Clustered Multi-task Linear Discriminant Analysis for View Invariant Color-Depth Action Recognition

Yan Yan[1], Elisa Ricci[2,3], Gaowen Liu[1], Ramanathan Subramanian[4], Nicu Sebe[1]
[1]University of Trento, Italy [2]University of Perugia, Italy
[3]Fondazione Bruno Kessler, Trento, Italy [4]Advanced Digital Sciences Center (ADSC), Singapore

***Abstract*—The widespread adoption of low-cost depth cameras has opened new opportunities to improve traditional action recognition systems. In this paper we focus on the specific problem of action recognition under view point changes and propose a novel approach for view-invariant action recognition operating jointly on visual data of color and depth camera channels. Our method is based on the unique combination of robust Self-Similarity Matrix (SSM) descriptors and multi-task learning. Indeed, multi-view action recognition is inherently a multi-task learning problem: images from a camera view can be modeled as visual data associated to the same task and it is reasonable to assume that the data of different tasks (camera views) are related to each other. In this work we propose a novel algorithm extending Multi-Task Linear Discriminant Analysis (MT-LDA) to enhance its flexibility by learning the dependencies between different views. Extensive experimental results on the publicly available ACT4^2 dataset demonstrate the effectiveness of the proposed method.**

I. Introduction

The problem of recognizing and understanding human actions in images and videos is probably one of the most important and challenging tasks which computer vision researchers are currently facing. Indeed, action recognition is fundamental in many applications ranging from robotics, human-computer interfaces, human behavior understanding, content-based video indexing, video surveillance, and ambient-assisted living.

There is a vast literature on visual-based action recognition systems (see *e.g.* [1, 2] for surveys on the topic). While the large majority of the methods are based on traditional cameras as sensors, recently the widespread adoption of low cost depth cameras has shifted the interest of research toward developing solutions specifically targeted to them.

In this paper we consider the problem of multi-view action recognition from RGB-D cameras. Having at disposal multiple sensors and images from both color and depth camera channels, issues related to self occlusions are greatly alleviated with respect to a single view RGB setting and in general improved recognition results can be obtained. However, how to address the effect of viewpoint changes on the recognition of human actions is still under investigation. While many works have considered this problem in the context of traditional cameras [3–6], very few approaches have been designed to operate specifically on a RGB-D setup [7].

Extracting view-invariant descriptors is a possible strategy for action recognition in the multi-view setting. Following this idea, some recent approaches have been developed: some are based on transferring information across views [4], other on computing view-invariant features [8]. In particular in [8] descriptors calculated from temporal Self-Similarity Matrices

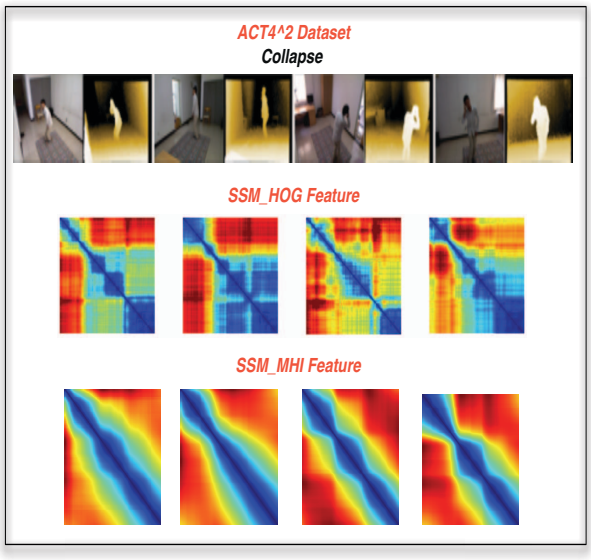

Fig. 1. Examples of SSMs computed from Histogram of Oriented Gradients (HOG) and Motion History Image (MHI) features on the ACT4^2 dataset. (best viewed in color)

are proposed. Temporal SSMs can be computed from different low-level features extracted on a frame basis (*e.g.* Histogram of Oriented Gradients, Histograms of Optical Flows, Motion History Images) and have been shown to be particularly robust to point of view changes. However, a careful analysis of SSMs reveals that, especially when the appearance changes considerably among different views, SSMs are similar only up to a certain extent. This effect can be observed in Fig.1, where SSMs computed for four sequences of the ACT4^2 dataset [7] are shown. It is worth noting that SSMs extracted from HOG descriptors and color frames are less stable with respect to those obtained computing Motion Histogram Images (MHIs) from depth images. In this paper we propose a novel approach for multi-view action recognition where SSM descriptors are used within a multi-task learning framework.

Multi-task learning [9] aims to simultaneously learn a classification model for a set of related tasks. The intuition is that a more accurate classifier can be obtained when taking into account task relationships. In this paper, we consider each camera view as a task and investigate how to share features across different views in order to boost the recognition performance. We present a novel multi-task learning framework which enhances the discriminative power of SSM descriptors by individuating view invariant information and by separating

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.601

it from view specific features. Inspired by previous works on Multi-task Linear Discriminant Analysis we propose a novel approach for MT-LDA [10] where the task relationships are not fixed *a priori* but are also learned during the training phase. In fact, while the relatedness of different tasks can be modeled by defining a similarity graph which reflects information about camera geometry or generally features similarity, refining such dependency graph during the learning process increases classifier's flexibility and generally improves its performance.

Our experiments show that sharing features among views is beneficial for multi-view and view invariant action recognition. On the ACT4^2 dataset, our approach achieves a recognition accuracy 10% higher than previous works based on SSMs descriptors.

A. Contributions

This paper is one of the first works to cope with the problem of multi-view action recognition when both color and depth camera channels are considered. Moreover, up to our knowledge, no previous works have addressed this problem within a multi-task learning framework. Our approach is inspired by the work in [10]. However, the proposed MT-LDA algorithm is novel since in [10] a fixed graph modeling tasks' dependencies is employed. While applied to the problem of action recognition, our algorithm is rather general and can be used in other applications, such as image annotation, pose estimation, *etc*.

II. Related Work

A. View Invariant Action Recognition

Understanding human daily activities is a very challenging problem. One of the main difficulties of action recognition is due to the viewpoint variations which are common in real conditions and create significant intra-class variability. Many previous works have tried to address this issue [8, 11, 12]. In [8] robust descriptors based on SSMs and on the traditional bag-of-words model are introduced. Rao *et al.* [12] presented a view-invariant representation of human action to capture the dramatic changes in the speed and direction of the trajectory using spatio-temporal curvature of 2D trajectory. In [11] the view invariant "Hankelet" descriptors are proposed being features which capture the dynamic properties of short tracklets. Other works [4, 13, 14] use transfer learning algorithms to learn a view invariant representation.

While addressing the view invariance issue is challenging, this is not the only problem when building an action recognition system. Additional difficulties arise due to illumination variations or cluttered scenes. While traditional approaches [1, 4, 11, 13] are based on the sole data recorded from color cameras, the advent of low cost RGB-D sensors have brought the opportunity to cope with these problems much more effectively. Only few works have considered the problem of action recognition combining both color and depth information [7, 15]. In [15] the problem of activity of daily living analysis from RGB-D data is addressed. Two multi-modality fusion schemes are proposed, developed from state-of-the-art features representation methods. However, a single camera setup is considered, thus no specific solutions are developed to cope with issues due to camera views variations. In [7] a database for multi-view multi-modal action recognition is made publicly available. Moreover, an approach based on combining color and depth descriptors is proposed. However, this method is not targeted to address the cross-view action recognition problem. Differently, in this paper we specifically aim to alleviate viewpoint variations by combining SSMs descriptors and a MTL approach.

B. Multi-task Learning

In the last few years multi-task learning approaches have become popular in the computer vision community and have been successfully applied to many problems such as image classification [16], image annotation [17] and head pose classification [18]. Multi-task learning methods develop from the intuition that, when taking multiple classification/regression problems associated to related tasks, learning a model which considers a shared component together with task-specific representations is convenient with respect to learning on each single task separately. In practice in many real world problems MTL typically leads to improved performance as compared to single task approaches.

Traditional MTL methods consider a single shared representation, assuming that all the tasks are related [9]. However, when some of the tasks are independent, this may lead to worse performance than single-task learning. Recently, more sophisticated approaches have been proposed to address this problem [18–20]. Multi-task extensions of Linear Discriminant Analysis have been introduced in [10, 21, 22]. However, the framework proposed in [21] is not flexible as no learning on the relationship between tasks is conducted. In [22] heterogeneous feature spaces among different tasks are considered, a scenario that is not appropriate in our application. In [10] MT-LDA is proposed for multi-view action recognition assuming that the camera view dependencies are known *a-priori* and specified in form of a graph. Differently, in this paper we propose to learn the task relationships simultaneously with task-specific parameters. Moreover while in [10] only traditional cameras are considered, in this paper we also employ depth information. Multi-task learning for multi-view action recognition is also proposed in [23] but the authors did not use RGB-D data.

III. Clustered Multi-task Linear Discriminant Analysis for View Invariant Color-Depth Action Recognition

In this section, we first present an overview of the proposed framework. Then we introduce the considered color and depth self-similarity matrix descriptors for RGB-D action recognition. Finally, we discuss our clustered multi-task linear discriminant analysis in detail.

A. Overview

In this section we describe the proposed approach. We first compute SSMs descriptors separately for color and depth images. Specifically we used different low level features, *e.g.* HOG for describing RGB data and MHIs and some of its variations for depth frames. Then, we adopt the standard bag-of-words paradigm to compute histograms for each video and learn a classification model for each camera view using the proposed MT-LDA algorithm. In the following we present our approach in detail.

B. Self-Similarity Matrix Descriptors

Given an image sequence $\mathcal{I} = \{I_1, I_2, ..., I_T\}$, an SSM is a symmetric matrix $E \in R^{T\times T}$, $E_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|^2$ is the Euclidean distance between low-level features $\mathbf{f}_i, \mathbf{f}_j \in R^d$ extracted from frames I_i, I_j.

In this paper, we use HOG descriptors [24] as low-level features computed on RGB video frames, while MHI [25] in the form of forward and backward MHIs [15] are adopted for depth images. Denoting the depth value corresponding to a pixel at location x, y, and at time t as $D(x, y, t)$, MHI is computed as:

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

where τ is the longest time window that the system considers (τ is equal to the number of frames in our experiments), and δD_{th} is the threshold for mask generation in the motion region. In order to exploit the depth information better, we also consider forward-MHIs H^{fD}_τ which encodes positive depth gradient and backward-MHIs H^{bD}_τ which models negative depth gradient [15], *i.e.*:

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

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

Once SSMs are computed for HOG, MHI, forward-DMHI and backward-DMHI features, the same strategy as described in [8] is adopted for calculating local descriptors. For each point on the unimodal SSM diagonal, three local descriptors are computed corresponding to different diameters in the log-polar domain (diameter of 28, 42 and 56 frames respectively). The bag-of-words model is then employed to obtain the final histogram representation for a video clip. Also, a codebook of 500 words is used in our experiments.

In Fig.2, we show an example of MHI, forward-MHI and backward-MHI features extracted for ACT4^2 dataset and the corresponding SSM descriptors. It is worth noting that SSMs are rather stable over different persons performing the same action under different viewpoints. However, this invariance is valid only to a certain extent. Therefore, in order to individuate common features among different views, clustered multi-task LDA is introduced as follows.

C. Clustered Multi-task Linear Discriminant Analysis

We consider a set of R related multi-class classification problems. We are given a training set $\mathcal{T}_t = \{(x^t_i, \ell^t_i)\}_{i=1}^{N_t}$ for each task $t = 1, 2, \ldots, R$, where $x^t_i \in I\!R^d$ is d-dimensional feature vector, $\ell^t_i \in \{1, 2, \ldots, C\}$ indicates the class membership. We denote with $N_{t,j}$ the sample size of j-th class in t-th task, $N_t = \sum_{j=1}^{C} N_{t,j}$ the total training samples in t-th tasks and $N = \sum_{t=1}^{R} N_t$. We define $\mathbf{x}_t \in I\!R^{N_t\times d}$, $\mathbf{x}_t = [x^t_1, ..., x^t_{N_t}]'$

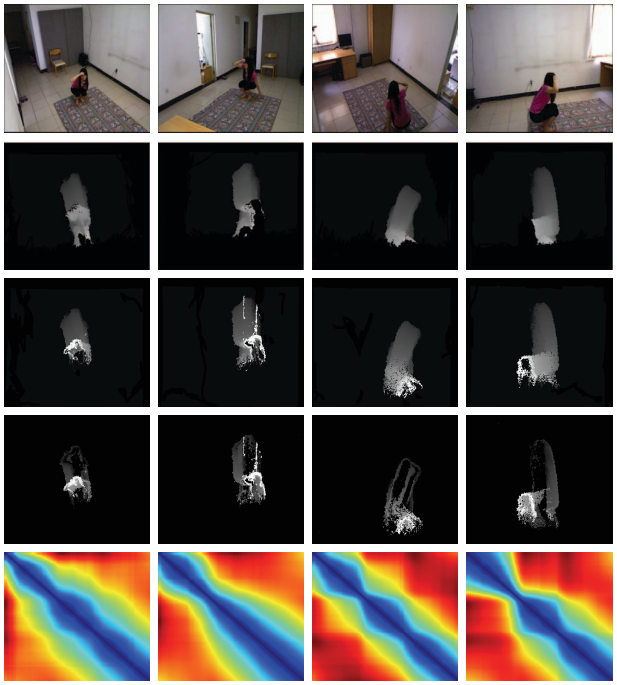

Fig. 2. ACT4^2 dataset and different types of features extracted. From top to bottom original frames, MHI, fMHI, bMHI, SSMs. (best viewed in color)

and the class indicator matrix $\mathbf{y}_t \in I\!R^{N_t\times C}$, $\mathbf{y}_t = [\ell^t_1, ..., \ell^t_{N_t}]'$ where:

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

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. In this paper we propose to learn a global weight matrix $\mathbf{W} = [\mathbf{w}'_1, \ldots, \mathbf{w}'_R]', \mathbf{W} \in I\!R^{d\times CR}$, $\mathbf{W} = \mathbf{P} + \mathbf{Q}$ by solving the following optimization problem:

$$\min_{\mathbf{P},\mathbf{Q}} \|\mathbf{Y} - \mathbf{X}(\mathbf{P} + \mathbf{Q})\|^2_F + \lambda\Omega(\mathbf{P}, \mathbf{Q}) \qquad (1)$$

The weight matrix $\mathbf{W}$ is obtained summing two terms, the matrix $\mathbf{P} = [\mathbf{p}'_1, \ldots, \mathbf{p}'_R]'$ modeling common features among tasks and the matrix $\mathbf{Q} = [\mathbf{q}'_1, \ldots, \mathbf{q}'_R]'$ which takes into account task specific features. The regularization term $\Omega(\cdot)$ is defined as:

$$\Omega(\mathbf{P}, \mathbf{Q}) = \ \ \|\mathbf{P}\|^2_F + \|\mathbf{Q}\|^2_F + \lambda_c \|\mathbf{G}\mathbf{P}'\|_1$$

In the regularization term, $\|\mathbf{P}\|^2_F$ regulates model complexity while $\|\mathbf{Q}\|^2_F$ penalizes large deviation of the common model $\mathbf{P}$ from the global model $\mathbf{W}$. The L_1 norm regularizer imposes the weights $\mathbf{p}_t$ of related tasks to be close together. The relatedness of the tasks is modeled specifying a matrix $\mathbf{G} \in I\!R^{|\mathcal{E}|\times CR}$:

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

Algorithm 1 Clustered Multi-task LDA

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

Initialize $\mathbf{P}_0$, $\mathbf{Q}_0$, $\alpha_0 = 1$.
LOOP:
$\alpha_n = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{n-1}^2})$

Updating $\mathbf{P}$:
$\hat{\mathbf{P}} = \mathbf{P}_n - 2\mathbf{X}^T(\mathbf{X}\mathbf{P}_n - \mathbf{Y})$
FOR $i = 1 : d$
Initialize $\mathbf{s}^{i,0}$, $\mathbf{z}^{i,0}$, $\mathbf{p}^{i,0}$
Compute Cholesky factorization of matrix $\mathbf{A}$.
LOOP:
Solve $\mathbf{A}\mathbf{p}^{i,k+1} = \mathbf{b}^k$
$\mathbf{s}^{i,k+1} = \Sigma_{\hat{\lambda}_1/\rho}(\mathbf{G}\mathbf{p}^{i,k+1} + \frac{1}{\rho}\mathbf{z}^{i,k})$
$\mathbf{z}^{i,k+1} = \mathbf{z}^{i,k} + \rho(\mathbf{G}\mathbf{p}^{i,k+1} - \mathbf{s}^{i,k+1})$
Until Convergence
END FOR
$\mathbf{P}_{n+1} = (1 + \frac{\alpha_{n-1}-1}{\alpha_n})\mathbf{P}_{n+\frac{1}{2}} - \frac{\alpha_{n-1}-1}{\alpha_n}\mathbf{P}_n$

Updating $\mathbf{Q}$:
$\hat{\mathbf{Q}} = \mathbf{Q}_n - 2\mathbf{X}^T(\mathbf{X}\mathbf{Q}_n - \mathbf{Y})$
$\mathbf{Q}_{n+\frac{1}{2}} = \frac{1}{1+\hat{\lambda}_\theta}\hat{\mathbf{Q}}$
$\mathbf{Q}_{n+1} = (1 + \frac{\alpha_{n-1}-1}{\alpha_n})\mathbf{Q}_{n+\frac{1}{2}} - \frac{\alpha_{n-1}-1}{\alpha_n}\mathbf{Q}_n$
Until Convergence
Output: $\mathbf{W} = \mathbf{P} + \mathbf{Q}$

Here, $\gamma_{ij} = (\sum_{i \neq j} \|SSM_i - SSM_j\|_2)^{-1}$, *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. It is worth noting that while the matrix $\mathbf{G}$ specifies some *a priori* knowledge about tasks'relatedness, the dependencies among tasks are learned by computing the matrix $\mathbf{Q}$, *i.e.* tasks with the same $\mathbf{q}_t$ are dependent.

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [26] is adopted to solve (1). Considering the function $\Pi(\mathbf{P}, \mathbf{Q}) = \|\mathbf{Y} - \mathbf{X}(\mathbf{P} + \mathbf{Q})\|_F^2$ which is convex and smooth and $\Theta(\mathbf{P}, \mathbf{Q}) = \lambda\|\mathbf{P}\|_F^2 + \lambda\|\mathbf{Q}\|_F^2 + \lambda\lambda_c \|\mathbf{G}\mathbf{P}'\|_1$ which is convex non smooth. FISTA solves the optimization problems in the form $\min_{\mathbf{U}} \Pi(\mathbf{U}) + \Theta(\mathbf{U})$ computing at each iteration a proximal step:

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

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 determined by line search. To solve the proximal step, the soft-thresholding operator $\Sigma_\lambda(x) = \text{sign}(x)\max(|x| - \lambda, 0)$ is adopted [27].

The proximal step in terms of $\mathbf{P}$, $\mathbf{Q}$ amounts into solving the following:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{Q}} \quad & \left\|\mathbf{P} - \hat{\mathbf{P}}\right\|_F^2 + \left\|\mathbf{Q} - \hat{\mathbf{Q}}\right\|_F^2 \\ & + \hat{\lambda}_c \|\mathbf{G}\mathbf{P}'\|_1 + \hat{\lambda}\|\mathbf{P}\|_F^2 + \hat{\lambda}\|\mathbf{Q}\|_F^2\end{aligned} \tag{2}$$

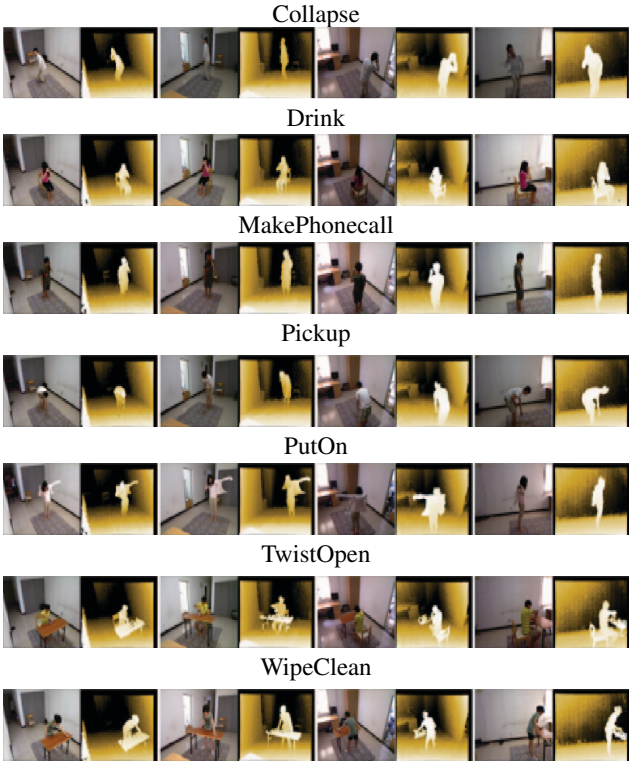

Fig. 3. Examples of different actions from the ACT4^2 dataset.

where $\hat{\mathbf{P}} = \mathbf{P} - 2\mathbf{X}^T(\mathbf{X}\mathbf{P} - \mathbf{Y})$ and $\hat{\mathbf{Q}} = \mathbf{Q} - 2\mathbf{X}^T(\mathbf{X}\mathbf{Q} - \mathbf{Y})$, $\hat{\lambda} = 2\lambda/L_k$ and $\hat{\lambda}_c = 2\lambda\lambda_c/L_k$. To solve (2) we consider $\mathbf{P}$ and $\mathbf{Q}$ separately. Solving (2) with respect to $\mathbf{Q}$ is straightforward. Solving (2) with respect to $\mathbf{P}$ is difficult because of the non-smooth L_1-norm term. We propose to solve d separate optimization problems, one for each row $\mathbf{p}^i$ of the matrix $\mathbf{P}$:

$$\min_{\mathbf{p}^i} \left\|\mathbf{p}^i - \hat{\mathbf{p}}^i\right\|_2^2 + \hat{\lambda}_1 \left\|\mathbf{G}\mathbf{p}^i\right\|_1 + \hat{\lambda}_c \left\|\mathbf{p}^i\right\|_2^2$$

and consider the equivalent constrained optimization problem (in the following the superscripts are removed for sake of clarity):

$$\begin{aligned}\min_{\mathbf{p},\mathbf{s}} \quad & \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2 + \hat{\lambda}_1 \|\mathbf{s}\|_1 + \hat{\lambda}_c \|\mathbf{p}\|_2^2 \\ \text{s.t.} \quad & \mathbf{G}\mathbf{p} - \mathbf{s} = 0\end{aligned} \tag{3}$$

The augmented lagrangian multipliers approach [27] is applied to solve the problem. The associated Lagrangian is:

$$\begin{aligned}L_\rho(\mathbf{p}, \mathbf{s}, \mathbf{z}) = & \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2 + \hat{\lambda}_1 \|\mathbf{s}\|_1 + \hat{\lambda}_c \|\mathbf{p}\|_2^2 \\ & + \mathbf{z}^T(\mathbf{G}\mathbf{p} - \mathbf{s}) + \frac{\rho}{2}\|\mathbf{G}\mathbf{p} - \mathbf{s}\|_2^2\end{aligned} \tag{4}$$

where $\mathbf{z}$ is the vector of augmented Lagrangian multipliers and ρ is the dual update step length. Three steps are alternated corresponding to solving (4) with respect to the three variables $\mathbf{p}$, $\mathbf{s}$ and $\mathbf{z}$. Solving (4) with respect to $\mathbf{s}$ has a closed form solution obtained by applying the soft-thresholding operator. The update step corresponding to solving with respect to $\mathbf{z}$ is straightforward. Solving with respect to $\mathbf{p}$ implies solving a linear system $\mathbf{A}\mathbf{p}^{k+1} = \mathbf{b}^k$ where $\mathbf{A} = \rho\mathbf{G}^T\mathbf{G} + (2 + 2\hat{\lambda}_c)\mathbf{I}$

and $\mathbf{b}^k = \rho \mathbf{G}^T \mathbf{s}^k - \mathbf{G}^T \mathbf{z}^k + 2\hat{\mathbf{p}}$. In this paper, Cholesky factorization is used to decompose $\mathbf{A}$ and solve the linear system efficiently. The resulting algorithm is shown in Algorithm 1.

IV. Experimental Results

In this section, we conduct extensive experiments to evaluate the performance of our proposed method. We also compare our method with other state-of-the-art methods.

A. Experimental Setup

In this paper the ACT4^2 dataset [7] is used for experimental evaluation. This is a very recent dataset which contains video sequences depicting 14 representative daily actions recorded through both RGB and depth channels simultaneously. Every action is recorded from four cameras. The daily actions considered 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*. Figure 3 shows some examples of different actions from ACT4^2 dataset. To evaluate the proposed approach we adopt the well known leave-one user-out strategy: videos of one actor are selected for testing while videos of the remaining people are used as training data. The optimal values of the regularization parameters λ and λ_c are determined using a separate validation set. We perform two series of experiments, to evaluate the benefit of our approach in the context of *Multi-view* and *View-invariant* action recognition.

B. Multi-view Action Recognition Results

In this series of experiments, all training samples from all camera views are used. According to multi-task learning theory, all related tasks are learned together in order to increase each individual task's performance. Specifically, once $\mathbf{P}, \mathbf{Q}$ are learned with our learning framework, for experiments in this series, the test sample x_{test} is projected into C dimensional output space by $x'_{test}(\mathbf{p}_t + \mathbf{q}_t)$ using $\mathbf{p}_t + \mathbf{q}_t$ according to the specific view t where a test sample belongs. The class label of the test sample is assigned using a k-nearest neighbor classifier.

Figure 4 shows the results of the comparison between our method and two baselines: a SVM classifier operating on the same features (*i.e.* a single task learning scenario) and the $\ell_{2,1}$-norm multi-task learning approach in [9] (assuming all the tasks related to each other). It is evident how sharing similarity information among different views using multi-task learning outperforms SVM by at least 5%. Moreover, it is clear that using a graph specifying some a-priori knowledge about the degree of similarity of different views is better than adopting a $\ell_{2,1}$-norm multi-task learning approach.

Figure 5 shows the confusion matrix on the ACT4^2 dataset. It is interesting to observe that for some actions such as *drink*, *thowaway* and *wipeclean*, our method achieves very high recognition accuracies. Even for some challenging actions (*e.g.*, *twistopen* and *takeoff*) having small and ambiguous motions, our method still guarantees quite accurate recognition.

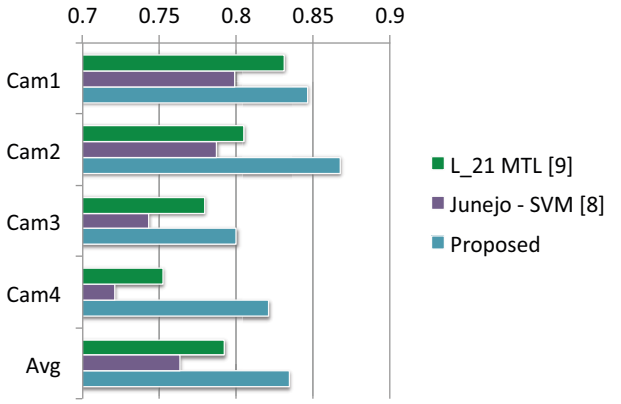

Fig. 4. Multi-view action recognition accuracy: comparison with baselines on the ACT4^2 dataset.

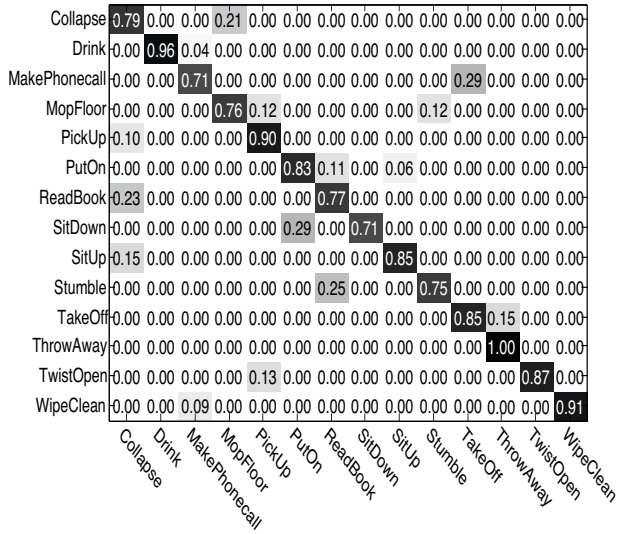

Fig. 5. Confusion matrix on the ACT4^2 dataset.

C. View-invariant Action Recognition Results

In this series of experiments, one camera view is missing in the training data and we use the model learned with data form the other views to perform prediction on the missing view. Specifically, once $\mathbf{P}, \mathbf{Q}$ are learned with our learning framework, for experiments in this series, the test sample x_{test} is projected into $(R-1)C$ dimensional output space by $x'_{test}(\mathbf{P} + \mathbf{Q})$ since only $R-1$ tasks are considered in this setting. The class label of the test sample is again assigned using a k-nearest neighbor classifier. The results are shown in Table I. Although there is some performance drop compared to the situation where all camera views are available at the training phase, our approach still achieves the best performance compared to a single task SVM and to $\ell_{2,1}$-norm multi-task learning.

TABLE I. CROSS-VIEW ACTION RECOGNITION ACCURACY ON ACT4^2 DATASET WHEN TRAINING IS PERFORMED WITH ONE VIEW MISSING.

	Missing View			
	Cam1	Cam2	Cam3	Cam4
Proposed	0.451	0.478	0.453	0.493
Junejo - SVM [8]	0.363	0.399	0.378	0.401
ℓ_{12} MTL [9]	0.415	0.453	0.448	0.462

V. CONCLUSIONS

In this paper, we proposed clustered multi-task LDA for view-invarinat human action recognition in the color-depth camera setup. Experimental results on the ACT4^2 datasets demonstrate the superior performance of our method compared to other SSM-based state-of-the-art methods. The proposed multi-task LDA algorithm is general and can be used in other applications, such as image annotation, pose estimation, *etc*. Future works include the integration of other features in combination with SSM descriptors and the investigation of a different strategy for graph construction.

VI. ACKNOWLEDGEMENTS

This work was partially supported by EU FP7 xLIME and PERTE project, and A*STAR Singapore under the Human Sixth Sense Program (HSSP) grant.

REFERENCES

[1] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, pp. 224–241, 2011.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, pp. 976–990, 2010.

[3] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *CVPR*, 2012.

[4] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *CVPR*, 2011.

[5] C.-H. Huang, Y.-R. Yeh, and Y.-C. F. Wang, "Recognizing actions across cameras by exploring the correlated subspace," in *ECCV*, 2012.

[6] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *ICCV*, 2007.

[7] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *ECCV Workshop on Consumer Depth Cameras for Computer Vision*, 2012.

[8] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *TPAMI*, vol. 33, no. 1, 2011.

[9] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *NIPS*, 2007.

[10] Y. Yan, G. Liu, E. Ricci, and N. Sebe, "Multi-task linear discriminant analysis for multi-view action recognition," in *ICIP*, 2013.

[11] B. Li, O. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *CVPR*, 2012.

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

[13] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point," in *ECCV*, 2008.

[14] G. Costante, V. Galieni, Y. Yan, M. Fravolini, E. Ricci, and P. Valigi, "Exploting transfer learning for personalized view invariant gesture recognition," in *ICASSP*, 2014.

[15] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *ICCV Workshops*, 2011.

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

[17] M.-H. Tsai, J. Wang, T. Zhang, Y. Gong, and T. S. Huang, "Learning semantic embedding at a large scale," in *ICIP*, 2011.

[18] 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 *ICCV*, 2013.

[19] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing, "Smoothing proximal gradient method for general structured sparse learning." in *UAI*, 2011.

[20] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *SIGKDD*, 2012.

[21] Y. Han, F. Wu, J. Jia, Y. Zhuang, and B. Yu, "Multi-task sparse discriminant analysis (mtsda) with overlapping categories." in *AAAI*, 2010.

[22] Y. Zhang and D.-Y. Yeung, "Multi-task learning in heterogeneous feature spaces," in *AAAI*, 2011.

[23] B. Mahasseni and S. Todorovic, "Latent multitask learning for view-invariant action recognition," in *ICCV*, 2013.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[25] A. Bobick and J. Davis, "The representation and recognition of action using temporal templates," *TPAMI*, vol. 23, no. 3, pp. 257–267, 2001.

[26] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2(1), pp. 183–202, 2009.

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