Max-Margin Multiattribute Learning With Low-Rank Constraint

Qiang Zhang, Lin Chen, *Student Member, IEEE*, and Baoxin Li, *Senior Member, IEEE*

***Abstract*— Attribute learning has attracted a lot of interests in recent years for its advantage of being able to model high-level concepts with a compact set of midlevel attributes. Real-world objects often demand multiple attributes for effective modeling. Most existing methods learn attributes independently without explicitly considering their intrinsic relatedness. In this paper, we propose max margin multiattribute learning with low-rank constraint, which learns a set of attributes simultaneously, using only relative ranking of the attributes for the data. By learning all the attributes simultaneously through low-rank constraint, the proposed method is able to capture their intrinsic correlation for improved learning; by requiring only relative ranking, the method avoids restrictive binary labels of attributes that are often assumed by many existing techniques. The proposed method is evaluated on both synthetic data and real visual data including a challenging video data set. Experimental results demonstrate the effectiveness of the proposed method.**

***Index Terms*— Multi-task learning, relative attribute, low rank, attribute learning, surgical skill.**

I. Introduction

IN VISUAL computing tasks involving modeling of visual objects, such as image-based object class recognition, it has been recognized that some mid-level visual properties, or "attributes", of the objects are not only helpful but even critical to solving the problem [1], [2]. Attributes of a visual object (or object class) characterize the object (or the class) in terms of semantically meaningful features such as "being blue in color", "having long legs" etc., and thus effectively help to bridge the gap between low-level visual features and high-level concepts like the object class. Learning classifiers based on such attributes has the potential advantage of being able to model a large number of categories using a compact set of attributes. Further, a well-defined attribute set may also be applied to unseen categories.

In practice, obtaining sufficient amount of labeled data for attribute learning is challenging, especially since many intuitively useful attributes are often subjective in nature. For example, for an attribute concerning the size of the bear in Fig. 1(a) (even if it is for a binary property of being big or small), there may not exist a single "correct" ground-truth label. Several recent efforts have attempted to address this issue. In [3], the concept of relative visual attributes was introduced to allow learning with only relative labels, which are presumably easier to obtain. Similar ideas have been applied to other applications such as distance metric learning [4], face verification [5], and human-machine interaction [6].

(a) (b)

Fig. 1. Describing the objects in the images with attributes: they are both animals; they both have four legs; (a) is larger than (b); (a) is more dangerous than (b); b is more likely to be found around human; etc.

For properly modeling objects in real-world problems, we typically need more than one attribute. For example, for images of animals illustrated in Fig. 1, we may utilize attributes concerning questions like, "is it an animal?", "is it wild?", "is it dangerous?", etc. There may be intrinsic relatedness among the attributes used to describe the same object if the attributes are indeed properties of the underlying object. For example, "being dangerous" is usually (negatively) correlated with "found around people". Learning these attributes independently, as is done in most existing work, cannot capture such intrinsic relatedness. We hypothesize that considering correlations among the attributes may contribute to improving the individual attribute learners.

In this paper, we explore approaches to learning multiple attributes jointly. We propose a novel formulation termed *Max-Margin Multi-attribute Learning with Low-rank Constraint* and develop an algorithm for obtaining a solution under this model. The proposed approach learns a set of attributes simultaneously under the multi-task learning framework, where learning each attribute is viewed as one task. By learning all the attributes simultaneously with low-rank constraint, the proposed approach is able to capture the intrinsic relatedness of the attributes. It also makes the proposed methods more robust when there are outliers or no sufficient data for certain attributes. In addition, instead of requiring absolute labeling of the training data, the proposed

Manuscript received August 27, 2013; revised February 10, 2014; accepted April 24, 2014. Date of publication May 7, 2014; date of current version May 27, 2014. The work was supported by the National Science Foundation under Grant 0904778. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gang Hua.

The authors are with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: qzhang53@asu.edu; lchen109@asu.edu; baoxin.li@asu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2322446

method utilizes the relative rankings between pairs of the inputs, which are more flexible and effective for describing the data. For evaluation, we first design synthetic experiments to systematically evaluate the model and the algorithm, and then perform experiments with two image datasets and one video dataset. The video dataset was from surgical training, which is a challenging example of intricate relatedness among attributes describing the object of interest. Improved performance of the proposed method over alternative solutions suggests that it is a promising solution to multi-attribute learning.

The key contribution of the work lies in the novel formulation of learning a set of attributes simultaneously only based on the relative ranking of the data and the proposed algorithm for obtaining solutions under the formulation. The rest of this paper is organized as follows: we first briefly review some related work in Section II; then Section III presents the proposed methods; the experimental results are described in Section IV; and the paper concludes with discussion in Section V. In this paper, we will use upper case bold font (e.g., $\mathbf{X}$) for matrices and lower case bold font (e.g., $\mathbf{x}$) for vectors.

II. Related Work

In this section, we briefly review some related work on multi-task learning and (relative) attribute learning. There is a huge amount of relevant work in the literature and our review focuses only on those we deem as closely related to the proposed method.

Multi-task learning, in which a set of tasks are learned simultaneously, has been applied in many applications including Web page categorization [7], Web image and video search [8], face verification [9] and disease prediction [10]. A typical multi-task learning method can be formulated as follows:

$$\min_{\{f_t\}} \frac{1}{T}\sum_{t}^{T}\frac{1}{n_t}\sum_{i}^{n_t} l(f_t(\mathbf{X}_{it}), \mathbf{L}_{it}) + \lambda\Omega(\{f_t\}) \tag{1}$$

where f_t is a classifying/regression function, $l(\cdot)$ a loss function (e.g., squared error), $\mathbf{L}_{it}$ the ground truth response for data $\mathbf{X}_{it}$ (e.g., data labels), and $\Omega(\cdot)$ a penalty term for encouraging common structures of f_t's.

Given different loss functions and penalty terms, many models have been proposed for multi-task learning. In [11] it was assumed that the classifying functions are close to each other, and thus hinge loss was used for the loss function, and the deviations of the classification functions from their means were used as the penalty. Although being intuitive, this assumption is too restrictive and may not be valid for real-world problems. In [12] the l_1/l_q mixture norm was used as the penalty term, where for $q \leq 2$, the sets of regression functions exhibit group sparsity, i.e., the regression functions will select a common sets of features. In real applications, different tasks would have their task-specific components besides the shared components, and thus [13] proposed to decompose the regression functions into two components, where the task-specific component is assumed to be sparse and regularized by a l_1 norm, and the shared common component is regularized by the l_1/l_q mixture norm. Similar idea was proposed in [14], where the l_1 norm was replaced by l_q/l_1 mixture norm to capture the irrelevant tasks. The relatedness among the tasks could also be captured by a low-rank structure. For example, [15] assumed the regression functions were linearly dependent and the trace norm was used as the penalty term. The trace norm has been also used in [16]–[18]. However, these methods require ground truth labeling (e.g., binary labels, real-valued scores) for the training data, which may be difficulty to obtain in many real-world applications.

Attribute learning has seen increasing application in visual processing in recent years, which is especially useful for large-scale dataset, where learning classifiers for data of each category is not practical [19]. In addition, often the attributes can be transferred to unseen categories, different datasets or even different applications (e.g., zero-shot learning [20]). However, most existing work utilizes binary labels or categorical labels, which is not only too restrictive but also unnatural. As a result, relative learning has been proposed. For example, in [4] the relative ranking of data points was used for learning a distance metric function, and the relative rank of some facial attributes was used for face verification in [5]. In [3], the ranking functions were learned from relative ranking of images and then used to describe the images; and in [6] a subject provides relative ranking as feedback to improve the performances of classifiers. There are also efforts on automatically extracting attributes [21]–[24]. However, most existing work learns each attribute independently, ignoring their intrinsic relatedness, which may be extremely helpful especially if the labeling is sparse.

The proposed method attempts to alleviate the requirement of knowing exact labels (through using only relative rankings in learning) while explicitly modeling intrinsic relatedness of the attributes in the learning task (through a multi-task learning framework with a low-rank constraint). As a result, the method achieves a few desired benefits that are not available in existing methods. Such benefits are demonstrated in experiments with both synthetic data and real images/videos, with comparison to typical existing solutions.

III. Proposed Method

The proposed method is capable of learning a set of attributes from only relative rankings. Given the ranking information $\mathbb{E}_t$ and $\mathbb{F}_t$, where $\mathbb{E}_t$ is the set of pairs (i, j) that Data i is better than Data j for Attribute t, and $\mathbb{F}_t$ for the set of pairs being similar for Attribute t, we want to learn a classifier $\mathbf{W}_t$, such that,

$$\mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \geq 1 \ \forall (i, j) \in \mathbb{E}_t$$
$$|\mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt})| \approx 0 \ \forall (i, j) \in \mathbb{F}_t$$

where $\mathbf{X}_{it}$ is the representation of Data i for Attribute t.

In many scenarios, e.g., image classification, we may need to learn multiple attributes and those attributes are likely to be correlated, as illustrated in the examples in Fig. 1. Conventional attribute learning approaches learn these attributes independently, and thus their intrinsic relatedness is

not utilized. We propose to learn the attributes simultaneously under the multi-task learning framework. One popular multi-task learning model assumes that the classifiers of different tasks are similar and their differences to their mean are small. By combining this idea with relative learning, we obtain a baseline approach termed Multi-Task Relative Learning (MTRL), which is formally defined as

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & \sum_t^T \frac{1}{2}|\mathbf{W}_t|_2^2 + \frac{\lambda}{2}|\mathbf{W}_t - \frac{1}{T}\sum_\tau \mathbf{W}_\tau|_2^2 \\ & +\rho_1 \sum_{(i,j)\in\mathbb{E}_t} \epsilon_{ij}^t + \rho_2 \sum_{(i,j)\in\mathbb{F}_t} \gamma_{ij}^t \\ \text{s.t.}\quad & \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 \\ & -\gamma_{ij}^t \leq \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \leq \gamma_{ij}^t \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \end{aligned} \quad (2)$$

where $\mathbf{X}_{it}$ is the representation of i_{th} data for Task t, $\mathbf{W}_t$ is the t_{th} column of $\mathbf{W}$ (i.e., classifier of Task t) and $|\mathbf{W}_t|_2^2$ is related to the margin of the classifier for Tasks t. This problem can be solved by quadratic programming in its dual form, and the details are included in Appendix A.

In the above baseline approach, the usage of a common component has limited the form of correlation that the formulation could model (e.g., when the two tasks are negatively correlated). To this end, we model the correlation among the tasks by linear dependence, which is more flexible than MTRL. If we put the classifiers into the columns of a matrix, the resultant matrix would be low-rank, i.e., its nuclear norm would be small. Thus, we can formulate this new solution as

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & \lambda|\mathbf{W}|_* + \sum_t^T \frac{1}{2}|\mathbf{W}_t|_2^2 + \rho_1 \sum_{(i,j)\in\mathbb{E}_t} \epsilon_{ij}^t + \rho_2 \sum_{(i,j)\in\mathbb{F}_t} \gamma_{ij}^t \\ \text{s.t.}\quad & \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 \\ & -\gamma_{ij}^t \leq \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \leq \gamma_{ij}^t \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \end{aligned} \quad (3)$$

where $|\cdot|_*$ is the nuclear norm or the sum of singular values of the matrix for casting the low-rank constraint. We refer the proposed solution in Eqn. 3 as Max-Margin Multi-Attribute Learning with Low-Rank Constraint.

This problem is equivalent to the following problem by introducing a slack variable $\mathbf{Z}$, which separates the low-rank constraint from the others:

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & \lambda|\mathbf{Z}|_* + \sum_t^T \frac{1}{2}|\mathbf{W}_t|_2^2 + \rho_1 \sum_{(i,j)\in\mathbb{E}_t} \epsilon_{ij}^t + \rho_2 \sum_{(i,j)\in\mathbb{F}_t} \gamma_{ij}^t \\ \text{s.t.}\ & \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 - \gamma_{ij}^t \leq \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \leq \gamma_{ij}^t \\ & \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \quad \mathbf{W} = \mathbf{Z} \end{aligned} \quad (4)$$

By applying the Augmented Lagrange Multiplier (ALM) method to the equality constraint $\mathbf{W} = \mathbf{Z}$, we have:

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & L(\mathbf{Z},\mathbf{W},\mathbf{b},\gamma,\epsilon,\mu,\mathbf{Y}) \\ \text{s.t.}\quad & \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 \\ & -\gamma_{ij}^t \leq \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \leq \gamma_{ij}^t\ \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \end{aligned} \quad (5)$$

Algorithm 1 Max-Margin Multi-Attribute Learning With Low-Rank Constraint

Input: $\mathbf{X}$,$\mathbf{L}$,λ,μ,ρ_1,ρ_2,σ
Output: $\mathbf{W}$,$\mathbf{b}$,ϵ,γ,$\mathbf{Z}$
Initialize $\mathbf{W}$ by solving T tasks independently and $\mathbf{Y} = \frac{\mathbf{W}}{|\mathbf{W}|_2}$;
while NOT converged **do**
  Solve the low-rank problem (Eqn. 7);
  Solve the ranking problem (Eqn. 8);
  Update $\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{W} - \mathbf{Z})$ and $\mu = \mu \times \sigma$;
  Check convergence;
end while

with

$$\begin{aligned} L(\mathbf{Z},\mathbf{W},\mathbf{b},\epsilon,\gamma,\mu,\mathbf{Y}) &= \lambda|\mathbf{Z}|_* + \langle \mathbf{Y}, \mathbf{W} - \mathbf{Z}\rangle \\ &+\frac{\mu}{2}|\mathbf{W} - \mathbf{Z}|_F^2 + \frac{1}{2}\sum_t |\mathbf{W}_t|_2^2 + \rho_1 \sum \epsilon_{ij}^t + \rho_2 \sum \gamma_{ij}^t \end{aligned} \quad (6)$$

where $\mathbf{Y}$ is the Lagrange multiplier, $\langle\cdot,\cdot\rangle$ is the inner product and μ is related to the Lipschitz constant of the primal problem $f(\mathbf{Z},\mathbf{W},\mathbf{b},\epsilon) = \lambda|\mathbf{Z}|_* + \frac{1}{2}\sum_t |\mathbf{W}_t|_2^2 + \rho_1 \sum \epsilon_{ij}^t + \rho_2 \sum \gamma_{ij}^t$.

The problem in Eqn. 5 can be solved via block coordinate descent, by considering the following two sub-problems:

Low-rank problem: fix $\mathbf{W}$, $\mathbf{b}$, ϵ, γ, μ and $\mathbf{Y}$ to solve $\mathbf{Z}$, i.e.,

$$\min_{\mathbf{Z}} \lambda|\mathbf{Z}|_* + \langle \mathbf{Y}, \mathbf{W} - \mathbf{Z}\rangle + \frac{\mu}{2}|\mathbf{W} - \mathbf{Z}|_F^2 \quad (7)$$

Ranking problem: fix $\mathbf{Z}$, μ and $\mathbf{Y}$ to solve $\mathbf{W}$, $\mathbf{b}$, ϵ and γ, i.e.,

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & \frac{\mu}{2}|\mathbf{W} - \mathbf{Z}|_F^2 + \langle \mathbf{Y}, \mathbf{W} - \mathbf{Z}\rangle + \sum_t^T \frac{1}{2}|\mathbf{W}_t|_2^2 \\ & +\rho_1 \sum_{(i,j)\in\mathbb{E}_t} \epsilon_{ij}^t + \rho_2 \sum_{(i,j)\in\mathbb{F}_t} \gamma_{ij}^t \\ \text{s.t.}\quad & \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 \\ & -\gamma_{ij}^t \leq \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \leq \gamma_{ij}^t\ \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \end{aligned} \quad (8)$$

In summary, the overall algorithm for solving the problem of Eqn. 3 can be described in Algorithm 1.

In the following subsections, we will present specific methods for solving the two sub-problems of Eqn. 7 and 8., and then analyze the overall algorithm. The convergence analysis of the algorithm is included in Appendix B, where we show the proposed problem is convex and the proposed algorithm will converge to its global optimum.

A. Solving the Low-Rank Problem

For the low-rank problem, we want to find the optimal $\mathbf{Z}$ for Eqn. 7, which is a convex problem. It has been shown in [25] that the optimal solution to the problem $\min_{\mathbf{X}} \lambda|\mathbf{X}|_* + \frac{1}{2}|\mathbf{X} - \mathbf{W}|_F^2$ can be computed via a singular value thresholding algorithm, i.e., $\mathbf{U}\mathcal{S}_\lambda(\Sigma)\mathbf{V}^T$, where $\mathbf{U}\Sigma\mathbf{V}^T \leftarrow \text{svd}(\mathbf{W})$ is the singular value decomposition and $\mathcal{S}_\cdot(\cdot)$ is the thresholding operator:

$$\mathcal{S}_a(b) = \begin{cases} b - a & b \geq a \\ 0 & a \geq b \geq -a \\ b + a & \text{otherwise} \end{cases} \quad (9)$$

Thus the optimal solution to Eqn. 7 is $\mathbf{Z}^* = \mathbf{U}\mathcal{S}_{\frac{\lambda}{\mu}}(\Sigma)\mathbf{V}^T$, where $\mathbf{U}\Sigma\mathbf{V}^T \leftarrow \text{svd}(\mathbf{W} + \frac{1}{\mu}\mathbf{Y})$.

B. Solving the Ranking Problem

By recognizing $|\mathbf{W} - \mathbf{Z}|_F^2 = \sum_t |\mathbf{W}_t - \mathbf{Z}_t|_2^2$ and $\langle \mathbf{Y}, \mathbf{W} - \mathbf{Z}\rangle = \sum_t \langle \mathbf{Y}_t, \mathbf{W}_t - \mathbf{Z}_t\rangle$, the problem in Eqn. 8 can be decomposed into T independent smaller problems, where each smaller problem is associated with only one attribute/task:

$$\begin{aligned} \min_{\mathbf{W},\epsilon,\gamma} \ & \frac{\mu}{2}|\mathbf{W}_t - \mathbf{Z}_t|_F^2 + \langle \mathbf{Y}_t, \mathbf{W}_t - \mathbf{Z}_t\rangle + \frac{1}{2}|\mathbf{W}_t|_2^2 \\ & + \rho_1 \sum_k \epsilon_{kt} + \rho_2 \sum_l \gamma_{lt} \\ \text{s.t.} \quad & \mathbf{W}_t^T \mathbf{E}_{kt} + \epsilon_{kt} \geq 1 \\ & -\gamma_{lt} \leq \mathbf{W}_t^T \mathbf{F}_{lt} \leq \gamma_{lt} \ \epsilon_{kt} \geq 0; \gamma_{lt} \geq 0 \end{aligned} \tag{10}$$

where we use $\mathbf{E}_{kt} = \mathbf{X}_{it} - \mathbf{X}_{jt} \forall (i, j) \in \mathbb{E}_t$, $\mathbf{F}_{lt} = \mathbf{X}_{it} - \mathbf{X}_{jt} \forall (i, j) \in \mathbb{F}_t$, k, l to re-index $(i, j) \in \mathbb{E}_t$ and $(i, j) \in \mathbb{F}_t$. By applying the Lagrange multipliers, for Eqn. 10 we can have:

$$\begin{aligned} \max_{\alpha,\beta,\delta,\eta,\zeta} \min_{\mathbf{w},\epsilon,\gamma} & \frac{\mu}{2}|\mathbf{W}_t - \mathbf{Z}_t|_F^2 \langle \mathbf{Y}_t, \mathbf{W}_t - \mathbf{Z}_t\rangle \\ & + \frac{1}{2}|\mathbf{W}_t|_2^2 + \sum_k \rho_1 \epsilon_k + \alpha_k(1 - \epsilon_k - \mathbf{W}_t^T \mathbf{Y}_k) - \eta_k \alpha_k \\ & + \sum_l \rho_2 \gamma_l + \beta_l(\mathbf{W}_t^T \mathbf{Z}_l - \gamma_l) + \delta_l(-\mathbf{W}_t^T \mathbf{Z}_l - \gamma_l) - \zeta_l \gamma_l \\ \text{s.t. } & \alpha, \beta, \delta, \eta, \zeta \geq 0 \end{aligned} \tag{11}$$

By checking the gradients, we have:

$$\mathbf{W}_t = \frac{\mu \mathbf{Z}_t - \mathbf{Y}_t + \sum_k \alpha_{kt} \mathbf{E}_{kt} + \sum_l (\delta_{lt} - \beta_{lt}) \mathbf{F}_{lt}}{1 + \mu} \tag{12}$$

$$0 \leq \alpha_{kt} \leq \rho_1 \tag{13}$$

$$0 \leq \beta_{lt} + \delta_{lt} \leq \rho_2 \tag{14}$$

Accordingly, we have the dual form for the problem in Eqn. 11, which is a quadratic programming problem:

$$\begin{aligned} \min_{\mathbf{u}_t} & \frac{1}{2}\mathbf{u}_t^T \mathbf{K}_t \mathbf{u}_t + \mathbf{f}_t^T \mathbf{u}_t \\ \text{s.t.} & \mathbf{lb}_t \leq \mathbf{u}_t \leq \mathbf{ub}_t \\ & \mathbf{A}_t^T \mathbf{u}_t = 0 \end{aligned} \tag{15}$$

with

$$\begin{aligned} \mathbf{u}_t &= [\alpha^T, \ -\beta^T, \ \delta^T]^T \\ \mathbf{K}_t &= \begin{bmatrix} \mathbf{E}_t^T \mathbf{E}_t & \mathbf{E}_t^T \mathbf{F}_t & \mathbf{E}_t^T \mathbf{F}_t \\ \mathbf{F}_t^T \mathbf{E}_t & \mathbf{F}_t^T \mathbf{F}_t & \mathbf{F}_t^T \mathbf{F}_t \\ \mathbf{F}_t^T \mathbf{E}_t & \mathbf{F}_t^T \mathbf{F}_t & \mathbf{F}_t^T \mathbf{F}_t \end{bmatrix} \\ \mathbf{f}_t &= [\mathbf{E}_t^T(\mathbf{Y}_t - \mu \mathbf{Z}_t) - 1, \ \mathbf{F}_t^T(\mathbf{Y}_t - \mu \mathbf{Z}_t), \ \mathbf{F}_t^T(\mathbf{Y}_t - \mu \mathbf{Z}_t)]^T \\ \mathbf{lb}_t &= [0\mathbf{e}_{|\mathbb{E}_t|}^T, \ -\rho_2 \mathbf{e}_{|\mathbb{F}_t|}^T, \ 0\mathbf{e}_{|\mathbb{F}_t|}^T]^T \\ \mathbf{ub}_t &= [\rho_1 \mathbf{e}_{|\mathbb{E}_t|}^T, \ 0\mathbf{e}_{|\mathbb{F}_t|}^T, \ \rho_2 \mathbf{e}_{|\mathbb{F}_t|}^T]^T \\ \mathbf{A}_t &= [0_{|\mathbb{F}_t| \times |\mathbb{E}_t|}, \ -\mathbf{I}_{|\mathbb{F}_t| \times |\mathbb{F}_t|}, \ \mathbf{I}_{|\mathbb{F}_t| \times |\mathbb{F}_t|}] \\ \mathbf{b}_t &= \rho_2 \mathbf{e}_{|\mathbb{F}_t|} \end{aligned}$$

where $\mathbf{e}_n \in \mathbb{R}^{n \times 1}$ is a all-1 vector, $0_{m \times n} \in \mathbb{R}^{m \times n}$ is all-0 matrix, $\mathbf{I}_{n \times n} \in \mathbb{R}^{n \times n}$ is the identity matrix. Thus the dimension of the dual form of the ranking problem is $|\mathbb{E}_t| + 2|\mathbb{F}_t|$. After we solve the problem in Eqn. 15, we can compute the classifier according to Eqn. 12.

C. Analysis of the Algorithm

The proposed algorithm involves two major sub-problems. For the low-rank problem, the most time consuming step is the singular value decomposition for a matrix of dimension $D \times T$ (D is the input dimension), where the typical complexity for an exact decomposition is $O(\min(TD^2, T^2D))$. However, we may not be interested in a full/exact decomposition, but only the singular vectors whose singular value are sufficiently large (e.g., PROPACK [26]). For the classification problem, we are solving T quadratic programming problems of dimension n_t, with n_t the number of data points for the $t - th$ task.

The proposed problem in Eqn. 3 is convex and the proposed algorithm will converge to its global optimum. The proof is given in Appendix B. For the stopping criterion, we compute $\frac{|\mathbf{W}-\mathbf{Z}|_F^2}{|\mathbf{W}|_F^2}$. If this value is sufficiently small (e.g., 10^{-6}), we will terminate the optimization. In our experiments, we observed the convergence was reached within 100 iterations.

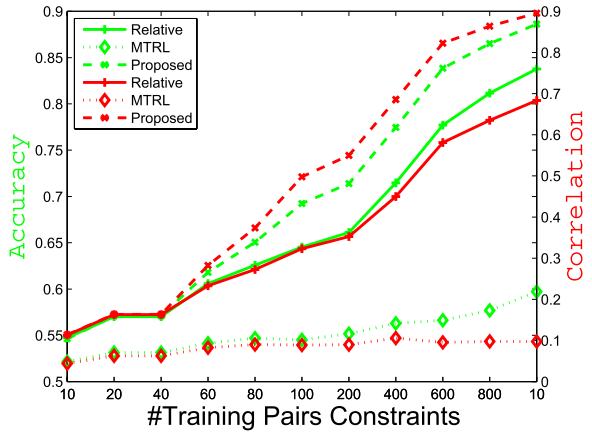
There are three parameters required for the proposed algorithm: λ (controlling the weight of the nuclear norm term), μ (controlling the weight of the term $|\mathbf{W} - \mathbf{Z}|_F^2$) and σ (controlling the increasing speed of μ). The selection of λ depends on the correlation among the tasks: if high correlation among the tasks is expected, we should use a large λ (i.e., $|\mathbf{W}|_*$ should be small); otherwise, we should set λ to a small value. When $\lambda = 0$, the proposed method is equivalent to the relative attribute learning method, where each task is solved independently. For μ, we utilizes the analysis in [27] and set it to $\frac{1.25\lambda}{|\mathbf{W}|_2}$. For ρ, we use $\rho = 1.2$.

IV. Experimental Results

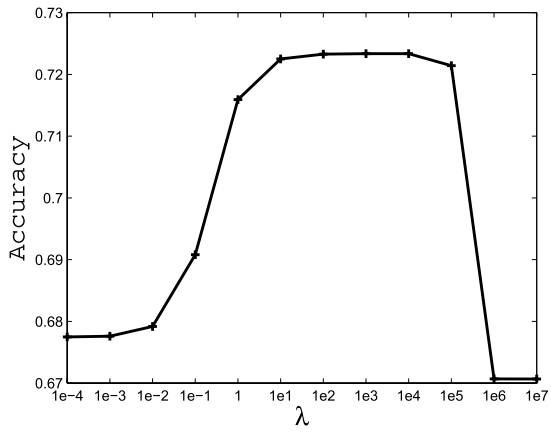
We evaluated the proposed approach on both synthetic data (Section IV-A) and real image/video data sets (Section IV-B, IV-C). The proposed method is compared with the relative attribute method of [3], where each attribute is learned independently, and with the multi-task relative learning method, where the classifying/ranking functions of the attributes are assumed to share a common component. Since no validation set is available for the real datasets (and they are too small to support creation of a validation set), we did not rely on cross-validation for parameter tuning. Instead, in the experiments we used the following fixed parameters for the proposed method and the multi-task relative attribute learning method: $\lambda = 10000$, $\rho_1 = 100$ and $\rho_2 = 100$. Default parameters were used for relative attribute learning.

A. Simulated Experiment

In this section, we evaluate the proposed method on synthetic data. We generate $T = 10$ tasks and the feature dimension of each tasks is $D = 1000$. The ground truth classification function (or ground truth ranking function) for Task t is $\mathbf{W}_t$, where $\mathbf{W}_t$ is the $t - th$ column of $\mathbf{W}$. $\mathbf{W}$ is



(a)



(b)

Fig. 2. The result for simulation experiments with varying P (a) ($\lambda = 10^4$) and λ (b) ($P = 100$), where the dashed curves correspond to the result of the proposed method, dot curve for the MTRL and solid curve for the relative attributes method. We compute accuracy (y axis of red curves) of the learned ranking functions and the correlation (y axis of green curves) between the learned ranking functions and ground truth ones. The X axis are the p (a) and λ (b) accordingly.

generated as:

$$\begin{aligned} \mathbf{W}_0 &= \text{rand}(D, T) - 0.5 \\ \text{svd}(\mathbf{W}_0) &\rightarrow \mathbf{U}\Sigma\mathbf{V}^T \\ \mathbf{W} &= \mathbf{U}(:, 1:r)\Sigma(1:r, 1:r)\mathbf{V}^T(:, 1:r) \end{aligned}$$

where $r = 2$ is the desired rank of $\mathbf{W}$. Note that by generating the ground truth classification function in this way, the classifiers are not necessarily similar or to share a common component. We uniformly draw the data $\mathbf{X}$ for each task, and each set of data contains 1000 data points. For Task t, we randomly select P pairs as the training pairs, i.e., $(i, j) \in \mathbb{E}$ if $\mathbf{W}_t^T\mathbf{X}_i - \mathbf{W}_t^T\mathbf{X}_j \geq \tau$; or $(j, i) \in \mathbb{E}$ if $\mathbf{W}_t^T\mathbf{X}_i - \mathbf{W}_t^T\mathbf{X}_j \leq -\tau$; otherwise $(i, j) \in \mathbb{F}$, where τ is the predefined margin. The proposed algorithm is applied to the training pairs to learn the ranking function for the tasks, with comparison with the relative attribute (refer as "Relative") method and also the baseline (i.e., Multi-Task Relative Learning or MTRL) method. We also test different combinations of λ (from 10^{-4}, i.e., low requirement of the low-rank constraint, to 10^7, i.e., high requirement on the low-rank constraint) and P (from 10 to 1000), where the results are shown in Fig. 2.

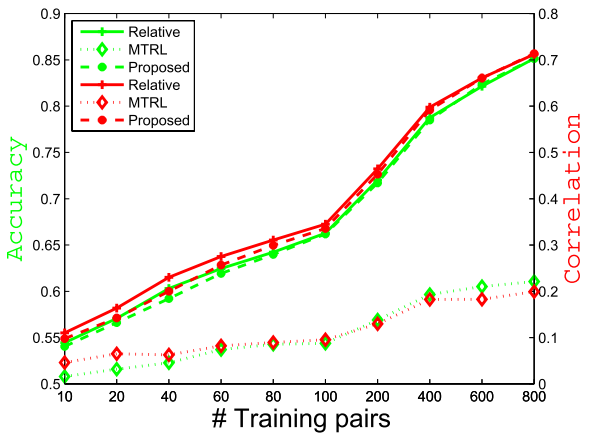


Fig. 3. The result for simulation experiment when the ground truth ranking functions of the tasks are linear independent. The matrix consisted of ground truth ranking functions as its columns has maximal singular value 0.7 and minimal singular value 0.6.

From Fig. 2(a), we can observe that, although the accuracy and the correlation increase with more training pairs, i.e., larger P, the proposed method consistently performs better than the other two competitors. Especially, when $P = 1000$, the correlation between the ranking functions learned by the proposed method and the ground truth ones is about 0.9, which is significantly better than 0.68 achieved by the relative attribute method. The results indicate that the proposed method is more likely to recover the ground truth ranking functions than the relative attribute method, when given the same number of training pairs. The performance of MTRL is significantly lower. This could be explained by the assumption made by its formulation: the classification functions of the tasks should be similar (or share a common component), which is not always true in the generation of the data (e.g., the classification functions can be negatively correlated).

Fig. 2(b) illustrates the performance of the proposed approach with different settings for the parameter λ, which controls the contribution of the low-rank constraint. From the plot, we can observe that the performance is stable for a wide range of $\lambda (\lambda \in [10, 10^4])$ and the best result is obtained when $\lambda = 10^4$.

We also performed simulations using data whose ground truth ranking functions are not correlated, i.e., the functions are linear independent by setting $r = 10$. The results are shown in Fig. 3, from which we can find that, the proposed method (dashed curve) obtained similar results as the relative attribute learning method (solid curve) in both accuracy and correlation. However, the MTRL (dotted curve) obtained obviously worse performances in both accuracy and correlation. This demonstrates that the proposed method is robust to different correlation levels of the tasks, and its performance is still comparable to that of the relative attribute learning method even when the tasks are totally linear independent. The performance of MTRL method, however, degrades dramatically when the assumption about the relatedness of the tasks does not hold.

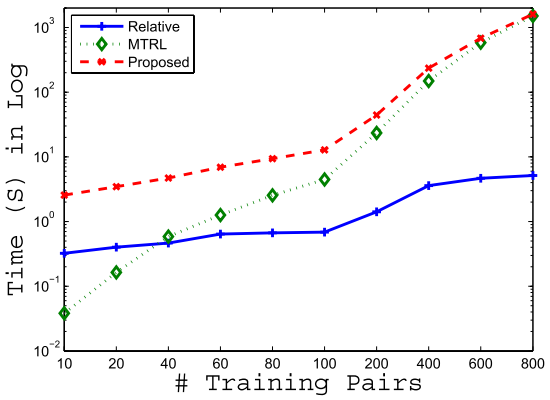


Fig. 4. The computation time of the proposed approach given different number of training pairs, with the comparison to the relative attribute learning methods and MTRL method. For the time axis (y-axis), we use logarithm.

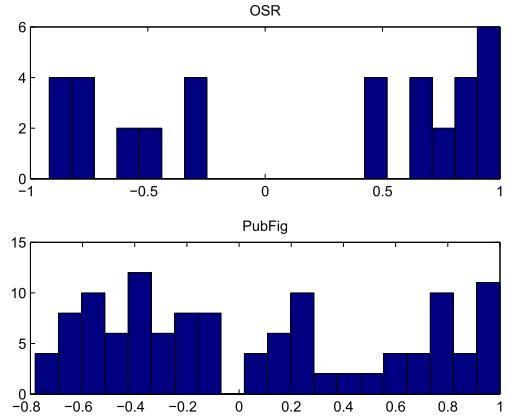


Fig. 5. The histogram of Pearson's correlation coefficients among the tasks for both two datasets. From these histograms, we can observe that the attributes are correlated, as there are non-trivial mass covering the regions towards -1 or 1. Note that, Pearson's correlation coefficient measures only linear dependency, and thus even if it is low, the tasks could still be highly dependent.

For understanding the computational efficiency of the proposed method, we note that its formulation as well as solutions bear similarity to MTRL, which is well-understood to have a polynomial complexity over the number of constraints. Hence the proposed method is expected to have the same order of complexity over the number of training pairs. To empirically verify this, we use Fig. 4 to depict the running times of the proposed approach under different numbers of training pairs, with the comparison to the relative attribute learning method and the MTRL method. It can be observed that the proposed method, while being more expensive than the basic relative attribute learning method, is indeed in par with the MTRL method in terms of asymptotic time complexity. Note that both axes of Fig. 4 are with logarithm for better illustration.

B. Learning Attributes for Images

To evaluate the performances of the proposed algorithm on real data, we utilize two datasets, (1) Outdoor Scene Recognition (OSR) Dataset [28] containing 2688 images from 8 categories; (2) A subset of the Public Figure Face Database (PubFig) [5] containing 800 images from 8 random identities (100 images each). We directly used the processed data[1] from [3] and the same experiment settings. To demonstrate that the attributes in these datasets indeed exhibit correlation, we first computed the histogram of the pairwise correlation coefficients among the tasks for each of the datasets, and the results are shown in Fig. 5. It is evident from these plots that the tasks are correlated. For example, one can observe that there is a non-trivial mass covering beyond the interval $[-0.5, 0.5]$ in either of the plots. Note that, both the rank of classifier matrix ($\mathbf{W}$) and the correlation coefficients between the tasks are some measurements of the dependency. For ideal case (perfectly dependent), the rank should be 1 and the correlation coefficient should be $+/-1$ cross different tasks.

Next we report the ranking accuracy of the proposed method and compare with the relative attribute method ("Relative" in short) and the multi-task relative attribute learning method ($MTRL$ in short). All the results on the two datasets are summarized in Table I. From Table I we can observe that the proposed method outperforms the other methods in both cases except that the Relative* row of (A). [3] has an insignificant gain over our method, even with much more training pairs (see also the caption of the Table). Additionally, we can observe that the performance gain of our method over Relative or $MTRL$ (when all trained under the same protocol with only 5% of the training pairs used in [3]) varies. This could be explained by possible varying degree of correlation among the tasks in the two datasets, as alluded by Fig. 5. However, we note that the correlation coefficient used in Fig. 5 measures only linear dependency and thus it is not proper to draw any quantitative conclusion. Additionally, the low-rank constraint would generally work better when there are many tasks considered jointly (comparing with the feature dimension) [15]. This is consistent with the results in the Table (e.g., better gain by the proposed in (B) than in (A)). The low-rank constraint used in the proposed method is more flexible than forcing the tasks to share common components in capturing the intrinsic relatedness of the attributes, which explains the gain of the proposed over $MTRL$.

C. Evaluating Surgical Skills From Videos

In this experiment, the data were videos collected from the Fundamentals of Laparoscopic Surgery (FLS) trainer box (www.flsprogram.org), which is a simulation-based training platform and has been widely used in many hospitals/training centers for minimally-invasive surgery training. The system has an on-board camera capturing a trainee's operation inside the box and the video is shown on a monitor. There are a set of standard operations defined for the FLS training system. Our experiment was based on data captured from the "Peg Transfer" operation, as illustrated in Fig. 6. In this operation, a trainee is required to lift one of the six objects with a grasper by his non-dominant hand, transfer the object midair to his dominant hand, and then place the object on a peg on the other side of the board. Once all six objects have been

[1] downloaded at http://filebox.ece.vt.edu/~parikh/relative.html#data

TABLE I

RANKING ACCURACY FOR OSR (A) AND PUBFIG (B) FROM DIFFERENT METHODS. IN [3], OVER 20,000 TRAINING PAIRS WERE USED AND THE RESULTS ARE REPORTED HERE AS "RELATIVE*". IN OUR EXPERIMENT, WE RANDOMLY PICKED ONLY 5% OF THOSE TRAINING PAIRS FOR EVALUATING THE THREE METHODS, AS SHOWN IN ROW 2 TO ROW 4 OF EACH TABLE. FOR THE PROPOSED METHOD, WE FIXED λ TO 10000

Attribute	1	2	3	4	5	6	Average
Proposed	94.81%	90.67%	86.61%	86.69%	88.63%	88.26%	89.28%
MTRL	90.50%	84.97%	82.04%	80.78%	83.15%	78.83%	83.38%
Relative	93.69%	91.05%	85.75%	86.78%	87.62%	87.82%	88.78%
Relative*	94.63%	91.28%	86.17%	86.90%	87.96%	89.07%	89.33%

A

Attribute	1	2	3	4	5	6	7	8	9	10	11	Average
Proposed	85.64%	81.18%	85.14%	84.13%	81.07%	89.08%	84.24%	83.11%	82.63%	84.46%	86.03%	**84.25**%
MTRL	84.08%	80.30%	84.08%	80.99%	79.45%	85.96%	83.82%	81.83%	82.21%	83.47%	83.53%	82.70%
Relative	81.32%	76.99%	81.86%	80.75%	77.54%	87.36%	79.51%	81.66%	75.57%	78.41%	81.59%	80.23%
Relative*	83.26%	79.91%	83.54%	83.04%	79.46%	89.65%	82.19%	82.93%	78.79%	81.09%	83.16%	82.46%

B

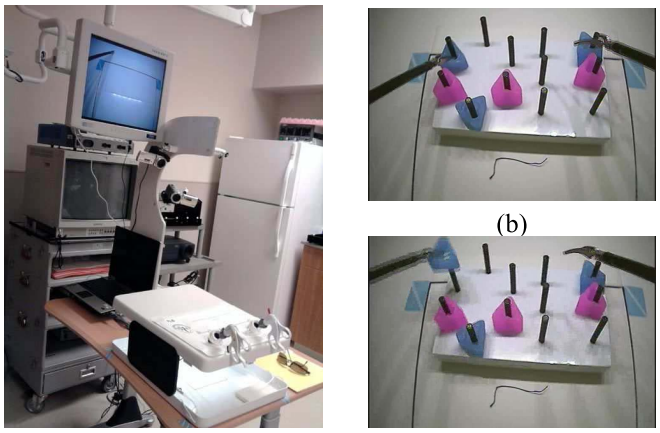


Fig. 6. Illustrating the FLS system: (a) the FLS system (white), (b) and (c) frames captured by onboard camera showing the operation within the FLS trainer box.

TABLE II

PRIMITIVE ACTIONS IN PEG TRANSFER

Name	Description
Lift	Grasp an object and lift it off a peg
Transfer	Object transfer from one hand to another
Place	Release an object and place it on a peg
Loaded Move	Move a grasper with an object
Unloaded Move	Move a grasper without any object

transferred, the process is reversed from one side to the other.[2] The Peg Transfer operation consists of several primitive actions (or therbligs [31]) as building blocks of manipulative surgical activities, which are defined in Table II. Ideally, these primitive actions are all necessary in order to finish one peg-transfer cycle. Since there are six objects to transfer left-to-right and then backwards, there are totally 12 cycles in one training session. Our experiment was based on video recordings from the FLS system on-board camera capturing training sessions of resident surgeons in their different residency years.

[2]For more details of the FLS trainer box and the "peg transfer" operation, we refer the readers to [29] and [30].

1) The Attribute-Learning Task: Given a video from an operation described above, we segment it into multiple clips, with each clip containing only one primitive action, e.g., lift. For providing automatic feedback to a trainee, we need to infer the motion skill from those clips, which is deemed a very difficult tasks, due to the semantic gap between the low-level visual feature and the high-level motion skill. We apply the proposed method to this challenging problem by first defining a set of attributes (Table III), which are designed according to domain knowledge on surgical skill evaluation [32]. These attributes describe varying aspects of motion skill and are easier to infer from the visual features. With these skill-defining attributes learned, we can provide a trainee with a more detailed rating of his/her performance, rather than a slim score. For example, when they find they have a low performance in attributes "instrument handling", they would spend more time in improving their handling of instruments.

2) Feature Extraction: Based on the attributes defined above, we design the following feature extraction scheme. We first utilize random forest (RF, learned from a training set) to segment the pixels of each frame into "tool" and "background", based on the color information. With the segmentation results, we perform morphological operation and blob analysis to extract the tool tips and orientations of the tools controlled by the left and right hands. After that, the motion features V used for skill attribute (Table III) learning are generated in 3 steps. In the first step, a few types of motion information are estimated to represent a trainee's operation, as summarized in Table IV. In the second step, we extract motion signatures from each of the motion features in Table IV. The motion signatures are 1-dimensional temporal signals (Table V) to further compact the motion information. In the last step, final motion features are extracted from each motion vector and its motion signatures as follows: in the time domain, we divide a signature into equal temporal bins; and we also divide the Fourier transform result into equal frequency bins. In each temporal or frequency bin, the maximal, minimal, and average values are kept.

3) Experiment Results: We selected 10 representative videos from trainees of different skill levels, where each video is a full training session consisting of 12 Peg Transfer cycles,

TABLE III

THE ATTRIBUTE USED IN THIS PAPER, WHICH ARE DEFINED ACCORDING TO [32]. NOTE, WE ONLY SELECT THE ATTRIBUTES WHICH ARE RELAVANT TO THE OPERATIONS IN OUR SIMULATED SURGICAL VIDEOS

Attributes	Bad manifestation	Good manifestation
Time and motion (T)	Many unnecessary moves	Clear economy of movement. Maximum efficiency
Flow of Operation (F)	Frequently stopped, seemed unsure of next move	Obviously planned course, effortless flow
Bimanual Dexterity (B)	Uses only one hand, poor coordination between hands	Expertly uses both hands to provide optimal exposure
Respect of tissues (R)	Frequent unnecessary force on tissues or caused damage by inappropriate use of instruments	Consistently handled tissue appropriately with minimal damage to tissues
Instrument handling (I)	Tentative/awkward moves or inappropriate use	Fluid moves with instruments. No awkwardness
Depth Perception (D)	Constantly overshoots, swings wide, slow correction	Accurately directs instruments in correct plane

TABLE IV

MOTION INFORMATION. ROI IS A REGION AROUND GRASPER TIPS TO INCLUDE OBJECT UNDER OPERATION

Name	Description
$V(t)$	The motion of grasper tip
$\hat{V}(t)$	Relative motion between grasper tip and its operation target
$A(t)$	The motion area of objects in ROI.
$M(x,t)$	The optical flow field in ROI

TABLE V

MOTION SIGNATURES, WHERE $Y(t)$ REPRESENTS ANY MOTION INFORMATION IN TABLE IV, E.G., $V(t)$, $\hat{V}(t)$ AND $M(x,t)$

Name	Description
$\bar{Y}(t) = \lvert Y(t) \rvert$	Instant velocity
$L(t)$	Length of trajectory
$J(t) = \lvert Y(t) - \bar{Y}(t) \rvert$	Motion smoothness metric
$R(t)$	Curl angular velocity

which leads to 12 video clips for each therblig. Thus we have in total 120 clips for each therblig. We manually label the relative rankings for 150 pairs of clips, following the guidelines provided by FLS (available on the FLS website). For each pair of clips, we label the attributes described in Table III as either "left is better than right", "right is better than left" or "unsure". Then five-fold random split (one fold for testing and remaining folds for training) is applied to evaluate the proposed method with the comparison to the other two methods. Due to space limitation, we only show the results of two therbligs "lift" and "transfer" in this paper, which are presented in Table VI.

From Table VI, we can find that, the proposed method (Col 3) and MTRL (Col 4) obtained significantly better result than the relative attribute method (Col 5), except for the attribute "Bimanual dexterity" for therblig "Lift" (Table VI(A) Row 4) and the attribute "Depth perception" for therblig "Transfer" (Table VI(B) Row 7). The improvement can be explained by the explicit consideration of intrinsic relatedness of those attributes in the proposed method and MTRL. The proposed method is on average better than MTRL, although MTRL achieves similar average accuracy in Table VI(B). This could be due to the fact that both the MTRL constraint and the proposed low-rank constraint did similarly well in capturing the correlation among the attributes for that particular action. However, as discussed earlier, the flexibility of the low-rank constraint in the proposed method would in general lead to a better performance, which is also evidenced by the overall

TABLE VI

THE EXPERIMENTAL RESULT IN EVALUATING MOTIONS SKILLS OF SURGICAL SIMULATIONS: (A) THERBLIG "LIFT" AND (B) THERBLIG "TRANSFER". COL 2 IS THE NUMBER OF DISSIMILAR PAIRS; COL 3 IS THE NUMBER OF SIMILAR PAIR. NOTE FOR ATTRIBUTE R AND I OF THERBLIGS "TRANSFER", WE DON'T ENOUGH GROUND TRUTH TO COMPUTE THE ACCURACY

Attribute	|**E**|	|**F**|	Proposed	MTRL	Relative
T	90	50	**78.89**%	72.22%	72.22%
F	77	63	**75.32**%	70.13%	67.53%
B	30	110	83.33%	**90.00**%	86.68%
R	62	78	**83.87**%	74.19%	61.29%
I	70	70	81.43%	**82.86**%	75.71%
D	29	111	**75.86**%	72.41%	62.07%
Overall	237	183	**79.61**%	75.70%	70.39%

(a)

Attribute	|**E**|	|**F**|	Proposed	MTRL	Relative
T	59	41	81.36%	**83.05**%	74.58%
F	46	54	78.26%	**82.61**%	73.91%
B	41	59	**65.85**%	58.54%	56.10%
R	1	99	N.A.	N.A.	N.A.
I	8	92	N.A.	N.A.	N.A.
D	41	59	80.49%	82.93%	**85.37**%
Overall	112	187	**77.55**%	77.04%	73.47%

(b)

better performance of the proposed method in Table VI (and in particular in (A)).

V. DISCUSSION AND CONCLUSION

In this paper we proposed a novel approach *Max-Margin Multi-Attribute Learning with Low-Rank Constraint*. Compared with existing methods in the literature, the proposed method learns a set of attributes simultaneously so that the intrinsic relatedness could be captured. In addition, it only require the relative ranking of the attributes instead of binary labels, leading to a more flexible solution to many learning applications in which absolute labels are difficult to obtain. We evaluated the proposed method on both simulated data and real image/video data, and compared its performance with the relative attribute method and the MTRL method, both being representative of typical alternative solutions. The experimental results demonstrated that the proposed method is more effective in capturing the intrinsic correlation among the tasks (or attributes) and delivers higher accuracy than the competing methods.

It is worth mentioning that, the proposed method is based on the assumption that the set of tasks are related. If indeed

the tasks are related, the proposed method is able to, as shown in our experiments, outperform the "relative attribute" method which treats each task independently. However, if the tasks are independent, the performance of the proposed method may degrade, as it would force a correlation model on the independent tasks. However, for real problems with multiple attributes describing the same underlying object of interest, it is reasonable to assume that completely independence among the attributes would be rare, and thus the proposed method is expected to able to deliver good performance in general. Nevertheless, it will be an interesting future direction to explicitly explore possible relationship between the performance of the method and the degree of relatedness among the tasks/attributes. In addition, considering that the current method relies on only the low-rank constraint, another future task is to investigate modeling of more complicated intrinsic relationship among the attributes with outlier handling.

APPENDIX A
ALGORITHM FOR MTRL

According to [11], Eqn. 2 is equivalent to the following problem with appropriate parameters $(\lambda, \rho_1, \rho_2)$:

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & \sum_t^T \frac{1}{2}|\mathbf{V}_t|_2^2 + \frac{\lambda}{2}|\mathbf{W}_0|_2^2 \\ & +\rho_1 \sum_{(i,j)\in\mathbb{E}_t} \epsilon_{ij}^t + \rho_2 \sum_{(i,j)\in\mathbb{F}_t} \gamma_{ij}^t \\ \text{s.t.}\quad & (\mathbf{V}_t + \mathbf{W}_0)^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 \\ & -\gamma_{ij}^t \leq (\mathbf{V}_t + \mathbf{W}_0)^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \\ & \leq \gamma_{ij}^t \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \end{aligned} \tag{16}$$

with $\mathbf{W}_t = \mathbf{V}_t + \mathbf{w}_0$. According to [11], we can also define the following mapping functions:

$$\Phi(\mathbf{X}_{it}) = [\sqrt{\frac{1}{T\lambda}}\mathbf{X}_{it}, 0, \ldots, 0, \mathbf{X}_{it}, 0, \ldots, 0] \tag{17}$$

$$\Phi(\mathbf{W}) = [\sqrt{T\lambda}\mathbf{w}_0, \mathbf{V}_1, \ldots, \mathbf{V}_t, \ldots, \mathbf{V}_T] \tag{18}$$

and get the following formulations:

$$\begin{aligned}\min_{\mathbf{W},\epsilon,\gamma}\ & \frac{1}{2}|\Phi(\mathbf{W})|_2^2 + \sum_t^T \rho_1 \sum_{(i,j)\in\mathbb{E}_t} \epsilon_{ij}^t + \rho_2 \sum_{(i,j)\in\mathbb{F}_t} \gamma_{ij}^t \\ \text{s.t.}\quad & \Phi^T(\mathbf{W})\Phi(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1 \\ & -\gamma_{ij}^t \leq \Phi^T(\mathbf{W})\Phi(\mathbf{X}_{it} - \mathbf{X}_{jt}) \\ & \leq \gamma_{ij}^t \epsilon_{ij}^t \geq 0;\ \gamma_{ij}^t \geq 0 \end{aligned} \tag{19}$$

Obviously $|\Phi(\mathbf{W})|_2^2 = \sum_t^T (|\mathbf{V}_t|_2^2 + \lambda|\mathbf{w}_0|_2^2)$ and $\Phi^T(\mathbf{W})\Phi(\mathbf{X}_{it} - \mathbf{X}_{jt}) = (\mathbf{V}_t + \mathbf{W}_0)^T(\mathbf{X}_{it} - \mathbf{X}_{jt})$.

By writing $(\mathbf{X}_i - \mathbf{X}_j) = \mathbf{Y}_k$ for $(i, j) \in \mathbb{E}$ and $(\mathbf{X}_i - \mathbf{X}_j) = \mathbf{Z}_l$ for $(i, j) \in \mathbb{F}$ and applying Lagrange multipliers we, can get the dual form of Eqn. 2:

$$\begin{aligned}\min_{\alpha,\beta,\lambda}\ & \frac{1}{2}|\sum_k \alpha_k \Phi(\mathbf{Y}_k) + \sum_l (\delta_l - \beta_l)\Phi(\mathbf{Z}_l)|_2^2 - \sum_k \alpha_k \\ \text{s.t.}\quad & 0 \leq \beta_l, \delta_l \leq \rho_2 \\ & 0 \leq \alpha_k \leq \rho_1 0 \leq \beta_l + \delta_l \leq \rho_2 \end{aligned} \tag{20}$$

which can be written as the following quadratic programming problem:

$$\begin{aligned}\min_{\mathbf{u}}\ & \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} + \mathbf{f}^T\mathbf{u} \\ \text{s.t.}\ & \mathbf{lb} \leq z \leq \mathbf{ub}\ \mathbf{A}\mathbf{u} \leq \mathbf{b} \end{aligned} \tag{21}$$

with

$$\begin{aligned}\mathbf{u} &= [\alpha^T,\ -\beta^T,\ \delta^T]^T \in \mathbb{R}^{T(|\mathbb{E}|+2|\mathbb{F}|)\times 1} \\ \mathbf{K} &= \begin{bmatrix} \mathbf{K}_{|\mathbb{E}|\times|\mathbb{E}|} & \mathbf{K}_{|\mathbb{E}|\times|\mathbb{F}|} & \mathbf{K}_{|\mathbb{E}|\times|\mathbb{F}|} \\ \mathbf{K}_{|\mathbb{F}|\times|\mathbb{E}|} & \mathbf{K}_{|\mathbb{F}|\times|\mathbb{F}|} & \mathbf{K}_{|\mathbb{F}|\times|\mathbb{F}|} \\ \mathbf{K}_{|\mathbb{F}|\times|\mathbb{E}|} & \mathbf{K}_{|\mathbb{F}|\times|\mathbb{F}|} & \mathbf{K}_{|\mathbb{E}|\times|\mathbb{F}|} \end{bmatrix} \\ \mathbf{f} &= [-\mathbf{e}_{|\mathbb{E}|}^T,\ 0\mathbf{e}_{|\mathbb{F}|}^T,\ 0\mathbf{e}_{|\mathbb{F}|}^T]^T \\ \mathbf{lb} &= [0\mathbf{e}_{|\mathbb{E}|}^T,\ -\rho_2\mathbf{e}_{|\mathbb{F}|}^T,\ 0\mathbf{e}_{|\mathbb{F}|}^T]^T \\ \mathbf{ub} &= [\rho_1\mathbf{e}_{|\mathbb{E}|}^T,\ 0\mathbf{e}_{|\mathbb{F}|}^T,\ \rho_2\mathbf{e}_{|\mathbb{F}|}^T]^T \\ \mathbf{A} &= [0_{|\mathbb{E}|\times|\mathbb{E}|},\ -\mathbf{I}_{|\mathbb{F}|\times|\mathbb{F}|},\ \mathbf{I}_{|\mathbb{F}|\times|\mathbb{F}|}] \\ \mathbf{b} &= \rho_2\mathbf{e}_{|\mathbb{F}|} \in \mathbb{R}^{T|\mathbb{F}|\times 1} \end{aligned}$$

where $\mathbf{e}_n \in \mathbb{R}^{n\times 1}$ is a all 1 vector, $0_{m\times n} \in \mathbb{R}^{m\times n}$ is all 0 matrix, $\mathbf{I}_{n\times n} \in \mathbb{R}^{n\times n}$ is identity matrix, $\mathbf{K}_{|\mathbb{E}|\times|\mathbb{E}|}(i,t;j,s) = \Phi^T(\mathbf{y}_{it})\Phi(\mathbf{y}_{js})$, $\mathbf{K}_{|\mathbb{E}|\times|\mathbb{F}|}(i,t;j,s) = \Phi^T(\mathbf{y}_{it})\Phi(\mathbf{z}_{js})$ and $\mathbf{K}_{|\mathbb{F}|\times|\mathbb{E}|}(i,t;j,s) = \Phi^T(\mathbf{z}_{it})\Phi(\mathbf{z}_{js})$. The mapping function $\Phi(\cdot)$ is defined in Eqn. 17.

After we solve the quadratic problem in Eqn. 21 with optimal solution $u^* = [\alpha^T, -\beta^T, \delta^T]^T$, we can compute $\Phi(\mathbf{W}) = [\frac{1}{\lambda}\mathbf{W}_0^T, \mathbf{V}_1^T, \ldots, \mathbf{V}_t^T]^T = \sum_t \sum_i \alpha_{it}\Phi(\mathbf{Y}_{it}) + \sum_j (\delta_{jt} - \beta_{jt})\Phi(\mathbf{Z}_{jt})$ and then recover classifier of each attribute as $\mathbf{W}_t = \mathbf{W}_0 + \mathbf{V}_t$.

As the proposed method can be formulated into a quadratic programming problem, the convergence and global optimality of the solution is guaranteed. The dimension of quadratic programming problem is $T(|\mathbb{E}| + 2|\mathbb{F}|) \times 1$ with T as the number of tasks, $|\mathbb{E}|$ and $|\mathbb{F}|$ as the number of constraints cast by relative rankings. The dimension of the problem and the computational cost could be high, when there are a lot of pairs of relative rankings. To solve this issue, we could utilize the idea of active constraints.

APPENDIX B
CONVERGENCE ANALYSIS

We will show that the proposed algorithm (Section III) will converge. In this section, we will use $\mathbf{Y}^k$ to represent the variable $\mathbf{Y}$ computed in k_{th} iteration. First, we can easily identify that, the two sub-problems, "low rank problem" and "classification" problem are convex. We define the space $\Omega = \{\mathbf{Z}, \mathbf{W}, \mathbf{b}, \epsilon, \gamma\,|\mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) + \epsilon_{ij}^t \geq 1\ \&\ -\gamma_{ij}^t \leq \mathbf{W}_t^T(\mathbf{X}_{it} - \mathbf{X}_{jt}) \leq \gamma_{ij}^t \&\ \epsilon_{it} \geq 0 \&\ \gamma_{it} \geq 0\ \forall i,t\}$, which is obvious convex, and the analysis will be within this space.

Lemma 1: $\mathbf{Y}^k$ is bounded.

Proof: Since $\mathbf{Z}^{k+1}$ is optimal for the low-rank problem with $\mathbf{W}^k$, $\mathbf{b}^k$, ϵ^k, γ^k, μ^k and $\mathbf{Y}^k$, we have $0 \in \frac{\partial L(\mathbf{Z},\mathbf{W}^k,\mathbf{b}^k,\epsilon^k,\mu^k,\mathbf{Y}^k)}{\partial \mathbf{Z}}$. That is $0 \in \frac{\partial\|\mathbf{Z}\|_*}{\partial\mathbf{Z}} - \mathbf{Y}^k + \mu^k(\mathbf{Z}^k - \mathbf{W}^k)$, so we have $\mathbf{Y}^{k+1} \in \frac{\partial\|\mathbf{Z}\|_*}{\partial\mathbf{Z}}$, where $\mathbf{Y}^{k+1} = \mathbf{Y}^k - \mu^k(\mathbf{Z}^k - \mathbf{W}^k)$. According to [27] Theorem 4 and Lemma 1, $\mathbf{Y}^{k+1}$ is bounded. This ends the proof of Lemma 1.

Lemma 2: the sequences $\mathbf{Z}^k$, $\mathbf{W}^k$, $\mathbf{b}^k$, ϵ^k, μ^k will converge to the optimal solution.

Proof: we define $f(\mathbf{W}, \mathbf{b}, \epsilon) = \lambda|\mathbf{W}|_* + \frac{1}{2}\sum_t |\mathbf{W}_t|_2^2 + \rho \sum_i \epsilon_{it}$ as the objective function of the primal problem (Eqn. 3). We have:

$$\begin{aligned} &L(\mathbf{Z}^{k+1}, \mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1}, \mu^k, \mathbf{Y}^k) \\ &= \min_{\mathbf{Z},\mathbf{W},\mathbf{b},\epsilon} L(\mathbf{Z}, \mathbf{W}, \mathbf{b}, \epsilon, \gamma, \mu^k, \mathbf{Y}^k) \\ &\leq \min_{\mathbf{Z}=\mathbf{W},\mathbf{b},\epsilon} L(\mathbf{Z}, \mathbf{W}, \mathbf{b}, \epsilon, \gamma, \mu^k, \mathbf{Y}^k) \\ &\leq \min_{\mathbf{Z}=\mathbf{W},\mathbf{b},\epsilon} f(\mathbf{W}, \mathbf{b}, \epsilon, \gamma) = f^* \end{aligned}$$

As $\mathbf{Z}^{k+1} - \mathbf{W}^{k+1} = \frac{1}{\mu^k}(\mathbf{Y}^{k+1} - \mathbf{Y}^k)$ and the boundedness of $\mathbf{Y}^k$, we have $\lim_{t\to\infty} \mathbf{Z}^k - \mathbf{W}^k = 0$. Thus $(\mathbf{W}^*, \mathbf{b}^*, \epsilon^*) = \lim_{t\to\infty}(\mathbf{W}^k, \mathbf{b}^k, \epsilon^k)$ is the feasible solution of the primal problem.

In addition, we have

$$\begin{aligned} &f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1}) \\ &\quad = L(\mathbf{Z}^{k+1}, \mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1}, \mu^k, \mathbf{Y}^k) \\ &\qquad - \frac{1}{2\mu^k}(|\mathbf{Y}^{k+1}|_F^2 - |\mathbf{Y}^k|_F^2) + |\mathbf{W}^{k+1}|_* - |\mathbf{Z}^{k+1}|_* \\ &\quad \leq f^* - \frac{1}{2\mu^k}(|\mathbf{Y}^{k+1}|_F^2 - |\mathbf{Y}^k|_F^2) - |\mathbf{W}^{k+1} - \mathbf{Z}^{k+1}|_* \\ &\quad \leq f^* - \frac{1}{2\mu^k}(|\mathbf{Y}^{k+1}|_F^2 - |\mathbf{Y}^k|_F^2) - \frac{1}{\mu^k}|\mathbf{Y}^{k+1} - \mathbf{Y}^k|_* \\ &\quad = f^* - O(\frac{1}{\mu^k}) \end{aligned} \tag{22}$$

where for the last step, we use the boundedness of $\mathbf{Y}^k$ (Lemma 1). Thus we have $\lim_{t\to\infty}[f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1})] = \lim_{t\to\infty} f^* - O(\frac{1}{\mu^k}) = f^*$.

Besides, by $|\mathbf{Z}|_* \geq |\mathbf{W}|_* - |\mathbf{Z} - \mathbf{W}|_*$, we have

$$\begin{aligned} &f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1}) \\ &\quad = L(\mathbf{W}^{k+1}, \mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \gamma^{k+1}, \epsilon^{k+1}) \\ &\quad \geq L(\cdots) - \lambda|\mathbf{Z}^{k+1} - \mathbf{W}^{k+1}|_* \\ &\quad \geq L(\cdots) - \frac{\lambda}{\mu}|\mathbf{Y}^{k+1} - \mathbf{Y}^k|_* \\ &\quad \geq L(\cdots) - O(\frac{\lambda}{\mu}) \geq f^* - O(\frac{\lambda}{\mu}) \end{aligned} \tag{23}$$

where we short $L(\mathbf{Z}^{k+1}, \mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1})$ by $L(\cdots)$.

Combining Eqn. 22 ($f^* - f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1}) \geq O(\frac{1}{\mu^k})$) and Eqn. 23 ($f^* - f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}, \gamma^{k+1}) \leq O(\frac{\lambda}{\mu^k})$), we have $|f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}) - f^*| \leq \max(\frac{1}{\mu^k}, \frac{\lambda}{\mu^k})$. As $\mu^{k+1} = \mu^k \times \sigma$ and if we choose $\sigma > 1$, we have $\lim_{t\to\infty}|f(\mathbf{W}^{k+1}, \mathbf{b}^{k+1}, \epsilon^{k+1}) - f^*| = 0$. This proves the convergence.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Proc. Adv. Neural Inform. Process. Syst.*, 2007, pp. 433–440.

[2] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1778–1785.

[3] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE ICCV*, Nov. 2011, pp. 503–510.

[4] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Proc. Adv. NIPS*, 2004, p. 41.

[5] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 365–372.

[6] D. Parikh, A. Kovashka, A. Parkash, and K. Grauman, "Relative attributes for enhanced human-machine communication," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012.

[7] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning a shared predictive structure from multiple tasks," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 137–144.

[8] O. Chapelle, P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang, and B. Tseng, "Multi-task learning for boosting with application to web search ranking," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1189–1198.

[9] X. Wang, C. Zhang, and Z. Zhang, "Boosted multi-task learning for face verification with applications to web image and video search," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 142–149.

[10] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 814–822.

[11] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.

[12] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statist. Comput.*, vol. 20, no. 2, pp. 231–252, 2010.

[13] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan, "A dirty model for multi-task learning," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 23. 2010, pp. 964–972.

[14] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 895–903.

[15] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[16] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, Nov. 2005.

[17] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, p. 22, 2012.

[18] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 42–50.

[19] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," in *Proc. ECCV 2010 Workshop Parts Attributes*, vol. 1.

[20] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 951–958.

[21] T. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proc. ECCV*, 2010, pp. 663–676.

[22] J. Wang, K. Markert, and M. Everingham, "Learning models for object recognition from natural language descriptions," in *Proc. British Mach. Vis. Conf.*, 2009.

[23] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1681–1688.

[24] S. Branson *et al.*, "Visual recognition with humans in the loop," in *Proc. ECCV*, 2010, pp. 438–451.

[25] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[26] R. M. Larsen, "Lanczos bidiagonalization with partial reorthogonalization," *DAIMI PB*, vol. 27, no. 537, pp. 1–101, 1998.

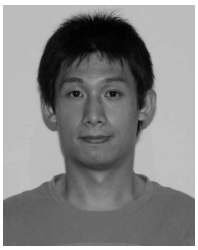
[27] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," in *Proc. NIPS*, 2011, doi: 10.1016/j.jsb.2012.10.010.

[28] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

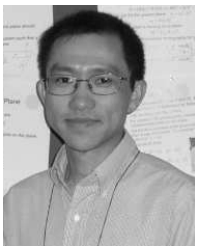
[29] Q. Zhang, L. Chen, Q. Tian, and B. Li, "Video-based analysis of motion skills in simulation-based surgical training," in *IS & T/SPIE Electron. Imag., Int. Soc. Opt. Photon.*, vol. 8667, pp. 86670A–86670A, Mar. 2013, doi: 10.1117/12.2005177.

[30] Q. Zhang and B. Li, "Relative hidden Markov models for evaluating motion skill," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 548–555.
[31] S. Jun *et al.*, "Robotic minimally invasive surgical skill assessment based on automated video-analysis motion studies," in *Proc. 4th IEEE RAS EMBS Int. Conf. BioRob*, Jun. 2012, pp. 25–31.
[32] J. Doyle, E. Webber, and R. Sidhu, "A universal global rating scale for the evaluation of technical skills in the operating room," *Amer. J. Surgery*, vol. 193, no. 5, pp. 551–555, 2007.

Qiang Zhang received the B.S. degree in electronic information and technology from Beijing Normal University, Beijing, China, in 2009, and the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA, in 2014. Since 2014, he has been with Samsung Semiconductor Inc., Pasadena, CA, USA, as a Senior Software Engineer in Computer Vision. His research interests include image/video processing, computer vision, and machine vision, specialized in sparse learning, face recognition, and motion analysis.



Lin Chen received the B.S. and M.S. degrees in computer science from Shandong University, Shandong, China, in 2008 and 2011, respectively. He is currently pursuing the Ph.D. degree in computer science and engineering with Arizona State University, Tempe, AZ, USA. His research interests include computer graphics, geometry processing, image/video processing, computer vision, and machine learning, specialized in attribute learning, multitask learning, and motion analysis.



Baoxin Li (S'97–M'00–SM'04) received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2000. He is currently an Associate Professor of Computer Science and Engineering with Arizona State University, Tempe, AZ, USA. From 2000 to 2004, he was a Senior Researcher with SHARP Laboratories of America, Camas, WA, USA, where he was the Technical Lead in developing SHARP's HiIMPACT Sports technologies. From 2003 to 2004, he was also an Adjunct Professor with Portland State University, Portland, OR, USA. He holds nine issued U.S. patents. His current research interests include computer vision and pattern recognition, image/video processing, multimedia, medical image processing, and statistical methods in visual computing. He was a recipient of the SHARP Laboratories' President Awards in 2001 and 2004, the SHARP Laboratories' Inventor of the Year Award in 2002, and the National Science Foundation's CAREER Award from 2008 to 2009.