MULTI-TASK LINEAR DISCRIMINANT ANALYSIS FOR MULTI-VIEW ACTION RECOGNITION

Yan Yan[⋆∗] *Gaowen Liu*[⋆] *Elisa Ricci*[†] *Nicu Sebe*[⋆]

[⋆] Department of Information Engineering and Computer Science, University of Trento, Italy
[†] Department of Electrical and Information Engineering, University of Perugia, Italy

ABSTRACT

Action recognition is a central problem in many practical applications, such as video annotation, video surveillance and human-computer interaction. Most action recognition approaches are currently based on localized spatio-temporal features that can vary significantly when the viewpoint changes. Therefore, the performance rapidly drops when training and test data correspond to different cameras/viewpoints. Recently, Self-Similarity Matrix (SSM) features have been introduced to circumvent this problem. To improve the performance of current SSM-based methods, in this paper we propose a multi-task learning framework for multi-view action recognition where discriminative SSM features are shared among different views. Inspired by the mathematical connection between multivariate linear regression and Linear Discriminant Analysis (LDA), we propose a novel learning algorithm, where a single optimization framework is defined for multi-task multi-class LDA by choosing an appropriate class indicator matrix. Experimental results on the popular IXMAS dataset demonstrate that our approach achieves accurate performance and compares favorably with state-of-the-art methods.

***Index Terms*—** Action Recognition, Multi-View, Self-Similarity Matrix, Multi-Task Learning, Linear Discriminant Analysis

1. INTRODUCTION

Over the past decades, recognizing and understanding human actions in images and videos have attracted considerable attention. Several approaches [1] have been proposed for action recognition in the last few years. ¿From the *representation* point of view, they can be classified mainly into methods computing the time evolution of human silhouettes, the action cylinders, the space-time shapes, and the local 3D patch descriptors. From the *feature extraction* point of view, the various approaches can be categorized into motion based, appearance based, space-time volume based, space-time interest points based, and SSMs-based.

∗Corresponding Author: Yan Yan, E-mail: yan@disi.unitn.it. This work has been funded by the EU project DALi and by FIRB project S-PATTERNS.

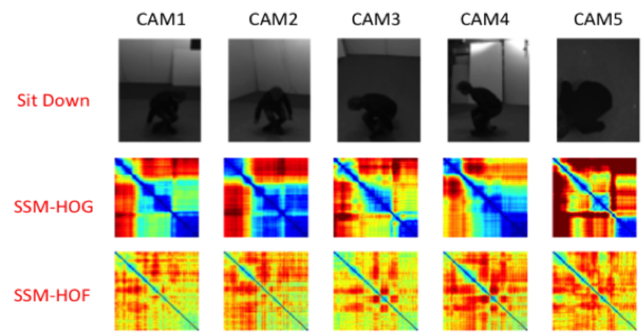


Fig. 1. Examples of SSMs computed from Histogram of Oriented Gradients (HOG) and Histograms of Optical Flows (HOF) features.

Multi-view action recognition has received much attention since a multi-camera setup can overcome the problem of self-occlusions and usually guarantees more robust action recognition compared to the single-view case. Extracting view-invariant information is an important step in the multi-view setting. Many recent approaches are based on transferring features across views [2–5] or using view-invariant features [6–8].

Intuitively, to perform multi-view and view-invariant action recognition, features which are stable across different viewpoints must be computed. Starting from this assumption, in [8] descriptors calculated from temporal Self-Similarity Matrices are proposed. Temporal SSMs can be computed from different low-level features (*e.g.* Histogram of Oriented Gradients, Histograms of Optical Flows) and are shown to be robust descriptors for view-invariant action recognition. 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 five sequences of the IXMAS dataset [9] are shown. Although the SSMs associated to all five cameras share some similar pattern, as the viewpoint from one camera (CAM5) is significantly different from the other four views (CAM1-CAM4), the associated SSM is also quite different.

Multi-task learning [10, 11] aims to simultaneously learn classification/regression models for a set of related tasks. This typically leads to better models as compared to a learner that does not account for 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 to enhance the discriminative power of SSM descriptors in multi-view action recognition. With the proposed algorithm the required level of similarity among different views can be easily controlled by defining a similarity graph reflecting some prior knowledge among relatedness between different views. Additionally, inspired by the equivalence relationship between multivariate linear regression and LDA [12], we cast our multi-class multi-task learning problem into a single optimization problem by choosing an appropriate class indicator matrix and we develop an efficient algorithm to solve it. Our experiments show that sharing features among views is beneficial for multi-view action recognition. On the IXMAS dataset, our approach achieves a recognition accuracy 10% higher than previous works based on SSMs descriptors. Also, our extensive experimental evaluation demonstrates that our method can be successfully used for view-invariant action recognition.

To summarize, the main contributions of this paper are: (i) It represents one of the first works to explore multi-task learning for multi-view action recognition. The proposed algorithm is shown to be effective and achieves improved performance with respect to classification methods based on view-invariant descriptors. (ii) We formulate a novel multi-task multi-class LDA learning problem, casting it into a single optimization problem by choosing an appropriate class indicator matrix. To our knowledge, no previous works have proposed a LDA framework for multi-task learning. Our learning algorithm is quite general and can also be used in other image processing and computer vision tasks.

2. MULTI-TASK LINEAR DISCRIMINANT ANALYSIS FOR ACTION RECOGNITION

The proposed approach for multi-view action recognition is illustrated in Fig.2. First, SSM descriptors are extracted from videos depicting actions at different viewpoints. Then, the bag-of-words model is used for encoding features into histograms and multi-task LDA is adopted to induce features sharing among different viewpoints. In this section we describe the proposed method in detail.

2.1. Self-Similarity Matrix

SSM descriptors have proved to be stable features under view changes [8]. For a sequence of images $\mathcal{I} = \{I_1, I_2, ..., I_T\}$, a SSM of $\mathcal{I}$ is a square symmetric matrix of size $T \times T$:

$$[e_{ij}]_{i,j=1,2,\ldots,T} = \begin{bmatrix} 0 & e_{12} & e_{13} & \cdots & e_{1T} \\ e_{21} & 0 & e_{23} & \cdots & e_{2T} \\ e_{31} & e_{32} & 0 & \cdots & e_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{T1} & e_{T2} & e_{T3} & \cdots & 0 \end{bmatrix} \quad (1)$$

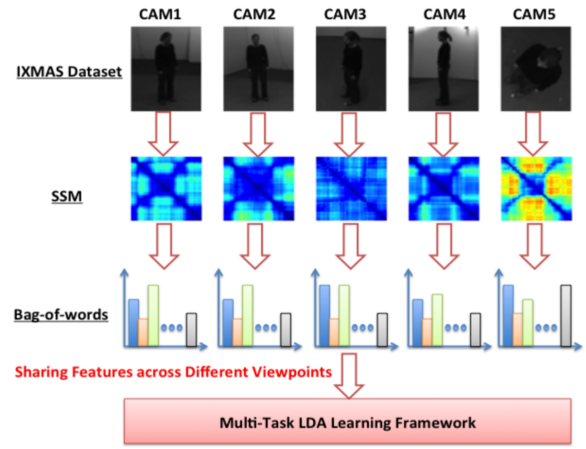


Fig. 2. Overview of the proposed action recognition approach.

where e_{ij} is the distance between certain low-level features extracted in frames I_i and I_j respectively. The diagonal corresponds to comparing a frame to itself (no dissimilarity), hence contains zeros. The exact structures of this matrix depends on the features and the distance measure used for computing the entries e_{ij}. In this paper we use euclidean distance and HOG, HOF and HOG+HOF features calculated at multiple temporal scales. Once SSMs have been computed, the same strategy described in [8] is adopted for calculating local descriptors. For each SSM diagonal point, three local descriptors are computed corresponding to three different diameters of the log-polar domain (respectively 28, 42 and 56 frames in diameter). The bag-of-words model with codebook length of 500 is used to compute a single histogram associated to each video sequence.

An example of SSMs is shown in Fig. 1. Obviously SSMs computed with different low-level features are different, since each feature captures specific properties of the action. Moreover, SSMs are rather stable over different people performing the same action under different viewpoints. However, as observed in Section 1, SSMs are only robust to view changes up to a certain extent. Therefore, in order to individuate features shared among different views, we propose a multi-task LDA learning approach.

2.2. Multi-task Linear Discriminant Analysis

Linear Discriminant Analysis is a well-known technique for dimensionality reduction and classification. Recently multi-class LDA has been shown to be equivalent to multivariate linear regression as long as appropriate class labels are provided [12]. Inspired by this recent result, in this paper an extension of LDA to a *multi-task* learning setting is proposed.

In this paper we assume a set of R related tasks. Each task is a multi-class classification problem and C classes are considered. We are given a training set $\mathcal{T}_t = \{(x_{t_n}, \ell_{t_n})\}_{n=1}^{N_t}$ for each task $t = 1, 2, \ldots, R$, where $x_{t_n} \in I\!R^d$ is d-dimensional feature vector, $\ell_{t_n} \in \{1, 2, \ldots, C\}$ is the label indicating the class membership. Let $(\cdot)'$ denotes the transpose operator.

For each task t we define $\mathbf{x}_t = [x_{t_1}, ..., x_{t_{N_t}}]' \in I\!R^{N_t \times d}$, $\mathbf{y}_t = [\ell_{t_1}, ..., \ell_{t_{N_t}}]' \in I\!R^{N_t \times C}$ which is the class indicator matrix defined as follow:

$$(\mathbf{y}_t)_{ij} = \begin{cases} \sqrt{\frac{N_t}{N_{t_j}}} - \sqrt{\frac{N_{t_j}}{N_t}} & if\ \ \ell_{t_i} = j \\ -\sqrt{\frac{N_{t_j}}{N_t}} & otherwise \end{cases} \quad (2)$$

where $(\cdot)_{ij}$ is the i-th row and j-column of matrix, N_{t_j} is the sample size of j-th class in t-th task, $N_t = \sum_{j=1}^{C} N_{t_j}$ is total training samples of all classes in t-th task. We concatenate $\mathbf{x}_t$ and $\mathbf{y}_t$ of R tasks as $\mathbf{X} = [\mathbf{x}_1', \ldots, \mathbf{x}_R']', \mathbf{X} \in I\!R^{N \times d}$, $\mathbf{Y} = [\mathbf{y}_1', \ldots, \mathbf{y}_R']', \mathbf{Y} \in I\!R^{N \times (CR)}$, where $N = \sum_{t=1}^{R} N_t$. We want 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:

$$\min_{\mathbf{U}} \frac{1}{2} \left\| (\mathbf{Y}\mathbf{Y}')^{-1/2}(\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \lambda_1 \left\| \mathbf{M}\mathbf{U}' \right\|_F^2 + \lambda_2 \left\| \mathbf{U} \right\|_1 \quad (3)$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius and the L_1 norm respectively, $(\mathbf{Y}\mathbf{Y}')^{-1/2}$ is a normalization factor which compensates for different number of samples per task. The matrix $\mathbf{M}$ is an edge-vertex incident matrix, $\mathbf{M} \in I\!R^{|\mathcal{E}| \times CR}$, $|\mathcal{E}|$ denotes graph set cardinality, $\mathbf{M}_{q=(i,j),h} = \gamma_{ij}$ if $i = h$, $\mathbf{M}_{q=(i,j),h} = -\gamma_{ij}$ if $j = h$, $\mathbf{M}_{q=(i,j),h} = 0$ otherwise. 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 between two different tasks and for the same action/class, averaged on the training data. γ_{ij} is normalized into 0-1 and the larger γ_{ij} indicates more similarity of specific action/class between views.

The proposed optimization objective function has three effects. 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 γ_{ij}. Sparsity is enforced in the learning process, performing a beneficial effect of feature selection, and de-emphasizing the contribution of less discriminative features.

We adopt the well-known accelerated gradient method FISTA [13] to solve (3). The key idea of is to solve the proximal operator associated to non-smooth term L_1 norm. We denote the smooth part of the objective function $f(\mathbf{U})$ and the non-smooth part $g(\mathbf{U})$ as:

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

We solve (3) with respect to $\mathbf{U}$ as described in Algorithm 1. L_k is the line search step length and $\hat{\lambda}_1 = 2\lambda_1/L_k$, $\hat{\lambda}_2 = 2\lambda_2/L_k$.

Algorithm 1 Accelerated Gradient Algorithm for solving (3)

INPUT: $\mathcal{T}_t = \{(x_{t_n}, y_{t_n})\}_{n=1}^{N_t}, \forall t = 1, \ldots, R, \lambda_1, \lambda_2, \mathbf{M}$
Initialize $\mathbf{U}_0$, $\alpha_0 = 1$.
LOOP:
$\alpha_k = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{k-1}^2})$
$\hat{\mathbf{U}} = \mathbf{U}_k - \frac{2}{L_k}\mathbf{X}'(\mathbf{Y}\mathbf{Y}')^{-1}(\mathbf{X}\mathbf{U}_k - \mathbf{Y})$
Solving $\mathbf{U}_{k+\frac{1}{2}} \leftarrow \min_{\mathbf{U}} \left\| \mathbf{U} - \hat{\mathbf{U}} \right\|_F^2 + \hat{\lambda}_1 \left\| \mathbf{M}\mathbf{U}' \right\|_F^2 + \hat{\lambda}_2 \left\| \mathbf{U} \right\|_1$ based on Soft Thresholding [14].
$\mathbf{U}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_k})\mathbf{U}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k}\mathbf{U}_k$
Until Convergence
Output: U

3. EXPERIMENTAL RESULTS

In this section, we show the results of our experiments where the performance of the proposed approach are assessed on a publicly available multi-view action recognition dataset.

3.1. Experimental Setup

The IXMAS dataset [9] 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 by five cameras observing the subjects from very different perspectives. The frame rate is 23fps and the frame size 390×291 pixels.

The leave-one-out strategy is employed in our experiments: videos of one subject are selected for testing while videos of the remaining subjects are used as training data. For all the methods, the optimal values of the regularization parameters are determined using a separate validation set and testing the values in the interval $[2^{-6}, 2^{-5}, ..., 2^{6}]$.

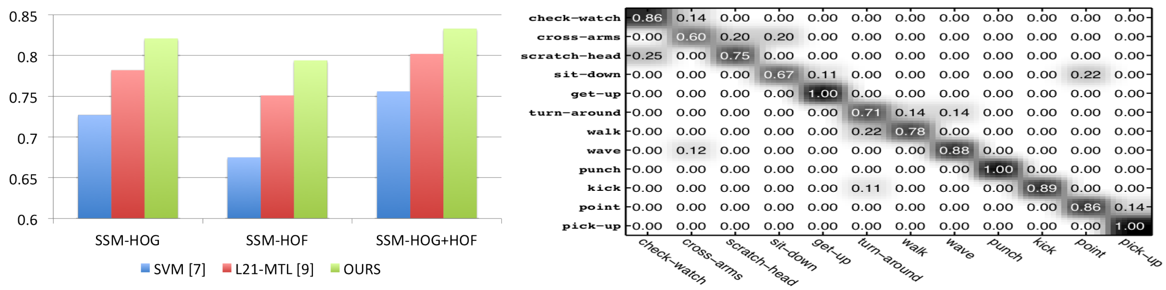
3.2. Quantitative Evaluation

We evaluate the effectiveness of our framework in two cases: (i) ***Multi-view Feature Sharing benefit***. All training samples from all camera views are used in this setting. According to multi-task learning theory, all correlated tasks are learned together. This should boost each individual task's performance. (ii) ***View-invariant Recognition benefit***. 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 $\mathbf{U}$ is learned with our learning framework, for experiments in the case (i), the test sample x_{test} is projected into C dimensional output space by $x_{test}'\mathbf{u}_t$ using $\mathbf{u}_t$ according to the specific view where a test sample belongs. For experiments in (ii), the test sample x_{test} is projected into $(R-1)C$ dimensional output space by $x_{test}'\mathbf{U}$ since only $R-1$ tasks are considered in this setting. The class label of the test sample is assigned using KNN.

We compare the proposed approach with a single SVM classifier [8] and the $\ell_{2,1}$-norm multi-task learning approach proposed in [10] which assumes all the tasks to be related to each other and no graph specifying their relationships is considered. In the SVM experiments, a radial basis kernel is chosen and the LIBSVM[1] software package is used. A

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/



(a) (b)

Fig. 3. (a) Recognition accuracy with different SSM features. (b) Confusion Matrix on the IXMAS dataset.

Table 1. Multi-view action recognition accuracy: comparison of different methods.

Training with All Cameras (Classification accuracy (0-1))						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
Proposed approach	0.900	0.854	0.812	0.793	0.763	0.825
Junejo - SVM [8]	0.748	0.745	0.748	0.706	0.612	0.727
ℓ_{12} MTL [10]	0.819	0.830	0.809	0.756	0.693	0.782
Li [3]	0.834	0.799	0.820	0.853	0.755	0.812
Liu [4]	0.790	0.747	0.752	0.764	0.712	0.753
Huang [15]	0.632	0.586	0.604	0.568	0.476	0.573
Weiland [9]	0.654	0.700	0.543	0.660	0.336	0.579
Reddy [16]	0.696	0.692	0.620	0.651	-	0.726
Farhadi [2]	-	-	-	-	-	0.581
Li [6]	0.910	0.919	0.911	0.906	0.871	0.905
Wu [17]	0.909	0.854	0.888	0.909	0.881	0.888

publicly available code[2] is also used for $\ell_{2,1}$-norm multi-task learning in our experiments. The top three lines in Table 1 show the results of such comparison. Here SSM descriptors with HOG+HOF features are used. It is evident how sharing similarity information among different views using multi-task learning outperforms SVM by at least 10%. 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. The viewpoint associated to CAM5 is significantly different from the other four views. However, also in this case, a multi-task learning approach is greatly beneficial as we improve the action recognition rate from 69.3% to 74.3%. That means CAM5 view is 'absorbing' some useful information from the other views. These observations show the benefit of *feature sharing* among different views achieved by our multi-task learning framework.

We also consider as baselines other action recognition methods which are not based on SSMs. A comparison of the performance is shown in Table 1. Our approach achieves higher recognition accuracy both on the single camera and on the average results when compared to most previous methods. On the other hand, the approaches proposed in [6] and [17] have a recognition rate higher than ours on the IXMAS dataset. However, the method in [17] is based on latent kernelized structural SVM which is intractable for inference in large-scale datasets. The feature extraction phase of the

[2] http://ttic.uchicago.edu/∼argyriou/code/index.html

Table 2. Cross-view action recognition accuracy: training is performed with one view missing.

	Missing View				
	Cam1	Cam2	Cam3	Cam4	Cam5
Proposed Method	**0.755**	**0.746**	**0.771**	**0.697**	**0.633**
Junejo - SVM [8]	0.666	0.655	0.650	0.624	0.496
ℓ_{12} MTL [10]	0.711	0.728	0.730	0.673	0.602

algorithm in [6] is also computationally demanding. Differently, our method is computationally efficient and easy to implement.

Figure 3(a) shows the results obtained using various SSM features. The best performance are achieved using SSM with HOG+HOF features. Figure 3(b) shows the confusion matrix on the IXMAS dataset. It is interesting to observe that for some actions such as 'get up', 'pick up' and 'punch', our method achieves very high recognition accuracies. Even for some challenging actions (*e.g.*, 'point', 'check watch' and 'wave') having small and ambiguous motions, our method still guarantees reasonable and promising results.

To evaluate the benefit of our approach on *view-invariant* action recognition, we evaluate its performance when one view is missing in the training data. The results are shown in Table 2. 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.

4. CONCLUSIONS

In this paper, we considered the problem of human action recognition in a multi-view scenario. We proposed a multi-task extension of multi-class LDA which operates by sharing SSMs features among different views. Experimental results on the IXMAS dataset demonstrate the superior performance of our method compared to other SSM-based state-of-the-art methods. Possible future works include the integration of other features in combination with SSM descriptors and the investigation of a different strategy for graph construction, eventually considering information about geometry and cameras'configuration.

References

[1] Daniel Weinland, Remi Ronfard, and Edmond 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] Ali Farhadi and Mostafa Kamali Tabrizi, "Learning to recognize activities from the wrong view point," in *ECCV*, 2008.

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

[4] Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese, "Cross-view action recognition via view knowledge transfer," in *CVPR*, 2011.

[5] Anoop Kolar Rajagopal, Subramanian Ramanathan, Radu L. Vieriu, Elisa Ricci, Oswald Lanz, Kalpathi Ramakrishnan, and Nicu Sebe, "An adaptation framework for head-pose classification in dynamic multi-view scenarios," in *ACCV*, 2012, pp. 652–666.

[6] Binlong Li, Octavia I. Camps, and Mario Sznaier, "Cross-view activity recognition using hankelets," in *CVPR*, 2012.

[7] Cen Rao, Alper Yilmaz, and Mubarak Shah, "View-invariant representation and recognition of actions," *IJCV*, vol. 50, no. 2, pp. 203–226, 2002.

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

[9] Daniel Weinland, Edmond Boyer, and Remi Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *ICCV*, 2007.

[10] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil, "Multi-task feature learning," in *NIPS*, 2007.

[11] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing, "Smoothing proximal gradient method for general structured sparse learning," in *UAI*, 2011.

[12] Jieping Ye, "Least squares linear discriminant analysis," in *ICML*, 2007.

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

[14] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan 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.

[15] Chun-Hao Huang, Yi-Ren Yeh, and Yu-Chiang Frank Wang, "Recognizing actions across cameras by exploring the correlated subspace," in *ECCV*, 2012.

[16] Kishore Reddy, Jingen Liu, and Mubarak Shah, "Incremental action recognition using feature tree," in *ICCV*, 2009.

[17] Xinxiao Wu and Yunde Jia, "View-invariant action recognition using latent kernelized structural SVM," in *ECCV*, 2012.