people" class has more samples than other classes), we report both the overall accuracy and mean accuracy per class. In all the experiments, the linear SVM is adopted as in [1].

For experiments on the UCI datasets, we evaluate the performance on each data set following the standard protocol whenever possible (*e.g.*, the specified splitting of the training set and test set for datasets 8vs9, heart, and valley), otherwise 2/3 randomized data are for training and the remaining 1/3 for testing (details are shown in Table 1), and report performance with the mean and standard variance on ten repeats of such a random process.

4.3. Experimental results

4.3.1. A toy problem

Before presenting the normal experimental results on the task of face verification and object recognition, we think that it will be useful to gain some intuitive understanding about the behavior of our algorithm. Therefore, we give some visualization of the learned classifiers on a simple two-dimensional toy data set. The toy data set D is generated by imposing an approximate linear relationship between two dimensions, *i.e.*, $y = a * x + \epsilon$, where $\epsilon = \mathcal{N}(0,1)$ is one-dimension Gaussian noise, and a controls the degree of correlation between the two dimensions (we set $a = 0.8$). Two groups of data are generated according to this model, as denoted with red circles and blue asterisks respectively in Fig. 4.

We can think of the two dimensions of the toy data as attributes with a strong correlation and due to this, they should have similar weights from our model (*cf.* Eq. (9)). In Fig. 4, we illustrate geometrically the weights learned before (W_1) and after (W_2) imposing the attribute-graph regularization on the linear SVM. It can be seen that the normal vector of the separation line after regularization is driven to a position more paralleling to the line of $x - y = 0$. So what's benefit of it? For this, we compare the sum of the sample

[2] http://archive.ics.uci.edu/ml/index.html.

Table 1
Detailed experimental settings for the UCI data sets. For each data set, the number of attributes (N_{atr}), the number of training samples (N_{tr}) and the number of test samples (N_{te}) are listed respectively.

	sonar	iono	8vs9	heart	spam	valley	wpbc	secom
N_{atr}	60	33	64	44	57	100	33	353
N_{tr}	138	234	554	80	333	606	130	134
N_{te}	70	117	562	187	167	606	64	66

Table 2
Comparative performance of our methods (and its kernelized version) and the baseline, *i.e.*, linear SVM (L.SVM), and comparing algorithm, *i.e.*, linear SVM with augmented features (L.SVM + Augm.Fea. [5]), on the LFW and the PubFig database. (Bold values indicate the best results.)

	L.SVM	L.SVM + Agum.Fea. [5]	Ours	Ours (kernelized)
LFW	83.4 ± 0.5	84.6 ± 0.6	85.5 ± 0.6	**85.9 ± 0.6**
PubFig	76.7 ± 0.9	77.6 ± 0.8	78.6 ± 0.8	**78.8 ± 0.8**

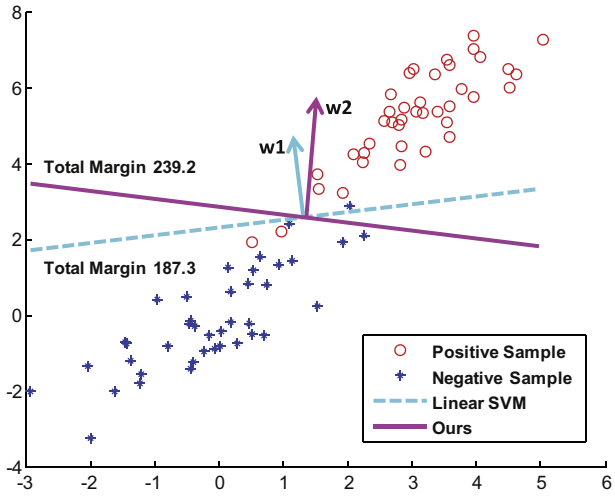

Fig. 4. The comparison of the learned separation lines of our method and the baseline.

margins of the two separators, and the results show that our model gets a score of 239.2, much larger than 187.3 of the traditional linear SVM. This indicates that our model has a better chance to achieve better generalization capability.

4.3.2. Face verification results

4.3.2.1. AUC comparison. Following the standard LFW evaluation protocol, Fig. 5(a) gives the overall performance of the proposed algorithm compared to the baseline. In particular, the AUC (Area Under the ROC Curve, the larger the better) value of our method is 0.925, compared to 0.913 of the baseline method [4] and 0.922 of the 'Augm.Fea.' approach [5], indicating that the proposed method is effective in exploiting the attribute relationship to enhance attribute-based recognition. Seen from the Fig. 5(b), the overall behavior of the compared algorithms on the PubFig database is consistent with that on LFW, and again our method performs best among the three in terms of the AUC value.

4.3.2.2. Accuracy comparison. Table 2 gives the overall performance comparison. We can see that both our method and the method of [5] consistently outperform the baseline on the two databases. This clearly demonstrates the benefit of exploiting the attribute-relationship. Specifically, on the LFW, the average performance using our method is 85.5 ± 0.6%, compared to 83.4 ± 0.5% of the baseline, showing more than 2% advantages. While the average performance for the method of [5] is 84.6 ± 0.6%, with nearly 1% advantage over the baseline. Similar observation can be made on the PubFig database. It is worth mentioning that the performance of our method is comparable to the state-of-the-art results of 85.2% in [4], without using more advanced techniques of feature combination. Kernelization only slightly improves the performance to 85.9 ± 0.6% on this (highly nonlinear) database.

Fig. 6(a) details the comparative performance on each of the ten cross-validation test sets defined in [10]. It can be seen that our method and the method of [5] perform much better than the baseline, and our method performs best among the three. It should be noted that although both our method and the method of [5] exploit the relationship between attributes to improve performance, the specific strategy for achieving this goal is different. In particular, we mainly use this as a prior for model regularization to reduce overfitting, while the method of [5] explores attribute-relationship as augmented features for face representation. However, the latter method may face the difficulty of high dimensionality – actually, the dimensionality of the feature space in [5] could be nearly twice as much as that in our method after augmenting all the pairwise features, which significantly increases the complexity of the model.

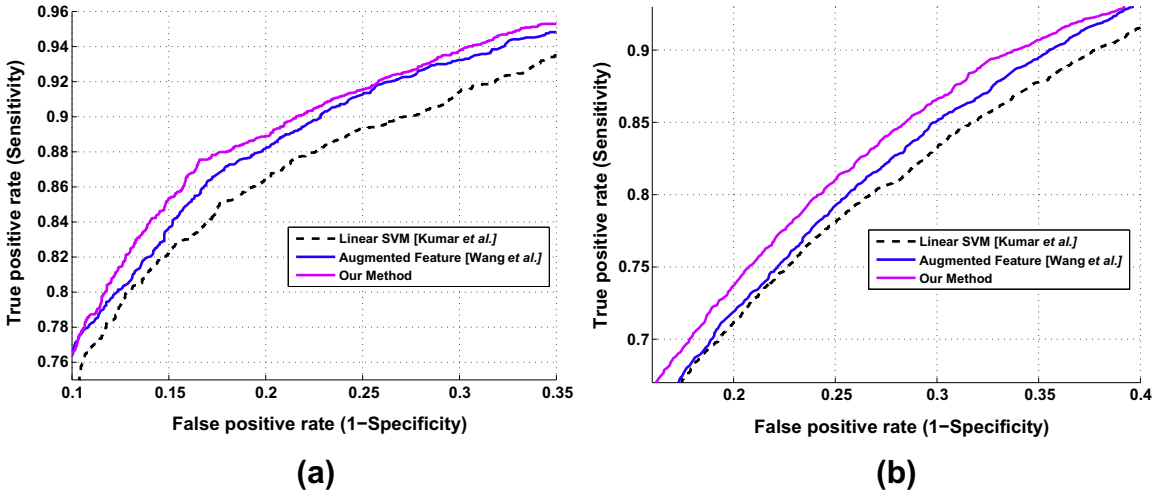

Fig. 5. Comparing overall ROC curve of our method with Kumar et al. [4] and Wang et al. [5] on (a) the LFW database and (b) the PubFig database.

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

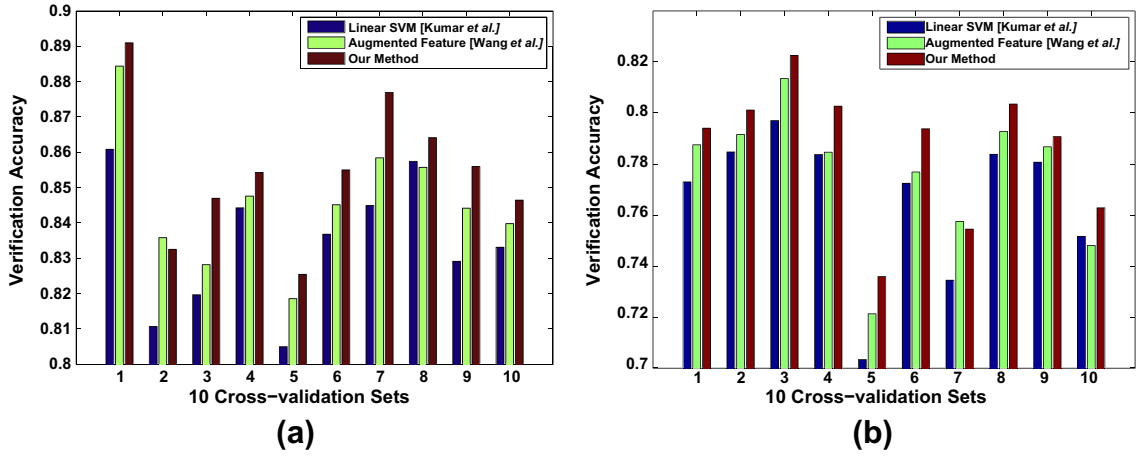

Fig. 6. Comparing the detailed performance on ten cross-validation test sets of our method with Kumar et al. [4] and Wang et al. [5] on (a) the LFW database and (b) the PubFig database.

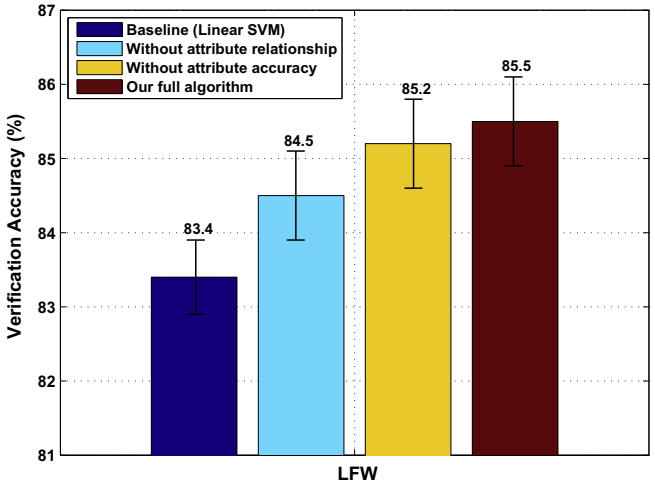

Fig. 7. Influence of the individual component of our algorithm. On the LFW face database, we compare the verification rates (%) with our full method to rates when each of the two major components is removed in turn, leaving the remaining one in place.

To handle this issue, one has to sparsify the full data matrix of pairwise attribute-relationship as did in [5], but useful information may be lost in this procedure, as well. By contrast, with our strategy of model regularization, the information of the full attribute-relationship graph can be exploited without imposing extra difficulties. In addition, the introduction of subject space (*cf.*, Fig. 1) in our method actually integrates the useful discriminative information into the model. We believe that the above factors are helpful to explain the performance difference between our method and the method of [5]. Despite this, with 72 pairs of attribute-relationship as augmented features, the method of [5] does perform better than ours on the second set of the LFW and the seventh set of the PubFig.

4.3.2.3. Results analysis. To understand how the two main components of our algorithm (*i.e.*, using attributes relationship and accuracy of attribute classifiers, respectively) contribute to the final performance, in Fig. 7 we illustrate the effect of removing each of them in turn while leaving the remaining component in place (the comparison is thus against our full algorithm, not against no knowledge about these two aspects). In general, each component is beneficial, and the results are cumulative over the two, but the benefits are much greater from exploiting the attribute relationship.

Fig. 8 gives some illustration of face pairs which are incorrectly recognized by the baseline classifier but correctly with our model. In particular, each pair of images in the leftmost three columns are respectively from the same subject but are misjudged as identity-unmatched face pairs with high confidence by the baseline classifier. However, our method does not make such mistakes. In the rightmost three columns, we show three pairs of face images from three different subjects, which are, unfortunately, incorrectly identified as identity-matched face pairs by the baseline classifier. However, in all these three cases, our method makes the correct decision. This shows that by taking the information of attribute relationship into account, our method effectively improves the generalization capability of the prediction model.

Fig. 9 lists some typical highly related attributes learned using the method in Section 3.1 on the LFW database. These attributes can be broadly divided into two categories: (1) with high semantic correlation (shown in the yellow rectangle) and (2) with high statistical co-occurrence (shown in the green rectangle). We can see that the learned attribute relationship is reasonable. For example, the semantic concept of "male" has high co-occurrence with male-specific attributes such as wearing the necktie, bushy eyebrows, while with negative correlations with things commonly used by females, such as lipstick, necklace, earrings. As another example, we see that an "attractive woman" usually has "heavy makeup" and being "youth". On the other hand, some concepts only have weak semantic connections but otherwise show strong co-occurrence property among them. As shown in the last row of Fig. 9, "color photo" is a general property of images with "non-baby", "non-sunglasses", *etc.*, which essentially reflects the statistical characteristics of images of this particular database.

4.3.2.4. Influence of parameter settings. We then investigate the effect of attribute-graph regularization on the model parameters w. Intuitively, the Laplacian constraint on the linear SVM (*cf.* Eq. (8)) will result in more consistent weights for highly related attributes while leaving those less related untouched. In this way, the structures among attributes can be exploited. Fig. 10 gives the comparative distribution of samples over the space of attribute-similarity and weight-difference before (left) and after (right) the attribute-graph regularization imposed (on the LFW database). It can be clearly seen that indeed the weights of similar attributes tend to

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

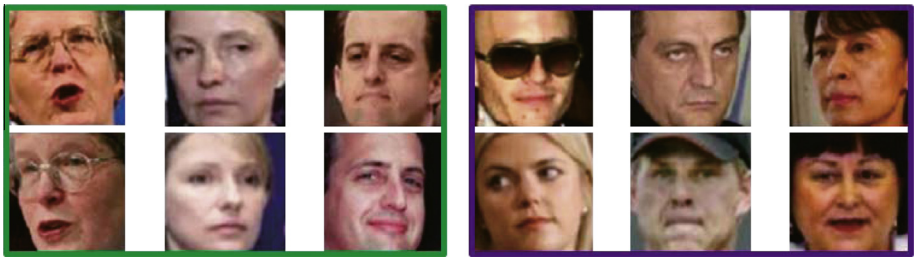

Fig. 8. Illustration of three pairs of face images from the same subject respectively (the leftmost three columns) and three pairs from different subjects (the rightmost three columns). All the six pairs are mistakenly identified by the baseline classifier but are correctly recognized by our method.

Fig. 9. Illustration of highly related attributes learned by our method on the LFW database. On the leftmost column, we show the typical semantic concepts in bold, and on the right we list the attributes correlated to those concepts.

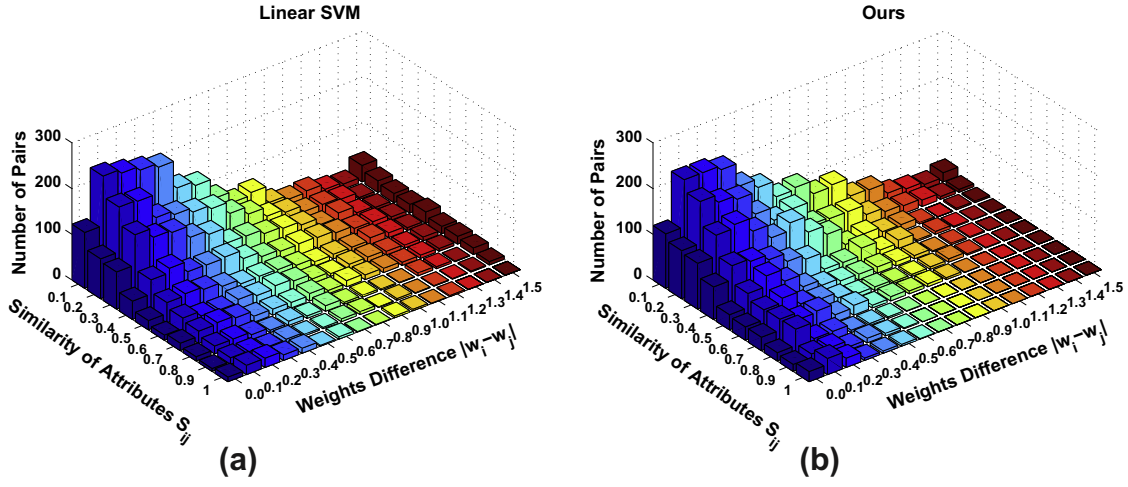

Fig. 10. Comparative distribution of samples over the space of attribute-similarity and weight-difference: (a) Linear SVM and (b) our attribute-graph regularized Linear SVM. Note the difference of distributions over the red region (bottom right of the plane) between (a) and (b). This figure is best viewed in the electronic form.

be similar after regularization (*cf.* the bottom right area of the plane in Fig. 10(b)).

We also study the influence of the dimensionality of the subject space on the verification performance on the LFW database by varying the number of subjects for the subject space and then measuring the corresponding verification performance. In particular, we obtain an average accuracy of 85.4 ± 0.06% over varying number of subjects (33, 60, 95, 155, 420, 606, 900). This shows that the performance of the method is not sensitive to the subject space size. Similar results are also observed on the PubFig database.

Finally, we investigate the influence of the two similarity measures (see Eq. (5)) on the face verification performance. As mentioned before in this work the Heat Kernel is adopted to evaluate the similarity between attribute vectors, and the parameter σ is set to be the mean of the L_2 distances of all the attribute pairs. This setting leads to a verification accuracy of 85.5 ± 0.6% on the LFW dataset. For comparison, we also test a version with a similarity measure replaced with Cosine similarity, which yields a slightly lower accuracy of 85.4 ± 0.6%. Fig. 11 gives the histograms of the similarity values from the attribute-relationship graph using these

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

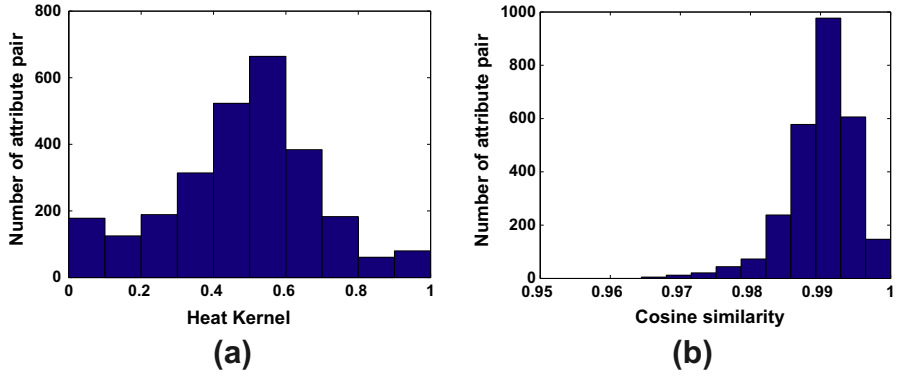

Fig. 11. Histograms of similarity values in the attribute-relationship graph using (a) Heat Kernel and (b) Cosine similarity, respectively, on the LFW dataset.

Table 3
Comparative performance of our method with the baseline method (L.SVM, [1]) on the a-Pascal dataset. Following [1], both overall accuracy (%) and mean accuracy (%) per class are reported. These results are based on two types of attribute data, *i.e.*, supervised attribute annotations and the responses of learned attribute classifiers. (Bold values indicate the best results.)

	Learned attributes		Annotated attributes	
	Overall	Mean per-cls.	Overall	Mean per-cls.
L.SVM [1]	56.1 ± 0.94	37.4 ± 1.24	82.1 ± 1.04	74.9 ± 1.37
Ours	**57.3** ± **0.74**	**40.6** ± **1.69**	**85.1** ± **0.85**	**75.1** ± **1.75**

two similarity measures, respectively. One can see from this figure that the Heat Kernel seems to yield more 'flat' distribution of similarity values (with entropy of 6.18 bits) than the Cosine similarity (with entropy of 1.51 bits), due to the 'normalization' effect of σ.

4.3.3. Object recognition results

Table 3 gives the overall experimental results on the a-Pascal dataset. We can see that again using attribute relationship improves the object recognition performance over the baseline method [1] consistently in terms of both overall accuracy and mean per class accuracy. Actually, we see that using learned attributes instead of ground truth attributes annotation as object representation leads to a significant deterioration in performance (see Fig. 12). This highlights the needs to improve the accuracy of attribute prediction.

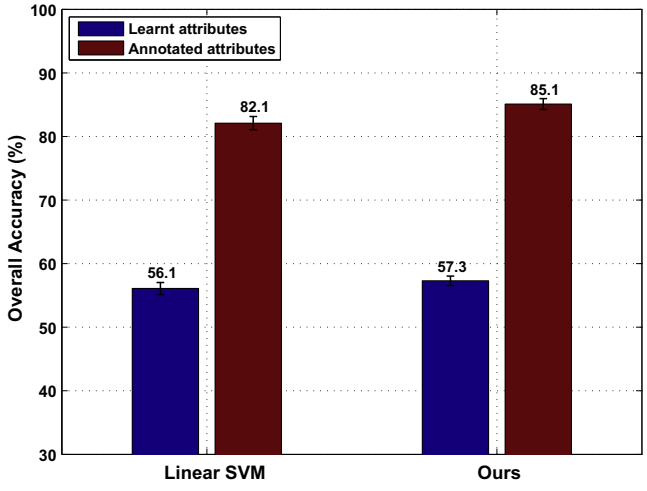

Fig. 12. Comparative overall performance of our method and the baseline [1] on the learned attributes and the ground truth attributes respectively, showing how the uncertainty of attributes influences the final recognition performance.

In addition, Table 3 indicates that exploiting the attributes relationship always improves the performance whether the attributes are manually annotated. In particular, it can be seen that the proposed method improves the overall accuracy of the baseline [1] from 82.1% to 85.1% in the case of annotated attributes.

Fig. 13 gives some typical highly correlated attributes learned by our method. The revealed relationship between attributes can be studied in different ways. Firstly, some attributes reflect a kind of part-whole relationship. For example, "Furniture legs" belong to some furniture which may have "Seat", "Back" and "Arms" as well, while "Wheel" is related to a vehicle which has "Taillight" and "Side mirror", and "Nose" is an important part of a human's body, so are "Mouth", "Face" and "Hair". Secondly, we see that some hidden negative correlations between attributes are also revealed by our algorithm. For example, "Furniture" usually do not have "Tail" and "Wool", "Snout" is not likely "Shiny", "Wing" and "Hair" are mutually exclusive, *etc.* Finally, it is worth noting that "Snout" and "Nose" have similar semantic meaning, but they are usually used to describe different animals; hence they have a different set of correlated attributes accordingly. We emphasize that classifiers may benefit from properly using such additional information.

4.3.4. Performance on continuous attributes

Fig. 15 summarizes the effects of using attribute-relationship graph constraints, while Fig. 14 gives the detailed comparative performance of different graph construction strategies on eight UCI real-world data sets. In Fig. 14, each sub-graph corresponds to one data set, and the horizontal axis indicates the graph construction strategies, while the vertical axis represents the performance obtained accordingly. We can see that applying attribute-graph constraints improves the recognition performance consistently over all the data sets tested. In particular, Fig. 15 shows that the average recognition performance of the baseline method (*i.e.*, linear SVM without graph prior) is boosted from about 77.0% to a level of higher than 80.0% by the LDA induced attribute graph prior. This improvement clearly indicates the benefits of incorporating pairwise discriminant prior information into the model.

The second best performer is the MI-induced feature graph, and it improves the performance by about 2.0% on average. Despite the improvements of the remaining three graph construction methods are not so evident as the previous two, in most cases they perform better than the baseline method. As we can see from Fig. 14, the var-graph gives very promising results on the inon and the heart data sets; the Relief-graph brings good improvement on the sonar and the spam data sets, while the PCA-graph outperforms all the other methods on the heart data set. From these, we conclude that it is helpful to take the relationship between attributes into

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

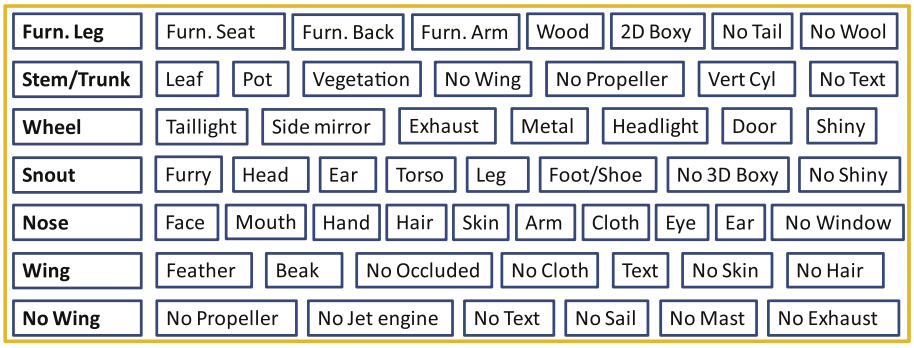

Fig. 13. Illustration of highly related attributes learned by our method on the a-Pascal dataset. On the leftmost column we show the typical semantic concepts in bold, and the related attributes correlated to those concepts are listed on the right.

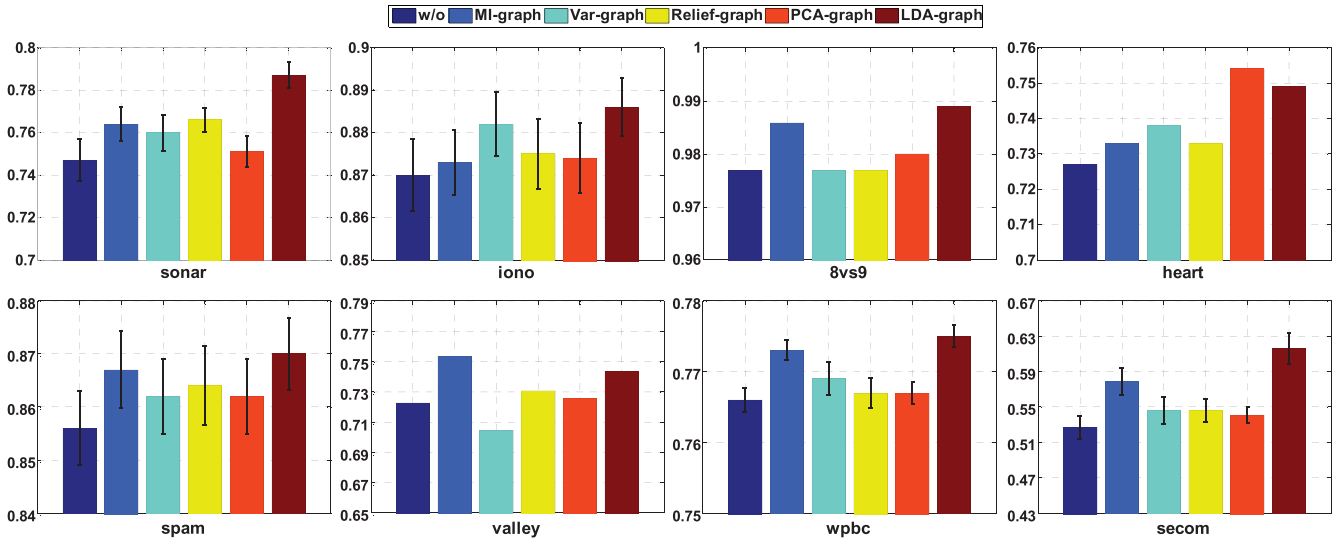

Fig. 14. Comparative performance of different attribute-relationship graphs on eight UCI real-world data sets. For datasets 8vs9, heart, and valley, we follow the standard protocol for training and testing data splitting and hence no variance is reported.

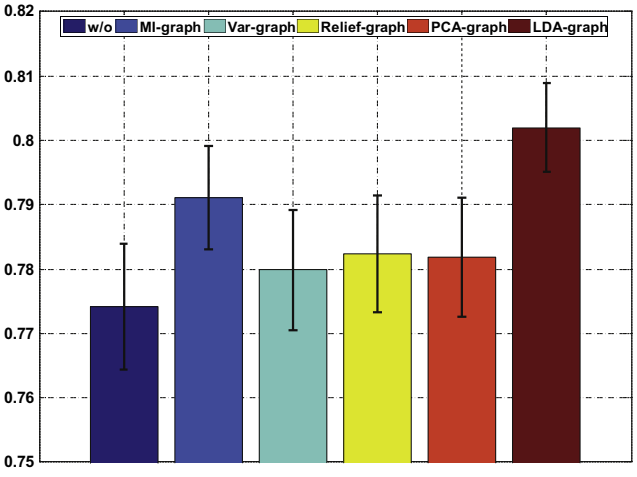

Fig. 15. Overall performance of different attribute-relationship graphs on the UCI data sets.

consideration. Although the best way to do this is still under investigation and depends on the characteristics of the data sets on hand, our experimental results indicate that LDA and MI induced attribute graphs perform well in practice.

5. Conclusions and future work

In this paper, we give a novel method to model the relationship between attributes, based on a discriminative distributed representation for attributes. This effectively allows our classifier to explore the hidden correlation between attributes in a general context of subjects. We show in this paper on the challenging face verification and object recognition databases that the mined attribute graph does reflect some real aspects of the semantic relationship among attributes existed in the real world; furthermore, such relationship is helpful to improve the accuracy and robustness of the face verification/object recognition system.

We also extend this method to more general scenarios where the values of attributes are continuous, and this leads to a new attribute-graph regularized SVM algorithm, which essentially opens a door to incorporate additional structural (or spatial) information at the level of attributes into the classifier. The effectiveness and feasibility of the proposed methodology are empirically verified in several application domains, showing that it can improve overall classification results, even when the available data are limited.

Although we do not focus on the problem of attribute extraction in this paper, our experiments on a-Pascal object dataset indicate that the accuracy of extracted attributes could have a great impact

Please cite this article in press as: F. Song et al., Exploiting relationship between attributes for improved face verification, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.02.010

on the final performance of an attribute-based recognition system (*cf.* Fig. 12). For accurate attribute extraction, we have to decide where to look at them. For example, Kumar et al. [4] performed a greedy forward selection for each attribute to find out the most discriminative local facial regions among the predefined nine regions. Ferrari and Zisserman [19] proposed to tackle the uncertain location of features by optimizing the likelihood ratio. Chen et al. [9] exploited pose information for attribute extraction, and they also considered the sensitivity of different attributes to different feature types. Alternatively, attributes can be treated as a kind of latent variables, and the task of attribute prediction simply boils down to assign values to latent variables [5,25]. These latter methods actually get around the problem encountered in the former ones, but the extracted attributes are less interpretable. Further study on this will be the focus of our future work.

Acknowledgements

The authors are grateful to the editors and reviewers for 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, 20133218110032) and the Fundamental Research Funds for the Central Universities (CXLX11_0204).

References

[1] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1778–1785.
[2] C. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 951–958.
[3] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of IEEE International Conference on Computer Vision, IEEE, 2009, pp. 365–372.
[4] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1962–1977.
[5] Y. Wang, G. Mori, A discriminative latent model of object classes and attributes, Proceedings of European Conference on Computer Vision, 6315, Springer, 2010, pp. 155–168.
[6] A. Farhadi, I. Endres, D. Hoiem, Attribute-centric recognition for cross-category generalization, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2352–2359.
[7] Y. Su, M. Allan, F. Jurie, Improving object classification using semantic attributes, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2010, pp. 26.1–26.10.
[8] T. Berg, A. Berg, J. Shih, Automatic attribute discovery and characterization from noisy web data, in: Proceedings of European Conference on Computer Vision, Springer, 2010, pp. 663–676.
[9] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: Proceedings of European Conference on Computer Vision, Springer, 2012, pp. 609–623.
[10] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Tech. rep. 07-49, University of Massachusetts, Amherst, October 2007.
[11] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[12] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and zero-shot learning in a large-scale setting, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1641–1648.
[13] P. Kankuekul, A. Kawewong, S. Tangruamsub, O. Hasegawa, Online incremental attribute-based zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3657–3664.
[14] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, B. Schiele, What helps where–and why? Semantic relatedness for knowledge transfer, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 910–917.
[15] R. Bhatt, C. Rovee-Collier, Infants' forgetting of correlated attributes and object recognition, Child Develop. 67 (1) (1996) 172–187.
[16] G. Wang, D. Forsyth, Joint learning of visual attributes, object classes and visual saliency, in: Proceedings of IEEE International Conference on Computer Vision, IEEE, 2009, pp. 537–544.
[17] L. Bourdev, S. Maji, J. Malik, Describing people: poselet-based attribute classification, in: Proceedings of IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1543–1550.
[18] D. Parikh, K. Grauman, Relative attributes, in: Proceedings of IEEE International Conference on Computer Vision, IEEE, 2011, pp. 503–510.
[19] V. Ferrari, A. Zisserman, Learning visual attributes, in: Proceedings of Advances in Neural Information Processing Systems, MIT Press, 2008, pp. 433–440.
[20] G. Hinton, Learning distributed representations of concepts, in: Proceedings of the Eighth Annual Conference of the Cognitive Science Society, 1986, pp. 1–12.
[21] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.
[22] F. Song, X. Tan, S. Chen, Exploiting relationship between attributes for improved face verification, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2012, pp. 27.1–27.11.
[23] J. Wang, K. Markert, M. Everingham, Learning models for object recognition from natural language descriptions, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2009, pp. 2.1–2.11.
[24] F.-F. Li, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2005, pp. 524–531.
[25] B. Siddiquie, R. Feris, L. Davis, Image ranking and retrieval based on multi-attribute queries, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 801–808.
[26] M. Douze, A. Ramisa, C. Schmid, Combining attributes and fisher vectors for efficient image retrieval, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 745–752.
[27] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 3337–3344.
[28] Q. Qiu, Z. Jiang, R. Chellappa, Sparse dictionary-based representation and recognition of action attributes, in: Proceedings of International Conference on Computer Vision, IEEE, 2011, pp. 707–714.
[29] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, B. Schiele, Script data for attribute-based recognition of composite activities, in: Proceedings of European Conference on Computer Vision, Springer, 2012, pp. 144–157.
[30] D. Parikh, K. Grauman, Interactively building a discriminative vocabulary of nameable attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1681–1688.
[31] A. Kovashka, S. Vijayanarasimhan, K. Grauman, Actively selecting annotations among objects and attributes, in: Proceedings of IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1403–1410.
[32] K. Duan, D. Parikh, D. Crandall, K. Grauman, Discovering localized attributes for fine-grained recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3474–3481.
[33] G. Patterson, J. Hays, Sun attribute database: Discovering, annotating, and recognizing scene attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2751–2758.
[34] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[35] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[36] L. Jacob, G. Obozinski, J. Vert, Group lasso with overlap and graph lasso, in: Proceedings of International Conference on Machine Learning, ACM, 2009, pp. 433–440.
[37] Y. Zhou, R. Jin, S. Hoi, Exclusive lasso for multi-task feature selection, in: Proceedings of International Conference on Artificial Intelligence and Statistics, 2010, pp. 988–995.
[38] S. Kim, E.P. Xing, Tree-guided group lasso for multi-task regression with structured sparsity., in: Proceedings of International Conference on Machine Learning, ACM, 2010, pp. 543–550.
[39] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft *k*-NN ensemble, IEEE Trans. Neural Networks 16 (4) (2005) 875–886.
[40] K. McRae, V.R. de Sa, M.S. Seidenberg, On the nature and scope of featural representations of word meaning., J. Exp. Psychol.: Gen. 126 (2) (1997) 99–130.
[41] D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, N. Ward, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, vol. 2, MIT Press, 2000.
[42] M. ApS, The MOSEK Optimization Software. <http://www.mosek.com>.
[43] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[44] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[45] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[46] H. Liu, H. Motoda, Computational Methods of Feature Selection, Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC, 2008.