Joint Learning of Semantic and Latent Attributes

Peixi Peng[1,4], Yonghong Tian[1,4(✉)], Tao Xiang[2], Yaowei Wang[3(✉)], and Tiejun Huang[1]

[1] National Engineering Laboratory for Video Technology, Peking University, Beijing, China
{pxpeng,yhtian,tjhuang}@pku.edu.cn
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
t.xiang@qmul.ac.uk
[3] Department of Electronic Engineering, Beijing Institute of Technology, Beijing, China
yaoweiwang@bit.edu.cn
[4] Cooperative Medianet Innovation Center, Beijing, China

Abstract. As mid-level semantic properties shared across object categories, attributes have been studied extensively. Recent approaches have attempted joint modelling of multiple attributes together with class labels so as to exploit their correlations for better attribute prediction and object recognition. However, they often ignore the fact that there exist some shared properties other than nameable/semantic attributes, which we call latent attributes. Basically, they can be further divided into discriminative and non-discriminative parts depending on whether they can contribute to an object recognition task. We argue that learning the latent attributes jointly with user-defined semantic attributes not only leads to better representation for object recognition but also helps with semantic attribute prediction. A novel dictionary learning model is proposed which decomposes the dictionary space into three parts corresponding to semantic, latent discriminative and latent background attributes respectively. An efficient algorithm is then formulated to solve the resultant optimization problem. Extensive experiments show that the proposed attribute learning method produces state-of-the-art results on both attribute prediction and attribute-based person re-identification.

Keywords: Attribute learning · Latent attributes · Person re-identification · Zero-shot learning · Dictionary learning

1 Introduction

Attributes are a type of mid-level semantic properties of visual objects that can be shared across different object categories. Typically, semantic attributes are

B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, pp. 336–353, 2016.
DOI: 10.1007/978-3-319-46493-0_21

Fig. 1. (a) Given only three user-defined attributes as representation, the two people are mis-matched. (b) When complemented by latent attributes, the representation is more discriminative and solving the person re-identification problem becomes easier.

nameable and often annotated based on a user-defined ontology. Attribute learning has been studied extensively recently [1–9]. Existing approaches vary drastically depending on the objectives of learning attributes. Specifically, attributes learning methods have been developed for three objectives: (1) attribute prediction for image search [10], where each image is indexed by a list of predicted attributes and can thus be searched by text queries; (2) learning mid-level representation from low-level features for object recognition, typically at the fine-grained [6] or instance-level [11]; (3) zero-shot learning where given an attribute 'prototype' [12], unseen classes can be recognised by comparing the prototypes with the predicted attributes.

Earlier attribute learning works often tried to learn a set of binary attribute classifiers for each attribute separately and independently, whilst ignoring the existence of correlations among them, e.g., 'female' and 'long-hair' are correlated. This has been rectified by recent approaches [4–9] which jointly learn multiple attributes together with the object class labels so as to exploit their correlations. However, all these joint modelling approaches focus on the semantic user-defined attributes only, whilst ignoring the factors that (1) semantic attributes are often not exhaustively defined; and (2) there are also other shareable but not nameable/semantic properties. We call these shareable but undefined properties *latent attributes* and argue that they should also be jointly modelled with the user-defined semantic attributes and object class labels.

Jointly learning semantic and latent attributes is important for both attribute prediction and object recognition. This is due to two reasons: First, these latent attributes can also be discriminative and thus useful for object recognition. For example, Fig. 1 shows that a limited list of user-defined semantic attributes are often inadequate for instance-level object recognition such as person re-identification [13]. However, when a set of complementary and discriminative latent attributes are learned to augment the user-defined semantic attributes, recognition can be made easier. Second, even if predicting the user-defined attributes is the only goal, discovering and learning these latent attributes is still useful – it is certain that shareable properties that do not belong to the user-defined attributes are accounted for the model rather than act as a distractor to corrupt the learned semantic attribute predictor. Furthermore, by modelling latent attributes together with class labels, we can identify two types of latent attributes: those that are related to class labels and thus are potentially useful

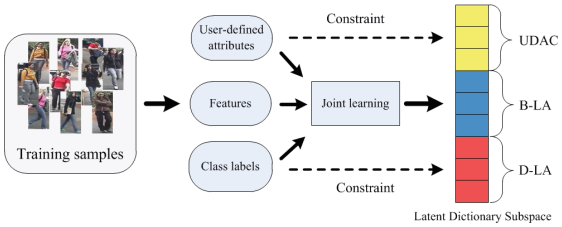

Fig. 2. Our framework for joint learning of user-defined-attribute-correlated (UDAC), discriminative latent attribute (D-LA), and background latent attribute (B-LA) dictionary subspaces.

for object recognition, and those that are not. The former is called discriminative latent attributes (D-LA), while the latter background latent attributes (B-LA) which could literally be object background that might appear in any object class.

To jointly learn both types of latent attributes as well as semantic attributes together with their correlations with the class labels, we propose a novel dictionary learning model with dictionary decomposition. Dictionary learning is naturally suited for learning a low-dimensional subspace corresponding to the latent attribute space. This is because by sharing the same dictionary with all object classes, it automatically discovers shareable properties. More importantly, we can easily decompose the learned dictionary into multiple parts and different parts are subject to different correlations with the available object annotations. Specifically, the learned dictionary subspace are decomposed into three parts: (1) The D-LA dictionary subspace part that is subject to the label correlation constraint so as to make sure that it is discriminative, (2) The B-LA dictionary subspace part that only helps data reconstruction and is subject to no constraint, and (3) The user-defined-attribute-correlated (UDAC) dictionary subspace part which is correlated to the user-defined attribute annotations. Note that in our framework, the user-defined attributes are learned through the latent attribute space. This is because a dictionary learning model aims to reconstruct the original signal using all dictionary atoms together, enforcing the learned three different types of attributes to be complementary to each other. Figure 2 illustrates the proposed dictionary learning framework.

2 Related Work

Learning Semantic Attributes. Earlier works on semantic attribution learning [1–3,14] consider predicting each attribute as a binary classification problem and solve them independently. Later works [4,5,7–9] realised that there exist correlations between different attributes, as well as between attributes and class labels, and proposed to learn different attributes jointly together with the class labels. For example, a unified multiplicative framework is proposed in [7] which projects images and category/class information into a shared feature space and the latent factors were disentangled and multiplied for attribute prediction. In [9,15], they learn the semantic attributes by incorporating class label information. Our model also learns user-defined semantic attributes and class labels

jointly. Different from existing jointly attribute modelling works, we additionally model discriminative latent attributes and background latent attributes to improve the learn of user-defined semantic attributes as well as making the learned attribute-based representation more discriminative for the object classification task.

Learning Latent Attributes. The method for learning discriminative latent attributes has been exploited before [6,16–21]. However, in theses works, the latent attributes are not learned jointly and thus are not necessarily complementary to the user-defined attributes. Comparing to the few exceptions which learn them jointly [22,23], there is a significant difference: by using an additive dictionary, we aim to reconstruct the original feature representation; we thus devise the third type of attributes: background latent attributes (B-LA) to explicitly account for non-discriminant part of the representation (e.g. scene background, or what a person looks like in general) that is useless for the targeted task but has to be learned to avoid corrupting the other two types of useful attributes. Experimental results demonstrate clearly the importance of learning all three jointly. This novel concept can also be applied to existing joint attribute learning models.

Attribute-Based Person Re-identification. Semantic attributes have been exploited as a mid-level representation for matching people across non-overlapping camera views, or the person re-identification (Re-ID) problem [13,24–26]. However, these attribute-based Re-ID representations are not competitive on the benchmark datasets. This is because (1) the user-defined attribute representations have very low dimensions (dozens vs. tens of thousands for the typical low-level feature based representations used by the state-of-the-art Re-ID methods [27]); and (2) no latent attributes are exploited. Recently, user-defined attributes and low-level feature are modelled jointly in [28] in a multi-task learning framework to learn a discriminative representation for Re-ID. However, the user-defined attributes are predicted independently and no latent attributes are used. In contrast, our model is flexible in that discriminative latent attributes can still be learned when no annotation on user-defined attributes is available. Another relevant work is [11] which deploys a generative model to transfer attribute annotations from auxiliary data (fashion clothing) to the target data (surveillance video). Again, as a generative model, it is weak in learning discriminative representation.

Dictionary Learning. Beyond attribute learning, dictionary learning [29,30] has been studied widely as a method for learning a low-dimensional subspace. Originally designed for unsupervised learning, it has been extended for supervised learning for tasks such as face verification/recognition [31] and person Re-ID [32–34]. Our model is related to these dictionary-learning-based Re-ID models in that all models learn discriminative latent attributes through the learned dictionary subspace. However, only our model is able to additionally learn user-defined attributes and background latent attributes for better representation learning.

Contributions. Our contributions are as follows: (1) A unified framework for learning both user-defined semantic attributes and discriminative latent attributes is proposed. (2) We further develop a novel dictionary learning model which decomposes the learned dictionary subspace into three parts corresponding to the semantic, discriminative latent as well as background latent attributes respectively. An efficient optimisation algorithm is also formulated. Extensive experiments are carried out on benchmark attribute prediction and person Re-ID datasets. The results show that the proposed unified framework generates state-of-the-art results on both tasks.

3 Methodology

3.1 Formulation

Assume that a set of training data are given which are labelled with some user-defined (semantic) attributes[1] and object classes. In this paper, we focus on the problem of learning user-defined semantic and latent attributes jointly by dictionary learning. Specifically, the learned dictionary are decomposed into following three parts (see Fig. 2): (1) D^u corresponding to the user-defined-attribute-correlated (UDAC) dictionary subspace part which is correlated to the user-defined attribute annotations, (2) D^d corresponding to the discriminative latent attributes (D-LA) dictionary subspace part which is correlated to the class labels and thus useful for the given classification/recognition task, and (3) D^b corresponding to the background latent attributes (B-LA) dictionary subspace part which captures all the residual information in the training data which is uncorrelated to either user-defined attributes or class labels and thus is learned without any supervision.

Formally, we assume $Y \in \mathbb{R}^{m \times n}$ is a data matrix where each column y_i corresponds to an m-dimensional feature vector representing the i^{th} object's appearance. n denotes the numbers of training samples. A is a $p \times n$ matrix where each column $a_i \in \{0,1\}^p$ indicates the absence or presence of all p binary user-defined attributes. The proposed method can be formulated as:

$$\begin{aligned}\left[D^u, D^d, D^b, W\right] = \arg\min & \left\|Y - D^u X^u - D^d X^d\right\|_F^2 + \left\|Y - D^u X^u - D^d X^d - D^b X^b\right\|_F^2 \\ & + \alpha \sum_{i,j=1}^{n} m_{i,j} \left\|x_i^d - x_j^d\right\|^2 + \beta^2 \left\|X^u - WA\right\|_F^2 . \\ s.t.\ & \left\|d_i^u\right\|_2^2 \leq 1,\ \left\|d_i^d\right\|_2^2 \leq 1,\ \left\|d_i^b\right\|_2^2 \leq 1,\ \left\|w_i\right\|_2^2 \leq 1\ \forall i,\end{aligned} \tag{1}$$

where matrices X^u, X^d and X^b are codes/coefficients corresponding to dictionaries D^u, D^d and D^b respectively; W is used to build correspondence between the codes obtained using D^u and the user-defined attribute annotation matrix A; d_i^u, d_i^d, d_i^b and w_i are the i^{th} columns of D^u, D^d, D^b and W respectively; x_i^d is the i^{th} column of X^d; α and β are free parameters controlling the strengths of

[1] We will show later that the requirement on the availability of user-defined attributes can be removed.

two regularisation terms to be explained later; M is an affinity matrix indicating the class-relationships (same/different class) among different training samples. Specifically, $m_{i,j} = 1$ if x_i^d and x_j^d are of same class, and $m_{i,j} = 0$ otherwise. The third term can be rewritten using the Laplacian matrix as:

$$\sum_{i,j=1}^{n} m_{i,j} \left\| x_i^d - x_j^d \right\|^2 = \mathrm{Tr}(X^d L X^{d\prime}), \tag{2}$$

where $L = Q - M$ and Q is a diagonal matrix whose diagonal elements are the sums of the row elements of M. There are four terms of three categories in the cost function which are now explained in detail:

1. The first two terms are reconstruction errors that make sure the learned dictionaries can encode the data matrix Y well. Note that the two reconstruction error terms are stepwise ordered. Specifically, the minimisation of the first reconstruction error term enables D^u and D^d to encode Y as much as possible, while the minimisation of the second reconstruction error term enables D^b to encode and align the residual part of Y that cannot be coded by D^u and D^d. This stepwise two reconstruction error term formulation is important to prevent the B-LA D^b from dominating the reconstruction error leading to trivial solutions for D^u and D^d.
2. The third term is a graph Laplacian regularisation term which dictates that the projections of columns of Y in the D-LA subspace, i.e., X^d are close to each other if the corresponding data points belong to the same class. The goal of this term is thus to make the D-LA subspace parametrised by D^d to be more discriminative (class-dependent).
3. The last term is the constraint for learning the UDAC subspace part D^u. Note that we attempt to establish a linear constraint W between the projection in that subspace, X^u and user-defined attribute annotations A, rather than simply setting them to be equal ($X^u = A$), because user-defined attributes are always not additive. As explained earlier, modelling user-defined attributes via the same dictionary subspace makes the learned other two types of latent attributes to be complementary to the user-defined attributes.

3.2 Optimisation

Here we detail how the optimisation problem in (1) is solved. The problem is divided into the following subproblems:

1. *Computing codes* X^u. Given fixed D^u, D^d, D^d, W, X^d and X^b, the coding problem of estimating X^u becomes:

$$\min \left\| \tilde{Y} - \tilde{D} X^u \right\|_F^2, \tag{3}$$

where

$$\tilde{Y} = \begin{bmatrix} Y - D^d X^d \\ Y - D^d X^d - D^b X^b \\ \beta W A \end{bmatrix}, \tilde{D} = \begin{bmatrix} D^u \\ D^u \\ \beta I \end{bmatrix},$$

and I is the identity matrix. Let the derivative of (3) equal to 0 and the analytical solution of X^u can be obtained with:

$$X^u = \left(\tilde{D}'\tilde{D}\right)^{-1}\tilde{D}'\tilde{Y}. \tag{4}$$

2. *Computing codes* X^d. Given the other variables fixed, the coding problem of X^d becomes:

$$\min \left\|\tilde{Y} - \tilde{D}X^d\right\|_F^2 + \alpha \mathrm{Tr}(X^d L X^{d'}), \tag{5}$$

where

$$\tilde{Y} = \begin{bmatrix} Y - D^u X^u \\ Y - D^u X^u - D^b X^b \end{bmatrix}, \tilde{D} = \begin{bmatrix} D^d \\ D^d \end{bmatrix}.$$

and the analytical solution of x_i^d (the i^{th} column of X^d) is:

$$x_i^d = \left(\tilde{D}'\tilde{D} + 2\alpha l_{t,i,i} I\right)^{-1}\left(\tilde{D}'\tilde{y}_i - 2\alpha \sum_{k \neq i} \tilde{y}_k l_{k,i}\right), \tag{6}$$

where $l_{k,i}$ is the (k,i) element of L and $\tilde{y}_i$ is the i^{th} column of $\tilde{Y}$.

3. *Computing code* X^b. Fix other terms and X^b can be solved by:

$$\min \left\|Y - D^u X^u - D^d X^d - D^b X^b\right\|_F^2. \tag{7}$$

Let the derivative of (7) equal to 0 and the analytical solution of X^b is:

$$X^b = \left(D^{b'} D^b\right)^{-1} D^{b'}\left(Y - D^u X^u - D^d X^d\right). \tag{8}$$

4. *Updating dictionaries.* First, when D^b, X^u, X^d and X^b are given, D^u and D^d are estimated by the following optimisation problem:

$$\min \|\mathcal{Y} - \mathcal{D}\mathcal{X}\|_F^2, \ s.t. \ \|d_i^u\|_2^2 \leq 1, \ \left\|d_i^d\right\|_2^2 \leq 1, \tag{9}$$

where

$$\mathcal{D} = [D^u, D^d], \ \mathcal{Y} = [Y, Y - D^b X^b], \ \mathcal{X} = \begin{bmatrix} X^u & X^u \\ X^d & X^d \end{bmatrix}. \tag{10}$$

(9) can be optimised with the Lagrange dual. Thus, the analytical solution of $\mathcal{D}$ is: $\mathcal{D} = (\mathcal{Y}\mathcal{X}')\left(\mathcal{X}\mathcal{X}' + \Lambda\right)^{-1}$, where Λ is a diagonal matrix constructed from all the dual variables. Second, we fix other variables and solve D^b with the following objective function:

$$\min \left\|Y - D^u X^u - D^d X^d - D^b X^b\right\|_F^2, \ \ s.t. \ \left\|d_i^b\right\|_2^2 \leq 1(\forall i), \tag{11}$$

(11) can be solved similar to (9).

Algorithm 1. The proposed algorithm

Input: X_t; initialise D^u, D^d, D^b and W randomly; $X^d \rightarrow \mathbf{0}$,$X^b \rightarrow \mathbf{0}$;
Output: D^u, D^d, D^b, X^u, X^d, X^b and W.
while *Non-convergence* **do**
  Coding problem:
    compute code X^u using (3),
    compute code X^d using (5),
    compute code X^b using (7).
  Updating dictionaries:
    update D^u and D^d using (9),
    update D^b using (11).
  Updating W:
    update W using (12),

5. *Updating* W. Similar to the dictionary updating procedure in Step 4, we fix other variables and solve W by:

$$\min \|X^u - WA\|_F^2, \quad s.t. \ \|w_i\|_2^2 \leq 1(\forall i). \tag{12}$$

(12) can be optimised using the Lagrange dual. The analytical solution of W is: $D^u = (X^u A')(AA' + \Lambda)^{-1}$, where Λ is a diagonal matrix constructed from all the dual variables.

Algorithm 1 summaries the whole algorithm. In practice, we found that it always converges after a few (<50) iterations in our experiments.

3.3 Application to Person Re-ID

In the Person Re-ID problem, we assume that the training images are represented by some feature representation denoted as Y, and labelled with identities encoded in the matrix M, and a set of user-defined attributes A. Once the three dictionaries are learned using the training set as described above, each test image y can be encoded as $\left[x^u, x^d, x^b\right]$ via D^u, D^d and D^b respectively. The encoding problem can be formulated as:

$$\left[x^u, x^d, x^b\right] = \arg\min \left\|y - D^u x^u - D^d x^d - D^b x^b\right\|_2^2 + \gamma \|x^u\|_2^2 + \gamma \left\|x^d\right\|_2^2 + \gamma \left\|x^b\right\|_2^2, \tag{13}$$

where x^u, x^d and x^b are the projections of y in the UDAC, D-LA and B-LA part of the learned dictionary subspaces respectively, and γ is a weight for the regularisation terms. (13) can be solved easily with a linear system. After we obtain x^u, the user-defined attribute vector a can be predicted via the linear constraint W:

$$a = \arg\min \|x^u - Wa\|_2^2 + \gamma \|a\|_2^2. \tag{14}$$

Now, the test sample y can be represented as the predicted user-defined attributes a and D-LA x^d. Simply treating the predicted attributes as features, Re-ID could be performed by score-level fusion of computing the cosine distance of a and x^d between the attribute vectors of a probe sample and a gallery one.

Note that the proposed method can still work without the user-defined attribute annotations A in the training data. In this case, D^u, W and X^u will be dropped and (1) becomes:

$$\begin{aligned}\left[D^u, D^b\right] = \arg\min \left\|Y - D^d X^d\right\|_F^2 + \left\|Y - D^d X^d - D^b X^b\right\|_F^2 + \alpha \sum_{i,j=1}^{n} m_{i,j} \left\|x_i^d - x_j^d\right\|^2, \\ s.t.\ \left\|d_i^d\right\|_2^2 \leq 1\ \left\|d_i^b\right\|_2^2 \leq 1,\ \forall i,\end{aligned} \tag{15}$$

(15) can be solved as a special case of (1). Consequently, the test sample y is represented only by its D-LA x^d, which can be obtained by solving an optimisation problem similar to (13).

3.4 Application to User-Defined Attribute Prediction

In this task, our only goal is to predict the user-defined attributes, hence having a separate D-LA D^d is unnecessary and D^b alone can be used to explain any information that cannot be explained by D^u. Consequently, D^d, X^d and the graph Laplacian regularisation from (1) can be removed, and the optimisation problem for dictionary learning becomes:

$$\begin{aligned}\left[D^u, D^b, W\right] = \arg\min \left\|Y - D^u X^u\right\|_F^2 + \left\|Y - D^u X^u - D^b X^b\right\|_F^2 + \beta^2 \left\|X^u - WA\right\|_F^2 \\ s.t.\ \left\|d_i^u\right\|_2^2 \leq 1,\ \left\|d_i^b\right\|_2^2 \leq 1,\ \left\|w_i\right\|_2^2 \leq 1\ \forall i.\end{aligned} \tag{16}$$

It can also be solved as a special case of (1) with a similar solver as described in Sect. 3.2. Once the model is learned using a training set, a test sample y can be encoded with D^u and D^b by solving an optimisation problem similar to (13). Finally, the user-defined attribute vector a is predicted via (14).

4 Experiments

The proposed attribute learning model is evaluated on three tasks: attribute-based person re-identification (Re-ID), user-defined attribute prediction and zero-shot learning[2].

4.1 Person Re-ID

For this task, our attribute learning model is used to learn a discriminative mid-level representation for matching people across camera views.

Datasets. Four widely used benchmark datasets are chosen for person Re-ID. **VIPeR** [35] contains 1,264 images of 632 individuals from two distinct camera

[2] The code can be downloaded at http://pkuml.com/resources/code.html.

views (two images per individual) featured with large viewpoint changes and varying illumination conditions. All individuals are randomly divided into two equal-sized subsets for training and testing respectively with no overlapping in identity between the two subsets. This random partition process is repeated 10 times, and the averaged performance is reported. For fair comparison, we use the same data splits as in [36]. **PRID** [37] consists of images extracted from multiple person trajectories recorded from two surveillance static cameras. Camera view A contains 385 individuals, camera view B contains 749 individuals, and 200 of them appearing in both the two views. The single shot version of the dataset is used in our experiments as in [36], and we use the same data splits as in [36]. In each data split, 100 people with one image from each view are randomly chosen from the 200 present in both camera views for the training set, while the remaining 100 of View A are used as the probe set, and the remaining 649 of View B are used as gallery. Experiments are repeated over the 10 splits. **iLIDS** [38] has 476 images of 119 individuals captured in an airport terminal from three cameras of distinct viewpoints. It contains heavy occlusions caused by a large number of people and luggages. As in [39], 119 identities are randomly divided into two equal halves, one for training and the other for testing. The reported results are obtained by averaging over 10 trials. **Market-1501** [40] is the biggest re-id benchmark dataset to date, containing 32,668 detected person images of 1,501 identities. Each identity is captured by six cameras at most, and two cameras at least. We use the provided fixed training and test sets in [40], under both the single-query and multi-query evaluation settings.

Attribute Annotation. The training sets of all three datasets have labels indicating the identities of the people. In addition, a total of 105 user-defined attributes have been annotated on each training images in VIPeR, PRID and iLIDs as in [14]. We remove the user-defined attributes which appear in each dataset rarely, and the numbers of the remaining attributes are 85, 56 and 73 for VIPeR, PRID and iLIDs respectively. Note that the attribution annotation is unavailable on Market-1501. As mentioned in Sect. 3.3, our model works with and without the user-defined attributes. For fair comparisons with existing methods which do not use additional attribute annotations, we report results of our model both with and without user-defined attributes.

Features and Evaluation Metric. The low-level feature representation in [36] is employed in our experiments. These include colour histogram, HOG and LBP features which are concatenated resulting in 5,138 dimensions. For evaluation metric, we compute Cumulated Matching Characteristics (CMC) curves. Due to space constraint as well as for easier comparison with published results, we only report the cumulated matching accuracy at selected ranks in tables rather than reporting the actual CMC curves. The only exception is the Market-1501 dataset. Since there are on average 14.8 cross-camera ground truth matches for each query, we additionally use mean average precision (mAP) as in [40].

Parameter Settings. On the VIPeR, PRID and iLIDs datasets, the sizes of D^u, D^d and D^b are set to 100. We found that the performance of our model is

insensitive to the dictionary size when it is between 100 to 200. The size of D^d is increased to 400 for Market-1501 due to the fact that the Market-1501 dataset is much bigger than the other three. The other free parameters in our model, α and β in (1) and γ in (13), are obtained using four-fold cross-validation.

Competitors. Twelve state-of-the-art Re-ID methods are selected for comparison. They fall into five categories: (1) Unsupervised: BoW features [40] based on Colour Names (CN) alone or in combination with Hue-Saturation Histograms (HS) are used to compute l_2 distance. (2) Distance metric leaning based methods: RPLM [41], Mid-level Filter [42], LADF [43], and Similarity Learning [44]. (3) Kernel-based Discriminative subspace learning methods: MFA [39], kLFDA [39], kCCA [36], XQDA [27], and MLAPG [45]; (4) Deep learning based: Improved Deep [46]; (5) Feature fusion based: Metric Ensembles [47]. Note that this method fuses more than one kind of features, which is known to be beneficial to all methods. (6) Attribute-based method: aMTL [28]. This is the most relevant to ours as it also utilises the user-defined attributes. Note that aMTL requires multiple images of each person for training, hence they apply data augmentation to generate more training samples on VIPeR and utilises the multi-shot setting of PRID rather than the single-shot one adopted by most other methods including ours. Furthermore, different from our model, aMTL cannot work without user-defined attributes. For fair comparison, we use the same features and the same training-test splits for the compared methods whenever possible (i.e. when the code is available we use the same features as ours). Three versions of our models are evaluated: "Ours_L" means only latent attributes are learned as representation, that is, the user-defined attribute annotation is not used as do most other compared methods. "Ours_U" means that only user-defined attributes are used to represent a person. "Ours_All" means both the user-defined and latent attributes are used.

Comparative Results. From the results shown in Table 1, we have the following key findings: (1) Even without using the additional attribute annotation, our method Ours_L outperforms all compared method particularly at low ranks. (2) If user-defined attributes are available, the results of Ours_U is very poor, showing that the user-defined attributes cannot represent a person discriminatively without latent attributes, because the user-defined attribute representations have very low dimensions as explained. Ours_All outperforms Ours_L and Ours_U on all datasets. That shows the learned user-defined attributes and discriminative latent attribute are indeed complementary to each other. (3) Compared to the alternative attribute-based Re-ID model aMTL, our model (Ours_All) is clearly better. In particular, the proposed method outperforms aMTL by a large margin even when they used more training data on PRID. In addition, aMTL can only be applied when there are user-defined attribute annotations, whilst our model is not restricted by that.

Table 1. Comparative results on four benchmark Re-ID datasets. '*' means we compare these methods with the same features using the author-provided code. '-' means no reported result is available.

(a) VIPeR

Rank	1	5	10	20
RPLM [41]	27.00	55.30	69.00	83.00
Mid-level [42]	29.11	52.34	65.95	79.87
Similarity[44]	36.80	70.40	83.70	91.70
LADF [43]	30.22	64.70	78.92	90.44
kCCA [36]	37.00	-	**85.00**	**93.00**
MFA* [39]	39.56	69.89	80.38	88.61
kLFDA* [39]	39.87	72.78	81.86	90.19
XQDA [27]	40.00	-	80.51	91.08
Deep [46]	34.81	63.61	75.63	84.49
MLAPG [45]	40.73	-	82.34	92.37
Ours_L	**41.25**	**73.60**	81.77	90.12
Metric Ensembles [47]	45.90	82.09	90.51	95.92
Ours_U	28.39	55.32	65.89	76.01
aMTL [28]	42.30	72.20	81.60	89.60
Ours_All	**45.03**	**74.11**	**83.13**	**90.51**

(b) PRID

Rank	1	5	10	20
RPLM [41]	15.00	32.00	42.00	54.00
kCCA [36]	15.00	-	47.00	60.00
MFA* [39]	20.90	50.30	57.90	68.10
kLFDA* [39]	21.60	51.50	60.00	68.10
Ours_L	**23.60**	**52.60**	**61.70**	**69.70**
Metric Ensembles [47]	17.90	39.00	50.00	62.00
Ours_U	16.30	27.60	34.80	43.20
aMTL [28]	18.00	37.40	50.10	66.60
Ours_All	**26.80**	**55.30**	**62.50**	**71.00**

(c) Market-1501

Query	singleQ		multiQ	
Evaluation metrics	Rank-1	mAP	Rank-1	mAP
BoW (CN) [40]	34.38	14.10	42.64	19.47
BoW(CN+HS) [40]	-	-	47.25	21.88
MFA* [39]	37.56	16.94	48.83	22.41
kLFDA* [39]	38.40	18.36	47.23	21.34
Ours_L	**47.39**	**21.06**	**56.17**	**26.85**

(d) iLIDs

Rank	1	5	10	20
MFA* [39]	49.20	80.41	87.90	94.28
kLFDA* [39]	48.41	78.40	87.73	**96.13**
Ours_L	**52.35**	**81.60**	**88.77**	94.91
Metric Ensembles [47]	50.34	72.50	81.50	91.00
Ours_U	39.19	65.38	75.49	83.52
Ours_All	**56.80**	**82.21**	**90.25**	**95.85**

4.2 User-Defined Attribute Prediction

Datasets and Settings. Three widely used benchmark datasets are chosen in this experiment. **AwA** is composed of 30,475 images from 50 animal categories and each category is annotated with 85 attributes. Following [2,3,9], we divide the dataset into two parts: 40 classes (24,295 images) for training and 10 classes (6,180 images) for testing. For fair comparison with the state-of-the-art methods, we adopt the same 4096-dimensional deep learning features DeCAF [48] provided by [3]. **CUB** contains 11,788 images of 200 bird classes. Each category is annotated with 312 attributes. We split the dataset following [8,9] to facilitate direct comparison with the state-of-the-art methods (150 classes for training and the rest 50 classes for testing). We also extract the same 4096-D DeCAF features as in [9]. **PETA** comprises 10 publicly available small-scale person image datasets totalling 19,000 images. Each image is labelled with 105 attributes. For fair comparison with [14], we follow the same setting and randomly select 9,500 images for training, 1,900 for validation and 7,600 for testing. We repeat 10 times and the average result is reported. As in [14], the same low-level color and texture features are extracted and the prediction results of the same selected 35 attributes are evaluated.

Competitors and Evaluation Metrics. Six state-of-the-art attribute learning approaches are compared. These include Direct Attribute Prediction (DAP)

Table 2. Comparative results on (a) predicting user-defined attributes and (b) zero-shot learning. "*" means same feature are used and "-" means no reported results.

(a) Attributes prediction

Approaches	AwA	CUB	PETA
DAP* [3]	72.80	61.80	69.50
IAP* [3]	72.10	-	-
ALE [8]	65.70	60.30	-
CSHAP* [9]	**74.30**	68.70	-
TbOs* [15]	67.55	68.37	70.20
MRF [14]	-	-	71.10
Ours	73.61	**74.85**	**73.12**

(b) Zero-shot learning

Approaches	AwA	CUB
DAP* [3]	57.23	-
UMF-IS [7]	48.60	18.20
CSHAP [9]	45.60	17.50
SSE-ReLU* [49]	76.33	30.41
Akata et al.* [50]	61.90	40.30
JLSE* [51]	80.46	42.11
Ours	**82.81**	**49.87**

[2,3], Indirect Attribute Prediction (IAP)[2,3], Attribute Label Embedding (ALE) [8], Class-Specific Hypergraph based Attribute Predictor (CSHAP) [9],"Two birds, One stone" (TbOs) [15] and Markov Random Field graph (MRF) [14]. For direct comparison with the reported results in the literature, the attribute prediction performance is measured by mean area under ROC curve (mAUC) on AwA and CUB, while the mean classification accuracy (mACC) is used on PETA.

Comparative Results. We report the user-defined attribute prediction performance in Table 2(a). The results show that the proposed method achieves state-of-the-art performance on CUB and PETA. In particular, on CUB, its mAUC is 6 % higher than the nearest competitor CSHAP. However, it is slightly inferior to CSHAP on AwA.

4.3 Zero-Shot Learning

Since images from different classes may share common attributes, we can recognize images from unseen classes based on transferred attribute concepts, which is referred as zero-shot learning [3]. Specifically, the user-defined attributes learned from seen classes are used to classify the images from unseen classes.

Datasets and Settings. Two benchmark datasets, **AwA** and **CUB**, are used in this experiment. For AWA, 40 classes are chosen as seen classes for training and the remanning 10 classes are chosen as unseen classes for testing. Also, we split CUB as 150 classes for training and 50 classes for testing. For both datasets, we utilize MatConvNet [52] with the "imagenet-vgg-verydeep-19" pretrained model [53] to extract a 4096-dim CNN feature vector for each image (or bounding box). The train-test split and features are as same as [49–51].

Comparative Results. In this experiments, we compare our methods with several state-of-the-art methods and the image classification accuracy is reported. As shown in Table 2(b), the performance of our method is significantly better than the state-of-the-art approaches on both datasets.

Table 3. Evaluation on the contributions of different model components for (a) user-defined attributes (att) prediction on AwA, (b) person Re-ID on VIPeR and (c) zero-shot learning (zsl) results on AWA. Note that D^d is not used for user-defined attributes prediction and zero-shot learning; there is thus no result under 'Without D^d' for AwA.

Dataset	AwA (att)	VIPeR (Re-ID)	AwA (zsl)
Evaluation metrics	mAUC	Rank 1	ACC
Without D^d	-	28.39	-
Without D^b	71.74	41.51	80.26
Without W	68.72	42.36	74.39
Ours_full	**73.61**	**45.03**	**82.81**

4.4 Further Evaluations

Contributions of Model Components. There are several key components in the proposed model (see (1)): (a) two types of latent attributes: D-LA (D^d) and B-LA (D^b) are learned together with the user-defined attributes; and (b) instead of learning it directly as part of the dictionary subspace, we model a linear transformation (W) from the user-defined attributes A to the UDAC dictionary subspace (D^u). In order to evaluate the effectiveness of these two components, we compare our full model (Ours_full) with various striped-down versions of our model. The results in Table 3 show clearly that all these components contribute positively to the final performance of the model.

Running Cost. All algorithms are implemented in Matlab and run on a server with 2.0 GHz CPU cores and 128 GB memory. For person Re-ID on the VIPeR dataset, our model takes 28.29 seconds to train and 0.35 seconds to match 312 images against 312 images. For predicting user-defined attributes on AwA, it takes 2,377 seconds to train and 0.33 seconds to predict 85 user-defined attributes on 6,180 images. It is thus extremely efficient during testing as a linear model.

5 Conclusions

We have proposed a novel attribute learning model which learns user-defined semantic attributes jointly with latent discriminative and background attributes. The model is based on dictionary learning with dictionary decomposition. An efficient algorithm is then formulated to solve the resultant optimization problem. Extensive experiments show that the proposed attribute learning method produces state-of-the-art results on attribute prediction, attribute-based person re-identification and zero-shot learning.

Acknowledgements. This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB351806, the National Natural Science Foundation of China under contract No. 61390515, No. 61425025 and No. 61471042, Beijing Municipal Commission of Science and Technology under contract

No. Z151100000915070 and the National Key Technology and Development Program of China under contract No. 2014BAK10B02. These authors are also supported by Microsoft Research Asia Collaborative Research Program 2016, project ID FY16-RES-THEME-034.

References

1. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1778–1785 (2009)
2. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 951–958, June 2009
3. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Machine Intell. **36**(3), 453–465 (2014)
4. Mahajan, D., Sellamanickam, S., Nair, V.: A joint learning framework for attribute models and object descriptions. In: IEEE International Conference on Computer Vision, pp. 1227–1234 (2011)
5. Jayaraman, D., Sha, F., Grauman, K.: Decorrelating semantic visual attributes by resisting the urge to share. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1629–1636 (2014)
6. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 155–168. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15555-0_12
7. Liang, K., Chang, H., Shan, S., Chen, X.: A unified multiplicative framework for attribute learning. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2506–2514, December 2015
8. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 819–826 (2013)
9. Huang, S., Elhoseiny, M., Elgammal, A., Yang, D.: Learning hypergraph-regularized attribute predictors. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 409–417 (2015)
10. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: Interactive image search with relative attribute feedback. Int. J. Comput. Vis. **115**(2), 185–210 (2015)
11. Shi, Z., Hospedales, T.M., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: Computer Vision and Pattern Recognition (CVPR) (2015)
12. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive multi-view zero-shot learning. IEEE Trans. Pattern Anal. Mach. Intell. **37**(11), 2332–2345 (2015)
13. Layne, R., Hospedales, T.M., Gong, S.: Attributes-Based Re-identification. Springer, London (2014)
14. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the ACM International Conference on Multimedia, pp. 789–792 (2014)
15. Li, Y., Wang, R., Liu, H., Jiang, H., Shan, S., Chen, X.: Two birds, one stone: jointly learning binary code for large-scale face image retrieval and attributes prediction. In: IEEE International Conference on Computer Vision, pp. 3819–3827 (2015)

16. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 771–778 (2013)
17. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7573, pp. 73–86. Springer, Heidelberg (2012)
18. Rifai, S., Bengio, Y., Courville, A., Vincent, P., Mirza, M.: Disentangling factors of variation for facial expression recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 808–822. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_58
19. Berg, T.L., Berg, A.C., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 663–676. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15549-9_48
20. Rastegari, M., Farhadi, A., Forsyth, D.: Attribute discovery via predictable discriminative binary codes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 876–889. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_63
21. Feng, J., Jegelka, S., Yan, S., Darrell, T.: Learning scalable discriminative dictionary with sample relatedness. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1645–1652 (2014)
22. Fu, Y., Hospedales, T.M., Tao, X., Gong, S.: Learning multimodal latent attributes. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 303–316 (2014)
23. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Augmented attribute representations. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 242–255. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33715-4_18
24. Layne, R., Hospedales, T.M., Gong, S.: Towards Person Identification and Re-identification with Attributes. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 402–412. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_40
25. N Hospedales, T., Layne, R., Gong, S.: Re-id: hunting attributes in the wild. In: British Machine Vision Conference (BMVC) (2014)
26. Layne, R., Hospedales, T.M., Gong, S.: Person re-identification by attributes. In: British Machine Vision Conference (2012)
27. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR, pp. 2197–2206 (2015)
28. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3739–3747, December 2015
29. Kenneth, K., Joseph, M., Bhaskar, R., Kjersti, E., Te-Won, L., Terrence, S.: Dictionary learning algorithms for sparse representation. Neural Comput. **15**(2), 349–396 (2003)
30. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Sig. Proces. **54**, 4311–4322 (2006)
31. Guo, H., Jiang, Z., Davis, L.S.: Discriminative dictionary learning with pairwise constraints. In: Proceedings of the 11th Asian conference on Computer Vision (2014)

32. Zheng, J., Jiang, Z.: Learning view-invariant sparse representations for cross-view action recognition. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3176–3183. IEEE (2013)
33. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
34. Karanam, S., Li, Y., Radke, R.J.: Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
35. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3. Citeseer (2007)
36. Lisanti, G., Masi, I., Del Bilmbo, A.: Matching people across camera views using kernel canonical correlation analysis. In: Proceedings of ICDSC (2014)
37. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21227-7_9
38. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
39. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_1
40. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124, December 2015
41. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 780–793. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_56
42. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Proceedings of CVPR (2014)
43. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.: Learning locally-adaptive decision functions for person verification. In: CVPR (2013)
44. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1565–1573 (2015)
45. Liao, S., Li, S.Z.: Efficient PSD constrained asymmetric metric learning for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), December 2015
46. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
47. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. arXiv preprint (2015). arXiv:1503.01543
48. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. University of California Berkeley, Brigham Young University, pp. 647–655 (2013)

49. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4166–4174, December 2015
50. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2927–2936, June 2015
51. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
52. Vedaldi, A., Lenc, K.: Matconvnet - convolutional neural networks for matlab. Eprint Arxiv (2016)
53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Computer Science (2014)