Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition

An-An Liu, *Member, IEEE*, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli, *Fellow, IEEE*

Abstract—This paper proposes a hierarchical clustering multi-task learning (HC-MTL) method for joint human action grouping and recognition. Specifically, we formulate the objective function into the group-wise least square loss regularized by low rank and sparsity with respect to two latent variables, model parameters and grouping information, for joint optimization. To handle this non-convex optimization, we decompose it into two sub-tasks, multi-task learning and task relatedness discovery. First, we convert this non-convex objective function into the convex formulation by fixing the latent grouping information. This new objective function focuses on multi-task learning by strengthening the shared-action relationship and action-specific feature learning. Second, we leverage the learned model parameters for the task relatedness measure and clustering. In this way, HC-MTL can attain both optimal action models and group discovery by alternating iteratively. The proposed method is validated on three kinds of challenging datasets, including six realistic action datasets (Hollywood2, YouTube, UCF Sports, UCF50, HMDB51 & UCF101), two constrained datasets (KTH & TJU), and two multi-view datasets (MV-TJU & IXMAS). The extensive experimental results show that: 1) HC-MTL can produce competing performances to the state of the arts for action recognition and grouping; 2) HC-MTL can overcome the difficulty in heuristic action grouping simply based on human knowledge; 3) HC-MTL can avoid the possible inconsistency between the subjective action grouping depending on human knowledge and objective action grouping based on the feature subspace distributions of multiple actions. Comparison with the popular clustered multi-task learning further reveals that the discovered latent relatedness by HC-MTL aids inducing the group-wise multi-task learning and boosts the performance. To the best of our knowledge, ours is the first work that breaks the assumption that all actions are either independent for individual learning or correlated for joint modeling and proposes HC-MTL for automated, joint action grouping and modeling.

Index Terms—Action recognition, multi-task learning, task grouping, task relatedness measure

1 Introduction

Human action recognition is a fundamental building block in diverse applications such as video surveillance, multimedia content analysis and retrieval, digital entertainment and natural human machine interaction [1], [2]. Being one of the most active areas of research in both computer vision and machine learning communities, multiple approaches towards this task have been developed [3].

1.1 Motivation and Overview

Although the current methods have demonstrated superior performance on this task, there still exist three challenges: 1) *Ignoring of the existence of latent relatedness among actions.* Most of current methods work in the single-task learning (STL) framework [4] as shown in Fig. 1a. They usually focus on designing features [5] but the human-crafted features are not discriminative enough to distinguish actions with similar motion patterns (such as walking, running, and jogging in KTH [5]). 2) *Ignoring of the grouping information among actions.* Recently several researchers have tried to employ multi-task learning (MTL) for human action recognition and leveraged the shared knowledge among them to improve the generalization ability of model learning [1], [2] as shown in Fig. 1b. One key assumption in these MTL-based methods is that all tasks are correlated so that the discovered knowledge can be transferred between them. However, it may not hold true for many complicated applications. In reality, it is often the case that a group of tasks are strongly related while the others are totally different. For example, boxing/handwaving/handclapping and running/walking/jogging in KTH [5] can be easily considered as two differing categories since they have salient upper-body motion and lower-body motion respectively. Therefore, joint learning all of the tasks together might have negative impact on the performance. 3) *Difficulty in grouping information discovery.* Although it seems easy to group actions in the constrained action datasets (KTH [5], IXMAS [6] etc.) as Liu et al. did based on the body regions with salient motion [7], it is usually challenging to heuristically cluster them just with human knowledge, especially for the realistic action datasets (Hollywood2 [8], UCF Sports [9], YouTube [10], UCF50 [11], etc.). Therefore, the automated action grouping method is mandatory before applying MTL on the related tasks.

To handle these problems, we propose a hierarchical clustering multi-task learning (HC-MTL) method for joint action relatedness discovery and action modeling as shown in Fig. 1c. We formulate the objective function with two latent variables, model parameters and grouping information, for joint optimization, which can effectively avoid: 1) the difficulty in heuristic action grouping based only on

- *A.-A. Liu, Y.-T. Su, and W.-Z. Nie are with the School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China. E-mail: anan0422@gmail.com, {ytsu, weizhinie}@tju.edu.cn.*
- *M. Kankanhalli is with the School of Computing, National University of Singapore. E-mail: mohan@comp.nus.edu.sg.*

Manuscript received 28 Nov. 2014; revised 11 Feb. 2016; accepted 24 Feb. 2016. Date of publication 1 Mar. 2016; date of current version 12 Dec. 2016. Recommended for acceptance by D. Ramanan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2537337

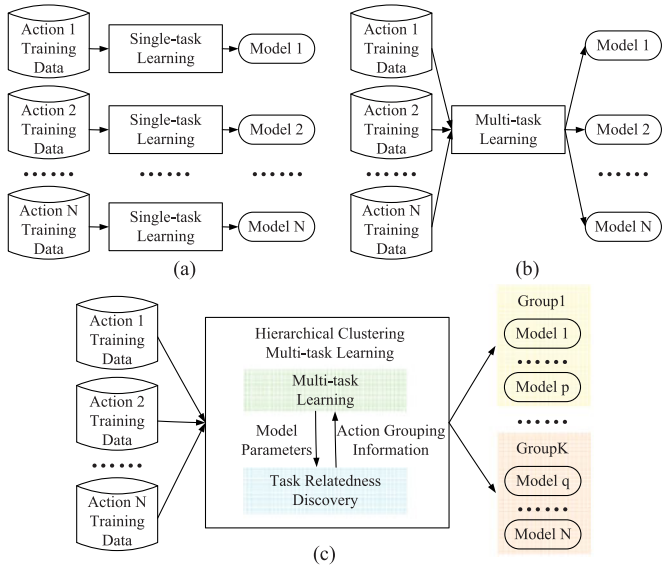

Fig. 1. Comparison of systematic frameworks. (a) The flowchart of the STL methods; (b) The flowchart of the MTL methods; (c) The flowchart of the proposed method with the shaded parts highlighting differences with respect to the state of the art.

human knowledge; 2) the possible inconsistency between the subjective action grouping depending on human knowledge and objective action grouping based on the feature subspace distributions of multiple actions. Differing from clustered multi-task learning (CMTL) [12] which performs task grouping and modeling in one shot, the proposed method decomposes this joint work into two sub-tasks: 1) multi-task learning: we formulate the objective function for model learning as the least square loss penalized by the "trace norm" and "group sparsity" terms to strengthen task correlation and discriminative feature learning; 2) task relatedness discovery: with the learned model parameters, we leverage sparse representation for the task relatedness measure and then utilize spectral clustering for task clustering; the obtained action grouping information is then fed back to the earlier sub-task to further induce group-wise multi-task learning. In this way, the proposed method can realize model learning and grouping (related set and unrelated set) in individual iterations and achieve the optimal solution by alternating between them iteratively.

1.2 Contributions

The main contributions are summarized as follows:

- To the best of our knowledge, our work is the first that breaks the assumption that all actions are either independent for individual learning [4] or correlated for joint modeling [1], [2] by proposing HC-MTL for joint action grouping and modeling. The experiments demonstrate the discovered latent relatedness greatly benefits in inducing the group-wise multi-task learning and consequently boosts the performance. Moreover, the hierarchical manner of HC-MTL helps model learning by implicitly augmenting the dataset since it repeatedly utilizes the dataset to optimize both of the sub-tasks. We notice that several methods [13], [14] have been developed to leverage task grouping information for model learning in the machine learning society. However, to the best of our knowledge, there is no previous work which particularly focuses on designing methods to integrate action grouping discovery and modeling together to benefit human action recognition.
- The proposed HC-MTL can aid in discovering both shared-action relatedness and action-specific feature subspaces. To preserve the characteristics of individual actions (e.g., different motion regions between walking and boxing) as well as to discover the common knowledge (e.g., similar motion pattern between walking and running), we design the objective function regularized by the trace norm and group sparsity terms for joint multiple action learning. Moreover, we suitably adapt the accelerated proximal algorithm to optimize this non-smooth objective function.
- The proposed method is evaluated for three senarios, including action recognition, action grouping, and direct comparison between HC-MTL and CMTL. Extensive experiments have been conducted on three kinds of datasets, including six realistic action datasets, two constrained action datasets, and two multi-view action datasets.

The rest of the paper is structured as follows. In Section 2, we introduce the related work. Then HC-MTL is presented in Section 3. The experimental method and results will be detailed in Sections 4 and 5. Section 6 concludes the paper.

2 Related Work

We mainly cover two topics which are closely related to our work.

1) *Single-task learning-based method*. Recently, the local feature-based methods have attracted increasing attention since the spatiotemporal features can capture local salient characteristics of appearance and motion [5], [15], [16]. Therefore, they are robust to spatiotemporal shifts and scales, background clutter and multiple motions present in realistic scenes. With these local salient points, the bag-of-word (BoW) approach [4], [17] can be leveraged to construct a histogram-based representation and lots of work have demonstrated that the BoW representation can work successfully with SVM classifier. Although much work has been done for this task, most of the current methods work in the single-task learning framework. Consequently, these methods usually ignore the relationship among multiple action categories and therefore they might not be discriminative enough to distinguish the actions with similar motion patterns (such as walking & running). Furthermore, these methods are highly restricted by the existing limited action datasets considering the complexity of human actions.

Unlike the single-task learning methods that model each action individually, we will leverage on multi-task learning for human action modeling by discovering and transferring the common knowledge across actions to boost the performance.

2) *Multi-task learning-based method*. Multi-task learning can help improve the generalization performance by jointly learning multiple tasks using the shared knowledge among them. Till now, only few MTL methods have been designed for human action recognition. Yuan et al. [1] represent each

sample with several kinds of features and consider learning the sparse representation under each feature modality as a single task. Since these features are inter-related, they formulate human action modeling as a multi-task sparse learning framework by constructing a test sample with multiple features from as few bases as possible. Similarly, Liu et al. [18] represent each action sample with pyramid part-wise BoW feature and consider each part-wise feature-based modeling as single task. Since different pyramid part-wise visual features are intrinsically correlated with each other during one action, individual part-wise feature-based tasks can be jointly learned in the multi-task learning framework. They further formulate this task as part-induced multi-task learning penalized by multiple graph structures to discover the latent correlation among multiple views and body parts and extend this method for both single and multi-view human action recognition [7]. In contrast to [1] which implements MTL to model different action categories, Zhou et al. [2] utilize MTL to learn latent tasks shared across categories, which correspond to basic motion patterns, and reconstruct each action classifier with these latent tasks, which can be understood as transferring of common knowledge between different action categories. Zhang et al. [19] formulate the action modeling together with penalty by the multiple attribute-induced learning tasks into one objective function and consequently convert this task into the multi-task learning framework. In this way, the learned action models can share the commonality while getting regularized based on the semantic constraints.

In contrast to the current methods that either consider all tasks as being related [1], [2] or heuristically impose explicit relationships among tasks [7], we focus on the automated exploration of the action grouping information to induce action modeling among the related tasks while avoiding the negative influence of non-relevant tasks.

3 Hierarchical Clustering Multi-Task Learning

We propose the use of hierarchical clustering multi-task learning to discover the latent task correlation as well as to learn the action model simultaneously. This can be achieved by iteratively alternating between multi-task learning and task relatedness discovery (TRD) as shown in Fig. 1c. We will formally present the HC-MTL framework in Section 3.1. Then, we will describe the proposed task relatedness discovery method in Sections 3.2. The notation used is presented in Table 1.

3.1 HC-MTL Framework

The ideal situation for multi-task learning is when joint learning is done on individual groups of tasks with known prior relationships among all the different tasks. However, it is extremely challenging to a priori identify the grouping information only with human knowledge, especially for the complicated actions in the realistic action datasets [11]. Unlike the previous methods either in the STL framework (Fig. 1a) or the MTL framework (Fig. 1b) without considering the latent task grouping information, we propose the hierarchical clustering multi-task learning method for joint action modeling and grouping.

TABLE 1
Notations and Definitions

Notation	Definition
$\mathbb{N}_m$	$\mathbb{N}_m = \{1, \ldots, m\}$, task index set
A	$A = [a_1, \ldots, a_m] \in \mathbb{R}^{d \times m}$
a_i	$a_i \in \mathbb{R}^d$ is the ith column of A
$\|\|a_i\|\|_2$	l_2 norm of a vector
$\|\|a_i\|\|_\infty$	l_∞ norm of a vector
$\|\|A\|\|_*$	trace norm (sum of singular values) of a matrix
$\|\|A\|\|_F$	Frobenius norm of a matrix
$\|\|A\|\|_{\infty,2}$	$\|\|A\|\|_{\infty,2} = \|\|a_j\|\|_2$, where $j = \arg\max_i \|\|a_i\|\|_2$
$\|\|A\|\|_{1,2}$	$\|\|A\|\|_{1,2} = \sum_{i=1}^m \|\|a_i\|\|_2$
$\mathcal{T}$	$\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots \mathcal{T}_t, \ldots, \mathcal{T}_T\}$, task grouping information
$\mathcal{T}_t$	task index set of the tth group
$\|\mathcal{T}\|$	group number in the task set $\mathcal{T}$
$\|\mathcal{T}_t\|$	task number in the group $\mathcal{T}_t$
e_+	$e_+ = \max(e, 0)$
$diag(C)$	vector consisting of diagonal elements of matrix C

3.1.1 Objective Function

Suppose there are totally m tasks (m action categories) and each task has n_i training samples $\{(x_j^i, y_j^i)\}_{j=1}^{n_i}$, where $x_j^i \in \mathbb{R}^d$ is the feature representation of each sample, $y_j^i \in \mathbb{R}$ is the action label, and $i \in \mathbb{N}_m$ denotes the task index. In our work, there are two closely correlated sub-tasks to be completed: 1) Action modeling: we aim to learn a linear predictive function f_i for an individual task, where $f_i(x_j^i; w_i) = w_i^\top x_j^i$ and w_i is the parameter of the corresponding action model. The union of w_i of m tasks can be denoted as a matrix $W = [w_1, \ldots, w_m] \in \mathbb{R}^{d \times m}$. 2) Action grouping: we aim to discover the latent task relatedness information contained in $\mathcal{T}$. For example, the proposed method aims to jointly learn the related actions (e.g., handwaving/handclapping/boxing can be regarded as related actions considering salient limb and hand motions) while identifying those non-relevant actions (e.g., jogging/walking/running can be considered as non-relevant actions to the aforementioned limb-wise actions). Consequently, the objective function can be formulated as:

$$(W^*, \mathcal{T}^*) = \arg\min_{W,\mathcal{T}} \mathcal{L}(\{(x_j^i, y_j^i)\}_{j=1}^{n_i}; W, \mathcal{T}, i \in \mathbb{N}_m) + Reg(W, \mathcal{T}). \tag{1}$$

This objective function consists of two terms:

a. Main tasks $\mathcal{L}(\{(x_j^i, y_j^i)\}_{j=1}^{n_i}; W, \mathcal{T}, i \in \mathbb{N}_m)$: The main tasks of multi-task learning minimize the joint empirical risk to learn the classifiers of all tasks (entire action models in our work) $F(x_j^i, W) = \{f_i(x_j^i; w_i)\}_{i \in \mathbb{N}_m}$ which can be utilized for multi-task classification. We formulate this term with the empirical evaluation of the least square loss of the m tasks with the grouping information $\mathcal{T}$ over the prescribed training data as

$$\mathcal{L}(\{(x_j^i, y_j^i)\}_{j=1}^{n_i}; W, \mathcal{T}, i \in \mathbb{N}_m) = \sum_{\mathcal{T}_t \in \mathcal{T}} \sum_{i \in \mathcal{T}_t} \sum_{j=1}^{n_i} \frac{1}{|\mathcal{T}_t| n_i} \left(w_i^\top x_j^i - y_j^i\right)^2. \tag{2}$$

b. Regularization $Reg(W, \mathcal{T})$: We design the regularization term with respect to model parameter set W and grouping information $\mathcal{T}$ for simultaneous action modeling and

grouping (related group and unrelated group). For human action recognition, the underlying action models may lie in a hypothesis space with a low rank structure, in which the multiple actions can be coupled using a set of shared factors, i.e., the basis of a low-rank subspace [20]. For example, boxing/handwaving/handclapping may share some basic action units, such as arm stretching out, arm drawing back, and arm moving horizontally. These similar action units will induce the shared feature subspace of a low-rank structure. Nevertheless, different actions can also have significant differences. For example, put-down and pick-up have similar action units (e.g., holding some object, arm/hand moving vertically, etc.) but the different transitions among action units can cause the different motion characteristics. Therefore, the prospective regularization term needs to penalize both aspects. Motivated by [13], we adopt the trace norm regularization since it can impose the low-rank structure to couple related tasks. Furthermore, we impose group sparsity regularization for group-induced discriminative feature learning.

Specifically, we decouple W into two components, a low-rank matrix $L = [l_1, \ldots, l_m] \in \mathbb{R}^{d\times m}$ and a column-wise group-sparse matrix $S = [s_1, \ldots, s_m] \in \mathbb{R}^{d\times m}$. The model parameter of the ith task can be expressed as

$$w_i = l_i + s_i,\ l_i \in \mathbb{R}^d,\ s_i \in \mathbb{R}^d,\ i \in \mathbb{N}^m, \tag{3}$$

where l_i and s_i are from the aforementioned low-rank structure and the group-sparse structure, respectively. If the ith task is from the related group, s_i is expected to be a zero-vector and hence w_i preserves the task relatedness constrained by the specified low-rank structure. Otherwise, if the ith task is from the unrelated group, s_i is expected to be non-zero and w_i, the direct sum of l_i and the non-zero s_i, is consequently expected to be classified into the unrelated group.

Therefore, the regularization on all groups can be formulated as

$$Reg(W, \mathcal{T}) = \sum_{\mathcal{T}_t\in\mathcal{T}} (\alpha\|L_{\mathcal{T}_t}\|_* + \beta\|S_{\mathcal{T}_t}\|_{1,2}), \tag{4}$$

where W can be represented in the group manner $W = \{W_{\mathcal{T}_t}\}_{\mathcal{T}_t\in\mathcal{T}}$; $W_{\mathcal{T}_t} = \{w_i\}_{i\in\mathcal{T}_t} = L_{\mathcal{T}_t} + S_{\mathcal{T}_t}$; $L_{\mathcal{T}_t} = \{l_i\}_{i\in\mathcal{T}_t}$; $S_{\mathcal{T}_t} = \{s_i\}_{i\in\mathcal{T}_t}$.

To sum up, the proposed objective function can be formulated as:

$$(W^*, \mathcal{T}^*) = \arg\min_{W,\mathcal{T}} \sum_{\mathcal{T}_t\in\mathcal{T}} \Bigg\{ \sum_{i\in\mathcal{T}_t} \sum_{j=1}^{n_i} \frac{1}{|\mathcal{T}_t| n_i} \left(w_i^\top x_j^i - y_j^i\right)^2 + \alpha\|L_{\mathcal{T}_t}\|_* + \beta\|S_{\mathcal{T}_t}\|_{1,2} \Bigg\}, \tag{5}$$

where α and β are non-negative trade-off parameters. The objective function in Eq. (5) denotes that the proposed HC-MTL will implement the group-wise model learning in terms of the latent grouping information $\mathcal{T}$, which is directly decided by the ground truth group number but not the task number, and jointly optimize all groups to achieve the optimal solution. Therefore, it is totally different from the current methods in the STL or MTL framework (Fig. 1).

3.1.2 Optimization

A close look into the above optimization problem reveals that it is nontrivial to directly optimize the objective function since it is not jointly convex with respect to both W and $\mathcal{T}$. Suppose if the task grouping information $\mathcal{T}$ is known, we can implement multi-task learning on individual group of tasks for joint model learning. On the other hand, if the models of all tasks are known, we can leverage the model parameters for task grouping. Consequently, to handle this challenging problem, our proposal is to decompose this joint task into two separate steps, multi-task learning and task relatedness discovery. By alternating both iteratively, the optimal solution can be achieved based on the coordinate descent theory.

- *Multi-task learning*: In the hth level (iteration), multi-task learning is first leveraged for model learning of the task set $\mathcal{T}^h$ ($\mathcal{T}^h \subseteq \mathcal{T}$). The task grouping information $\mathcal{T}^h$ is fixed by treating all the tasks in $\mathcal{T}^h$ equally and independently, i.e., each group $\mathcal{T}_t^h$ in $\mathcal{T}^h$ corresponds to an individual task. In this way, the latent variable of the unknown grouping information can be eliminated. Moreover, the trace norm regularization is utilized to couple the related tasks by imposing the low-rank structure and the group sparsity regularization is utilized to induce group-wise discriminative feature learning. Consequently, in the hth level, the objective function in Eq. (5) can be converted into:

$$(W^h)^* = \arg\min_{L^h,S^h} \sum_{i\in\mathcal{T}^h} \sum_{j=1}^{n_i} \frac{1}{|\mathcal{T}^h| n_i} \cdot ((l_i^h + s_i^h)^\top x_j^i - y_j^i)^2 + \alpha||L^h||_* + \beta||S^h||_{1,2}. \tag{6}$$

  This objective function is convex with respect to W. We adapt the accelerated proximal method (APM) [21] for solution.
- *Task relatedness discovery*: As discussed in the Regularization part of Section 3.1.1, S^h, the group sparsity component of W^h, can be utilized to measure the group similarity and further cluster tasks in $\mathcal{T}^h$. In this step, we fix S^h and propose the method of task relatedness discovery based on subspace-sparse clustering, which will be detailed in Section 3.2. To avoid the difficulty in balancing between multiple clusters when CMTL works, we propose to identify the tasks with strong relationships as the related task set $\mathcal{T}_R^h$ and leave all the others as the outlier task set $\mathcal{T}_O^h$. The task number in $\mathcal{T}_R^h$ is adaptively decided by the similarity among tasks but not heuristically fixed as the traditional hierarchical clustering, which only merges two in each iteration. In this way, we can cluster the easiest tasks in $\mathcal{T}^h$ as $\mathcal{T}_R^h$ and keep the corresponding models $w_i^h = l_i^h + s_i^h (i \in \mathcal{T}_R^h)$ to update the corresponding columns of W^*. The tasks in $\mathcal{T}_O^h$ will be utilized for $\mathcal{T}^{h+1}$. $\mathcal{T}_R^h$ is required to be removed from $\mathcal{T}^{h+1}$. Otherwise, the same models and grouping information will be achieved by preserving the same task set and training data. No updated grouping information can be obtained to further benefit model learning in the next level.

To sum up, in each level, we leverage multi-task learning for model learning and then the achieved model parameters

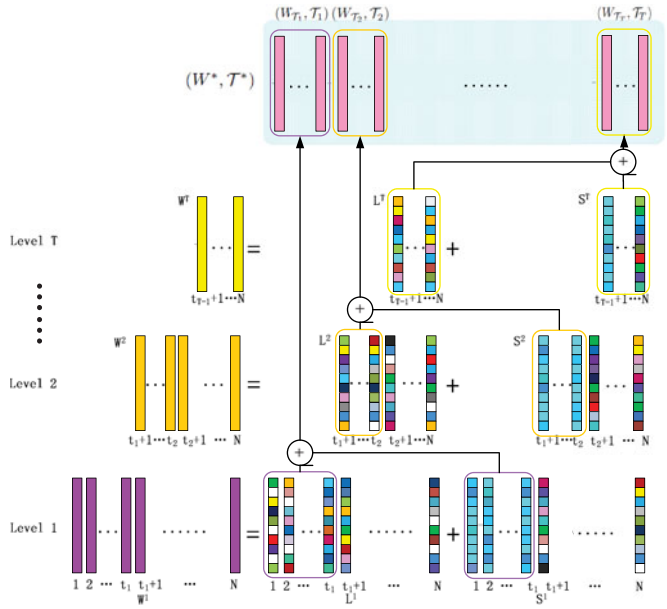

Fig. 2. HC-MTL. W^* denotes the optimal models and each column vector depicts the parameter of each model; $W^h (W^h = L^h + S^h)$ denotes the obtained models in the hth level; L^h is the low-rank part; S^h is the group-sparse part; the elements of the blue vector in S^h denote the near 0 value; each box denotes individual group of actions, $\mathcal{T}_t$, and $\mathcal{T}^*$ is the optimal grouping information; $\oplus$ denotes the column-wise vector addition in Eq. (3).

will be fed into the module of task relatedness discovery for task grouping. We only keep the model of the related task set and treat the others for optimization in the next level, which will be fed back to the module of multi-task learning for model updating. Consequently, the task number of $\mathcal{T}^h$ for the next iteration will decrease until the stopping criteria is reached. This hierarchical manner of working is shown in Fig. 2 and is also outlined in Algorithm 1.

Algorithm 1. HC-MTL Algorithm

Input: entire task set $\mathcal{T}$ with m tasks, $n_i (i \in \mathbb{N}_m)$ training samples $\{(x_j^i, y_j^i)\}_{j=1}^{n_i}$ for ith task, α, β

Output: optimal model parameter set W^*, optimal grouping information $\mathcal{T}^*$

Initialize

related task set $\mathcal{T}_R^0 \leftarrow \emptyset$, outlier task set $\mathcal{T}_O^0 \leftarrow \emptyset$, task set for model learning and clustering $\mathcal{T}^1 \leftarrow \mathcal{T}$, model parameter set $W^0 \leftarrow 0$, max iteration number H_{max}

end

// main loop

for $h \leftarrow 1 \ldots H_{max}$ **do**

1) Multi-task learning for W^h (L^h and S^h) by optimizing Eq. (6) with APM
2) Task relatedness discovery by clustering S^h into $\mathcal{T}_R^h$ & $\mathcal{T}_O^h$ by TRD (Algorithm 2)
3) Update W^* with $w_i^h = l_i^h + s_i^h (i \in \mathcal{T}_R^h)$
4) Update $\mathcal{T}^*$ with group information in $\mathcal{T}_R^h$
5) Update $\mathcal{T}^{h+1}$ with $\mathcal{T}_O^h$

end

3.2 TRD Algorithm

In our work, we propose the TRD method, which consists of the task relatedness measure and task clustering. It can benefit HC-MTL optimization due to two reasons: 1) TRD implements clustering on the group-sparse component, S, which is discriminative enough for task grouping. Penalized by the group sparsity regularization in Eq. (6), S has the nature of group structure since s_i will approximate zero to preserve w_i obeying the low-rank structure constraint when the ith task belongs to the related group and otherwise s_i will become non-zero when the ith task belongs to an unrelated group. 2) TRD is more robust for clustering by leveraging self-expressiveness property for similarity measure and utilizing spectral clustering to overcome the negative influence by noise and outliers. Both task relatedness measure and task clustering will be separately introduced below.

3.2.1 Task Relatedness Measure

Suppose that all s_i in S lie in the union of n subspaces, $\{\xi_i\}_{i=1}^n$. Since the structure of each subspace is unknown, the similarity between two models (s_i and s_j) cannot be directly computed in terms of specific distance metrics. Motivated by the sparse representation, s_i, the parameter of the ith model lying in a union of subspaces, can be reconstructed by a combination of other model parameters $\{s_j\}_{j \in N_m, j \neq i}$ as

$$s_i = Sc_i \quad s.t. \;\; c_{ii=0}, \tag{7}$$

where $c_i \triangleq [c_{i1}\ c_{i2} \ldots c_{im}]^\top$ and the constraint $c_{ii} = 0$ avoids the trivial solution of reconstructing s_i by itself. In Eq. (7), the model parameter s_i can be written as a linear combination of the parameters of other models and c_{ik}, the kth dimension of coefficient c_i, denoting the weight of the kth model. Therefore, c_{ik} denotes the relatedness between the ith and kth tasks.

In general, the representation of $\boldsymbol{s}_i$ by the bases in S is not unique. Motivated by the theory of subspace-sparse representation [22], sparse constraints can be imposed to achieve a sparse solution. Ideally, the L_0-normal of C_i can be intuitively imposed to induce sparsity. However, this will lead to the general NP-hard problem of finding the sparsest representation of the given point. Therefore, we consider minimizing the tightest convex relaxation of the L_0-norm of C_i with L_1-norm. The objective functions can be formulated as:

$$min \|C\|_1 \quad s.t. \quad S = SC,\ diag(C) = 0, \tag{8}$$

where $C \triangleq [c_1\ c_2 \ldots c_m]$ is the matrix, whose ith column corresponds to the sparse representation of s_i. Eq. (8) can be efficiently solved using convex programming tools [23].

3.2.2 Task Clustering

With the solution of Eq. (8), we can obtain the sparse representation for every individual model parameter, whose nonzero elements ideally correspond to the tasks from the same subspace. Consequently, the sparse coefficients can be used for task clustering.

To address this problem, we build a similarity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ denotes the set of m nodes of the graph corresponding to m tasks and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges between nodes. $\mathcal{W} \in \mathbb{R}^{m \times m}$ is a symmetric nonnegative similarity matrix representing the weights of the edges, i.e., node i is connected to node j by an edge whose weight is equal to $\mathcal{W}_{ij}$. In an ideal similarity graph $\mathcal{G}$, the ideal similarity matrix $\mathcal{W}$ is one in which nodes that

correspond to points from the same subspace are connected to each other and there are no edges between nodes that correspond to points in different subspaces. Therefore, the similarity matrix can be obtained by $\mathcal{W} = |C| + |C|^\top$ to preserve the symmetrization. Then, task clustering can be accomplished by spectral clustering [24] on $\mathcal{G}$.

The proposed task relatedness discovery algorithm is outlined in Algorithm 2.

Algorithm 2. TRD Algorithm

Input: Model parameters $S \in \mathbb{R}^{d \times m}$ of m tasks lying in a union of n linear subspaces
Output: Category indexes of individual tasks
// main step
1: Compute C with S by Eq. (8).
2: Normalize the columns of C by $c_i \leftarrow \frac{c_i}{\|c_i\|_\infty}$.
3: Construct $\mathcal{G}$ with m tasks by setting $\mathcal{W} \leftarrow |C| + |C|^\top$.
4: Cluster tasks by implementing spectral clustering [24] on $\mathcal{G}$.

4 Experimental Method

4.1 Data

The proposed method is firstly evaluated on the School data [12], [25], which has been widely used for MTL evaluation. Then it is extensively evaluated on three kinds of action datasets. We will briefly introduce them first.

4.1.1 Realistic Action Dataset

We utilized 6 datasets for the evaluation of HC-MTL.

- UCF Sport [9] contains 10 actions with 150 samples. We followed the leave-one-sample-out (LOSO) strategy as [4].
- YouTube [10] contains 11 actions with 1,168 samples. For each category the videos are divided into 25 groups. We followed the leave-one-group-out (LOGO) strategy as [10].
- Hollywood2 [8] contains 12 actions. In our experiments, we used the clean dataset with 1,707 samples. We followed the "split" strategy as [8].
- HMDB51 [26] contains 51 actions with 6,766 videos. All the videos are obtained from movies, YouTube, and so on. We used the original three train-test splits on the stabilized data as [26].
- UCF50 [11] contains 50 actions with 6,618 videos. Each category is divided into 25 groups with at least 100 videos. In our experiments, we followed the leave-one-group-out strategy as [10].
- UCF101 [27] is an extension of UCF50 and has 101 actions with 13,320 samples. To our best knowledge, it has been the largest dataset so far. We followed the three train-test splits released in Thumos'13 challenge.

4.1.2 Constrained Action Dataset

Two datasets were used for cross-dataset scenario.

- KTH [5] contains six categories by 25 subjects in four different scenarios. We followed the "split" strategy as [4].
- TJU [7] contains 22 actions with 1,760 samples in both light and dark environment. We selected the RGB data of six actions, identical to KTH, in our experiment. We followed the "split" strategy as [7].

4.1.3 Multi-View Action Dataset

Two datasets were used for cross-view scenario.

- MV-TJU [7] contains 22 actions with 7,040 samples. We selected the RGB/depth data of all classes from both views for our experiment. We followed the "split" strategy as [7].
- IXMAS [6] is the well-known multi-view action dataset. It contains 11 actions with 1,148 sequences. We selected four of five views (excluding the top-down view) in our experiment. The leave-one-person-out strategy was implemented.

4.2 Implementation Details

For RGB data, STIP with the Harris3D detector and the HoG/HoF descriptor [28] was implemented for space-time interest point extraction and description since it has been demonstrated that this configuration can outperform other detectors and descriptors [4]. For depth data, the popular DSTIP [29] was used for interest point detection and description. We utilized the original implementation and parameter settings of both for fair comparison. When evaluating on each dataset, we implemented K-means for codebook generation. Then the popular BoW framework can be leveraged for feature representation. The dimensions of BoW features for UCF Sport, YouTube, Hollywood2, and UCF101 were set as 4,000 as in [4], [27] and the dimension of BoW for UCF50 and HMDB51 was set as 2,000 as in [11], [26]. The dimensions of BoWs for KTH and TJU were set as 4,000 as in [4]. The dimensions of BoWs for MV-TJU (RGB/depth) and IXMAS were set as 1,000. The best α & β were selected by cross validation. The best iteration number was selected when the difference of accuracy between two adjacent iterations is less than a pre-set threshold. The average accuracy over all classes were computed for evaluation.

5 Experimental Results

In our experiments, we first evaluated HC-MTL on the School dataset. Second, it is evaluated on the 6 realistic datasets since the action categories in these realistic datasets are too complicated to be grouped based on human knowledge before action modeling. Therefore, the proposed method is expected to boost the performance by discovering the task relatedness. Specifically, we implemented the sensitivity study on the parameter selection for low rank structure and group sparsity structure as well as the iteration number of HC-MTL. Third, we evaluated HC-MTL on two cross-domain scenarios: 1) cross-dataset: we selected the identical action categories from KTH and TJU to test whether the proposed method can achieve the same grouping results on two different datasets; 2) cross-view: we selected the multi-view datasets, MV-TJU and IXMAS, to test whether the proposed method can achieve the same grouping results on different views. Finally, to explicitly show the superiority of HC-MTL, we directly compared HC-MTL against CMTL on both action grouping and modeling.

TABLE 2
Performance Comparison

Measure	Training Ratio	Ridge	Lasso	Trace	CMTL1	CMTL2	RMTL	HC-MTL
nMSE	10%	1.0398±0.0038	1.0261±0.0132	0.9359±0.0370	1.0850±0.0206	0.9413±0.0021	0.9130±0.0039	**0.8838±0.0021**
	20%	0.8773±0.0043	0.8754±0.0194	0.8211±0.0032	0.8864±0.0094	0.8327±0.0039	0.8055±0.0103	**0.7726±0.0015**
	30%	0.8171±0.0090	0.8144±0.0091	0.7870±0.0012	0.8006±0.0081	0.7922±0.0052	0.7600±0.0032	**0.7378±0.0041**
aMSE	10%	0.2713±0.0023	0.2682±0.0036	0.2504±0.0102	0.2831±0.0050	0.2552±0.0032	0.2330±0.0018	**0.2173±0.0011**
	20%	0.2303±0.0003	0.2289±0.0051	0.2156±0.0015	0.2322±0.0022	0.2131±0.0071	0.2018±0.0025	**0.1871±0.0033**
	30%	0.2165±0.0021	0.2137±0.0012	0.2089±0.0012	0.2101±0.0016	0.1922±0.0102	0.1822±0.0014	**0.1696±0.0067**

The smaller values of nMSE and aMSE represent better performances. The performances of Ridge, Lasso, Trace, CMTL1, CMTL2, and RMTL are cited from [12], [25].

5.1 Performance on School Dataset

We randomly selected 10, 20, and 30 percent of the samples from each task to form the training set and used the rest of the samples as the test set as [12], [25]. We compared HC-MTL to six representative methods from [12], [25], which include ridge regression (Ridge), least squares with L1-norm regularization (Lasso), least squares with trace norm regularization (Trace), clustered multi-task learning (CMTL1) [12], convex multi-task feature learning (CMTL2) [30], and robust multi-task learning (RMTL) [25]. The evaluation was performed with the normalized mean squared error (nMSE) and the averaged means squared error (aMSE). For each training ratio, both nMSE and aMSE were averaged over 15 random splittings of training and test sets. The parameters of HC-MTL were selected by cross validation. The comparison in Table 2 demonstrates that HC-MTL can consistently outperform all competing methods in terms of nMSE and aMSE by discovering the latent relatedness among multiple tasks to benefit task modeling and prediction.

5.2 Performance on Realistic Datasets

As shown in Table 3, HC-MTL regularized by both low rank structure and group sparsity structure can achieve 78.5, 95.0, 89.7, 93.2, 51.4, and 76.3 percent overall accuracy on Hollywood2, UCF Sport, YouTube, UCF50, HMDB51, and UCF101 respectively. The corresponding α, β, and the optimal iteration numbers are shown in Table 3. Sections 5.3 and 5.4 will detail the selection of α, β, and the optimal iteration number. From Table 3, there are three important observations made:

1) HC-MTL+L/S Reg can outperform all the other methods. Especially, the proposed method can achieve notably large gains of 20.3, 7.0, 5.6, 8.7, 29.4, and 32.4 percent over "BoW+SVM" on Hollywood2, UCF Sports, YouTube, UCF50, HMDB51, and UCF101 respectively. The superior performance can be explained by two reasons: 1) The proposed method can leverage the low rank and group sparsity regularizations to jointly discover the relatedness among multiple tasks and learn the discriminative features in each iteration and then group the tasks into several clusters with strong relationships for joint learning iteratively. Therefore, it can discover more latent common knowledge and avoid transferring the negative information between groups in the hierarchical manner; 2) The proposed MTL framework indirectly increases the amount of training data by multiple task learning and the knowledge sharing mechanism. Comparing to the traditional MTL (like CMTL [12]), the proposed method can further augment the training data by the hierarchical approach. The performance can be improved by discovering the reasonable task clusters iteratively.

2) HC-MTL with three kinds of regularizations can work better than HC-MTL+No Reg since the three kinds of regularizations can aid relatedness discovery and/or feature learning and avoid transfer of negative information in between, which might exist in the traditional MTL without regularization by treating each task equally.

3) The comparison between the fourth and fifth columns in Table 3 shows the results obtained by HC-MTL+L Reg outperform those obtained by HC-MTL+S Reg. It implies that the low rank regularization is usually more important than the group sparsity regularization and consequently the discovery of task relationship is usually more essential than discriminative feature learning to augment the performance.

To demonstrate the superiority of HC-MTL, we further compare it against the state of the arts. There are two important observations made:

1) From Table 4, the proposed method can outperform most of the competing methods on Hollywood2, UCF Sports, YouTube, and UCF50 except [49] with the improved dense trajectory and the novel encoding method of stacked Fisher vectors. The improvement of the proposed method over the others is quite reasonable. The state-of-the-art

TABLE 3
Performance Comparison on Six Realistic Datasets (#OI:Optimal Iteration Number; No Reg: no Regularization; L Reg: Low Rank Structure; S Reg: Group Sparsity Structure; L/S Reg: Both Regularizations)

Dataset	BoW+SVM	HC-MTL+No Reg	HC-MTL+L Reg	HC-MTL+S Reg	HC-MTL+L/S Reg
Hollywood2	58.2 [31]	58.7 ($\alpha=0,\beta=0$)	71.6 ($\alpha=10^{-1},\beta=0$)	67.4 ($\alpha=0,\beta=10^{-3}$)	**78.5** ($\alpha=10^{-1},\beta=10^{-3}$, # OI=3)
UCF Sport	88.0 [31]	88.1 ($\alpha=0,\beta=0$)	91.4 ($\alpha=10^{-4},\beta=0$)	89.5 ($\alpha=0,\beta=10^{-3}$)	**95.0** ($\alpha=10^{-4},\beta=10^{-3}$, # OI=3)
YouTube	84.1 [31]	83.7 ($\alpha=0,\beta=0$)	87.4 ($\alpha=10^{-2},\beta=0$)	85.5 ($\alpha=0,\beta=10^{-2}$)	**89.7** ($\alpha=10^{-2},\beta=10^{-2}$, # OI=2)
UCF50	84.5 [31]	84.2 ($\alpha=0,\beta=0$)	88.6 ($\alpha=5\times10^{-2},\beta=0$)	86.2 ($\alpha=0,\beta=10^{-3}$)	**93.2** ($\alpha=5\times10^{-2},\beta=10^{-3}$, # OI=3)
HMDB51	22.0 [26]	24.6 ($\alpha=0,\beta=0$)	46.1 ($\alpha=10^{-2},\beta=0$)	44.2 ($\alpha=0,\beta=10^{-3}$)	**51.4** ($\alpha=10^{-2},\beta=10^{-3}$, # OI=4)
UCF101	43.9 [27]	44.6 ($\alpha=0,\beta=0$)	73.8 ($\alpha=10^{-2},\beta=0$)	71.5 ($\alpha=0,\beta=10^{-3}$)	**76.3** ($\alpha=10^{-2},\beta=10^{-3}$, # OI=4)

TABLE 4
Accuracy (Acc) Comparison on Hollywood2, UCF Sports, YouTube, UCF50, KTH, IXMAS (%)

Hollywood2	Acc	UCFSports	Acc	YouTube	Acc	UCF50	Acc	KTH	Acc	IXMAS	Acc
Sun [32]	48.1	Sun [32]	86.6	Wang [31]	84.1	Wang [31]	84.5	Ji [33]	90.2	Junejo [34]	79.6
Zhang [35]	50.9	Zhang [35]	87.5	Wu [36]	87.0	Wu [37]	86.0	Wang [4]	92.1	Tran [38]	80.2
Wang [39]	64.3	Wu [40]	90.7	Yang [41]	88.0	Wu [36]	88.0	Sun [32]	93.1	Wu [42]	85.6
Hoai [43]	73.6	Narayan [44]	92.8	Xu [45]	89.3	Narayan [44]	89.4	Zhang [35]	94.8	Wu [46]	88.2
Fernando [47]	73.7	Sandanand [48]	95.0	Peng [49]	93.8	Wang [39]	91.2	Wu [40]	97.0	Wang [31]	93.5
Proposed	78.5	Proposed	95.0	Proposed	89.7	Proposed	93.2	Proposed	94.3	Proposed	94.7

literatures mentioned in Section 2 usually focus on feature representation or model learning. However, not much work has been done to share knowledge for multiple related tasks to improve the generalization ability. Although Zhou et al. [2] designed a multi-task learning method to learn latent tasks shared across categories and reconstructed an action classifier for each category with these latent tasks, they did not discriminate the groups of multiple actions and only achieved 80.2 percent overall accuracy on UCF50. In comparison, the proposed method can achieve action grouping as well as model learning simultaneously and consequently the grouping information can induce model learning of tasks, which are strongly correlated with each other and avoid the negative information transferring between unrelated tasks. Therefore, the proposed method can effectively augment the classification performance.

2) Since HC-MTL with BoW visual representation aims to discover the latent grouping information in the framework of multi-task learning to benefit action modeling, its main counterpart includes the single task learning methods [26,47-51] as well as BoW+SVM. For fair comparison, we also utilized STIP for BoW+SVM (STIP+BoW). We notice that in recent years the performances on HMDB51 and UCF101 have been significantly improved with the discriminative feature of improved dense trajectory (iDT) and the powerful feature encoding method of Fisher vector (FV) [39], [49]. Moreover, the recent work on deep learning further augments the performance on large-scale realistic action recognition [57]. For fair comparison, we mainly compared HC-MTL to those without using iDT or FV and only list the performances of three representative cases (STIP+FV, iDT+BoW, iDT+FV) for comparison. From Table 5, HC-MTL can outperform the competing methods [26,47-51] and STIP+BoW. It can even beat STIP+FV and produce competing performances to iDT+BoW. However, HC-MTL still does worse than iDT+FV. Therefore, it will be promising to integrate both CNN features and iDT [57] for HC-MTL learning in our future work.

TABLE 5
Comparison on HMDB51 & UCF101 (%)

HMDB51	Acc	UCF101	Acc
STIP+BoW [26]	22.0	STIP+BoW [27]	43.9
STIP+FV [50]	38.8	STIP+FV [50]	69.9
iDT+BoW [39]	52.1	iDT+BoW [50]	76.8
iDT+FV [39]	57.2	iDT+FV [51]	85.9
Wang [52]	42.1	Wang [53]	54.7
Jiang [54]	40.7	Burghouts [55]	63.5
Wang [31]	48.3	Cho [56]	65.9
Proposed	51.4	Proposed	76.3

We quantitatively analyze and evaluate the performance of realistic action grouping. We refer to the action categories of HMDB51 [26] and UCF101 [27] to set the ground truth of action grouping since there is no such ground truth for the realistic datasets. In our work, we mainly clustered actions into two categories, Only Body Movement & Body Movement with Objection Interaction, due to three reasons: 1) it is extremely challenging to come up with a reasonable ground truth with more action groups since there exists complicated human-object interaction and significant intra-class variation with different motion direction, intensity, speed etc.; 2) it is extremely challenging to preserve the consistence between the subjective action grouping with human knowledge and the objective action grouping based on the feature distribution; 3) The current action categories in either UCF101 or HMDB51 have overlaps to some extent. For example, in UCF 101, Diving & Punch from "Sports" belong to Only Body Movement while all actions from "Playing Musical Instruments" belong to Body Movement with Objection Interaction. Therefore, we selected Only Body Movement & Body Movement with Objection Interaction, to cover all the categories in HMDB51 and UCF101. In our experiment, we respectively achieved 75, 70, 64, 68, 71, 64 percent on Hollywood2, UCF Sports, YouTube, UCF50, HMDB51, and UCF101. We further implemented HC-MTL on UCF101 with two settings to test how the task number affects task grouping and modeling: 1) HC-MTL was implemented on 101 actions of UCF101; 2) HC-MTL was implemented on 50 actions of UCF101, which are identical to UCF50. Then we analysis the grouping results with these 50 actions. The experiment shows that more actions can increase the chances to discover the relationships to benefit grouping and consequently boost the performance 72.3 percent by the second setting to 74.5 percent by the first setting.

Three representative group results were observed - (1) Consistent Grouping, (2) Not Consistent Grouping, (3) False Grouping:

1) The grouping results are consistent with the ground truth. The proposed method can discover the common motion patterns and benefit inducing the correct grouping results. For example, RideBike/RideHorse (Fig. 3a) are correctly clustered into one group mainly due to the similar motion pattern while PlayGuitar/PlayViolin (Fig. 3b) are correctly clustered into one group mainly due the similar gesture and human-object interaction.

2) The grouping results are not consistent with the ground truth while the ground truth is not quite reasonable. The proposed method can discover the common motion patterns (facial actions of Chew/Drink and human-human interaction of ShavingBeard/HeadMassage in (Figs. 3c and

Fig. 3. Grouping results. The samples in the first & second lines are from HMDB51 and UCF101 respectively.

3d)) and respectively group them together. However, depending on the ground truth, Chew/Drink (ShavingBeard/HeadMassage) are from different groups due to the existence of human-object interaction or not. Actually, the ground truth is not reasonable since the local interactive action between the hand and the cup/hair clipper is not salient enough comparing to the global facial/body motion. Comparatively, the proposed method can not discover the common motion patterns (human-object interaction of ShootBow/ShootBall and Hammering/HammerThrow (Figs. 3e and 3f)) and respectively group them into different groups. However, depending on the ground truth, ShootBow/ShootBall (Hammering/HammerThrow) are from the same groups due to the existence of human-object interaction. Again, the ground truth is not reasonable since ShootBow and Hammering mainly focus on the local motion of arms and hands while ShootBall and HammerThrow mainly focus on the salient motion of entire human body. Therefore, the grouping results by the proposed method are more reasonable from our viewpoint. Both cases show that it is extremely challenging to come up with a reasonable grouping ground truth for realistic action datasets as we mentioned above. More related knowledge is required for proper ground truth setting.

3) The proposed method fails to discover the common knowledge and leads to the false grouping. For example, the proposed method clustered Smile/Laugh in Fig. 3g (ShaveBeard/BrushTeeth in Fig. 3h) into different groups. However, they should be grouped into the same group based on the ground truth since Smile/Laugh have the similar facial action and ShaveBeard/BrushTeeth have similar human-object interaction although there may exist significantly different motion intensity, orientation, and so on. This failure case indicates that it is essential to develop more discriminative visual features and more powerful modeling methods for human action grouping and modeling.

5.3 Sensitivity Study on Regularization Parameters

For this experiment, we first fix β with 10^{-4} to minimize the effect of the group sparsity regularization and vary α within different ranges for six realistic datasets to study how the weight of the low rank regularization affects the performance in terms of different iteration numbers. From Figs. 6a, 6c, 6e, 6g, 6i, and 6k, the optimal performances of six datasets can be respectively achieved when $\alpha = 10^{-1}, 10^{-4}, 10^{-2}, 5\times 10^{-2}, 10^{-2}, 10^{-2}$. Then we fix α with the corresponding optimal α and vary β within $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$ to study how the weight of the group sparsity regularization affects the performance in terms of different iteration numbers. From Figs. 6b, 6d, 6f, 6h, 6j, and 6l, the optimal performances of six datasets can be respectively achieved when $\beta = 10^{-3}, 10^{-3}, 10^{-2}, 10^{-3}, 10^{-3}, 10^{-3}$. The performance will decrease when either α or β become too large or too small. We can consistently get this trend in terms of any iteration number. Moreover, when the optimal performance is obtained, α is usually larger than or equal to β on most datasets (except UCF Sports), which denotes that the low rank regularization is more important than the group sparsity regularization. This conclusion is theoretically acceptable since the low rank regularization forces the tasks with strong relationship to obey the low rank structure for joint learning, which will directly affect the performance of model learning. It is also experimentally consistent with the comparison results between HC-MTL+L Reg and HC-MTL+S Reg in Table 3. UCF Sports contains complicated motion patterns of sport actions and action grouping is not explicit enough to be discovered simply by task relatedness discovery. Comparatively, group-wise discriminative feature learning becomes fundamental and plays significant role for action grouping and learning. Therefore, β is a little larger than α for UCF Sports.

5.4 Sensitivity Study on Iteration Number

Fig. 4 shows the optimal iterations for YouTube, Hollywood2/UCF Sports/UCF50, HMDB51/UCF101 are 2, 3, 4 respectively. Since the proposed method can cluster the tasks into sub-groups with strong relatedness iteratively, the performance can increase before achieving the upper bound, which corresponds to the optimal iteration. However, the performance will decrease after the peak value is obtained. This trend is quite understandable. HC-MTL in the first iteration works in the classic MTL framework (Fig. 1b) to jointly learn all tasks. By increasing the iteration numbers, more reasonable task relatedness will be discovered to boost the performance. When the iteration number approaches infinity, HC-MTL will over-divided the tasks by treating each task as individual group and work in the classic STL framework (Fig. 1a) for model learning individually. We implemented HC-MTL in the STL manner and BoW +SVM on HMDB51 and UCF101. HC-MTL achieved 21.7%/

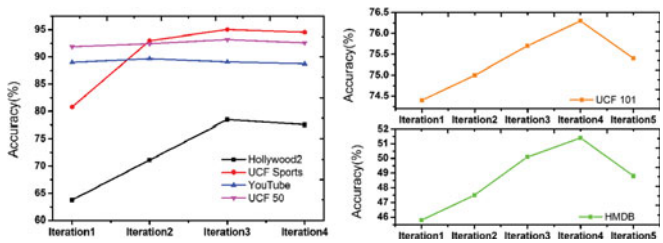

Fig. 4. Comparison by varying the iteration number.

TABLE 6
Comparison on Cross-Domain Scenarios (%)

Dataset	BoW+SVM	Proposed
KTH	91.8	94.3
TJU	94.8	96.9
MV-TJU-View1 (RGB/Depth)	84.8/78.1	85.2/80.3
MV-TJU-View2 (RGB/Depth)	85.8/79.2	86.6/81.3
IXMAS-Camera0	93.0	96.1
IXMAS-Camera1	93.0	95.6
IXMAS-Camera2	88.2	92.4
IXMAS-Camera3	90.0	94.7

37.4% on HMDB51/ UCF101, which is lower than the performances by BoW+SVM (22.0%/43.9% on HMDB51/ UCF101). Therefore, HC-MTL in the STL manner can output the lower bound since it may lose the transferable knowledge for model learning caused by over-dividing tasks.

5.5 Performance on Constrained Datasets

5.5.1 Cross-Dataset Scenario

Six identical action categories are selected from both KTH and TJU for the evaluation of cross-dataset scenario. We respectively implemented HC-MTL on KTH and TJU to classify the six actions into two clusters. To construct the objective ground truth, we adopted the attribute-based action grouping method to avoid subjective classification simply based on human knowledge. Specifically, we kept 40 distinct attributes from the attribute lists of the MIXED action dataset and the UIUC action dataset [58]. Then the 40-D attribute annotation of individual action can be considered as the semantic representation of this action for action grouping. The grouping output can be considered as the ground truth for task clustering. From Table 7, the action grouping results by HC-MTL on both KTH and TJU are consistent with the ground truth. Consequently, the proposed method can augment the performance on both KTH and TJU, when comparing to the single task method (BoW +SVM) as shown in Table 6. The performance by HC-MTL on KTH is also compared against the state of the arts in Table 4. Note that Wu et al. [40] achieved 97 percent by adopting the leave-one-subject-out strategy, which is higher than the performance of HC-MTL (94.3 percent) with the standard split strategy.

5.5.2 Cross-View Scenario

We first tested the proposed method on MV-TJU. We implemented HC-MTL on the RGB/depth modality from both views to classify the 22 actions into two clusters and constructed the ground truth by clustering the attribute-based action representation as stated above. From Table 7, the RGB/depth grouping results of both views of MV-TJU are consistent with the ground truth. Consequently, HC-MTL can effectively discover the task grouping information in the cross-view scenario and improve the performance on both views of MV-TJU comparing to the BoW+SVM method as shown in Table 6.

TABLE 7
Comparison on Action Grouping

KTH&TJU	HC-MTL		CMTL	
	KTH	TJU	KTH	TJU
Accuracy	**6/6**	**6/6**	4/6	3/6
MV-TJU	HC-MTL		CMTL	
	View1 (RGB&D)	View2 (RGB&D)	View1 (RGB&D)	View2 (RGB&D)
Accuracy	**22/22 & 22/22**	**22/22 & 22/22**	11/22 & 10/22	11/22 & 12/22

IXMAS	HC-MTL				CMTL			
	C0	C1	C2	C3	C0	C1	C2	C3
Accuracy	**9/11**	**9/11**	**5/11**	**9/11**	6/11	7/11	4/11	5/11

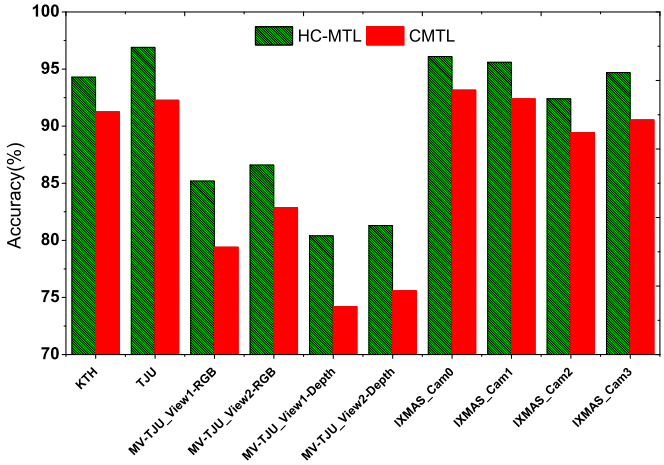

Fig. 5. Comparison between HC-MTL & CMTL (%).

To increase the challenge, we utilized the popular IXMAS dataset with multiple views to evaluate the cross-view scenario.[1] We implemented the same procedure for action grouping and ground truth construction. From Table 7, it can be seen that the proposed method can achieve satisfactory results by correctly classifying 9/11 actions in Camera-0, Camera-1, and Camera-3 respectively except only correctly classifying 5/11 actions in Camera-2. The comparison results in Table 6 further demonstrated the proposed method can consistently outperform BoW +SVM on all of the views by leveraging on the grouping information discovered by the proposed method. Table 4 shows that HC-MTL can outperform the competing methods on IXMAS.

5.6 Comparison between HC-MTL and CMTL

Both methods are tested on the cross-dataset and cross-view scenarios to evaluate action grouping and recognition. As shown in Fig. 5 and Table 7, HC-MTL can consistently outperform CMTL in terms of both action recognition and grouping due to three main reasons: 1) HC-MTL implements clustering on the group sparse component, S, which is more discriminative for task grouping comparing to the task model, W, directly used by CMTL for task grouping; 2) HC-MTL adopts spectral clustering for action grouping, which is more robust to noise and outliers of natural data than K-means utilized in CMTL [59]; 3) The proposed method for the optimization of HC-MTL has analytical solution while the objective function of CMTL is non-convex and the convex relaxation during optimization can cause false correlation [12]. The comparison between the action

1. We only utilized the four side views (Camera0-3) and excluded the top-down view (Camera-4) since it has obviously different visual characteristics and is not discriminative enough for reasonable action grouping.

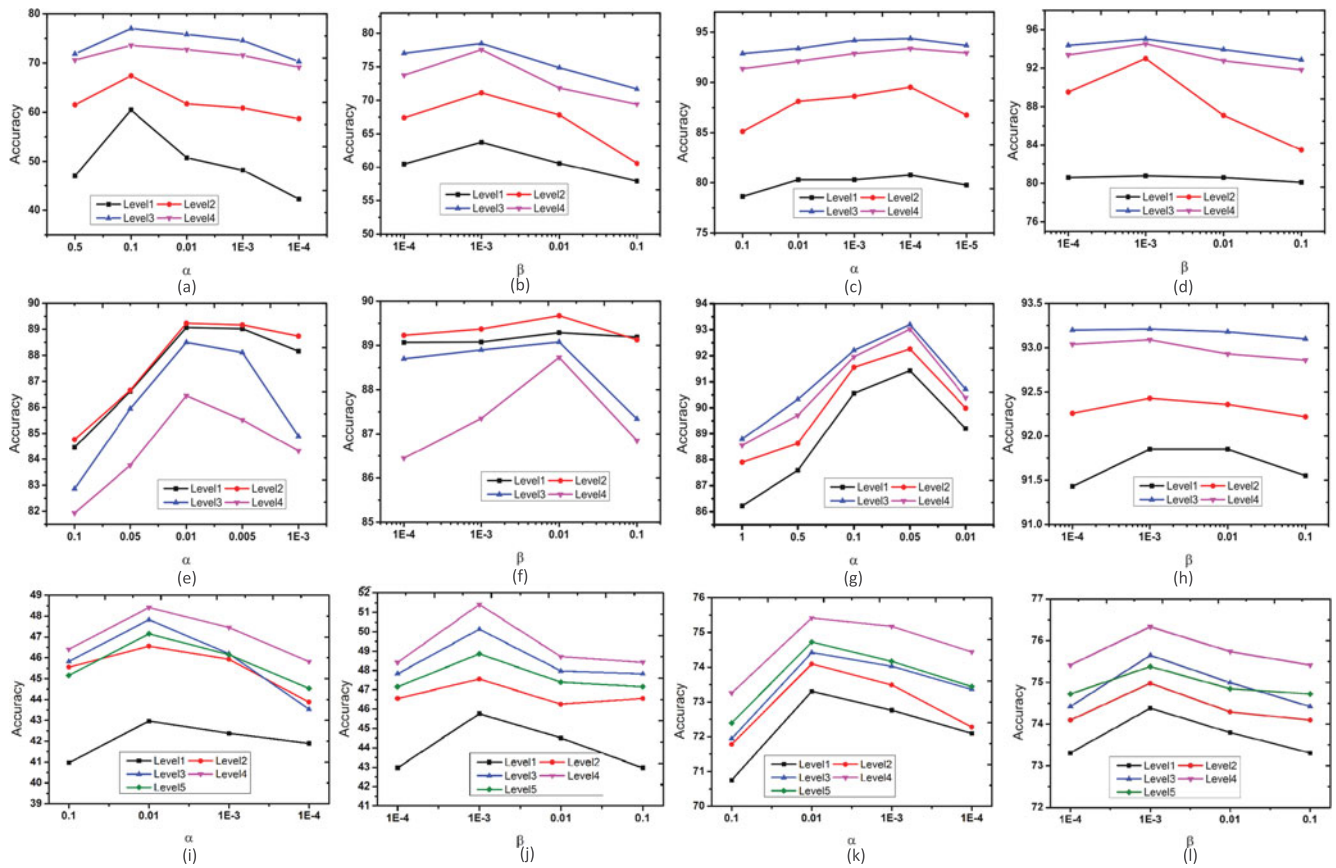

Fig. 6. Performance comparison by varying α & β on Hollywood2(a-b), UCF Sports(c-d), YouTube(e-f), UCF50(g-h), HMDB51(i-j), UCF101(k-l) (%).

grouping results in Table 7 leads to three important observations:

- In the cross-dataset scenario (KTH & TJU), HC-MTL clustered actions in both datasets into the limb-based actions (Action 1-boxing/2-handwaving/3-handclapping) and the leg-based actions (Action 4-walking/5-jogging/6-running) while CMTL clustered the KTH actions into two categories (Action 1,3,4; Action 2,5,6) and clustered the TJU actions into two categories (Action 1,2,4,6; Action 3,5), which are unexplainable based on intuitive human knowledge. Comparing against CMTL, not only can HC-MTL achieve consistent clustering on both datasets but also discover the latent knowledge of salient motion region for the task relatedness measure and further leverage the grouping information to boost the performance.
- In the cross-view scenario (MV-TJU & IXMAS), HC-MTL can cluster actions in RGB/depth modality from both datasets into the limb-based actions and the leg-based actions while CMTL can not achieve such reasonable and consistent groupings. As compared to CMTL, HC-MTL can effectively avoid the negative transferred knowledge by unreasonable task grouping (e.g., it is not at all clear why Turn-around and Punch in IXMAS are classified into the same group by CMTL).
- Both HC-MTL and CMTL fail to achieve satisfactory performance for both action recognition and grouping on Camera-2 of IXMAS although HC-MTL can work slightly better than CMTL as shown in Tables 6 and 7. Camera-0 and Camera-1 almost capture the front view of multiple actions and Camera-3 almost records the side view of multiple actions. Little occlusion happens around the hand/arm/leg/foot regions, which are directly related to the part-wise attributes. Therefore, the extracted visual feature in individual camera setting can effectively represent the part-wise motion characteristics (part-wise attributes) for action classification. Comparatively, the Camera-2 setting in the top-left of an actor leads to the drastic affine variation of human body and especially this view setting will lead to the occlusion of part-wise motion and indirectly have negative influence to represent the related attributes. Therefore, the extracted visual feature in this view cannot properly represent the motion pattern and the task grouping results by HC-MTL are not as good as those in Camera-0,1,3. Consequently, the unsatisfactory grouping information discovered by HC-MTL on Camera-2 cannot improve the performance of action recognition much as it does in case of Camera-0,1,3.

6 Conclusion

In this paper, we propose a hierarchical clustering multi-task learning method for joint human action grouping and recognition. We formulate the objective function into the group-wise least square loss regularized by low rank and sparsity with respect to two latent variables, model parameters and grouping information, for joint optimization.

Furthermore, we decompose it into two correlated subtasks, multi-task learning and task relatedness discovery and utilize the APM and TRD methods to solve them respectively. By alternating between them iteratively, the proposed method can achieve the optimal action models as well as group discovery. The extensive experimental results on six realistic action datasets, two constrained datasets, and two multi-view datasets show that this method can outperform the competing methods for both action grouping and recognition. Moreover, the direct comparison against the popular CMTL further shows its superiority.

In the current experiments, we utilized the BoW feature for visual representation similar to most of previous methods. The experiments on large-scale realistic action datasets (HMDB51 & UCF101) reveal that it is necessary to develop discriminative visual feature with the success of deep learning to benefit action learning in our future work. Furthermore, with the novel multimodal and multiview dataset (MV-TJU), it would be useful and challenging to extend the proposed method to jointly leverage the multimodal and multiview information to discover even more latent correlations among different actions to augment the performance.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (61472275, 61100124, 61502337), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of China Scholarship Council (201506255073), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046). Mohan's research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. An-An Liu is the corresponding author of this paper.

REFERENCES

[1] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-task sparse learning with beta process prior for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 423–429.

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

[3] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and S. J. Maybank, "Learning human actions by combining global dynamics and local appearance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2466–2482, Dec. 2014.

[4] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

[5] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 432–439.

[6] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[7] A. Liu, Y. Su, P. Jia, Z. Gao, T. Hao, and Z. Yang, "Multipe/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, Jun. 2015.

[8] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 2929–2936.

[9] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[10] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1996–2003.

[11] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vis. Appl.*, vol. 24, no. 5, pp. 971–981, 2012.

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

[13] J. Chen, J. Zhou, and J. Ye, "Low-rank and sparse multi-task learning," *Low-Rank and Sparse Modeling for Visual Analysis*. New York, NY, USA: Springer, 2014, pp. 151–180.

[14] X. Zhang, "Convex discriminative multitask clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 28–40, Jan. 2015.

[15] Y. Gao, R. Ji, W. Liu, Q. Dai, and G. Hua, "Weakly supervised visual dictionary learning by harnessing image attributes," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5400–5411, Dec. 2014.

[16] Z. Gao, H. Zhang, G. P. Xu, Y. B. Xue, and A. G. Hauptmann, "Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition," *Signal Process.*, vol. 112, pp. 83–97, 2015.

[17] R. Ji, L. Duan, J. Chen, T. Huang, and W. Gao, "Mining compact bag-of-patterns for low bit rate mobile visual search," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3099–3113, Jul. 2014.

[18] A. Liu, Y. Su, Z. Gao, T. Hao, and Z. Yang, "Single/multi-view human action recognition via regularized multi-task learning," *Neurocomputing*, vol. 151, pp. 544–553, 2015.

[19] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute regularization based human action recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1600–1609, Oct. 2013.

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

[21] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. New York, NY, USA: Springer, 2014.

[22] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.

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

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[27] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

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

[29] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2834–2841.

[30] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[31] H. Wang, A. Klaer, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, 2013.

[32] L. Sun, K. Jia, T. Chan, Y. Fang, G. Wang, and S. Yan, "DL-SFA: Deeply-learned slow feature analysis for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2625–2632.

[33] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

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

[35] H. Zhang, W. Zhou, C. M. Reardon, and L. E. Parker, "Simplex-based 3d spatio-temporal feature description for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2067–2074.

[36] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural SVM," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1422–1431, Aug. 2013.
[37] J. Wu and D. Hu, "Learning effective event models to recognize a large number of human actions," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 147–158, Jan. 2014.
[38] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 548–561.
[39] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.
[40] B. Wu, C. Yuan, and W. Hu, "Human action recognition based on context-dependent graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2609–2616.
[41] X. Yang and Y. Tian, "Action recognition using super sparse coding vector with spatio-temporal awareness," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 727–741.
[42] D. Wu and L. Shao, "Multi-max-margin support vector machine for multi-source human action recognition," *Neurocomputing*, vol. 127, pp. 98–103, 2014.
[43] M. Hoai and A. Zisserman, "Improving human action recognition using score distribution and ranking," in *Proc. 12th Asian Conf. Comput. Vis.*, 2014, pp. 3–20.
[44] S. Narayan and K. R. Ramakrishnan, "A cause and effect analysis of motion trajectories for modeling actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2633–2640.
[45] X. Xu, I. Tsang, and D. Xu, "Soft margin multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 749–761, May 2013.
[46] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 489–496.
[47] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5378–5387.
[48] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 1234–1241.
[49] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked Fisher vectors," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 581–595.
[50] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *arXiv:1405.4506*, 2014.
[51] H. Wang and C. Schmid, "LEAR-INRIA submission for the THUMOS workshop," in *Proc. ICCV Workshop THUMOS*, 2013, pp. 1–3.
[52] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3d parts for human motion recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2674–2681.
[53] F. Wang, X. Li, and W. Shu, "Experimenting motion relativity for action recognition with a large number of classes," in *Proc. ICCV Workshop THUMOS*, 2013, pp. 1–2.
[54] Y. Jiang, Q. Dai, X. Xue, W. Liu, and C. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 425–438.
[55] G. Burghouts, P. Eendebak, H. Bouma, and R.-M. ten Hove, "Action recognition by layout, selective sampling and soft-assignment," in *Proc. ICCV Workshop THUMOS*, 2013, p. 1.
[56] H. Cho, H. Lee, and Z. Jiang, "Evaluation of lc-ksvd on ucf101 action dataset," in *Proc. ICCV Workshop THUMOS*, 2013, pp. 1–7.
[57] Z. Xu, L. Zhu, Y. Yang, and A. G. Hauptmann, "UTS-CMU at THUMOS 2015," *CVPR'15 Int. Workshop Competition Action Recog. Large No. Classes*, pp. 1–3, 2015.
[58] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 3337–3344.
[59] S. Jones and L. Shao, "Unsupervised spectral dual assignment clustering of human actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 604–611.

An-An Liu (M'10) received the PhD degree in electronic engineering from Tianjin University, China. He is an associate professor with the School of Electronic Engineering, Tianjin University. His current research interests include computer vision and machine learning. He is a member of the IEEE.

Yu-Ting Su received the PhD degree in electronic engineering from Tianjin University, China. He is a professor with the School of Electronic Engineering, Tianjin University. His research interests include computer vision and machine learning.

Wei-Zhi Nie received the PhD degree in electronic engineering from Tianjin University, China. He is an assistant professor with the School of Electronic Engineering, Tianjin University. His research interests include computer vision and machine learning.

Mohan Kankanhalli (F'14) received the BTech degree from IIT Kharagpur, Kharagpur, India, and the MS and PhD degrees from the Rensselaer Polytechnic Institute, Troy, NY. He is a professor with the School of Computing, National University of Singapore, Singapore. He is the director with the SeSaMe Centre and also the Vice Provost for Graduate Education at NUS. His research interests include multimedia systems and multimedia security. He is active in the Multimedia Research Community. He was the ACM SIGMM director of Conferences from 2009 to 2013. He is on the editorial boards of several journals. He is a fellow of the IEEE.