Multitask Linear Discriminant Analysis for View Invariant Action Recognition

Yan Yan, *Student Member, IEEE*, Elisa Ricci, *Member, IEEE*, Ramanathan Subramanian, *Member, IEEE*, Gaowen Liu, and Nicu Sebe, *Senior Member, IEEE*

***Abstract*— Robust action recognition under viewpoint changes has received considerable attention recently. To this end, self-similarity matrices (SSMs) have been found to be effective view-invariant action descriptors. To enhance the performance of SSM-based methods, we propose multitask linear discriminant analysis (LDA), a novel multitask learning framework for multiview action recognition that allows for the sharing of discriminative SSM features among different views (i.e., tasks). Inspired by the mathematical connection between multivariate linear regression and LDA, we model multitask multiclass LDA as a single optimization problem by choosing an appropriate class indicator matrix. In particular, we propose two variants of graph-guided multitask LDA: 1) where the graph weights specifying view dependencies are fixed *a priori* and 2) where graph weights are flexibly learnt from the training data. We evaluate the proposed methods extensively on multiview RGB and RGBD video data sets, and experimental results confirm that the proposed approaches compare favorably with the state-of-the-art.**

***Index Terms*— Multi-view action recognition, self-similarity matrix, multi-task learning, linear discriminant analysis.**

I. Introduction

HUMAN action recognition and understanding from image and video content has attracted considerable attention in computer vision due to its critical role in surveillance, behavior analysis, human-computer interaction, robotics and content-based retrieval. Several solutions have been proposed for action recognition over the years–readers may refer to [1], [2] for extensive surveys. From the *representation* point of view, the approaches can be mainly classified into methods computing the time evolution of human silhouettes [3], action cylinders [4], space-time shapes [5], covariance features [6] and local 3D patch descriptors [7]. From the *feature descriptor* point of view, the various approaches can be categorized into motion-based [8], appearance-based [9], space-time volume-based [5], space-time interest point-based [10], and Self Similarity Matrices-based [11].

Recently, multi-view action recognition methods have gained in popularity. Since self-occlusion problems can be tackled effectively by employing multiple cameras, multi-view frameworks can achieve more robust action recognition than monocular methods. However, as actions are typically recognized based on the actor's motion trajectories with respect to the camera viewpoint, viewpoint changes significantly impact action understanding. Therefore, extracting view-invariant information is an important step in multi-view settings but relatively few works have addressed the effect of viewpoint changes on action recognition. Some recent approaches have achieved view-invariant recognition of actions by transferring features across views [13]–[15] or by using view-invariant features [11], [16], [17].

A possible methodology for achieving view-invariant action recognition is to compute features which are stable across different viewpoints. Temporal self-similarity matrices (SSMs) [11], computed from different low-level features such as Histogram of Oriented Gradients (HOG) and Histogram of Optical Flows (HOF), are shown to be robust descriptors for view-invariant action recognition. However, a careful analysis of SSMs reveals that they are also sensitive to large viewpoint-related appearance changes. This effect can be observed in Fig. 1, where SSMs corresponding to five views for action sequences from the IXMAS [12] and NIXMAS [3] datasets are shown. Although the SSMs associated to all five views share some similarities, it is easy to note that the HOG-based SSM corresponding to the last view (CAM5) is significantly different from the remaining views (CAM1-CAM4) for both datasets. To arrive at an action representation in the presence of large view-related appearance changes, one approach is to find those camera views in which the motion patterns for that action are highly correlated. Multi-task learning (MTL) [18], [19], which simultaneously learns classification/regression models for a set of related tasks, represents an attractive solution to this end. By learning latent relationships between tasks, MTL typically enables the synthesis of models superior to a learner that models each task independently.

In this paper, we present **Multi-task LDA**, a novel multi-task learning framework to enhance the discriminative power of SSMs for multi-view action recognition, and

Manuscript received December 5, 2013; revised June 4, 2014 and September 2, 2014; accepted October 21, 2014. Date of publication October 29, 2014; date of current version November 18, 2014. This work was supported in part by the Italian Ministry for Education Universities and Research through the Cluster Project Active Ageing at Home and in part by the FIRB S-PATTERNS Project and Agency for Science, Technology and Research, Singapore, through the Human Sixth Sense Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nilanjan Ray.

Y. Yan, G. Liu, and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy (e-mail: yan@disi.unitn.it; gaowen.liu@unitn.it; sebe@disi.unitn.it).

E. Ricci is with the Department of Engineering, University of Perugia, Perugia 06123, Italy, and also with Fondazione Bruno Kessler, Trento 38122, Italy (e-mail: elricci@fbk.eu).

R. Subramanian is with the Advanced Digital Sciences Center, Singapore 138632, and also with the University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: subramanian.r@adsc.com.sg).

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.2014.2365699

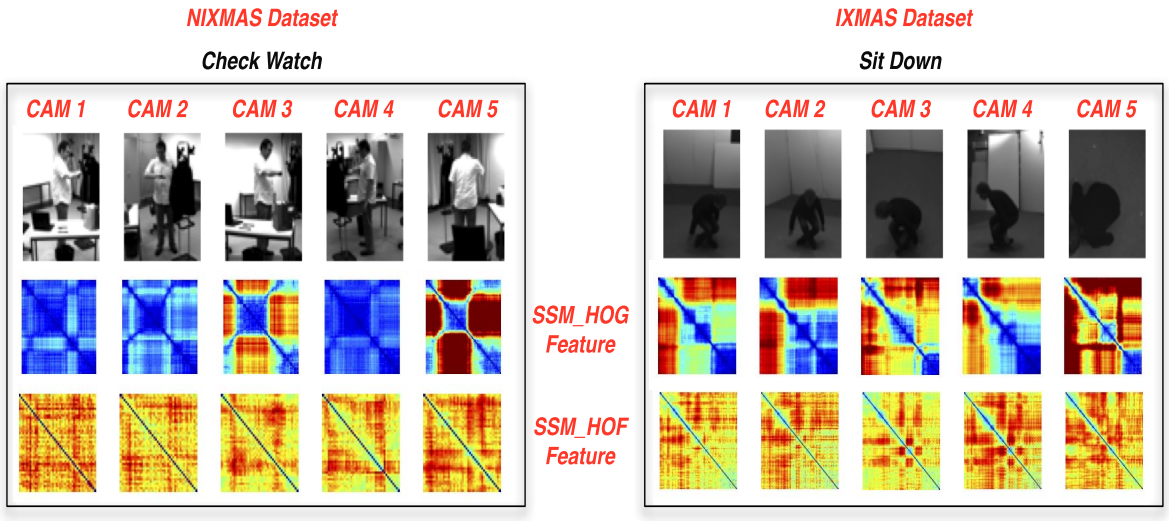


Fig. 1. Exemplar SSMs computed from HOG and HOF features for the NIXMAS [3] and IXMAS [12] datasets. Note the large discrepancy between the SSMs corresponding to CAM 5 and others for both datasets.

demonstrate how sharing of features across views (tasks) leads to improved recognition performance. Inspired by the equivalence relationship between multivariate linear regression and linear discriminant analysis (LDA) [20], we cast multi-task multi-class LDA as a single optimization problem by choosing an appropriate class indicator matrix, and develop an efficient algorithm to solve it. Also, by defining a graph reflecting prior knowledge on the similarity among different views, the degree of relatedness of the corresponding view features can be controlled using the proposed approach. We describe two variants of graph-guided multi-task LDA and evaluate their performance: (1) Multi-task sparse graph-guided LDA, where the graph weights specifying view dependencies are defined *a priori*, and (2) Multi-task flexible graph-guided LDA where the graph weights are flexibly learned from (or iteratively refined based on) training features.

Our experiments demonstrate how our methods can be successfully employed for view invariant action recognition, *i.e.* considering the case where images corresponding to the test view are not available in the training data. The obtained results also confirm that sharing features among views is indeed beneficial for multi-view action recognition–our methods outperform competing SSM-based approaches that do not consider task relationships by 10% on the IXMAS dataset. Overall, the proposed approaches achieve efficient recognition of actions on both RGB (video) and RGBD (depth image) data, and compare favorably with the state-of-the-art.

To summarize, the main contributions of this paper are:

- It represents one of the first works to explore a multi-task learning framework for multi-view action recognition. While other recent methods such as [21] also adopt MTL, a unique aspect of our framework is that, by effectively combining SSMs descriptors, sparsity and MTL, it easily allows for accurate action classification on missing views, for which no examples are available in the training data.
- The proposed approach is shown to be highly effective and achieves improved action recognition performance with respect to other classification methods based on SSM descriptors. While competing works have typically evaluated their algorithms on multi-view video data, we also demonstrate how our framework is applicable to multi-view depth images as in the ACT4^2 [22] dataset.
- The proposed multi-task LDA framework is novel, and is modeled as a single optimization problem through the use of a class indicator matrix. The described graph-guided learning algorithms can be generically applied to other computer vision tasks as well.

II. Related Work

We review related work on multi-view action recognition, linear discriminant analysis and multi-task learning.

A. Multi-View Action Recognition

Multi-view action recognition has received much attention recently, since a multi-view setup can overcome the problem of self-occlusions and can enable more robust action recognition as compared to monocular methods. Both 3D and 2D-based approaches have been proposed for multi-view action recognition as detailed below.

Knowing the 3D scene geometry enables the adaptation of action features from one view to another through the use of geometric transformations. For example, Weinland *et al.* [12] use 3D occupancy grids synthesized from multiple viewpoints to model actions using an exemplar-based HMM. Yen *et al.* [23] employ a 4D action feature model for recognizing actions from arbitrary views. This model encodes shape and motion of actors observed from multiple views, and requires the reconstruction of 3D visual hulls of actors at each time instant. Both approaches lead to computationally intensive algorithms as finding the best match between a 3D model and a 2D observation requires searching over a large model parameter space. Weinland *et al.* [3] develop a hierarchical classification method based on

3D Histogram of Oriented Gradients (HOG) to represent a test sequence. Robustness to occlusions and viewpoint changes are achieved by combining training data from all viewpoints to train hierarchical classifiers.

A successful approach to tackle the problem of viewpoint-related appearance differences on action recognition in 2D approaches involves the design of view-invariant features. Rao *et al.* [17] present a view-invariant representation of human actions by capturing changes in the speed and direction of action trajectories using spatio-temporal trajectory curvature. Parameswaran *et al.* [24] propose modeling actions in terms of view-invariant canonical body poses and trajectories in 2D invariance space, which enables recognition of human actions from a generic view-point. Junejo *et al.* [11] introduce temporal SSMs descriptors as features robust to viewpoint changes.

Farhadi and Tabrizi [13] explicitly address correlations between actions observed from different views. They use a split-based representation to describe clusters of codewords in each view. The transfer of these splits between views is learned from multi-view action sequences. In [25], Farhadi *et al.* model view as a latent parameter, and learn features that can discriminate between both views and actions. Liu *et al.* [15] use a bipartite graph to model the relationship between two codebooks generated by k-means clustering of videos acquired for each view. A bipartite partition is used to co-cluster two view-dependent codebooks into shared visual-word clusters, and a codebook composed of these shared clusters is used to encode videos from both views. However, this approach only exploits codebook-to-codebook correspondence at video-level, which cannot guarantee that a pair of videos corresponding to two different views has similar feature representations based on the shared codebook. In addition, it uses a fusion method to combine prediction outputs, which requires the clustering of test videos in the target view. Zheng and Jiang [26] present an approach to jointly learn a set of common and view-specific dictionaries for cross-view action recognition. However, their main focus is on transferring information from one view to another, and not on jointly modeling the relations among multiple views.

B. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is widely used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. This makes LDA a very practical tool for classification and dimensionality reduction. The optimal projection or transformation in classical LDA is obtained by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination.

LDA has been applied successfully in many computer vision applications such as face recognition [27] or head pose estimation [28]. Multi-task extensions of LDA have been proposed in [29] and [30]. However in [29], the proposed framework is not flexible as no learning of the relationship between tasks is conducted. In [30], the problem of designing a multi-task LDA algorithm when the different tasks correspond to heterogeneous feature spaces is addressed. However, we do not consider this scenario as it is not appropriate for our application.

C. Multi-Task Learning

Many real-world applications involve related classification/regression tasks. Multi-task learning methods aim to simultaneously learn models for a set of related tasks. By learning tasks in parallel while using a shared representation, performance is typically improved.

Traditional MTL methods consider a single shared model, assuming that all the tasks are related [18], [19]. However, when some of the tasks are unrelated, this may lead to negative transfer and the performance can be even worse than single-task learning. Recently, more sophisticated approaches have been proposed to counter this problem. These methods assume some *a-priori* knowledge (*e.g.* in the form of a graph) defining task dependencies [31] or learning task relationships simultaneously with task-specific parameters [32]. For example, Jalali *et al.* [33] assume that the data follows a dirty model. Zhou *et al.* [34] prove that the clustered MTL approach is equivalent to alternating structure optimization that assumes the tasks share a low-dimensional structure. The approach proposed in [35] assumes that tasks are clustered, and that clustering structure can be inferred automatically during learning.

Multi-task learning has received considerable attention from the vision community, and has been successfully applied to many problems such as image classification [36], visual tracking [37], daily activity recognition from first-person videos [38], image-based indoor localization [39] and head pose classification under motion [40]–[42]. An MTL approach to monocular action recognition is proposed in [43], where the authors exploit relatedness of action categories to learn latent tasks (motion patterns) shared across actions.

This paper is an extension of previous work presented in [44], where Multi-task LDA guided by a graph with fixed edge weights is proposed. To our knowledge, multi-view action recognition using multi-task learning has not been considered by other works with the exception of [21], which is contemporaneous to ours. Also, our approach is different with respect to [21] in the following respects: (1) While [21] seeks to learn latent action groups, so that within-group feature sharing is allowed but between-group feature sharing is prohibited, we explore learning of latent and discriminative SSM features across views; (2) A part-based action representation is used in [21], while we use the bag-of-words model for encoding SSM features; (3) A large-margin framework is used for LMTL formulation in [21], while we propose LDA-based MTL, and (4) While in [21] the main focus is multi-view action recognition, we also consider the problem of action recognition with missing view, *i.e.* on a novel camera view for which no examples are available in the training set. Furthermore, we show action classification results on the ACT4[2] multi-view depth image dataset, in addition to traditional action video datasets. A description of the

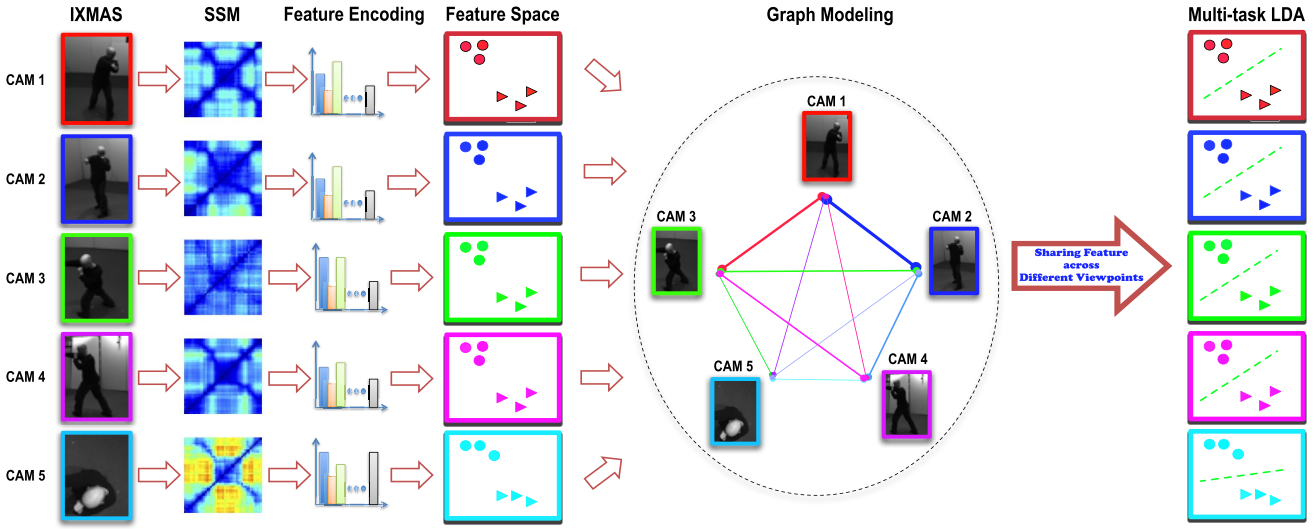


Fig. 2. Overview of Multi-task LDA-based multi-view action recognition. Note that the viewpoint associated to CAM5 is significantly different from the other four views and correspondingly, the decision boundary for CAM5 is remarkably distinct from the others.

proposed Multi-task LDA framework is presented in the following section.

III. Multi-Task LDA for Multi-View Action Recognition

In this section, we further discuss the motivation behind the proposed framework and present an overview of our approach. Then, Self-Similarity Matrix (SSM) descriptors are introduced followed by the analysis of the equivalence between LDA and linear regression. Finally, our multi-task LDA algorithm and its application to the problem of view-independent action recognition are described.

A. Motivation and Overview

Among the several view-invariant descriptors proposed throughout the years in the context of multi-view action recognition, SSMs have been particularly successful, mainly due to their simplicity and robustness, even in the challenging situation of recognition with missing views. However, since a careful analysis of SSMs reveals that these descriptors are sensitive to large viewpoint-related appearance changes (Fig. 1), we propose to improve the discriminative power of SSM features using a novel MTL approach. Our MTL framework permits to individuate a subset of features which truly possess the view-invariance property and must be shared across different views, thus leading to more robust cross-view recognition with respect to previous approaches based on SSMs. Importantly, our method is very flexible, and relies on a graph structure modeling the degree of similarity among different views. Note that actions observed from neighboring cameras may be more similar as compared to other far-away sensors, as discussed earlier.

The proposed approach for robust view independent action recognition is illustrated in Fig. 2. First, different types of low-level features are extracted from videos on a per-frame basis. The type of computed low-level features depends on the considered data: we use Histogram of Oriented Gradients (HOG), Histogram of Optical Flows (HOF) and their combination to describe RGB data, while Motion History Images (MHIs) and their variations are adopted to encode information from depth images. Once the SSM descriptors for these low-level features are computed, the standard bag-of-words model is employed for encoding features into histograms. Finally, the proposed multi-task LDA is adopted to learn a set of linear classifiers (one for each view), imposing the constraint that the weight vectors of related camera views should be similar. Relatedness among different views is modeled by a graph structure. Our approach is described in detail in the following subsection.

B. Self-Similarity Matrix Descriptors

Junejo *et. al.* [11] introduced SSM descriptors as features robust to viewpoint changes. Given a sequence of images $\mathcal{I} = \{I_1, I_2, \ldots, I_T\}$ a SSM is a square symmetric matrix:

$$\text{SSM}(\mathcal{I}) = \begin{bmatrix} 0 & e_{12} & e_{13} & \cdots & e_{1T} \\ e_{21} & 0 & e_{23} & \cdots & e_{2T} \\ e_{31} & e_{32} & 0 & \cdots & e_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{T1} & e_{T2} & e_{T3} & \cdots & 0 \end{bmatrix}, \quad (1)$$

where $e_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|^2$ is the Euclidean distance between low-level features $\mathbf{f}_i, \mathbf{f}_j$ extracted at frames I_i and I_j, respectively. Obviously, as the diagonal corresponds to comparing a frame to itself, it contains all zeros. As low level features in this paper, we use HOG and HOF descriptors on RGB frames, while MHIs are adopted in the case of depth images. The chosen features are described in detail in the experimental section. Once SSMs have been computed (separate SSMs are obtained for each type of low level features), the strategy described in [11] is adopted for calculating local descriptors. For each point on the diagonal of a single SSM, three local descriptors are computed corresponding to

three different diameters in the log-polar domain (28, 42 and 56 frames respectively in diameter). The bag-of-words model is then employed to obtain the final histogram representation of a video clip. A codebook size of 500 words is used in our experiments.

An example of SSMs computed on a sequence extracted from the IXMAS dataset is shown in Fig. 1. Obviously, SSMs obtained with different low-level features are different, since each feature captures specific properties of an action. Moreover, SSMs are rather stable over different people performing the same action under multiple viewpoints. However, as noted earlier, SSMs are robust to view changes only up to a certain extent. Therefore, in order to individuate common features from different views, multi-task LDA learning is proposed in this work.

C. Linear Discriminant Analysis

Linear Discriminant Analysis is a popular technique for dimensionality reduction and classification. We consider a dataset of N samples, $\mathcal{T} = \{(x_i, \ell_i)\}_{i=1}^{N}$, where $x_i \in I\!R^d$ and $\ell_i \in \{1, 2, ..., k\}$ denote respectively the feature vector and the associated class label for the i-th sample, d is the data dimensionality, and k the number of classes. Let $(\cdot)'$ denote the transpose operator. In discriminant analysis [45], three scatter matrices are defined as follows:

$$S_w = \frac{1}{N}\sum_{j=1}^{k}\sum_{\{x\in\mathcal{T},\ x:\ell=j\}}(x-c_j)(x-c_j)', \tag{2}$$

$$S_b = \frac{1}{N}\sum_{j=1}^{k} N_j (c_j - c)(c_j - c)', \tag{3}$$

$$S_t = \frac{1}{N}\sum_{i=1}^{N} (x_i - c)(x_i - c)', \tag{4}$$

where N_j and c_j denote the number of points and the centroid for the j-th class, while c is the computed centroid of the entire data. It follows from the definition that trace (S_w) and trace (S_b) measure the within-class cohesion and between-class separation respectively. The total scatter matrix is then obtained as $S_t = S_b + S_w$. LDA computes a linear transformation $U \in I\!R^{l\times d}$, mapping the vector $x_i \in I\!R^d$ to a vector $x_i^l \in I\!R^l, x_i^l = Ux_i$, $(l < d)$. In the low dimensional space resulting from the linear transformation U, the scatter matrices become:

$$S_w^l = U'S_wU, \quad S_b^l = U'S_bU, \quad S_t^l = U'S_tU. \tag{5}$$

The optimal transformation U^{LDA} is computed by solving the following optimization problem [45]:

$$U^{LDA} = \max_U \ \text{trace}(S_b^l (S_t^l)^{-1}). \tag{6}$$

The matrix U^{LDA} is represented by the eigenvectors of $S_t^{-1}S_b$ corresponding to the largest k-1 eigenvalues. In the specific case of a binary-class problems, the optimal transformation [46] is given by:

$$U^{LDA} = S_t^{+}(c_1 - c_2), \tag{7}$$

where c_1 and c_2 are the centroids of the the negative and positive classes, respectively.

D. Linear Regression and LDA

The objective of linear regression is to learn the optimal weight vector $w \in I\!R^d$ such that the function $f(x) = x'w$ can be used to obtain a good estimate of the desired output value ℓ_i, given as input the associated vector x_i. A popular technique for estimating w is the least squares approach, in which the following objective function is minimized:

$$L(w) = \frac{1}{2}\left\|\mathbf{X}'w - \mathbf{y}\right\|^2, \tag{8}$$

where $\mathbf{X} = [x_1, x_2, \ldots, x_N]$ is the data matrix and $\mathbf{y} = [\ell_1, \ldots, \ell_N]$ is the vector of class labels. Considering a binary classification problem and assuming that both the data vectors and labels have been centered (*i.e.* $\sum_i^N x_i = 0$ and $\sum_i^N \ell_i = 0$), it follows that $\ell_i \in \{-2N_2/N, 2N_1/N\}$ where N_1 and N_2 denote the number of samples from the negative and positive classes respectively. The optimal w is given by $w = (\mathbf{X}\mathbf{X}')^{+}\mathbf{X}\mathbf{y}$ [47]. Noticing that $\mathbf{X}\mathbf{X}' = NS_t$ and $\mathbf{X}\mathbf{y} = \frac{2N_1N_2}{N}(c_1 - c_2)$ it follows that:

$$w = \frac{2N_1N_2}{N^2}S_t^{+}(c_1 - c_2) = \frac{2N_1N_2}{N^2}U^{LDA}, \tag{9}$$

where U^{LDA} is the optimal solution to LDA in (7). Hence linear regression with the class labels as output values is equivalent to LDA, as the projection in LDA is invariant to scaling [46].

Recently, similar results have been proven for multi-class LDA [20] showing that it is equivalent to multivariate linear regression if an indicator matrix $\bar{\mathbf{Y}} \in I\!R^{N\times k}$ is defined as follows:

$$(\bar{\mathbf{Y}})_{ij} = \begin{cases} \sqrt{\dfrac{N}{N_j}} - \sqrt{\dfrac{N_j}{N}} & \text{if } \ \ell_i = j \\ -\sqrt{\dfrac{N_j}{N}} & \text{otherwise,} \end{cases} \tag{10}$$

where $(\cdot)_{ij}$ is the element in the i-th row, j-th column of the matrix. The optimal projection matrix $\mathbf{W} \in I\!R^{k\times d}$ is obtained by solving the following optimization problem:

$$\min_{\mathbf{W}} \quad \frac{1}{2}\left\|\mathbf{X}'\mathbf{W} - \bar{\mathbf{Y}}\right\|_F^2, \tag{11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Further details about this equivalence can be found in [20].

E. Multi-Task Linear Discriminant Analysis

In this paper, an extension of multiclass LDA [20] to a multi-task learning setting is proposed.

1) Definition and Notation: We consider a set of R camera views (*i.e.* related tasks). Each task is a multi-class classification problem with C categories. For each task $t = 1, 2, \ldots, R$, a training set $\mathcal{T}_t = \{(x_n^t, \ell_n^t)\}_{n=1}^{N_t}$ is given, where $x_n^t \in I\!R^d$ is a d-dimensional feature vector, and $\ell_n^t \in \{1, 2, \ldots, C\}$ is the label indicating the class membership. For each task t we define the matrices $\mathbf{x}_t \in I\!R^{N_t\times d}$, $\mathbf{x}_t = [x_1^t, \ldots, x_{N_t}^t]'$ and

$\mathbf{y}_t \in I\!R^{N_t \times C}$, with:

$$(\mathbf{y}_t)_{ij} = \begin{cases} \sqrt{\frac{N_t}{N_{t,j}}} - \sqrt{\frac{N_{t,j}}{N_t}} & \text{if } \ell_i^t = j \\ -\sqrt{\frac{N_{t,j}}{N_t}} & \text{otherwise,} \end{cases} \quad (12)$$

where $N_{t,j}$ is the sample size of the j-th class in the t-th task, $N_t = \sum_{j=1}^{C} N_{t,j}$ is the total number of training samples of all classes in the t-th task. Concatenating $\mathbf{x}_t$ and $\mathbf{y}_t$ of all the R tasks, the matrices $\mathbf{X} = [\mathbf{x}_1', \ldots, \mathbf{x}_R']'$, $\mathbf{X} \in I\!R^{N \times d}$ and $\mathbf{Y} = [\mathbf{y}_1', \ldots, \mathbf{y}_R']'$, $\mathbf{Y} \in I\!R^{N \times CR}$ are obtained, where $N = \sum_{t=1}^{R} N_t$. We further define a graph which models the similarity among different views (tasks) specifying the edge-vertex incident matrix $\mathbf{M}$, $\mathbf{M} \in I\!R^{\frac{R(R-1)}{2} \times CR}$, where:

$$(\mathbf{M})_{q=(i,j),h} = \begin{cases} \gamma_{ij} & \text{if } i = h \\ -\gamma_{ij} & \text{if } j = h \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\gamma_{ij} = (\sum_{i \neq j} \frac{1}{pq} \sum_p \sum_q \left\| x_p^i - x_q^j \right\|_2)^{-1}$, *i.e.*, γ_{ij} is set by calculating the inverse of the normalized Euclidean distance of SSMs descriptors between two different views (tasks) for the same action/class, averaged on the training data. γ_{ij} is normalized into the interval [0, 1] and a large γ_{ij} indicates high similarity of specific action/class between views. In practice, the similarity between two views is defined according to the corresponding observed data. However, other approaches are possible as well, such as considering camera geometry to specify γ_{ij}.

2) Proposed Approach: In our multi-task LDA approach we propose to learn a global weight matrix $\mathbf{U} = [\mathbf{u}_1', \ldots, \mathbf{u}_R']'$, $\mathbf{U} \in I\!R^{d \times CR}$ by solving the following optimization problem:

$$\min_{\mathbf{U}} \quad \Lambda(\mathbf{U}) = \frac{1}{2} \left\| (\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \Omega_{\mathbf{M}}(\mathbf{U}). \quad (13)$$

In (13), the loss term minimizes the training errors on each view (task) separately, while the regularization term $\Omega_{\mathbf{M}}(\mathbf{U})$ ensures that related views (according to the graph specified by $\mathbf{M}$) have similar classifiers. We propose two variant approaches to multi-task LDA, corresponding to different regularization terms $\Omega_{\mathbf{M}}(\mathbf{U})$. We present them in the following subsections.

F. Multi-Task Sparse Graph Guided LDA (MT-SGG-LDA)

The intuition behind our first approach, Multi-task Sparse Graph Guided LDA, is simple: we want to learn a projection matrix $\mathbf{U}$ which optimally separates data from different classes, is sparse (thus filtering out noisy features) and has a structure which reflects view-similarity (*i.e.* related views should have similar classifiers). To achieve this goal, we propose to solve the following optimization problem:

$$\min_{\mathbf{U}} \quad \frac{1}{2} \left\| (\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \lambda_1 \left\| \mathbf{M}\mathbf{U}' \right\|_F^2 + \lambda_2 \left\| \mathbf{U} \right\|_1, \quad (14)$$

where $\| \cdot \|_1$ denote the L_1 norm, which enforces sparsity on the learned weights matrix. In the proposed objective function all tasks are related thanks to the graph regularization term, and therefore knowledge from one task can be utilized by the other tasks. Prior knowledge about the required level of sharing feature is embedded in the learning framework through $\mathbf{M}$. To solve (14) we adopt the accelerated gradient method Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [48], as described in Algorithm 1.

Algorithm 1 Multi-Task Sparse Graph Guided LDA

Input: $\mathcal{T}_t = \{(x_n^t, \ell_n^t)\}_{n=1}^{N_t}$, $\forall t = 1, \ldots, R$, λ_1, λ_2, $\mathbf{M}$

Initialize $\mathbf{U}_0$, $\alpha_0 = 1$, $\hat{\lambda}_1 = 2\lambda_1 / L_k$, $\hat{\lambda}_2 = 2\lambda_2 / L_k$.
LOOP:
$\alpha_k = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{k-1}^2})$
$\hat{\mathbf{U}} = \mathbf{U}_k - \frac{2}{L_k}[\mathbf{X}'(\mathbf{Y}\mathbf{Y}')^{-1}(\mathbf{X}\mathbf{U}_k - \mathbf{Y}) + \hat{\lambda}_1 \mathbf{U}_k \mathbf{M}'\mathbf{M}]$
Solve $\mathbf{U}_{k+\frac{1}{2}} \leftarrow \min_{\mathbf{U}} \left\| \mathbf{U} - \hat{\mathbf{U}} \right\|_F^2 + \hat{\lambda}_2 \left\| \mathbf{U} \right\|_1$ using $\Sigma_\lambda(x) = \text{sign}(x) \max(|x| - \lambda, 0)$.
$\mathbf{U}_{k+1} = (1 + \frac{\alpha_{k-1} - 1}{\alpha_k})\mathbf{U}_{k+\frac{1}{2}} - \frac{\alpha_{k-1} - 1}{\alpha_k}\mathbf{U}_k$
Until Convergence
Output: $\mathbf{U}$

1) Optimization: The objective function in (14) can be decomposed into two parts, *i.e.* a smooth term $\Pi(\cdot)$ and a non smooth term $\Omega(\cdot)$,

$$\Pi(\mathbf{U}) = \frac{1}{2} \left\| (\mathbf{Y}\mathbf{Y}')^{-1/2}(\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \lambda_1 \left\| \mathbf{M}\mathbf{U}' \right\|_F^2$$
$$\Omega(\mathbf{U}) = \lambda_2 \left\| \mathbf{U} \right\|_1.$$

The term $(\mathbf{Y}\mathbf{Y}')^{-1/2}$ has been integrated as a normalization factor, to compensate for the different number of samples per class. FISTA solves the optimization problems in the form $\min_{\mathbf{U}} \Pi(\mathbf{U}) + \Omega(\mathbf{U})$, where $\Pi(\mathbf{U})$ is convex and smooth, $\Omega(\mathbf{U})$ is convex but non-smooth. In each FISTA iteration, a proximal step is computed [48]:

$$\min_{\mathbf{U}} \left\| \mathbf{U} - \hat{\mathbf{U}} \right\|_F^2 + \frac{2}{L_k} \Omega(\mathbf{U}),$$

where $\hat{\mathbf{U}} = \tilde{\mathbf{U}}_k - \frac{1}{L_k} \nabla \Pi(\tilde{\mathbf{U}}_k)$, $\tilde{\mathbf{U}}_k$ is the current iterate and L_k is a stepsize found by line search. To solve the proximal step, the soft-thresholding operator $\Sigma_\lambda(x) = \text{sign}(x) \max(|x| - \lambda, 0)$ is adopted [49].

G. Multi-Task Flexible Graph Guided LDA (MT-FGG-LDA)

The second multi-task LDA approach we propose also develops from (13), but instead of fixing the graph weights modeling task dependencies as in (14), it learns them from training data. Intuitively, we expect that this increased flexibility leads to improvements in terms of classification accuracy. We define $\mathbf{U} = \mathbf{C} + \mathbf{S}$, $\mathbf{C}, \mathbf{S} \in I\!R^{d \times CR}$, $\mathbf{C} = [\mathbf{c}_1', \ldots, \mathbf{c}_R']'$, $\mathbf{S} = [\mathbf{s}_1', \ldots, \mathbf{s}_R']'$ *i.e.*, we consider the weight matrix as the matrix obtained on summing two terms, the matrix $\mathbf{C}$ modeling common features among tasks (views), and the matrix $\mathbf{S}$ which accounts for task/view-specific features. We formulate the following optimization problem:

$$\min_{\mathbf{C},\mathbf{S}} \left\| \mathbf{Y} - \mathbf{X}(\mathbf{C} + \mathbf{S}) \right\|_F^2 + \lambda \Omega(\mathbf{C}, \mathbf{S}), \quad (15)$$

Algorithm 2 Multi-Task Flexible Graph Guided LDA

Input: $\mathcal{T}_t = \{(x_n^t, \ell_n^t)\}_{n=1}^{N_t}$, $\forall t = 1, \ldots, R$, λ, λ_c, $\mathbf{M}$.

Initialize $\mathbf{C}_0$, $\mathbf{S}_0$, $\alpha_0 = 1$.
OUTER LOOP:
$\alpha_k = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{k-1}^2})$
$\hat{\mathbf{C}} = \mathbf{C}_k - 2\mathbf{X}^T(\mathbf{X}\mathbf{C}_k - \mathbf{Y})$
FOR $i = 1 : D$
Initialize $\mathbf{q}^{i,0}$, $\mathbf{z}^{i,0}$, $\mathbf{c}^{i,0}$
Compute $\mathbf{A} = \rho\mathbf{M}^T\mathbf{M} + (2 + 2\hat{\lambda}_c)\mathbf{I}$.
Compute Cholesky factorization of matrix $\mathbf{A}$.
INNER LOOP:
$\mathbf{b}^n = \rho\mathbf{M}^T\mathbf{q}^{i,n} - \mathbf{M}^T\mathbf{z}^{i,n} + 2\hat{\mathbf{c}}^i$
Solve $\mathbf{A}\mathbf{c}^{i,n+1} = \mathbf{b}^n$
$\mathbf{q}^{i,n+1} = \Sigma_{\hat{\lambda}_1/\rho}(\mathbf{M}\mathbf{c}^{i,n+1} + \frac{1}{\rho}\mathbf{z}^{i,n})$
$\mathbf{z}^{i,n+1} = \mathbf{z}^{i,n} + \rho(\mathbf{M}\mathbf{c}^{i,n+1} - \mathbf{q}^{i,n+1})$
Until Convergence
END FOR
$\mathbf{C}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_k})\mathbf{C}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k}\mathbf{C}_k$
$\hat{\mathbf{S}} = \mathbf{S}_k - 2\mathbf{X}^T(\mathbf{X}\mathbf{S}_k - \mathbf{Y})$
$\mathbf{S}_{k+\frac{1}{2}} = \frac{1}{1+\lambda}\hat{\mathbf{S}}$
$\mathbf{S}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_n})\mathbf{S}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k}\mathbf{S}_k$
Until Convergence
Output: $\mathbf{U} = \mathbf{C} + \mathbf{S}$

where $\Omega(\cdot)$ is an appropriate regularization term defined as:

$$\Omega(\mathbf{C}, \mathbf{S}) = \|\mathbf{C}\|_F^2 + \|\mathbf{S}\|_F^2 + \lambda_c \left\|\mathbf{M}\mathbf{C}'\right\|_1.$$

In the regularization term, similarly to what was proposed for Multi-task Sparse Graph Guided LDA, the L_1 norm regularizer imposes the weights $\mathbf{c}_t$ of related tasks to be close together. However, with respect to (14), they become identical as $\lambda_c \to \infty$, leading to task clusters. Importantly this effect is feature-specific, *i.e.*, the cluster structure can vary from feature to feature, thus taking into account the different discriminative power of features. Moreover, $\|\mathbf{C}\|_F^2$ regulates model complexity and $\|\mathbf{S}\|_F^2$ penalizes large deviation of $\mathbf{c}_t$ from $\mathbf{u}_t$ $\forall t$. The procedure we propose to solve (15) is outlined in Algorithm 2.

1) Optimization: To apply FISTA to our optimization problem we define:

$$\Pi(\mathbf{C}, \mathbf{S}) = \|\mathbf{Y} - \mathbf{X}(\mathbf{C} + \mathbf{S})\|_F^2$$
$$\Omega(\mathbf{C}, \mathbf{S}) = \lambda\|\mathbf{S}\|_F^2 + \lambda\|\mathbf{C}\|_F^2 + \lambda\lambda_c \left\|\mathbf{M}\mathbf{C}'\right\|_1.$$

The proximal step amounts to solving the following:

$$\min_{\mathbf{C},\mathbf{S}} \left\|\mathbf{C} - \hat{\mathbf{C}}\right\|_F^2 + \left\|\mathbf{S} - \hat{\mathbf{S}}\right\|_F^2 + \hat{\lambda}_c \left\|\mathbf{M}\mathbf{C}'\right\|_1 + \hat{\lambda}\|\mathbf{C}\|_F^2 + \hat{\lambda}\|\mathbf{S}\|_F^2, \quad (16)$$

where $\hat{\lambda} = 2\lambda/L_k$ and $\hat{\lambda}_c = 2\lambda\lambda_c/L_k$, $\hat{\mathbf{S}} = \mathbf{S} - 2\mathbf{X}^T(\mathbf{X}\mathbf{S} - \mathbf{Y})$ and $\hat{\mathbf{C}} = \mathbf{C} - 2\mathbf{X}^T(\mathbf{X}\mathbf{C} - \mathbf{Y})$. To solve (16), we consider $\mathbf{C}$, $\mathbf{S}$ separately. While solving with respect to $\mathbf{S}$ is straightforward, solving with respect to $\mathbf{C}$ is more challenging due to the presence of the L_1 norm. However, since in our approach each feature dimension is considered independently, the update of the weight vectors $\mathbf{C}$ can be made very efficient by solving d separate optimization problems (one for each row $\mathbf{c}^i$ of the matrix $\mathbf{C}$) as:

$$\min_{\mathbf{c}^i} \left\|\mathbf{c}^i - \hat{\mathbf{c}}^i\right\|_2^2 + \hat{\lambda}_c \left\|\mathbf{M}\mathbf{c}^i\right\|_1 + \hat{\lambda}\left\|\mathbf{c}^i\right\|_2^2.$$

In this paper we propose to apply the augmented Lagrangian multipliers approach [49], and consider the equivalent optimization problem (in the following the superscripts are removed for sake of clarity):

$$\min_{\mathbf{c},\mathbf{q},\mathbf{z}} \left\|\mathbf{c} - \hat{\mathbf{c}}\right\|_2^2 + \hat{\lambda}_c \|\mathbf{q}\|_1 + \hat{\lambda}\|\mathbf{c}\|_2^2 + \mathbf{z}^T(\mathbf{M}\mathbf{c} - \mathbf{q}) + \frac{\rho}{2}\|\mathbf{M}\mathbf{c} - \mathbf{q}\|_2^2, \quad (17)$$

with ρ being the dual update step-length [49]. Three steps are alternated (see Algorithm 2), for solving (17) with respect to $\mathbf{c}$, $\mathbf{q}$ and $\mathbf{z}$. Solving with respect to $\mathbf{c}$, while keeping $\mathbf{q}, \mathbf{z}$ fixed, implies solving a linear system with Cholesky factorization $\mathbf{A}\mathbf{c}^{k+1} = \mathbf{b}^k$ where $\mathbf{A} = \rho\mathbf{M}^T\mathbf{M} + (2 + 2\hat{\lambda}_c)\mathbf{I}$ and $\mathbf{b}^k = \rho\mathbf{M}^T\mathbf{q}^k - \mathbf{M}^T\mathbf{z}^k + 2\hat{\mathbf{c}}$. Solving with respect to $\mathbf{q}$ produces a closed form solution, obtained by applying the soft-thresholding operator. The update step for solving with respect to $\mathbf{z}$ is straightforward.

IV. Experimental Results

In this section, we assess the proposed MT-SGG-LDA and MT-FGG-LDA approaches on three publicly available multi-view action recognition datasets, namely IXMAS [12], NIXMAS [3] and ACT4^2 [22].

A. Datasets

We consider three different datasets. The *IXMAS dataset* [12] consists of 12 action classes (*e.g. check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point* and *pick up*). Each action is executed three times by 12 subjects and is recorded with five cameras observing the subjects from very different perspectives. The frame rate is 23 fps and the frame size 390×291 pixels.

The *NIXMAS dataset* [3] contains new videos showing the same actions as in the IXMAS dataset. The dataset is recorded with different actors, cameras, and viewpoints, and about 2/3 of the videos have objects which partially occlude the actors. Overall, it contains 1148 sequences.

The *ACT4^2 dataset* [22] contains RGB+depth video sequences depicting 14 representative daily actions. The considered daily actions are: *collapse, drink, make phone call, mop floor, pick up, put on, read book, sit down, sit up, stumble, take off, throw away, twist open* and *wipe clean*.

Fig.3 shows some sample frames extracted from the IXMAS, NIXMAS and ACT4^2 datasets observed from different camera viewpoints. In our experiments, we use all of the IXMAS and NIXMAS data, and a subset (videos corresponding to 10 actors) of the ACT4^2 data.

B. Feature Representation

As discussed in Section III-B, our approach is based on SSM descriptors computed using low-level features extracted

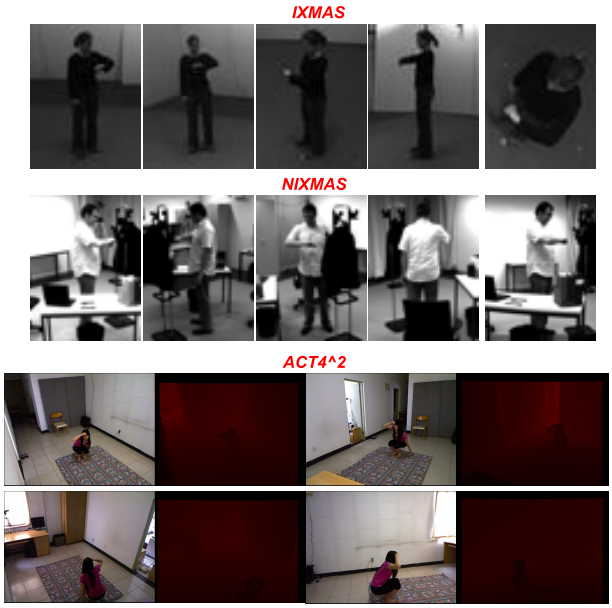


Fig. 3. Sample frames extracted from the three considered datasets. A majority of the action sequences in the NIXMAS dataset involve occlusions due to other scene objects. Figure is best viewed in color.

from individual frames. To obtain features from RGB image sequences, we used HOG and HOF features [7] in the case of IXMAS and NIXMAS datasets, while only HOG features were used for the $ACT4^2$ dataset. Moreover, in case of the $ACT4^2$ dataset, MHIs [50] were used to compute features on depth images. In an MHI, each pixel intensity is a function of the motion history at that location, and a brighter pixel corresponds to more recent motion. Denoting by $D(x, y, t)$, the depth value corresponding to a pixel at location x, y and at time t, the MHI is computed as:

$$H_\tau^D(x,y,t) = \begin{cases} \tau, & if \ \ |D(x,y,t) - D(x,y,t-1)| > \delta D_{th} \\ \max(0, H_\tau^D(x,y,t-1) - 1), & otherwise \end{cases}$$

where τ is the longest time window that the system considers (τ is set equal to the number of video frames in our experiments) and δD_{th} is the threshold value for generating the mask for a motion region.

Moreover, in order to benefit to the highest degree from the depth information, two other MHI descriptors, namely forward-MHIs $H_\tau^{fD}(x, y, t)$ (encoding information about the increase of depth) and backward-MHIs $H_\tau^{bD}(x, y, t)$ (decrease of depth) [51], are defined:

$$H_\tau^{fD}(x,y,t) = \begin{cases} \tau, & if \ \ D(x,y,t) - D(x,y,t-1) > \delta D_{th} \\ \max(0, H_\tau^{fD}(x,y,t-1) - 1), & otherwise \end{cases}$$

$$H_\tau^{bD}(x,y,t) = \begin{cases} \tau, & if \ \ D(x,y,t) - D(x,y,t-1) < -\delta D_{th} \\ \max(0, H_\tau^{bD}(x,y,t-1) - 1), & otherwise \end{cases}$$

To represent each video of the $ACT4^2$ dataset, we computed separate SSM descriptors for HOG, MHI, forward-MHI and backward-MHI features and, applying a bag-of words model (using 500 words), we constructed a 2000-bin histogram corresponding to the final descriptor. In Fig.4, the extracted MHI, forward-MHI and backward-MHI features and the corresponding SSM descriptors are shown.

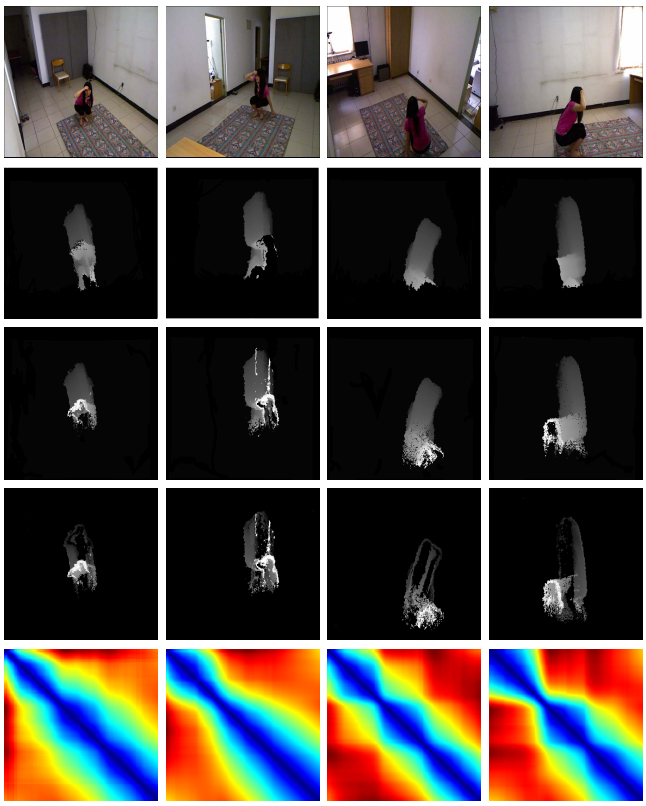


Fig. 4. $ACT4^2$ dataset and different types of features extracted: (from top to bottom) original RGB frames, Motion History Images, forward Motion History Images, backward Motion History Images, SSM descriptors, respectively.

C. Experimental Setup

A leave-one-user-out strategy was employed in our classification experiments: videos of one actor were selected for testing, while videos of the remaining actors were used as training data. For all the methods, the optimal values of the regularization parameters were determined by testing values in the interval $[2^{-6}, 2^{-5}, \ldots, 2^6]$ on a separate validation set. Mean action recognition accuracies are reported for all experiments. We evaluated the effectiveness of our algorithms in two cases:

- *Multi-View Feature Sharing Benefit:* Training samples from all camera views were used in this setting. According to the MTL theory, all correlated tasks are learned together. This should consequently boost each individual task's performance. Specifically, once $\mathbf{U}$ is learned for MT-SGG-LDA and $\mathbf{C}, \mathbf{S}$ are learned for MT-FGG-LDA, the test sample x_{test} is projected into C dimensional output space through the operation $x'_{test}\mathbf{u}_t$ for MT-SGG-LDA, and through $x'_{test}(\mathbf{c}_t + \mathbf{s}_t)$ for MT-SGG-LDA using $\mathbf{u}_t = \mathbf{c}_t + \mathbf{s}_t$ corresponding to the test view. The class label of the test sample is assigned using k-nearest neighbor classification ($k = 5$ in our experiments).
- *View-Invariant Recognition Benefit:* Images corresponding to one camera view were missing in the training data, and we used the model learned with images from other views to perform prediction on the missing view. In practice, the test sample x_{test} is projected into

TABLE I
MULTI-VIEW ACTION RECOGNITION ACCURACY: COMPARING SINGLE AND MULTI-TASK LEARNING ON THE IXMAS DATASET

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg (std)
MT-SGG-LDA	0.900	0.854	0.812	0.793	0.763	0.825 (0.014)
MT-FGG-LDA	**0.912**	**0.877**	**0.821**	**0.815**	**0.791**	**0.843** (0.021)
Junejo - SVM [11]	0.748	0.745	0.748	0.706	0.612	0.727 (0.013)
ℓ_{12} MTL [18]	0.819	0.830	0.809	0.756	0.693	0.782 (0.017)

TABLE II
MULTI-VIEW ACTION RECOGNITION ACCURACY: COMPARING SINGLE AND MULTI-TASK LEARNING ON THE NIXMAS DATASET

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg (std)
MT-SGG-LDA	0.874	0.834	0.791	0.769	0.733	0.800 (0.022)
MT-FGG-LDA	**0.888**	**0.841**	**0.799**	**0.785**	**0.764**	**0.815** (0.013)
Junejo - SVM [11]	0.712	0.721	0.708	0.693	0.633	0.693 (0.025)
ℓ_{12} MTL [18]	0.803	0.794	0.768	0.763	0.672	0.760 (0.019)

a $(R-1)C$ dimensional output space through the $x'_{test}\mathbf{U}$ operation for MT-SGG-LDA and through $x'_{test}(\mathbf{C}+\mathbf{S})$ for MT-FGG-LDA since only $R-1$ views are considered in this setting. The label of the test sample is again assigned using k-nearest neighbor classification.

D. Quantitative Evaluation

A first set of experiments was devoted to demonstrate the advantage of using an MTL approach for multi-view action recognition. To this end, we compared the proposed methods with a single SVM classifier [11], and the $\ell_{2,1}$-norm multi-task learning approach [18] which assumes all the tasks to be related (no graph explicitly specifying task relationships is considered). In the SVM experiments, an RBF kernel was chosen and the LIBSVM[1] software package was used. A publicly available code[2] was used for $\ell_{2,1}$ multi-task learning.

Table 1 shows the comparison results for the IXMAS dataset. Evidently, sharing similarity information among different views using MTL is beneficial as the proposed approaches outperform SVM by at least 10%. Moreover, the fact that graph-guided MTL outperforms $\ell_{2,1}$-norm MTL confirms the benefit of modeling view similarity using a graph, as against assuming that features across all views are related. As noted earlier, the viewpoint corresponding to CAM5 is significantly different from the other four views. However, even in this case, MTLDA is greatly beneficial as action recognition accuracy improves from 69.3% to 76-79%, implying that CAM5-view features are 'enhanced' with discriminative information from the other views. Overall, these results demonstrate the benefit of *feature sharing* among different views achieved by our MTL framework.

Similar results were also obtained for the other two datasets as shown in Tables 2 and 3. Comparing the two proposed approaches, we observe similar performances with MT-SGG-LDA and MT-FGG-LDA on the RGBD ACT4^2 dataset, while superior accuracies are achieved using the latter on IXMAS and NIXMAS video data. We believe that the superiority of MT-FGG-LDA on video data is due to the greater flexibility of the model achieved with graph-based learning. However, this improvement is not observed for RGBD data, possibly because of the different features used in this case and/or due to the noisy nature of depth images. More specifically, with the leave-one-actor-out classification procedure adopted, we found that MT-FGG-LDA outperforms MT-SGG-LDA for 11 out of 12 targets on IXMAS, and for all targets on NIXMAS. In contrast, similar accuracies were observed over all targets using the two approaches on the ACT4^2 dataset. Fig. 6(a-b) also reports the performance at varying k in the k-nearest neighbor classification respectively for MT-SGG-LDA and MT-FGG-LDA.

TABLE III
MULTI-VIEW ACTION RECOGNITION ACCURACY: COMPARING SINGLE AND MULTI-TASK LEARNING ON THE ACT4^2 DATASET

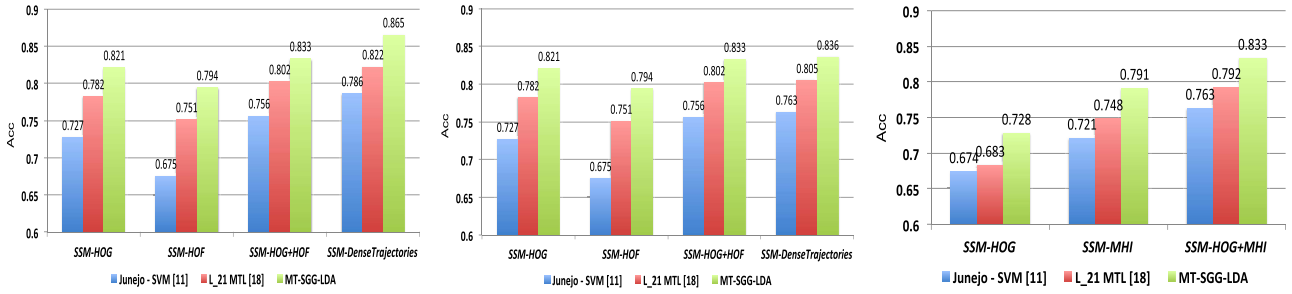
Training with All Cameras					
	Cam1	Cam2	Cam3	Cam4	Avg (std)
MT-SGG-LDA	**0.867**	0.853	**0.804**	0.808	0.833 (0.011)
MT-FGG-LDA	0.846	**0.867**	0.800	**0.821**	**0.834** (0.015)
Junejo - SVM [11]	0.799	0.787	0.743	0.721	0.763 (0.023)
ℓ_{12} MTL [18]	0.831	0.805	0.779	0.752	0.792 (0.021)

We also evaluated the effectiveness of different features– Fig. 5(a-b) show results on IXMAS and NIXMAS videos obtained using SSM descriptors computed with HOG, HOF and HOG+HOF features. As expected, the best performance on RGB video data is achieved with the combination of HOG and HOF features. For the RGBD ACT4^2 dataset, combining HOG and MHI features produces highest recognition accuracy (Fig. 5(c)), implying that having access to both color and depth information improves multi-view action recognition performance. This is in accordance with the findings in [22], where different features extending LBP descriptors to depth images are employed for action recognition.

To demonstrate that our SSM-based MTL framework is generalizable in terms of features, we also report the performance obtained with dense trajectories [52] on IXMAS and NIXMAS videos (Fig. 5(a-b)). With trajectory features, we observe a further improvement in recognition accuracy using MT-SGG-LDA on the IXMAS dataset, while the improvement is modest for NIXMAS. As clearly seen from Fig. 5, there is further scope for improving action recognition performance with our MTL framework using more sophisticated features. However, since the primary focus of this paper is to show the advantages of combining MTL with SSMs, we did not further analyze the impact of low-level features on recognition performance, and leave this investigation for future research.

Fig. 7 shows the confusion matrices obtained with MT-SGG-LDA for the multi-view feature sharing experiments on the IXMAS, NIXMAS, ACT4^2 datasets, respectively. By observing the matrices for the IXMAS and NIXMAS datasets, it is interesting to notice that for some actions such as 'get up', 'pick up' and 'punch', our method achieves very

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
[2] http://ttic.uchicago.edu/∼argyriou/code/index.html

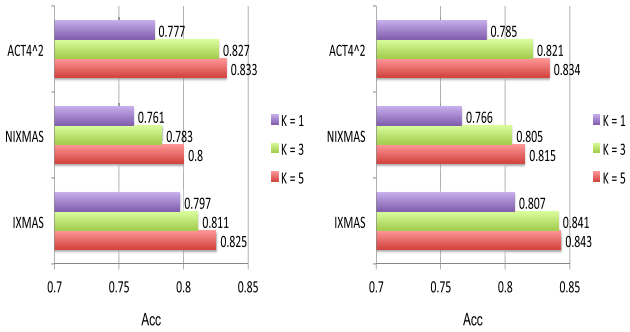


(a) (b) (c)

Fig. 5. Recognition accuracy with different SSM features on the (a) IXMAS, (b) NIXMAS and (c) ACT4^2 datasets.

TABLE IV
MULTI-VIEW ACTION RECOGNITION ACCURACY: COMPARISON OF DIFFERENT METHODS ON IXMAS DATASET

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
MT-SGG-LDA	0.900	0.854	0.812	0.793	0.763	0.825
MT-FGG-LDA	0.912	0.877	0.821	0.815	0.791	0.843
Li [14]	0.834	0.799	0.820	0.853	0.755	0.812
Liu [15]	0.790	0.747	0.752	0.764	0.712	0.753
Huang [53]	0.632	0.586	0.604	0.568	0.476	0.573
Weinland [12]	0.654	0.700	0.543	0.660	0.336	0.579
Reddy [54]	0.696	0.692	0.620	0.651	-	0.726
Farhadi [13]	-	-	-	-	-	0.581
Li [16]	0.910	0.919	0.911	0.906	0.871	0.905
Wu [55]	0.909	0.854	0.888	0.909	0.881	0.888
Mahasseni MTL (g=3) [21]	0.811	0.828	0.825	0.804	0.776	0.809
Mahasseni LMTL (g=1) [21]	0.864	0.855	0.802	0.841	0.768	0.826
Mahasseni LMTL (g=3) [21]	**0.968**	**0.956**	**0.947**	**0.965**	**0.921**	**0.951**



(a) (b)

Fig. 6. Performance at varying k for (a) MT-SGG-LDA and (b) MT-FGG-LDA

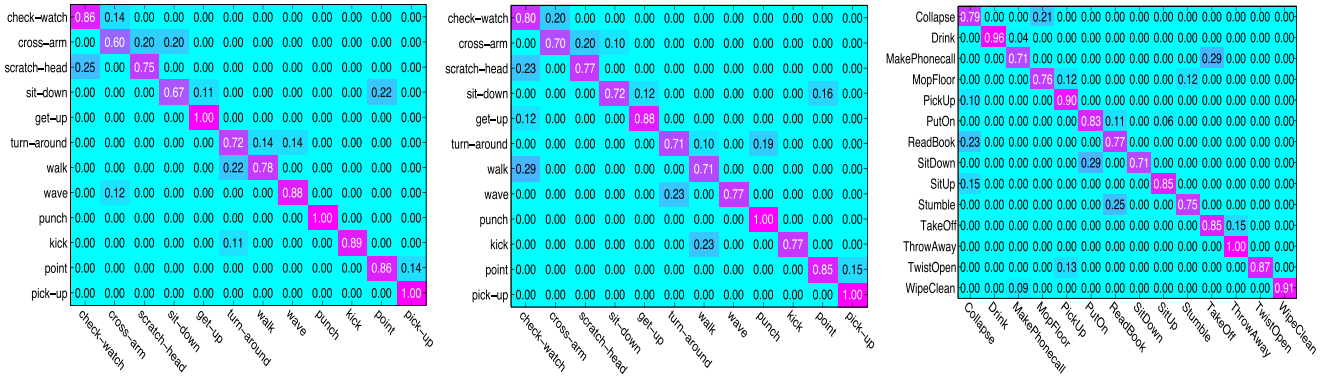
high recognition accuracies. Even for some challenging actions (*e.g.*, 'point', 'check watch' and 'wave') having small and ambiguous motions, our method still produces reasonable and promising results.

We also compared the proposed methods with other action recognition algorithms which are not based on SSMs. The results of such comparison on the IXMAS dataset are shown in Table 4. Our approach achieves higher recognition performance, in terms of both single-view and (average) multi-view accuracies, as compared to most previous methods. While the approaches proposed in [16], [21], and [55] achieve higher recognition as compared to our methods, they suffer from other limitations. The algorithm in [55] is based on latent kernelized structural SVM which is intractable for inference on large-scale datasets. The feature extraction phase of the algorithm in [16] is computationally demanding. Differently, our method is computationally efficient and also easy to implement. It is particularly interesting to compare our approach with [21], as this is also based on multi-task learning. The best performance reported in [21] is achieved by combining MTL with a feature representation which takes into account the layout of body parts. This representation is very powerful and generally superior to the bag-of-words approach, which we adopt in this paper. However, if the parts-based representation is not used, the accuracy of [21] degrades (from 0.951 drops to 0.809 as stated in their work) to a value lower than ours, thus demonstrating the effectiveness of our proposed MT-LDA approach.

TABLE V
MULTI-VIEW ACTION RECOGNITION ACCURACY: COMPARISON OF DIFFERENT METHODS ON NIXMAS DATASET

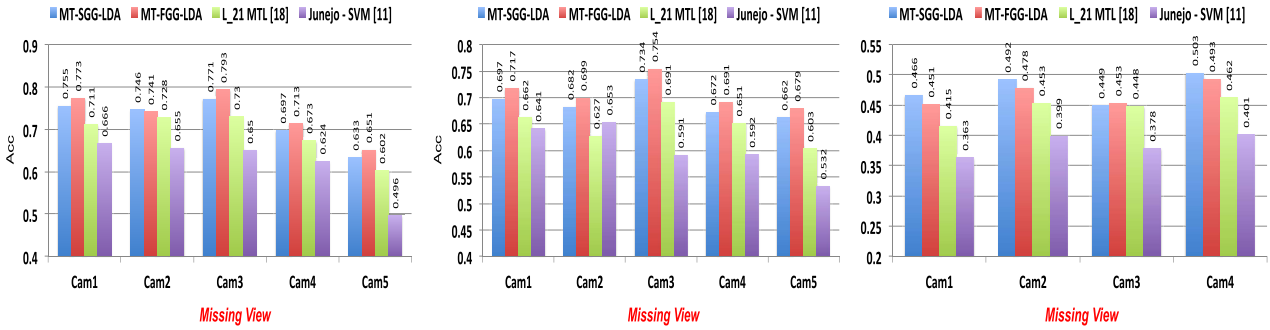
Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
MT-SGG-LDA	0.874	0.834	0.791	0.769	0.733	0.800
MT-FGG-LDA	0.888	0.841	0.799	0.785	0.764	0.815
Weinland [3]	-	-	-	-	-	0.767
Mahasseni MTL (g=3) [21]	0.688	0.701	0.665	0.706	0.644	0.688
Mahasseni LMTL (g=1) [21]	0.782	0.816	0.807	0.776	0.761	0.788
Mahasseni LMTL (g=3) [21]	**0.902**	**0.914**	**0.887**	**0.881**	**0.844**	**0.886**

More generally, an accurate comparison between our method and [21] would require careful analysis of the impact of different features. However, we believe that both our work and [21] clearly demonstrate the advantages of multi-task learning over single-task methods for multi-view action recognition. Moreover, as confirmed by Tables 1-3 and the last two rows of Table 4, assuming that all views are related (*i.e.* $g = 1$ as opposite to $g = 3$ in [21]) is not optimal, and more flexible algorithms which model more accurately the relations among multiple views are required. This further confirms the importance of the graph structure employed in this work. Finally, we should point out that [21] performs very poorly in the missing view setting (see Fig.4 in [21]) while our approach naturally extends to this setting as discussed below. Similar observations can be made observing the results in Table 5, which presents a comparison of the different methods on the NIXMAS dataset. Corresponding results are not reported for the ACT4^2 dataset as the set-up used for the



(a) (b) (c)

Fig. 7. Confusion matrices (MT-SGG-LDA) on the (a) IXMAS, (b) NIXMAS and (c) ACT4^2 datasets.



(a) (b) (c)

Fig. 8. Cross-view action recognition accuracy: training is performed with one view missing on (a) IXMAS, (b) NIXMAS, (c) ACT4^2 datasets.

experiments in [22] cannot be exactly reproduced (the videos of an unspecified subset of actors are used for evaluation in [22]).

Finally, to demonstrate the benefits of our approach on *view-invariant* action recognition, we evaluated its performance when one view was missing in the training data. Results on the IXMAS, NIXMAS and ACT4^2 datasets are shown in Fig.8(a), (b) and (c) respectively. Although there is some performance drop compared to the situation where all camera views are available in the training phase, our approach still achieves better performance than the single-task SVM and $\ell_{2,1}$ multi-task learning methods. The recognition accuracies of both our approaches are similar, with MT-SGG-LDA outperforming MT-FGG-LDA on the ACT4^2 dataset. This may be due to the importance of sparsity when the feature dimensionality increases.

V. Conclusions

In this paper, we propose a multi-task extension of multi-class LDA to effectively improve SSM-based multi-view action recognition. In particular, we propose two variants of graph-guided multi-task LDA: (i) where the graph weights specifying view dependencies are fixed *a priori* and (ii) where graph weights are flexibly learnt from the training data. The intuition is that multi-task learning can share view-invariant SSM features across different views for better multi-view action recognition. Extensive experiments on the IXMAS, NIXMAS and ACT4^2 datasets demonstrate the superior performance of our method compared to other SSM-based state-of-the-art methods.

Overall, the proposed multi-task LDA solutions are novel in the context of view-invariant action recognition, which is a relevant and important research problem in applications such as human behaviour understanding and surveillance. Possible future works include (i) the integration of other view-invariant features in combination with SSM descriptors; (ii) the investigation of a different strategy for graph construction based on camera geometry information; and (iii) the use of deep structures instead of a shallow, single-layer model for the considered problem, since deep learning has achieved considerable success in solving many computer vision and image processing problems.

References

[1] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224–241, 2011.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[3] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 635–648.

[4] T. Mahmood, A. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," in *Proc. IEEE Workshop Detection Recognit. Events Video*, Jul. 2001, pp. 64–72.

[5] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 984–989.

[6] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2479–2494, Jun. 2013.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[8] A. A. Efros, A. C. Berg, E. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 726–733.

[9] M. Grundmann, F. Meier, and I. Essa, "3D shape context and distance transform for action recognition," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[10] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 444–451.

[11] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[12] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.

[13] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 154–166.

[14] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2855–2862.

[15] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3209–3216.

[16] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using Hankelets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1362–1369.

[17] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.

[18] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Neural Inf. Process. Syst.*, 2007, pp. 41–48.

[19] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. Conf. Knowl. Discovery Data Mining (SIGKDD)*, 2004, pp. 109–117.

[20] J. Ye, "Least squares linear discriminant analysis," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1087–1093.

[21] B. Mahasseni and S. Todorovic, "Latent multitask learning for view-invariant action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3128–3135.

[22] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *Proc. ECCV Workshop Consum. Depth Cameras Comput. Vis.*, 2012, pp. 52–61.

[23] P. Yan, S. M. Khan, and M. Shah, "Learning 4D action feature models for arbitrary view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[24] V. Parameswaran and R. Chellappa, "Human action-recognition using mutual invariants," *Comput. Vis. Image Understand.*, vol. 98, no. 2, pp. 294–324, 2005.

[25] A. Farhadi, M. K. Tabrizi, I. Endres, and D. A. Forsyth, "A latent model of discriminative aspect," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 948–955.

[26] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3176–3183.

[27] P. N. Belhumeour, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[28] D. Huang, R. S. Cabral, and F. De la Torre, "Robust regression," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 616–630.

[29] Y. Han, F. Wu, J. Jia, Y. Zhuang, and B. Yu, "Multi-task sparse discriminant analysis (MtSDA) with overlapping categories," in *Proc. AAAI Conf. Artif. Intell.*, 2010, pp. 469–474.

[30] Y. Zhang and D.-Y. Yeung, "Multi-task learning in heterogeneous feature spaces," in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 574–579.

[31] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing, "Smoothing proximal gradient method for general structured sparse learning," in *Proc. Uncertainity Artif. Intell.*, 2011, pp. 105–114.

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

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

[34] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 702–710.

[35] L. W. Zhong and J. T. Kwok, "Convex multitask learning with flexible task clusters," in *Proc. Int. Conf. Mach. Learn.*, pp. 49–56, 2012.

[36] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 523–536, Feb. 2013.

[37] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 367–383, 2013.

[38] Y. Yan, E. Ricci, G. Liu, and N. Sebe, "Recognizing daily activities from first-person videos with multi-task clustering," in *Proc. Asian Conf. Comput. Vis.*, 2014.

[39] G. Lu, Y. Yan, N. Sebe, and C. Kambhamettu, "Knowing where I am: Exploiting multi-task learning for multi-view indoor image-based localization," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[40] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe, "No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1177–1184.

[41] A. K. Rajagopal *et al.*, "An adaptation framework for head-pose classification in dynamic multi-view scenario," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 652–666.

[42] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe, "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 3–10.

[43] Q. Zhou, G. Wang, K. Jia, and Q. Zhao, "Learning to share latent tasks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2264–2271.

[44] Y. Yan, G. Liu, E. Ricci, and N. Sebe, "Multi-task linear discriminant analysis for multi-view action recognition," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 2842–2846.

[45] K. Fukunaga, *Introduction to Statistical Pattern Classification*. New York, NY, USA: Academic, 1990.

[46] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.

[47] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," *Math. Intell.*, vol. 27, no. 2, pp. 83–85, 2005.

[48] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[49] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[50] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

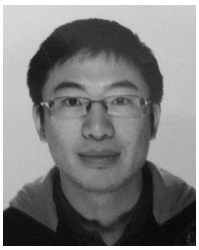
[51] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Proc. IEEE ICCV Workshops*, Nov. 2011, pp. 1147–1153.

[52] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3169–3176.

[53] C.-H. Huang, Y.-R. Yeh, and Y.-C. F. Wang, "Recognizing actions across cameras by exploring the correlated subspace," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 342–351.

[54] K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1010–1017.

[55] X. Wu and Y. Jia, "View-invariant action recognition using latent kernelized structural SVM," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 411–424.



Yan Yan received the Ph.D. degree from the University of Trento, Trento, Italy, in 2014, where he is currently a Post-Doctoral Researcher with the Multimedia and Human Understanding Group. His research interests include machine learning and its application to computer vision and multimedia analysis.

Elisa Ricci is currently an Assistant Professor with the University of Perugia, Perugia, Italy, and a Researcher with Fondazione Bruno Kessler, Trento, Italy. She received the Ph.D. degree from the University of Perugia, in 2008. She was a Visiting Student with the University of Bristol, Bristol, U.K. She has been a Post-Doctoral Researcher with Idiap Research Institute, Martigny, Switzerland, and Fondazione Bruno Kessler. Her research interests are mainly in the areas of computer vision and machine learning.

Ramanathan Subramanian received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2008. He is currently a Research Scientist with the Advanced Digital Sciences Center, Singapore. His research interests span human-centered computing, human behavior understanding, computer vision, and multimedia processing.

Gaowen Liu received the B.S. degree in automation from Qingdao Univerisity, Qingdao, China, in 2006, and the M.S. degree in system engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2008. She is currently pursuing the Ph.D. degree with the Multimedia and Human Understanding Group, University of Trento, Trento, Italy. Her research interests include machine learning and its application to computer vision and multimedia analysis.

Nicu Sebe (M'01–SM'11) received the Ph.D. degree from Leiden University, Leiden, The Netherlands, in 2001. He is currently with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human behavior understanding. He was the General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference in 2008 and the ACM Multimedia Conference in 2013, and the Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010, and the ACM Multimedia Conference in 2007 and 2011. He will be the Program Chair of the European Conference on Computer Vision in 2016 and the International Conference on Computer Vision in 2017. He is a Senior Member of the Association for Computing Machinery and a fellow of the International Association for Pattern Recognition.