Exploiting Relationship between Attributes for Improved Face Verification

Fengyi Song
f.song@nuaa.edu.cn
Xiaoyang Tan
x.tan@nuaa.edu.cn
Songcan Chen
s.chen@nuaa.edu.cn

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

Abstract

Recent work has shown the advantages of using attribute-based representation over low-level feature descriptors in face verification, due to its capability to explicitly encode high-level semantic meaning with economical coding bits. However, most work assumes that the attributes of a given face is independent to each other. In this paper we present a novel method to show how to model the relationship between attributes and exploit such information in the task of face verification, while 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. Using this, 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 the challenging Labeled Face in the Wild (LFW) database 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 descriptor [2, 5, 6, 12, 13, 14, 19, 21]. N.Kumar et al. [12, 13] have recently shown that using the outputs of a series of component classifiers with each tailored to some particular aspect of human face images, called visual attribute, they are able to achieve close to state-of-the-art performance of face verification on the challenging Labeled Faces in the Wild (LFW, [9]). 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, SIFT, etc., 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

© 2011. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
http://dx.doi.org/10.5244/C.26.27

not least, such a descriptor is generalizable, which makes it particularly suitable for such problems as zero-shot learning [18] or between-class transfer learning [14].

In this work, we focus on the first and the second property of attribute descriptors mentioned above, which allow us to explicitly investigate the similarity relationship between attributes and to see how such relationship could be exploited to improve the performance of face verification. Actually, research in the field of cognitive discovery has shown the usefulness of the relationship between feature sets. For example, Bhatt and Rovee-Collier [3] experimentally show that infants as young as 3 months of age gain the capability to encode the relations among object features, and use such 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 is very difficult for a human being to understand what exactly such a big feature configuration means. Fortunately, both aforementioned problems can be addressed by the attribute descriptor due to its high level and compactness in object description. Indeed, despite of the partial success of using attribute descriptors by treating them statistically independent to each other [5, 12, 13, 20] or conditionally independent given the class label [12], recent work has shown that it is beneficial to exploit the relationship between attributes under various contexts [21][4][17][8]. Some of them will be discussed in the next section.

In this paper, we proposed a novel method to model the relationship between attributes and exploit such information to improve the performance of face verification. In particular, 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, which enable us to conveniently construct the corresponding attribute-relationship graph based on the distribution of attributes in that space (c.f., Fig.1). The resulting attribute-relationship can be thought of as a way to encode 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 (c.f.,Fig.4). To exploit such information, we propose to integrate the attribute-relationship graph into a linear classifier to constrain the searching space of its parameters, based on the idea that similar attributes should have similar weights. This is helpful to avoid over-fitting and improve the generalization capability of the learnt classifier. We also extend the model to handle uncertainty in attribute responses. Encouraging experimental results of the proposed method is shown on the challenging LFW database.

2 Related work

As mentioned in the previous section, the low level features is commonly treated as a whole, while the attribute descriptor is processed individually. In other words, each attribute value is just a real number (e.g., using a binary bit to denote having the attribute or not), which means that it is not trivial to model the attribute relationship and then to exploit such information. In what follows we discuss the related work concerning these two aspects.

In [8] the concept of binary attributes is introduced to describe the spatial relationship between a pair of attributes corresponding to two image segments respectively. Such relationship is shown to be very effective in describing simple geometric patterns like strips. In [21], Wang et al. give a method which tries to exploit more general relationship between attributes to improve the performance of object recognition. For this they treat the corre-

lations among attributes as augmented feature sets in a latent SVM framework. However, one unwelcome consequence of this is that the number of possible combinations between attributes will grow quadratically with the number of attributes. To address this they have to sparsify the undirected graph which encodes the attribute correlations to a tree by keeping only highly related attributes while pruning others. Recently Parikh et al. [17] 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, for example, you don't have to explicitly describe whether a man is smiling or not when it is difficult to make such judgement, but only need to say that his expression is roughly between smiling and not smiling. In [4], even higher-order relationship between attributes are explored. They build for each attribute a regressor from all the other attribute responses, and use the output of each regressor as the corresponding attribute value. In this way, the attribute response is effectively "denoised".

Our method is different from the aforementioned ones in several ways. Firstly, all the above methods have shown its effectiveness in their particular context, e.g., object category recognition [8, 21] or scenarios analysis [17], but few 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, although it is most closed to [21]. In particular, instead of learning a pairwise relationship between attributes independently as in [8, 17], we try to model a attribute-relationship graph based on the understanding of the meaning of 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 [17, 21] 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 kind of prior constraints on the searching space of model parameters.

In the field of machine learning, the graph-based prior is commonly adopted to control model complexity of a learner. Typical example is the Laplacian SVM method proposed by Belkin et al. [1]. In this method, an instance-graph is organized to constrain the label value of neighboring instance, based on the manifold assumption that similar instance 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 attribute-graph is that its complexity is controllable since its size will not grow with the number of instances as in [1] but only with the number of attributes, which is usually not too large in practice as mentioned in the previous section. Furthermore, our graph is not meant to constrain the output space of instances but the searching space of model parameters, based on the simple idea that similar attributes should play similar roles in the learnt classifier.

In this sense, our method can also be thought of as a mechanism to automatically regularize the coefficients of 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 literatures [10, 11, 23], 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. The overall pipeline of our algorithm is presented in Fig.1.

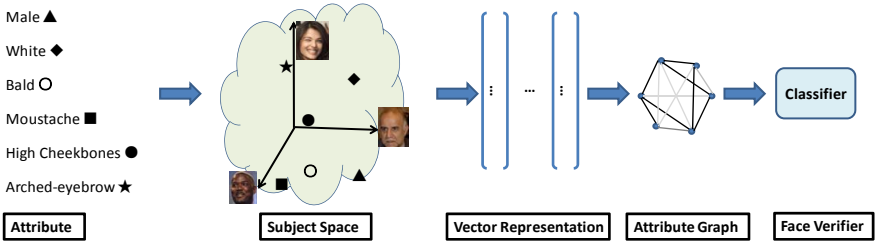


Figure 1: The overall pipeline of the proposed algorithm. Each attribute descriptor is first projected into a common subject space to obtain a high-dimensional vector representation, which are then used to construct an attribute graph. The graph is finally exploited to regularize the objective of a linear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

3.1 Modeling the Attribute Relationship

Assuming that we are given a set of M attribute descriptors $A = \{A_i \in R\}_{i=1}^{M}$ 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 real number is too simple. 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 more advanced inference.

One commonly used trick in computer vision for this purpose is to think each face of as a document which is described by words (attributes) [7, 22]. 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 attributes. One particular way we choose is the so called featural representation [15], 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 subject space (see Fig.1). Hence for K subjects, we have a subject space 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 attribute. For several images from the same subjects, the value of the corresponding entry is accumulated and then normalized with the total number of images of 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 theory framework. In particular, we first compute the point-wise mutual information $I(A_i, y_j)$ of each attribute A_i with each subject with label y_j, which are then collected as an another vector AK_i,

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

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

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

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

$$\begin{pmatrix} & y_1 & y_2 & \cdots & y_j & \cdots & y_K \\ A_1 & I(A_1,y_1) & I(A_1,y_2) & \cdots & I(A_1,y_j) & \cdots & I(A_1,y_K) \\ A_2 & I(A_2,y_1) & I(A_2,y_2) & \cdots & I(A_2,y_j) & \cdots & I(A_2,y_K) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_i & I(A_i,y_1) & I(A_i,y_2) & \cdots & I(A_i,y_j) & \cdots & I(A_i,y_K) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_M & I(A_M,y_1) & I(A_M,y_2) & \cdots & I(A_M,y_j) & \cdots & I(A_M,y_K) \end{pmatrix} \tag{3}$$

Before proceeding, we briefly discuss how to calculate the mutual information E.q.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,

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

To improve the reliability of the MLE estimation for subjects with only a few face images, we use the following Laplace smoothing strategy,

$$\begin{aligned} p_{smoothed}(A_i,y_j) &= p(A_i,y_j) + p(A_i) \\ p_{smoothed}(A_i) &= 2 * p(A_i) \end{aligned} \tag{5}$$

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

$$S_{ij} = \frac{AK_i^T AK_j}{||AK_i|| \cdot ||AK_j||} \quad \text{or} \quad S_{ij} = e^{\frac{1}{\sigma}||AK_i - AK_j||_2^2}, \quad i,j = 1,2,\cdots,M \tag{6}$$

Note that the size of our attribute 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 posterior probability (MAP), we have $p(w|D) \propto p(D|w)p(w)$, where $p(D|w)$ is the likelihood while $P(w)$ is the prior on the distribution of w. This formulation has an equivalent form,

$$log(p(w|D)) \approx log(p(D|w)) + log(p(w)) \tag{7}$$

And 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 hinge loss, the objective energy function of linear SVM is,

$$\min_{w} \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T x_i + b)\} + \frac{\lambda_1}{2} w^T w \tag{8}$$

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,

$$\min_{w} \sum S_{ij}(w_i - w_j)^2 \tag{9}$$

where $w = (w_1, w_2, \cdot, w_M)$ are the model parameters and S_{ij} is defined in E.q.6. 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 E.q.9 can be reformulated as $w^T L w$, and we add this to the standard SVM objective function,

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

Sometimes the uncertainty in the attribute response is available to us (e.g.,[12]), and we may take this into account. Suppose that we are given the accuracy π_i of each attribute classifier. We organized them as a diagonal matrix P with $P_{ii} = e^{-\pi_i}$, based on the intuition that the less accuracy the attribute classifier the more punishment it should receive. By adding this to E.q.10, we have,

$$\min_{w} \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T x_i + b)\} + \frac{\lambda_1}{2} w^T P w + \frac{\lambda_2}{2} w^T L w \tag{11}$$

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

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T (\lambda_1 P + \lambda_2 L)^{-1} x_j - \sum_{i=1}^{N} \alpha_i \\ s.t. \ & 0 \leqslant \alpha_i \leqslant 1, i = 1, 2, \cdots, N \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{12}$$

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 used the Mosek Optimization Toolbox [16] 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 loss term, and 2) trade-off between the two regularization terms. For this, 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 data set are $\lambda_1 = 0.16$ and $\lambda_1 = 0.8$, where the larger value of λ_2 emphasizes more importance of attribute correlation constraints.

4 Experiments

To verify the effectiveness of the proposed method, we conducted a series of experiments on the LFW dataset [9], which is a de fact standard dataset to test 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. Thanks to Kumar et al.[12], the attribute descriptors of faces in LFW can be obtained freely through internet. In particular, there are totally 73 facial attributes for each face (c.f., Fig.2), which can be roughly categorized into four types: (1) appearance description of key facial parts, such as shape, size and style of 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 environment; and (4) personal specific traits like bald, goatee, and attractiveness, see Fig.2 for details.

Male, Non-Asian, White, Non-Black, Non-Baby, Non-Child, Non-Youth, Non-Middle Aged, Non-Senior, Non-Black Hair, Non-Blond Hair, Non-Brown Hair, Bald, No Eyewear, Non-Eyeglasses, Non-Sunglasses, Non-Mustache, Smiling, Non-Frowning, Non-Chubby, Non-Blurry, Non-Harsh-Lighting, Non-Flash, Soft Lighting, Non-Outdoor, 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-Wearing Hat, Non-Oval-Face, Non-Square Face, Non-Round Face, Color Photo, Posed Photo, Non-Attractive Man, Non-Attractive Woman, Non-Indian, Non-Gray Hair, Bags Under Eyes, Non-Heavy Makeup, Non-Rosy Cheeks, Non-Shiny Skin, Non-Pale-Skin, 5 o'Clock Shadow, Strong Nose-Mouth-Lines, Non-Wearing Lipstick, Non-Flushed Face, High-Cheekbones, Brown Eyes, Non-Wearing Earrings, Wearing Necktie, Non-Wearing Necklace

Figure 2: Illustration of the 73 attribute descriptors used for face verification [13].

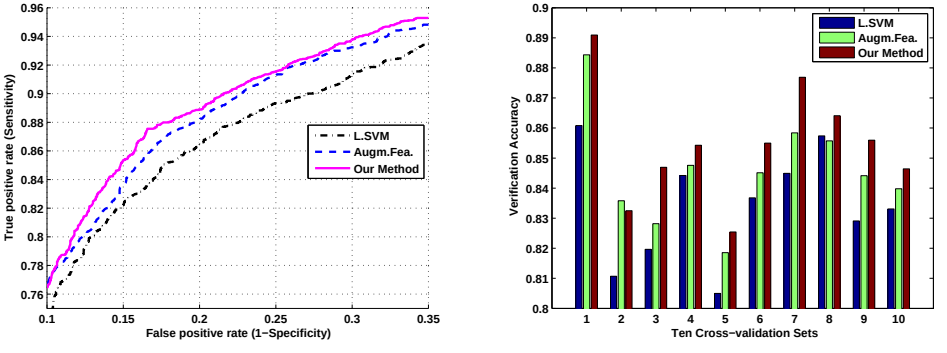
In Kumar et al.'s original paper [12, 13], a SVM with RBF kernel is used as classifier. The input for this, however, involves two parts, one is the absolute difference between the attribute features of two face images to be verified, i.e., $|A_i - A_j|$, while the other part is the bitwise product of these two attributes, i.e., $A_i A_j$. Although adding the second part increases the performance by about 2%, in our experiment, we did not use this since it is not so natural for us - commonly we don't take the product of two feature vectors as new features since this will double the dimension of 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 or not. For the above reasons, we use the scheme of '$|A_i - A_j|$ + Linear SVM' as our baseline classifier [1].

We also compared our method with the strategy of [21], where 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 augmented the original feature sets with the product of correlated attributes and named this approach 'Aug.Fea'.

Following the standard LFW evaluation protocol, Fig. 3(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 and 0.922 of the 'Aug.Fea' approach, indicating that the proposed method does improve on the baseline performance. Fig.3(b) details the comparative performance on

[1] We also test the '$|A_i - A_j|$ + RBF SVM' scheme but this only leads to very slightly improvement (about 0.3%)

each of the ten cross-validation test sets defined in [9]. We can see that both our method and the method of [21] consistently outperform the baseline over all of the test sets, while our method performs best among the three. This clearly demonstrates the benefit of exploiting the attribute-relationship constrains. In fact, the average performance using our method is $85.5\% \pm 0.6\%$, compared to $83.4\% \pm 0.5\%$ for the baseline algorithm and $84.6\% \pm 0.6\%$ for the method of [21]. It is worth mentioning that the performance of our method is comparable to the state of the art results of 85.2% in [13] , without using more advanced techniques of feature combination and kernel-based classifier, although we plan to extend our method to its kernel-version in the near future.



(a) (b)

Figure 3: Comparison of our method with [13][21] on the LFW dataset: a) overall ROC curve, b) detailed performance on ten cross-validation test sets.

Table.1 gives the comparative performance with respect to whether using attribute accuracy prior and attribute-relationship prior respectively for model constraints. It can be seen that incorporating information from the accuracy of attribute classifiers improves the performance of baseline classifier from 83.4% to 84.5%, which is further improved to 85.5% by adopting attribute relationship constraints.

Table 1: Comparative average performance among four methods: linear SVM (L.SVM), linear SVM with augmented features (L.SVM + augm.Fea), linear SVM with attribute accuracy prior (L.SVM + acc.), and linear SVM with both attribute accuracy and attribute-relationship prior (L.SVM + acc.+ graph).

L.SVM	L.SVM + agum.Fea.	L.SVM + acc.	L.SVM + acc.+ graph
83.4±0.5	84.6±0.6	84.5±0.6	85.5±0.6

Fig.4 lists some typical highly related attributes learnt using the method in Section 3.1. These attributes can be broadly divided into two categories: 1) with high semantic correlation (shown in yellow rectangle), and 2) with high statistical co-occurrence (shown in green rectangle). We can see that the learnt attribute relationship is reasonable. For example, the semantic concept of "male" has high co-occurrence with male-specific attributes such as wearing necktie, bushy eyebrows, while with negative correlations with things commonly used by female, such as lipstick, necklace, earrings, and so on. As another example, we see that an "attractive woman" is usually "heavily markuped" 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.4, '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 dataset.

Figure 4: Illustration of highly related attributes learnt by our method. On the left-most column we show the typical semantic concepts in bold, and on the right we list the related attributes correlated to those concepts.

Fig.5 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 left-most three columns are respectively from the same subject but are misjudged as from different subjects with high confidence by the baseline classifier. However, our method does not make such mistakes. On the other hand, in the right-most three columns, we show three pairs of face images from three different subjects, which are unfortunately incorrectly identified as each pair from the same subject by the baseline classifier. However, in all these three cases our method is able to make 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.

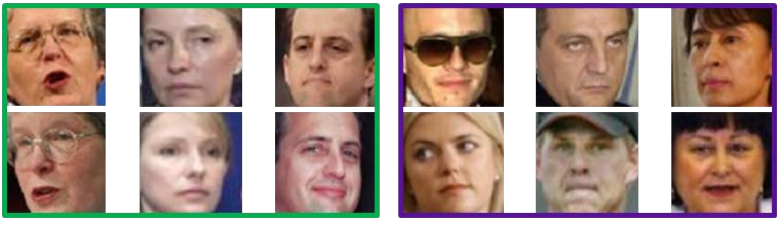


Figure 5: Illustration of three pairs of face images from the same subject respectively (the left-most three columns) and three pairs from different subjects (the right-most three columns). All the six pairs are mistakenly identified by the linear SVM classifier but are correctly recognized with our method.

5 Conclusion

In this paper we give a novel method to model the relationship between attributes, which effectively allows our classifier to aware the hidden correlation among attributes in a general context of subjects. We show in this paper on the challenging LFW database that the mined attribute graph does reflect some aspects of the semantic relationship among attributes existed in real world, and furthermore, such relationship is beneficial to improve the accuracy and robustness of the face verification system. The proposed method is general and can be used beyond the task of face verification, which will be the focus of our future research.

Acknowledgements: The work was financed by the (key) National Science Foundation of China (61073112, 61035003) and the Fundamental Research Funds for the Central Universities (CXLX11_0204). We thank the anonymous reviewers for their helpful comments.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. *European Conference on Computer Vision(ECCV 2010)*, pages 663–676, 2010.

[3] R.S. Bhatt and C. Rovee-Collier. Infants' forgetting of correlated attributes and object recognition. *Child development*, 67(1):172–187, 1996.

[4] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: Poselet-based attribute classification. In *International Conference on Computer Vision (ICCV 2011)*, 2011.

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1778–1785, 2009.

[6] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2352–2359, 2010.

[7] Li. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, pages 524–531, 2005.

[8] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, 2007.

[9] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[10] L. Jacob, G. Obozinski, and J.P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

[11] Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, pages 543–550, 2010.

[12] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision (ICCV 2009)*, pages 365–372, 2009.

[13] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (99):1–1, 2011.

[14] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition (CVPR 2009)*, pages 951–958, 2009.

[15] Ken McRae, Virginia R. de Sa, and Mark S. Seidenberg. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99–130, 1997.

[16] A.S. MOSEK. The mosek optimization software. *Online at http://www. mosek. com*.

[17] D. Parikh and K. Grauman. Relative attributes. In *International Conference on Computer Vision (ICCV 2011)*, pages 503–510, 2011.

[18] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, pages 1641–1648, 2011.

[19] Y. Su, M. Allan, and F. Jurie. Improving object classification using semantic attributes. In *British Machine Vision Conference (BMVC 2010)*, pages 26–1, 2010.

[20] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *International Conference on Computer Vision (ICCV 2009)*, pages 537–544, 2009.

[21] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. *European Conference on Computer Vision(ECCV 2010)*, pages 155–168, 2010.

[22] X.Tan, S.Chen, Z.-H. Zhou, and F. Zhang. Recognizing partially occluded, expression variant faces from single training image per person with som-based knn ensemble. *IEEE Transactions on Neural Networks*, pages 875–886, 2005.

[23] Y. Zhou, R. Jin, and S.C.H. Hoi. Exclusive lasso for multi-task feature selection. *JMLR*, pages 988–995, 2010.