On the Exploration of Joint Attribute Learning for Person Re-identification

Joseph Roth, Xiaoming Liu

Department of Computer Science and Engineering, Michigan State University,
{rothjos1,liuxm}@cse.msu.edu

Abstract. This paper presents an algorithm for *jointly* learning a set of mid-level attributes from an image ensemble by locating clusters of dependent attributes. Human describable attributes are an active research topic due to their ability to transfer between domains, human understanding, and improvement to identification performance. Joint learning may allow for enhanced attribute classification when there is inherent dependency among the attributes. We propose an agglomerative clustering scheme to determine *which* sets of attributes should be learned jointly in order to maximize the margin of performance improvement. We evaluate the joint learning algorithm on a set of attributes for the task of person re-identification. We find that the proposed algorithm can improve classifier accuracy over both independent or fully joint attribute classification. Furthermore, the enhanced classifiers also improve performance on the person re-identification task. Our algorithm can be widely applicable to a variety of attribute-based visual recognition problems.

1 Introduction

Person re-identification seeks to locate the same individual across multiple non-overlapping cameras within a short time frame [1]. As an enabling technique for video surveillance [2, 3], it has many applications such as tracker linking, person retrieval, searching missing children in public spaces, etc. Depending on the applications, person re-identification can be posed in different scenarios. For example, classic *person re-identification* is image-to-image matching where one image is the occurrence of the person of interest in one of the cameras. *Zero-shot identification* is description-to-image matching where the only prior knowledge is a verbal description by an eyewitness.

While many prior work of person re-identification rely on low-level visual feature based image matching [4–8], recently human describable, mid-level attributes have become a promising approach for both re-identification [9] and zero-shot identification [10] scenarios. This is especially true for the latter where describable attributes are the *only* source of input information. These attributes have a number of advantages over low-level visual features. First, they enable the possibility of human-in-the-loop to assist decision making. Second, they can improve the system performance by fusion with low-level features. Third, human understanding of the attributes allows for their use as evidences in a courtroom.

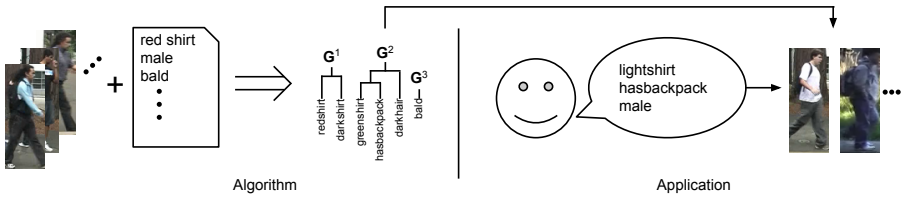


Fig. 1. Given an image ensemble with labels on a set of attributes, our *algorithm* automatically partitions the attribute set into various clusters and jointly learns a classifier for multiple attributes within each cluster. This leads to superior performance in both attribute classification and person re-identification *application* (e.g., zero-shot identification).

In order to detect multiple attributes from an image, normally an array of classifiers are *independently* learned from training data - one classifier per attribute [9]. However, there are various potential *dependencies* among attributes that may enable a better approach to learning. Correlation in attribute occurrence may exist. For example, knowledge that a person is a male will impact the prior probability about the hair length. Another potential dependency is the subset of low-level features that define the attributes. Attributes about the same local area (e.g., wearing jeans or skirt) will likely share a common set of low-level features. Recent works recognize there exist dependencies [9, 11], and few [12, 13] seek to leverage them to *jointly* learn attribute classifiers from the features.

This paper aims to explore *whether* and *how* joint learning can improve attribute classification performance. As shown in Fig. 1, we propose an approach for jointly learning attributes by leveraging their dependencies. Given a set of images labeled with a set of attributes, we recognize that *not all* attributes have strong dependencies and therefore it is desirable to identify clusters of attributes for which joint learning will have greater impact. We propose a data-driven, agglomerative clustering scheme where each attribute begins in a separate cluster and we iteratively combine clusters based on the expected improvement from joint classification. This scheme efficiently partitions the attributes into K clusters, where K is estimated in a data-driven manner. We then train a set of classifiers, one for each cluster of related attributes. To predict the attributes for an unseen image, the image is given to the set of classifiers, which collectively assign all attribute labels. Using the person re-identification datasets, VIPeR [14] and PRID [15], we evaluate the joint learning on a challenging set of human labeled attributes [9] with little inter-attribute correlation. We demonstrate superior attribute classification performance of the proposed algorithm, and also improvement on zero-shot person identification using the predicted attributes.

In summary, this paper has two main contributions:

– We develop a joint attribute classification algorithm that leverages attribute dependencies to learn a set of attribute classifiers. Our algorithm can automatically determine the attribute combinations for joint learning.

– We demonstrate that joint learning improves classification accuracy of human labeled attributes for person re-identification and also improves zero-shot identification.

2 Related Work

Attribute-based visual analysis is a popular research topic. Computer vision has increasing interest in describing objects by a rich set of human describable attributes [16–18]. For example, [19] presented a system with 65 attributes for unconstrained face recognition, which performs well on the labeled faces in the wild dataset. Soft biometrics have been used to improve commercial face matchers [20]. For person re-identification, [9, 10] explored the use of human describable attributes either computed directly from the low-level features or provided by a human operator. Most recently, [11] used human annotated soft biometrics as ancillary information to improve face recognition at a distance. In all of these works, the performances of the attribute classifiers are crucial to the overall performance of the problem at hand because attribute classification errors will propagate throughout the entire system.

All aforementioned applications have a separate, independently learned classifier for each attribute. Both [19, 9] used an SVM to classify each attribute. Using independent classifiers is the naive approach for multi-attribute classification and may be inefficient since potentially different features need to be extracted for each classifier. With this insight, [21, 22] proposed techniques to find optimal common sets of features in order to make multi-attribute classifications *computationally* efficient. These techniques find a subset of features, which jointly predict different classes where only one class will be present at a time. If $\mathbf{x}$ is the set of features used for classification and $\mathbf{y}$ is the output, these approaches seek to minimize $|\mathbf{x}|$, but do not place any criteria on the classification performance on $\mathbf{y}$. Also, in these works, $\mathbf{y}$ has only one attribute present in a given image whereas our work allows any number of attributes to be present.

A few multi-task learning works try to take advantage of dependencies in order to improve overall performance. These works have the same motivation for joint learning as us and try to maximize the classification performance on $\mathbf{y}$. Most notably [23] presented an approach for using support vector regression to jointly predict multidimensional output. They claimed this exploits the dependencies between variables and reduces the effects of noise in the input. In [24], image based regression (IBR) is proposed to use boosting to predict multiple outputs. One very recent work [25] automatically determines attribute dependency and learns a *single* classifier to jointly predict *all* attributes. These works differ from ours in that they assume *all* attributes should be learned jointly. In contrast, we recognize that *not all* attributes should be combined in order to improve performance, i.e., attributes without dependencies may hurt performance when learned jointly. Indeed, our experiment (Tab. 2) shows that *fully* joint learning does degrade performance.

A set of works [26–28] took advantage of the object labels along with the attribute labels in the joint learning framework to further improve performance for attribute learning. These object labels can act as latent variables or side information [29] for attribute classification since they are not directly in the feature set, but are known during training. For the task of person re-identification, we could use the identity information during training in the same manner if the database includes sufficient images of the same person. Unfortunately, the databases we use only have two images per person so we choose not to use the identity information while training.

There are a few recent works recognizing that fully joint learning may lead to overfit in attribute classification. The authors in [13] clustered attributes manually based on human understanding of relatedness. Features were encouraged to be shared among attributes within the same cluster, while attributes from different clusters were encouraged to use different features. In contrast, we use a data-driven approach to both clustering and feature selection. A data-driven approach was used in [12] to create attribute clusters and learn a separate classifier for each cluster. We use a data-driven approach to estimate the performance margin of clustering, whereas [12] used a regularization of the selected features to cluster attributes that reside in a low rank subspace of selected features.

3 Joint Attribute Learning

This section presents our approach for joint attribute learning. We start by formally defining the problem and objectives. Then we analyze one specific means of simultaneously predicting a set of multiple attributes. Finally, we present our hierarchical clustering scheme to efficiently identify the sets of attributes for joint learning, in order to best improve the attribute classification accuracy.

3.1 Problem Definition

Let us assume there are Q user-defined mid-level attributes to describe "person" in videos, and one example of such attributes is shown in Tab. 1. We denote the collection of attributes as $\mathbb{A} = \{1, 2, \cdots, Q\}$, where each integer corresponds to a particular attribute. The training data of the joint attribute learning includes the low-level visual features of N images $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathbb{R}_D$ and their corresponding attribute labels $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$, $\mathbf{y}_n \in \mathbb{R}_Q$. Here D is the feature dimension of the visual features. Each element of the Q-dim vector $\mathbf{y}_n$ can be either 0 or 1 for binary attributes such as gender and bald, or a real number scaled within $[0, 1]$ for ordinal attributes such as age and weight.

The first objective of joint attribute learning is to learn one classifier $\mathbf{G}(\mathbf{x}) : \mathbb{R}_D \rightarrow \mathbb{R}_Q$ that minimizes attribute classification errors:

$$J(\mathbf{G}) = \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{G}(\mathbf{x}_n)\|^2. \tag{1}$$

While this is a basic objective, it assumes that all attributes should be learned together, but in Sec. 3.3 we will show that K, rather than one, attribute classifier(s) should be learned to minimize $J()$. We will present an algorithm on how to estimate the optimal K value and partition Q attributes into K clusters. Note that when $K = Q$, this degenerates to the conventional approach where one classifier is trained for each individual attribute. For the clarity of presentation, in the next section we first present the fully joint attribute learning where $K = 1$.

3.2 Learning via Image based Boosted Regression

Given a set of attributes, we seek to learn a classifier that predicts all attributes simultaneously. For this task, we use IBR, which has shown success in various vision applications [24] and its regressor formulation is also suitable for predicting both binary and ordinal attributes. We present a brief overview of the basic IBR algorithm. Given training data $\mathbf{X}$ and $\mathbf{Y}$, it learns a classifier in the form,

$$\mathbf{G}(\mathbf{x}) = \sum_{t=1}^{T} \alpha_i \mathbf{h}_t(\mathbf{x}), \tag{2}$$

where α_i is the weight, $\mathbf{h}_t(\mathbf{x})$ is the weak classifier that predicts all Q attributes simultaneously, and is comprised of Q 1-dim weak learners, i.e., $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_Q(\mathbf{x})]^\intercal$.

We use the 1-dim decision stump weak learner $h(\mathbf{x})$, which has a low-level feature $g(\mathbf{x})$, a parity indicator $\tilde{p} \in \{-1, 1\}$, and a threshold θ. That is,

$$h(\mathbf{x}) = \begin{cases} +1 : \tilde{p}g(\mathbf{x}) \geq \tilde{p}\theta, \\ -1 : \text{otherwise.} \end{cases} \tag{3}$$

The low-level feature $g(\mathbf{x})$ may be from the color or texture of a localized region, or commonly used local descriptors.

During each boosting iteration, a weight α and weak classifier $\mathbf{h}(\mathbf{x})$ are chosen by minimizing the cost function,

$$J(\mathbf{G}) = \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{G}(\mathbf{x}_n)\|^2_{\mathbf{B}_1} + \lambda \sum_{n=1}^{N} \|\mu - \mathbf{G}(\mathbf{x}_n)\|^2_{\mathbf{B}_2}. \tag{4}$$

While the first term of this function is similar to Eq. 1, a regularization term with μ equal to the sample mean of $\mathbf{Y}$ is used to diminish overfitting. The matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ are for normalization and are naturally related to the covariance matrix of the attribute set. For details on how to select α and $\mathbf{h}(\mathbf{x})$ from a pool of features and weak classifiers, we refer the readers to [24].

From this formulation, we see that joint learning occurs in part based on the choice of $\mathbf{B}_1$ and $\mathbf{B}_2$. If either is a non-diagonal matrix, it is computationally unfeasible to select the optimal weak classifier $\mathbf{h}(\mathbf{x})$. In contrast, if both $\mathbf{B}_1$ and $\mathbf{B}_2$ are identity matrices, each weak learner $h(\mathbf{x})$ can be optimally chosen independent of the other $Q - 1$ weak learners, and thus an incremental feature

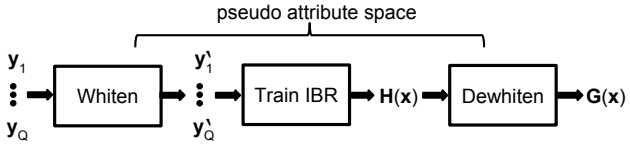


Fig. 2. Flow chart demonstrating a three-step approach to IBR. These steps correspond to Lines 24-26 in Algorithm 1.

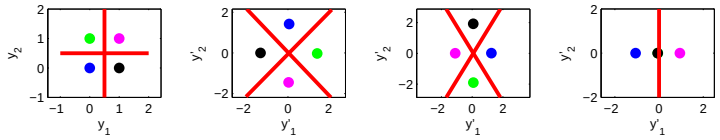


Fig. 3. Example demonstrating the effects of attribute correlation on their whitened attribute space. Red lines are decision boundaries. Figures are initial attribute space $\mathbf{Y}$ (a), and then transformed space $\mathbf{Y}'$ with 0 (b), 0.5 (c), and 1 (d) correlation.

selection scheme can be employed. Based on this observation, [24] suggests a three-step approach to IBR, as shown in Fig. 2. In the first step, the multi-attribute labels $\mathbf{Y}$ are decorrelated via whitening. Specifically, let $\mathbf{D}$ and $\mathbf{V}$ be the eigenvalue and eigenvector matrices of the covariance matrix of $\mathbf{Y}$. We generate uncorrelated pseudo-attribute labels by

$$\mathbf{y}'_n = \mathbf{D}^{-1/2}\mathbf{V}^{\intercal}(\mathbf{y}_n - \mu). \tag{5}$$

The second step learns the regressor $\mathbf{H}(\mathbf{x}) = \sum \alpha_i \mathbf{h}_t(\mathbf{x})$ to predict the uncorrelated labels $\mathbf{Y}'$, by setting $\mathbf{B}_1 = \mathbf{B}_2 = \mathbf{I}$. In the third step, the final attribute classifier $\mathbf{G}(\mathbf{x})$ can be obtained by dewhitening the estimated uncorrelated labels, $\mathbf{G}(\mathbf{x}) = \mu + (\mathbf{V}^{\intercal})^{-1}\mathbf{D}^{1/2}\mathbf{H}(\mathbf{x})$.

It is interesting to note that the joint learning is achieved in this implementation mainly because the whitening and dewhitening *share* the selected features across the attributes. Thus, we hypothesize that any standard regression technique can be applied to predict each one of Q attributes of $\mathbf{Y}'$ independently. In Fig. 3 we demonstrate the effects of whitening on the $\mathbf{Y}'$ attribute space ($Q = 2$) for different amounts of initial correlation.

While the learned regressor $\mathbf{G}(\mathbf{x})$ directly classifies ordinal attributes such as age and height, for binary attributes we must also find a threshold τ to perform classification. We select τ_q for attribute a_q that minimizes the error on the training data. Note that even in a case with all binary attributes, regression is still necessary because the $\mathbf{Y}'$ space is ordinal.

3.3 Attribute Clustering

The aforementioned joint attribute learning assumes that on average the jointly learned classifier can achieve superior classification performance for Q attributes

than Q independently learned attribute classifiers. However, it is important to note that this assumption may not always hold true. Let us consider two simple scenarios when $Q = 2$. First, the two attributes have dependencies and there is a difference between their independent classifier performances. Thus, the dependency between the attributes may be exploited by a joint classifier to improve the performance of the harder-to-classify attribute. Second, the two attributes have no dependency and they both have high independent classifier performance. If we apply joint learning in this scenario, the labeling noise in the training samples or bias in their sampling from the population may cause the joint classifier to assume there to be dependence when there is none. Hence, the joint learning may actually hurt the performance of attribute classification.

Therefore, our goal is to identify the contributing factors and predict *when* **joint learning will improve the attribute classification performance over independent learning, and to use this knowledge to** *partition* **the set of attributes into multiple clusters where the attributes within each cluster may be jointly learned to best improve performance.**

Mathematically we define this process as follows. Let a partitioning $\mathbb{C}$ split all Q attributes into K non-overlapping clusters, i.e., $\mathbb{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$, where $\mathbf{c}_k \subseteq \mathbb{A}$, $\cup_{k=1}^{K} \mathbf{c}_i = \mathbb{A}$, and $\mathbf{c}_{k_1} \cap \mathbf{c}_{k_2} = \emptyset$, $\forall 1 \leq k_1 \neq k_2 \leq K$. The objective of jointly learning K classifiers $\mathbb{G} = \{\mathbf{G}^1, \cdots, \mathbf{G}^K\}$ is to minimize the classification error,

$$J(\mathbb{G}, \mathbb{C}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \|\mathbf{y}_n^k - \mathbf{G}^k(\mathbf{x}_n)\|^2, \tag{6}$$

where both $\mathbf{G}^k(\mathbf{x}_n)$ and $\mathbf{y}_n^k$ are the estimated and true labels of the attributes in the $\mathbf{c}_k$ cluster. As an extension of Eq. 1, this objective function is difficult to optimize since it depends on both $\mathbb{G}$ and $\mathbb{C}$. Therefore we propose a sub-optimal solution by estimating $\mathbb{C}$ and $\mathbb{G}$ sequentially. Most of the remaining section will present our approach to estimate $\mathbb{C}$ since learning $\mathbf{G}^k$ can be easily done by using the IBR approach in Sec. 3.2 or any other multi-attribute predictor.

The estimation of $\mathbb{C}$ is nontrivial due to the large solution space. The number of partitions for a Q-attribute set is equal to the Q^{th} Bell number [30], which grows exponentially and is computationally unfeasible to enumerate as Q increases. Therefore, we propose a greedy approach similar to agglomerative hierarchical clustering. We start by placing each attribute in its own cluster and then iteratively merge the two clusters that are expected to benefit most from joint learning. The merging process stops when we arrive at a single cluster or more likely when the merging of any two clusters no longer has expected improvement.

Specifically, we denote a an attribute of the cluster $\mathbf{c}$, $p(a)$ the classification accuracy of a when learned independently, and $\widehat{p}(a, \mathbf{c})$ the accuracy when learned jointly as a part of $\mathbf{c}$. The performance margin of an attribute cluster is defined

Algorithm 1: Joint attribute learning via attribute clustering.

Data: Attributes $\mathbb{A}$, training samples and labels $\mathbb{D} = \{\mathbf{X}, \mathbf{Y}\}$, validation samples and labels $\mathbb{D}^v = \{\mathbf{X}^v, \mathbf{Y}^v\}$, a flag *useReg*.
Result: The partitioning $\mathbb{C}$, cluster classifiers $\mathbb{G}$, and thresholds $\mathbb{T}$.
`/* Find partitioning */`
Initialize clusters $\mathbf{c}_1 = 1, \mathbf{c}_2 = 2, \cdots, \mathbf{c}_K = Q$ and $K = Q$;
Initialize margins $m(\mathbf{c}_1) = \cdots = m(\mathbf{c}_K) = 0$;
Train Q classifiers from $\mathbb{D}$ and compute $p(a)$ $\forall a \in \mathbb{A}$ on $\mathbb{D}^v$;
if *useReg* **then**
  Learn $R() : \mathbf{f} \rightarrow m(\mathbf{c})$ via $\mathbf{f}$ from $\mathbb{D}$, $\mathbb{D}^v$ and $m(\mathbf{c})$ from $\mathbb{D}^v$;
repeat
  $bestGain = 0$;
  foreach $k_1 = 1, \cdots, K-1$ **do**
    foreach $k_2 = k_1 + 1, \cdots, K$ **do**
      if *useReg* **then**
        Compute $\mathbf{f}$ from $\mathbb{D}$,$\mathbb{D}^v$;
        Compute $m(\{\mathbf{c}_{k_1}, \mathbf{c}_{k_2}\})$ via $R(\mathbf{f})$;
      else
        Train joint classifier on $\mathbb{D}$ via IBR;
        Evaluate $m(\{\mathbf{c}_{k_1}, \mathbf{c}_{k_2}\})$ on $\mathbb{D}^v$;
      if $s(\mathbf{c}_{k_1}, \mathbf{c}_{k_2}) > bestGain$ **then**
        $bestGain = s(\mathbf{c}_{k_1}, \mathbf{c}_{k_2})$;
        $\mathbf{c}^t = \{\mathbf{c}_{k_1}, \mathbf{c}_{k_2}\}$; ▷ Remember the best cluster
  if $bestGain > 0$ **then**
    Merge two clusters into one, $\mathbf{c}_{k_1} = \mathbf{c}^t$, $\mathbf{c}_{k_2} = \emptyset$;
    $K = K - 1$; ▷ One less total number of clusters
until $bestGain \leq 0$;
`/* Train cluster-specific classifiers */`
foreach $k = 1, \cdots, K$ **do**
  Whiten attribute labels from cluster $\mathbf{c}_k$ via Eq. 5;
  Train regressor $\mathbf{H}^k(\mathbf{x})$ as Eq. 2 via IBR on $\mathbb{D}$,$\mathbb{D}^v$;
  $\mathbf{G}^k(\mathbf{x}) = \mu + (\mathbf{V}^\intercal)^{-1}\mathbf{D}^{1/2}\mathbf{H}^k(\mathbf{x})$;
  Compute decision boundaries τ_q for binary attributes on $\mathbb{D}$, $\mathbb{D}^v$;
return $\mathbb{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$, $\mathbb{G} = \{\mathbf{G}^1(\mathbf{x}), \cdots, \mathbf{G}^K(\mathbf{x})\}$, $\mathbb{T} = \{\tau_1, \cdots, \tau_Q\}$.

as the average margin of each attribute in the cluster,

$$m(\mathbf{c}) = \frac{1}{|\mathbf{c}|} \sum_{a_i \in \mathbf{c}} (\widehat{p}(a_i, \mathbf{c}) - p(a_i)). \tag{7}$$

Our objective is to find a partitioning $\mathbb{C}$ that maximizes the average performance margins across all clusters,

$$\hat{\mathbb{C}} = \underset{\mathbb{C}}{\operatorname{argmin}}\, J_1(\mathbb{C}) = \frac{1}{K} \sum_{k=1}^{K} (m(\mathbf{c}_k)|\mathbf{c}_k|), \tag{8}$$

where $|\mathbf{c}_k|$ is the cardinality of $\mathbf{c}_k$.

Our greedy approach starts with Q clusters, each with one distinct attribute. In each iteration, we search for a pair of clusters, which have the maximal expected improvement when combining them into one cluster compared to leaving them as two clusters. Hence, the expected improvement is computed by

$$s(\mathbf{c}_{k_1}, \mathbf{c}_{k_2}) = m(\{\mathbf{c}_{k_1}, \mathbf{c}_{k_2}\}) - \frac{|\mathbf{c}_{k_1}|m(\mathbf{c}_{k_1}) + |\mathbf{c}_{k_2}|m(\mathbf{c}_{k_2})}{|\mathbf{c}_{k_1}| + |\mathbf{c}_{k_2}|}. \tag{9}$$

The iteration continues until all $s(\mathbf{c}_{k_1}, \mathbf{c}_{k_2}) < 0$, which means that there is no expected performance improvement by combining any two existing clusters. Algorithm 1 summarizes our clustering and learning algorithm. In theory the complexity of each iteration is $O(K^2)$. However, with memoization of the $s(\mathbf{c}_{k_1}, \mathbf{c}_{k_2})$, each iteration after the first iteration has a complexity of $O(K)$, which becomes fairly efficient to compute.

The only thing remaining to implement Algorithm 1 is in Lines 11-15, i.e., how to estimate the performance margin of an attribute cluster, $m(\mathbf{c})$. We propose two methods for this estimation: a learned regressor based on the properties of the attribute set $\mathbf{c}$, and an empirical estimation based on its performance on a validation set, as detailed below.

Regression To learn the regressor, we hypothesize that dependencies among attributes may impact the performance margin of joint learning, and these dependencies will be the features for regressor learning. First, the correlation among attributes may help since not all attributes are equally classifiable and the result of the "easier" attribute can benefit the prediction of the "harder" attribute. Thus we define f_0 to be the average of pair-wise correlation coefficients of any two attribute labels $\mathbf{Y}^k$ in the cluster $\mathbf{c}_k$, Second, this line of reasoning also suggests that the performance difference among the independently learned attributes could be indicative, which leads to $f_1 = \text{var}(p(a)), \forall a \in \mathbf{c}$. Third, we note that the independent accuracy $p(a)$ can place restrictions on the joint performance. For example, if an attribute is 99% accurate independently, it is less likely to be improved via joint learning. Hence we define $f_2 = \text{mean}(p(a)), \forall a \in \mathbf{c}$. Finally the correlation of the top selected features from the independently boosted classifiers also matters. If two attributes share many discriminative features, they will be less likely to help improve each other. But if their top features differ, they may potentially help create a more robust joint classifier. We define $\mathbf{f}_3$ to be a 11-dim histogram of the pair-wise correlation coefficients of the top five selected features from each independently learned attribute's classifier. The collection of features $\mathbf{f} = [f_0, f_1, f_2, \mathbf{f}_3^\intercal]^\intercal$ becomes the input variable for the regressor $R(\mathbf{f})$, whose output variable is the expected performance margin $m(\mathbf{c}_k)$. Specifically we train a simple linear regressor by using $\mathbf{f}$ computed on $\mathbb{D}$, $\mathbb{D}^v$ and $m(\mathbf{c}_k)$ computed on $\mathbb{D}^v$, based on a small set of random clusters $\mathbf{c}$.

Validation For the empirical estimation, we evaluate $p(a)$ and $\widehat{p}(a, \mathbf{c})$ on a separate validation set $\mathbb{D}^v$ for every attribute in the cluster, and then compute $m(\mathbf{c})$ directly. This scheme has a computational burden as it has to train a joint classifier each time it examines a potential cluster merge, but it will give us an accurate estimate of the performance margin. Using $R(\mathbf{f})$ to estimate the performance margin requires an upfront computational investment to train the regressor, but after it is trained, it efficiently examines potential merges. Ultimately, we seek to find an accurate prediction $R(\mathbf{f})$ of the margin through the regressor, which is one of the future research directions.

Fig. 4. Sample images from the VIPeR (top) and PRID (bottom) dataset of the 21 attributes used for person re-identification.

Table 1. 21 Mid-level attributes.

redshirt	blueshirt	lightshirt	darkshirt	greenshirt
nocoats	notlightdarkjeanscolour	darkbottoms	lightbottoms	
hassatchel	barelegs	shorts	jeans	male
skirt	patterned	midhair	darkhair	
bald	hashandbagcarrierbag	hasbackpack		

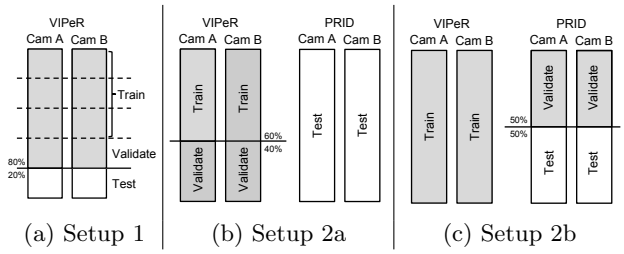
4 Experiments

Our experimental goal is to determine *whether* joint learning improves upon independent learning and *how* it works by exploring the aspects of the regressor that best predict the performance improvement for a set of attributes. We do present the results as reported by [9] solely for reference to show how our baseline (independent learning) performance compares to the state of the art, but the most important question is if the joint learning technique can improve upon the independent formulation.

4.1 Experimental Setup

Datasets We conduct experiments on the classic person re-identification datasets, VIPeR [14] and PRID [15], which contain 632 and 200 subjects respectively with a pair of images taken from different cameras (Cam A and Cam B) with arbitrary poses. The images are cropped and scaled to 128×48 pixels in size for VIPeR and 128×64 for PRID. Layne *et al.* [9] provided 21 manually labeled attributes for each image. Sample images from the databases are shown in Fig. 4 and the list of attributes is in Tab. 1.

Feature representation We extract the same low-level visual features as [5, 9]. Images are split into 6 equal sized horizontal sections. For each section we compute 8 color channels and responses from 19 texture channels of Gabor and Schmid filters. Each channel is quantized by a 16-bin histogram. A total of 2592 low-level features $\mathbf{x}$ are extracted from each image, i.e., $D = 2592$.



(a) Setup 1 (b) Setup 2a (c) Setup 2b

Fig. 5. Experimental setups. Shaded region is the data used to train final $\mathbb{G}$ used for classification. Vertical is across the subject space for each camera.

Table 2. Comparison of independent and joint learning techniques.

	Indep.	Full	Reg.	Valid.
Setup 1: $Q = 16$	74.75%	74.28%	75.18%	**75.31**%
Setup 2a: $Q = 15$	65.63%	64.98%	67.83%	**67.87**%
Setup 2b: $Q = 15$	68.60%	68.20%	**70.43**%	69.17%

Experimental setup Figure 5 displays the two different setups for our experiments. Setup 1 is intra-dataset and examines attribute classification within a single dataset. We elect to only examine VIPeR since it contains sufficient images to make decisions for the attributes, whereas PRID has multiple attributes with only a few positive examples. VIPeR also covers a wider range of camera viewpoints. This setup splits the subjects in five folds and we repeat the experiment using one fold for testing each time. The validation set is used to learn the regressor and to predict the cluster margins for classification purposes. Setup 2 is inter-dataset and examines the ability to transfer attribute classifiers. This setup has two variations. Setup 2a uses completely different datasets for training and testing. Setup 2b uses half of PRID as the validation set to help training.

Evaluation metrics For attribute classification, we report *accuracy*, which is the number of correctly classified samples over the total number of samples. When evaluating person re-identification, we report the *expected rank*, which is the mean rank of genuine matches and provides an estimate of how many images a manual operator will have to examine to find the genuine match. We also report performance at rank n for re-identification, which is the probability that the genuine match appears within the top n matching results.

4.2 Attribute Classification

We evaluate joint attribute learning and report the accuracies for four different techniques in Tab. 2. 1) *Independent*, where each attribute is learned by a separate classifier, i.e., $K = Q$. 2) *Fully joint*, where all attributes are jointly learned in *one* cluster, i.e., $K = 1$. 3) *Regressor clustering*, where we use the regressor

Table 3. Attribute accuracy for clustering in Fig. 6 on the testing set. For each cluster, we also report the predicted margin on the validation set and the actual margin on the test set $\mathbb{D}^t$.

	Indep.	Valid.	$m(\mathbf{c})$ on $\mathbb{D}^v$	$m(\mathbf{c})$ on $\mathbb{D}^t$
male	55.16%	53.97%		
hasbackpack	48.02%	47.62%	4.41%	−0.66%
midhair	66.27%	65.87%		
lightshirt	76.98%	80.95%	2.17%	2.18%
greenshirt	83.73%	84.13%		
darkbottoms	76.98%	75.00%	1.88%	0.79%
lightbottoms	62.30%	65.87%		
redshirt	93.65%	93.65%	1.68%	−1.38%
blueshirt	86.51%	83.73%		
nocoats	67.86%	73.41%		
darkhair	67.46%	67.86%	1.04%	2.68%
notjeanscolor	80.16%	82.54%		
shorts	76.19%	78.57%		
darkshirt	85.71%		—	
barelegs	75.00%		—	
jeans	75.40%		—	

$R(\mathbf{f})$ to form the clustering. 4) *Validation clustering*, where the validation set is used for clustering.

Our hypothesis is that adaptively combining the attributes for joint learning will increase the performance of attribute classification. We evaluate using Setup 1 with the same 16 attributes as used by [9] where the remaining five attributes are not used due to extreme imbalanced samples. Our independent result, 74.75% is better than the 66.9% as reported in [9]. For Setup 2, we can only use 15 attributes because the other six attributes of PRID have very few samples. We make a few comments based on the results reported in Tab. 2. First, fully joint learning has a negative impact on the classifier accuracy. Note that this observation is different to the very recent work [25] where fully joint learning has shown improvement. Part of the reason is that the chosen datasets have little overall correlation among attributes making joint learning difficult. We also note that conceptually [25] is a special case of our algorithm, because it is possible that all attributes are clustered into *one* cluster by our algorithm as long as the expected improvement s is positive. Second, both proposed clustering techniques improve over independent classifiers for all experiments. In Setup 1 we report the average accuracies from using each one of the five folds as the test set. Even though the relative improvement from the "Indep." to "Valid." is small, the p-value from one-tailed paired t-test is less than 0.05, which demonstrates this is a *statistically significant* performance increase.

We also examine a reduced subset of only five attributes in order to compare our greedy clustering with the global optimal clustering, which is obtained through brute force all possible attribute partitions. In this example, clustering with validation achieves 78.4% accuracy barely below the global optimal clustering at 78.5%.

4.3 Regression versus Validation based Clustering

We propose two means of estimating the performance margin of a cluster in the agglomerative clustering scheme, regression and validation. In Fig. 6 and Tab. 3,

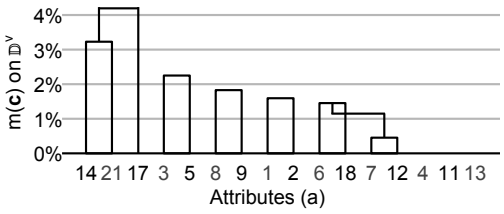


Fig. 6. Clustering of attributes as defined by the validation set and their expected margin for each cluster. Attributes 4, 11, and 13 are determined to be individual clusters by our algorithm.

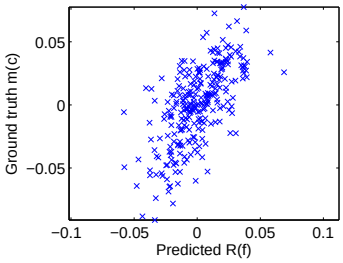


Fig. 7. Predicted margin from clustering regressor versus ground truth margin. The Pearson correlation coefficient is 0.66.

we show the chosen clusters as well as the per attribute accuracy improvement using the clustering with the validation set for one fold of the Setup 1 experiment. It can be observed that some of the clustering results are consistent with human intuition. For instance, darkbottoms and lightbottoms are negatively correlated and hence clustered together. Also, the inconsistency between the $m(\mathbf{c})$ on $\mathbb{D}^v$ and on $\mathbb{D}^t$ indicates the insufficient validation samples, and hence a more representative validation set will help us improve performances in the future.

Using a validation set to find the clusters can take several hours for $Q = 16$. We train $R(\mathbf{f})$ using less time than the validation approach, and once trained it takes a few minutes to cluster $\mathbb{A}$. In our experiments, we only train $R(\mathbf{f})$ on Setup 1 and use the same regressor for Setup 2a and 2b.

As discussed in Sec. 3.3, we define a regressor using dependencies between attributes to predict the performance margin of a cluster. We train the regressor with Setup 1, where we learn joint classifiers for all pairs of attributes and a random selection of sets of three attributes. The ground truth performance margins are computed on the validation set. Figure 7 displays the ability of the regressor to predict the actual margin. Of the four types of attribute dependencies modeled, the descending order of importance as defined by weights of the regressor $R(\mathbf{f})$ is feature correlation ($\mathbf{f}_3$), variance of independent classifier performance (f_1), mean independent classifier performance (f_2), and the attribute correlation (f_0). It might be counter-intuitive that attribute correlation would have the least impact on joint learning, but this is mainly an anomaly caused by the attributes having little correlation by design. There are too few attribute pairs with high correlation to impact the regressor. For example, the average inter-attribute correlation is only 0.07 in VIPeR. This low correlation makes joint attribute learning a very challenging problem, so if we can improve on this dataset, it is likely to improve other problems where more inter-attribute correlation exists. Using a regressor to predict the performance margin demonstrates promise, but further exploration to improve the regressor prediction accuracy is still necessary.

Table 4. Zero-shot identification performance on VIPeR.

	ExpRank	Rank 1	Rank 5	Rank 10
Layne [9]	50.1	6.0%	17.1%	26.0%
Ind.	27.11	6.1%	24.8%	37.8%
Valid.	26.13	7.7%	26.3%	38.1%

4.4 Zero-shot Identification

Improving attribute classification is good, but does joint learning also improve the applications of the attributes? We examine the zero-shot identification scenario [9], where only attribute descriptions of an eyewitness are available. Following the same experimental setup as [9], we use the provided human labels for the subjects $\mathbf{y}$ as the probe and use the raw regression output $\mathbf{G}(\mathbf{x})$ for VIPeR Cam B (Setup 1 test set) as the gallery. The distance metric is computed as the weighted sum of errors between $\mathbf{y}$ and $\mathbf{G}(\mathbf{x})$, i.e., $\tilde{s} = \sum_{k=1}^{K} \mathrm{e}^{\widehat{p}(\cdot,\mathbf{c}_k)} \cdot |\mathbf{y}^k - \mathbf{G}^k(\mathbf{x})|$.

Table 4 reports the results averaged across all five folds for independent and the proposed clustering scheme with the validation set. Joint learning of the attributes improves both the expected rank (smaller is better) and the rank n accuracy (lager is better) at low ranks. To calibrate the zero-shot identification performance of our independently learned attributes, we show the performance of [9] as reported in the paper.

5 Conclusions

We have shown that joint learning of attributes can increase the average attribute classification performance. Our main contribution is the clustering scheme that identifies which sets of attributes should be jointly learned for maximum performance increase. For joint learning, we used IBR, but any multi-output classification algorithm can be substituted. We demonstrated the effectiveness of this joint attribute learning approach on the task of person re-identification and improved zero-shot identification performance.

A common characteristic of exploratory paper is to raise interesting questions and present opportunities for further work. This exploratory work is no exception. How will this translate to other multi-attribute problems such as the face attributes [19, 31]? What other dependencies impact joint learning and can we leverage to improve the clustering prediction? Would accurate body alignment [32] have a positive impact on joint learning? Would we see a larger performance gain from joint learning if there is more inherent correlation among the user-defined attributes, especially on a larger dataset such as [33]? On the contrary, it is interesting to note that even a lack of correlation has the potential to improve joint classifier performance because that knowledge may lead to a selection of different low-level features to predict each attribute. Finally and most importantly, because our approach is independent to the definitions or types of attributes, we believe it is widely applicable to many attribute-based visual recognition problems, which warrants future research on this topic.

References

1. Cai, Q., Aggarwal, J.K.: Tracking human motion using multiple cameras. In: ICPR. (1996) 68–72
2. Liu, X., Tu, P., Rittscher, J., Perera, A., Krahnstoever, N.: Detecting and counting people in surveillance applications. In: AVSS. (2005)
3. Tu, P.H., Doretto, G., Krahnstoever, N.O., Perera, A.G.A., Wheeler, F.W., Liu, X., Rittscher, J., Sebastian, T.B., Yu, T., Harding, K.G.: An intelligent video framework for homeland protection. In: Proc. of SPIE Defense & Security Symposium, Conference on Unattended Ground, Sea, and Air Sensor Technologies and Applications IX, Orlando, Florida (2007)
4. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR. (2006) 1528–1535
5. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV. (2008) 262–275
6. Madden, C., Cheng, E.D., Piccardi, M.: Tracking people across disjoint camera views by an illumination-tolerant appearance representation. Mach. Vision and Appl. **18** (2007) 233–247
7. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC. (2010) 21.1–21.11
8. Yang, Y., Yan, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: ECCV. (2014) 536–551
9. Layne, R., Hospedales, T., Gong, S.: Attributes-based re-identification. In Gong, Cristani, Yan, Loy, eds.: Person Re-Identification. Springer (2013)
10. Thornton, J., Baran-Gale, J., Butler, D., Chan, M., Zwahlen, H.: Person attribute search for large-area video surveillance. In: HST. (2011) 55–61
11. Tome, P., Fierrez, J., Vera-Rodriguez, R., Nixon, M.S.: Soft biometrics and their application in person recognition at a distance. IEEE Trans. Information Forensics and Security **9** (2014) 464–475
12. Kang, Z., Grauman, K., Sha, F.: Learning with whom to share in multi-task feature learning. In: ICML. (2011) 521–528
13. Jayaraman, D., Sha, F., Grauman, K.: Decorrelating semantic visual attributes by resisting the urge to share. In: CVPR. (2014)
14. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE Int. Work. Performance Evaluation Tracking and Surveillance. (2007)
15. Hirzer, M., Beleznai, C., Roth, P., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Proc. Scandinavian Conf. Image Analysis. (2011) 91–102
16. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009) 1778–1785
17. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: CVPR. (2011) 1657–1664
18. Parikh, D., Grauman, K.: Relative attributes. In: ICCV. (2011) 503–510
19. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: ICCV. (2009)
20. Park, U., Jain, A.K.: Face matching and retrieval using soft biometrics. IEEE Trans. Information Forensics and Security **5** (2010) 406–415
21. Shalev-Shwartz, S., Wexler, Y., Shashua, A.: Shareboost: Efficient multiclass learning with feature sharing. In: NIPS. (2011) 1179–1187

22. Torralba, A.B., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE T-PAMI **29** (2007) 854–869
23. Tuia, D., Verrelst, J., Alonso, L., Pérez-Cruz, F., Camps-Valls, G.: Multioutput support vector regression for remote sensing biophysical parameter estimation. IEEE J. Geoscience Remote Sensing Letters **8** (2011) 804–808
24. Zhou, S.K., Georgescu, B., Zhou, X.S., Comaniciu, D.: Image based regression using boosting method. In: ICCV. Volume 1. (2005) 541–548
25. Liu, M., Zhang, D., Chen, S.: Attribute relation learning for zero-shot classification. Neurocomputing **139** (2014) 34–46
26. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV. (2010) 155–168
27. Hwang, S.J., Sha, F., Grauman, K.: Sharing features between objects and their attributes. In: CVPR. (2011) 1761–1768
28. Wang, X., Ji, Q.: A unified probabilistic approach modeling relationships between attributes and objects. In: ICCV. (2013) 2120–2127
29. Chen, J., Liu, X., Lyu, S.: Boosting with side information. In: ACCV. Springer (2013) 563–577
30. Bell, E.T.: Exponential polynomials. Annals of Mathematics **35** (1934) 258–277
31. Klare, B.F., Klum, S., Klontz, J., Taborsky, E., Akgul, T., Jain, A.K.: Suspect identification based on descriptive facial attributes. In: ICJB. (2014)
32. Liu, X., Yu, T., Sebastian, T., Tu, P.: Boosted deformable model for human body alignment. In: CVPR. (2008) 1–8
33. Liao, S., Mo, Z., Hu, Y., Li, S.Z.: Open-set person re-identification. (2014) arXiv:1408.0872v1 [cs.CV].