Multi-Domain and Multi-Task Learning for Human Action Recognition

An-An Liu, *Member, IEEE*, Ning Xu, Wei-Zhi Nie, Yu-Ting Su, and Yong-Dong Zhang, *Senior Member, IEEE*

***Abstract*— Domain-invariant (view-invariant and modality-invariant) feature representation is essential for human action recognition. Moreover, given a discriminative visual representation, it is critical to discover the latent correlations among multiple actions in order to facilitate action modeling. To address these problems, we propose a multi-domain and multi-task learning (MDMTL) method to: 1) extract domain-invariant information for multi-view and multi-modal action representation and 2) explore the relatedness among multiple action categories. Specifically, we present a sparse transfer learning-based method to co-embed multi-domain (multi-view and multi-modality) data into a single common space for discriminative feature learning. Additionally, visual feature learning is incorporated into the multi-task learning framework, with the Frobenius-norm regularization term and the sparse constraint term, for joint task modeling and task relatedness-induced feature learning. To the best of our knowledge, MDMTL is the first supervised framework to jointly realize domain-invariant feature learning and task modeling for multi-domain action recognition. Experiments conducted on the INRIA Xmas Motion Acquisition Sequences data set, the MSR Daily Activity 3D (DailyActivity3D) data set, and the Multi-modal & Multi-view & Interactive data set, which is the most recent and largest multi-view and multi-model action recognition data set, demonstrate the superiority of MDMTL over the state-of-the-art approaches.**

***Index Terms*— Domain-invariant Learning, multi-task learning, human action recognition.**

I. Introduction

HUMAN action recognition [1]–[4] is attracting increasing attention in the field of computer vision because of its various real-world applications, such as human-computer interaction, intelligent video surveillance, and multimedia content understanding and management [5]–[8]. Many approaches have been proposed to advance the development of this field.

A. Motivation

Despite decades of extensive research on related topics, action recognition, especially by fusing multi-view and multi-modality information, remains an active research area [9]–[12]. There are two main challenges: (1) *how to discriminatively represent the multi-domain (multi-view & multi-modality) action patterns* and (2) *how to learn robust classifiers to identify action categories with multi-domain data.* We explain the motivations of this paper in two aspects.

1) Multi-Domain Representation: Since actions contain strong spatiotemporal patterns of appearance or motion, most state-of-the-art approaches rely on discriminative visual representations [13]–[17]. In particular, [15] used a deep learning-based unsupervised feature learning method that performed well on the Hollywood2, UCF, KTH and YouTube datasets. However, the same action looks quite different in different domains (view/modality), as shown in Fig. 1. Thus, action models learned from a single view and a single modality become less discriminative for recognition tasks when view or modality variance exists.

We collectively define *multi-view* and *multi-modality* data as *multi-domain* data, as shown in Fig. 1. Particularly, we consider only the most common visual modalities for actions, including RGB and depth, in this paper. Most existing works on action recognition are based on the RGB modality and are negatively affected by varying illumination conditions, complex backgrounds and camera motion. With the recent advancements in cost-effective depth sensors (e.g., Microsoft Kinect, Asus Xtion and Prime Sense), considerable attention has been paid on the use of depth data for computer vision tasks [18]. Compared with RGB images, depth images are less sensitive to changes in illumination, occlusions, and background clutter. Moreover, RGB information and depth information are complementary [10], [19].

To date, only limited works on multi-domain human action recognition for fusing RGB-D information have been carried out [11], [20], [21]. Kong and Fu [20] factorized the feature matrixes of RGB-D sequences and enforced the same semantics to learn shared features from multi-modal data. Rahmani *et al.* [21] focused on how to detect and describe

Manuscript received September 4, 2017; revised June 19, 2018; accepted September 19, 2018. Date of publication September 28, 2018; date of current version October 22, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61772359, Grant 61472275, Grant 61525206, Grant 61872267, and Grant 61502337, in part by the National Key Research and Development Program of China under Grant 2017YFC0820600, in part by the National Defense Science and Technology Fund for Distinguished Young Scholars under Grant 2017-JCJQ-ZQ-022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gustavo K. Rohde. *(Corresponding authors: An-An Liu; Ning Xu; Wei-Zhi Nie.)*

A.-A. Liu, N. Xu, W.-Z. Nie, and Y.-T. Su are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: anan0422@gmail.com; ningxu@tju.edu.cn; weizhinie@tju.edu.cn).

Y.-D. Zhang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China, and also with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.

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

Digital Object Identifier 10.1109/TIP.2018.2872879

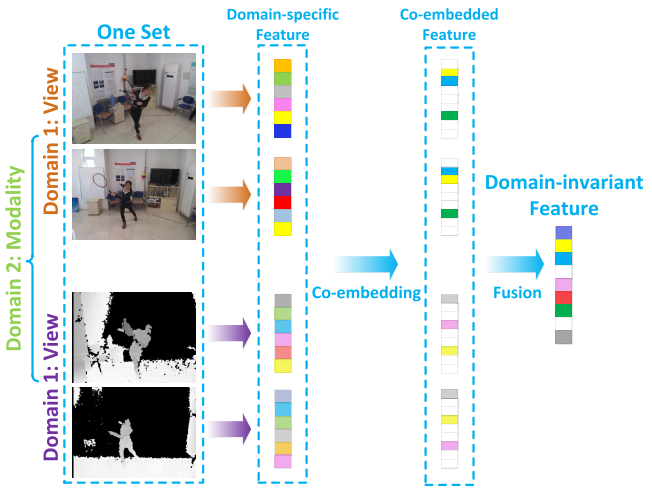


Fig. 1. *Multi-view* and *multi-modality* data are collectively defined as *multi-domain* data. This figure shows the flowchart of *multi-domain* fusion for an example of *Tennis Swing*, from M^2I. We consider all instances (the first column) of one identical action as one set. With the domain-specific feature representations (the second column) of the instances in one set, *multi-domain* information is co-embedded into a common sparse space (the third column), where the white rectangular regions represent the value of "0" in the sparse features. We fuse the co-embedded features and obtain a domain-invariant representation for human action representation.

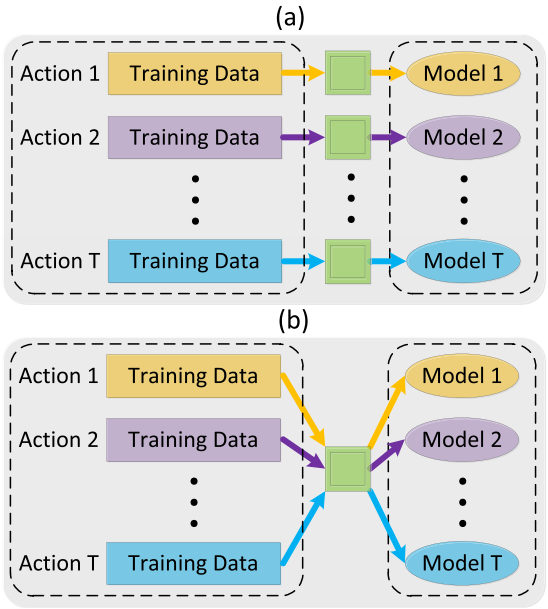


Fig. 2. Comparison of two systematic frameworks: (a) the flowchart of single-task learning (STL) and (b) the flowchart of multi-task learning (MTL). ☐ denotes model learning in either STL or MTL architecture.

spatio-temporal keypoints from 3D pointcloud videos for cross-view action recognition. However, these approaches ignore the domain-invariant information across multiple domains. Overall, the extraction and fusion of information shared across multi-domain sources remain a challenging problem [22], [23].

2) Multi-Task Learning: Most current approaches focus on the single-task learning (STL) problem [17], as shown in Fig. 2(a). The objective is to learn multiple actions (tasks) independently since these methods ignore the action relatedness in between. Therefore, STL is not discriminative enough to enable action classification with similar motion patterns. Recently, several researchers have attempted to employ multi-task learning (MTL) for action recognition [24]–[26]. MTL can leverage the associated knowledge shared among multiple actions for model learning and further improve the generalizability [27], as shown in Fig. 2(b). In reality, it is often the case that there exist some groups of strongly related tasks, while other tasks are weakly correlated [24].

MTL cannot be forwardly utilized for multi-domain action recognition. Multi-domain information can be induced by view variations and modality variations, which can significantly increase the difficulty in action recognition. For example, given a single modality (RGB or depth) and a single view (frontal view or side view), *boxing* and *running* can be easily recognized as two distinct action categories by MTL since they can be represented by salient upper-body motion and lower-body motion, respectively. However, in the case of multiple modalities (RGB and depth), in which each action identification task consists of two modality-variant subtasks (e.g., *boxing* in RGB and depth), confusion can easily arise in the attempt to simultaneously handle both modalities by MTL. Moreover, in the case of multiple views (frontal and side views), in which each task consists of two view-variant subtasks (e.g., *boxing* in the RGB or depth modality with respect to the frontal and side views, respectively), the same problem also exists. Meanwhile, the RGB and depth modalities have their own discriminative characteristics (e.g., illumination, backgrounds, and appearances). Therefore, multi-domain learning based on the classic MTL theory with multi-domain data will have negative impact on performance. Overall, it remains difficult to incorporate the relatedness among actions in multi-domain scenarios.

B. Overview

Motivated by [9], [24], [28], and [29], we propose a multi-domain & multi-task learning (MDMTL) framework to (1) extract domain-invariant information for multi-view and multi-modal action representation and (2) explore the relatedness among multiple action categories. Most current works focus on supervised learning, and each example (instance) is labeled with one action tag. In our framework, an action set is labeled with one action tag. In particular, all instances in the action set are captured from different views & modalities for the same action sample. Each set is associated with a group of feature vectors, which are extracted from multi-domain action instances in this set, as shown in Fig. 1. An ideal learning algorithm should generate classifiers that are capable of classifying previously unseen sets correctly. The key challenge of this framework is to jointly realize domain-invariant feature representation and model learning.

To deal with this problem, we formulate MDMTL into an objective function with two latent components: domain-invariant feature representation and action modeling. First, MDMTL simultaneously learns a group of co-embedding matrixes from individual domains to facilitate direct

transformation of each instance into a common domain-invariant space. We propose an instance-level fusion method for an individual set based on the ensemble-learning strategy. Second, the Frobenius-norm regularization term and the sparse constraint term are incorporated into the MDMTL framework for combined feature learning and relatedness discovery.

We evaluate MDMTL with the multi-view and/or multi-modality human action recognition dataset. Extensive quantitative experiments on the INRIA Xmas Motion Acquisition Sequences (IXMAS), MSR Daily Activity 3D (DailyActivity3D), and Multi-modal & Multi-view & Interactive (M^2I) datasets demonstrate the effectiveness of MDMTL in action recognition with multi-domain data.

C. Contributions

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, MDMTL is the first supervised framework to jointly realize domain-invariant feature learning and latent task relatedness discovery for view-invariant & modality-invariant action recognition.
- We propose a specific objective function for this problem and decompose the solution of this non-convex formulation into two consecutive steps: domain-invariant feature learning and action modeling.
- Comprehensive experiments on the IXMAS [30], DailyActivity3D [31], and M^2I [32] datasets demonstrate the superiority of MDMTL. In particular, we evaluate MDMTL against deep learning methods.

The remainder of the paper is organized as follows. Section II briefly introduces related works. In Section III, we detail the MDMTL framework. Section IV reports the experimental results on three benchmark datasets. Section V concludes the paper.

II. Related Works

The related works can be roughly grouped into two categories: multi-domain representation and multi-task learning (MTL). The former focuses on extracting and fusing information shared across multi-view/multi-modal sources. The latter focuses on leveraging the associated knowledge shared among multiple actions for model learning.

A. Multi-Domain Representation

Researchers have developed various methods for achieving view-variant action recognition, and these methods can be broadly categorized into three types.

Methods of the first type adopt a view-independent approach to determine an appropriate classification scheme. Specifically, classification is performed either by training multiple classifiers [33], [34] or by training a universal classifier using training data of all available views [35]–[39]. Methods of the second type rely on cross-view action recognition, and they learn action classes in one view (often called the reference view) and recognize actions in another (target) view. Several techniques have been adopted for this purpose, including transfer learning [9], [40], [41], information maximization [42], and methods that exploit appropriately designed features [21] and the scene geometry [43]. Methods of the third type exploit view-invariant action representations [12], [44]–[48], which are built based on 2D images acquired by multiple cameras. Some works were proposed, including trajectory extraction [49], [50], the self-similarity matrix [45], and a method [12], [51] that exploits the connection between the source and target views.

At the same time, many researchers have focused on the joint usage of modality-variant information. In this paper, we consider only the most common visual modalities for actions, including RGB and depth, and categorize modality-variant fusion methods into three basic schemes.

First, the heuristic fusion scheme, which is the original framework, is proposed for the integration of multi-modal sources. For instance, Cruz *et al.* [52] proposed the use of the Kinect gaming system, in which different algorithmic modules are selected to extract meaningful information from both the RGB and depth modalities. Second, the assembled fusion scheme attempts to feed multi-modal information into existing formulations in parallel. For instance, Ni *et al.* [53] developed two RGB-D fusion techniques based on two state-of-the-art feature representation methods for action recognition. Third, the adaptive fusion scheme focuses on capturing the relationship between pieces of multi-modal information. For instance, Kong and Fu [10], [20] projected RGB and depth features into a shared space and learned the cross-modal features shared between them for action recognition.

Although the aforementioned fusion schemes have been successfully applied to solve some vision problems, they are not sufficiently sophisticated and adaptive to simultaneously consider view-invariant and modality-invariant feature distributions. Therefore, it is highly desirable to develop a fusion methodology that can leverage a large number of domain-invariant sources. In our framework, view-invariant and modality-invariant instances are collectively defined as multi-domain instances, which are leveraged to form action sets. We proposed the ensemble-learning strategy to quantize the contribution of each instance in one set.

B. Multi-Task Learning

Recently, there has been a growing interest in MTL, which can improve the generalization performance of models by enabling the joint execution of multiple learning tasks using the knowledge shared among them. The effectiveness of MTL has been demonstrated both theoretically [54], [55] and empirically [56], [57]. Furthermore, different MTL methods differ in how the relatedness among tasks is modeled, and these methods can be broadly categorized into two types.

Methods of the first type are based on the assumption that all tasks are related. Evgeniou *et al.* [58] proposed the regularized MTL method, in which the models for all tasks are heuristically constrained to be close to each other. Furthermore, task relatedness can be modeled by constraining multiple tasks to share a common underlying structure [59].

TABLE I
NOTATIONS AND DEFINITIONS

Notation	Definition
X	a group of instance-level feature matrixes
Y	instance-level binary label matrix
P	a group of co-embedding matrixes
S	a group of co-embedded feature matrixes of X
R	weight matrix for multi-domain instances
B	a group of set-level co-embedded set
F	domain-invariant feature matrix

Ando and Zhang [54] proposed a structural learning formulation in which it is assumed that the multiple predictors for various tasks share a common structure in the underlying predictor space. However, tasks may exhibit a more sophisticated group structure. Methods of the second type assume that models for tasks from the same group are more similar to each other than those for tasks from a different group. Many previous works have pursued this direction of research, known as clustered multi-task learning (CMTL) [60], [61]. In [60], the mutual relatedness of tasks was estimated, and knowledge of one task could be transferred to other tasks in the same cluster. Bakker and Heskes [61] used clustered multi-task learning in a Bayesian setting by considering a mixture of Gaussians instead of single Gaussian priors. Furthermore, Chen *et al.* [62] proposed a robust multi-task learning (RMTL) algorithm that captures the relationship among multiple related (inlier) tasks using a low-rank structure and identifies irrelevant (outlier) tasks using a group-sparse structure. Moreover, Gong *et al.* [63] proposed a robust multi-task feature learning (rMTFL) algorithm that simultaneously captures a common set of features among inlier and outlier tasks.

Based on the current methods, we explore the generation of multi-domain classifiers. Different from most of the aforementioned MTL mehtods, our work contributes by studying not only capturing the relatedness among multi-domain action categories but also jointly realizing the domain-invariant feature learning and multi-task modeling for boosting human action recognition.

III. MULTI-DOMAIN & MULTI-TASK LEARNING

The goal of this work is to jointly learn domain-invariant feature representation and discover latent relatedness among multiple actions for multi-domain human action recognition. The notation is presented in Table I, and the framework is illustrated in Fig. 3.

We consider V domains represented by $\{X_i, Y\}_{i=1}^{V}$, where $X_i = \{X_i^{[train]}, X_i^{[val]}\}$ consists of a training feature matrix $X_i^{[train]}$ and a validation feature matrix $X_i^{[val]}$ and Y is the corresponding binary label matrix. Here, each domain corresponds to one specific view and modality. For example, each column of the action sample in Fig. 3 is from one domain. X_i can also be represented by $\{x_{ij}\}_{j=1}^{N}$, where x_{ij} denotes the feature of the j-th instance in the i-th domain.

We define an instance set by $A_j = \{x_{ij}\}_{i=1}^{V}$, which collects the instances of the j-th action from V domains, where j ranges from 1 to N. For example, each row of action samples

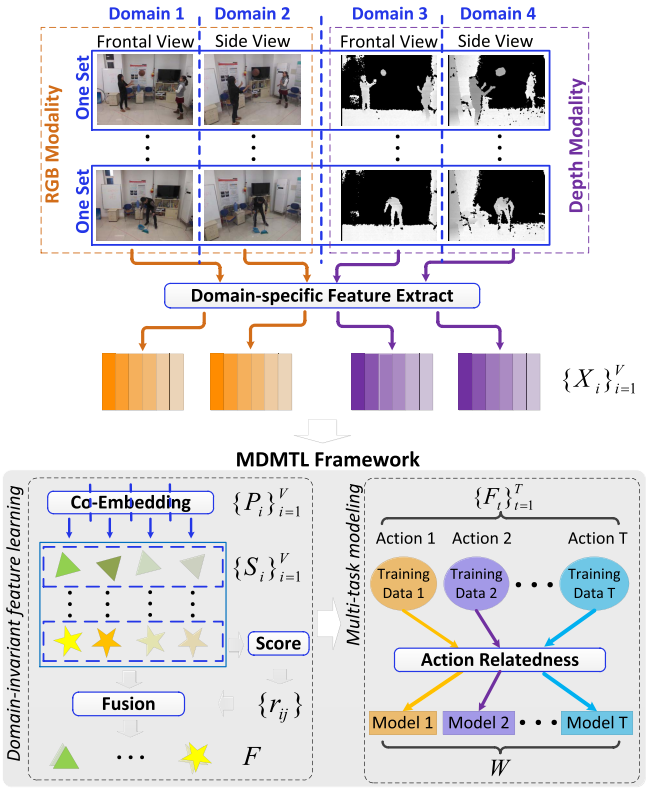


Fig. 3. Illustration of the MDMTL framework. We consider four domains from multi-view & multi-modality sources. Based on domain-specific features $\{X_i\}_{i=1}^{V}$, a group of co-embedding matrixes P are simultaneously learned by forcing multi-domain instances of the same action to have similar sparse representations S. Then, we compute the score R of individual instances and fuse the co-embeded features to obtain domain-invariant representations F. It is fed into the MTL architecture to obtain the multi-domain classifiers.

is an instance set, as shown in Fig. 3. The major challenge in extracting the domain-invariant feature of one instance set is that they are drawn from different domains and thus have different visual characteristics. To solve this problem, we learn a group of co-embedding matrixes $P = \{P_i\}_{i=1}^{V}$ for the feature set of multiple domains X. Each co-embedding matrix P_i maps the corresponding feature matrix X_i to the sparse matrix $S_i = \{s_{ij}\}_{i=1,2,\cdots,V;j=1,2,\cdots,N}$. $S = \{S_i\}_{i=1}^{V}$ constructs the sparse space $\mathbb{O}$ that is shared across multiple domains. At this point, A_j is transformed into $B_j = \{s_{ij}\}_{i=1}^{V}$, where s_{ij} is the corresponding sparse feature of x_{ij}. X is consequently transformed into $B = \{B^{[train]}, B^{[val]}\}$.

The weight matrix $R = \{r_{ij}\}_{i=1,2,\cdots,V;j=1,2,\cdots,N}$ is learned to score each instance with respect to an action category by base classifiers (introduced in Section III-B). We then obtain the domain-invariant feature matrix $F = \{f_j\}_{j=1}^{N}$ by an ensemble-learning strategy with R and B_j.

A. Objective Function

In our framework, there are two closely related components: 1) domain-invariant feature learning, which aims to learn a fusion function that extracts domain-invariant information from multiple domains. For example, MDMTL aims to extract discriminative visual characteristics (e.g., the interactive motion between two persons in *Handshake*) from

view-invariant & modality-invariant domains, and 2) multi-task relatedness learning, which aims to discover the latent relatedness among tasks so that the models of tasks from the same group are more correlated than those from different groups. For example, MDMTL jointly learns related actions (e.g., *Call Cellphone* and *Drink* can be regarded as relevant actions in the sense that both involve salient limb and hand motions) while simultaneously identifying nonrelated actions (e.g., *Chat* and *Handshake* can be considered irrelevant actions based on the aforementioned limb-wise actions). The objective function can be formulated as follows:

$$\begin{aligned} &\arg\min_{W,P,R} \mathcal{L}(\{X_i, Y\}_{i=1}^{V}; W, P, R) + Reg(W) \\ &s.t.\ Cons(W) \end{aligned} \tag{1}$$

where $W = \{w_t\}_{t=1}^{T}$ represents the classifiers for T tasks; w_t is the t-th column of W; P is a group of co-embedding matrixes; R is the weight matrix; and $Reg(\cdot)$ and $Cons(\cdot)$ are the regularization term and the constraint term, respectively, which together aim to impose the task relatedness. This objective function consists of three terms:

1) Domain-Invariant Feature Learning: We design the function $G(X, P, R)$ to realize joint multi-domain feature co-embedding and fusion. A group of co-embedding matrixes P is learned to enable the transformation of a group of multi-domain features X into a group of co-embedded features S. We adopt the ensemble-learning strategy to score each instance and then fuse each instance set into the domain-invariant representation. It can be formulated as follows:

$$G(X, P, R) = \sum_{i=1}^{V} r_{ij}(P_i x_{ij}) \tag{2}$$

where P_i is the co-embedding matrix of the i-th domain and r_{ij} is the weight for instance x_{ij}. The output of $G(\cdot)$ is the domain-invariant feature matrix F.

2) Multi-Task Learning: The main task of MDMTL is to train the classifiers for all tasks by minimizing the joint empirical loss $f(W)$. In particular, least-squares regression is one popular method for classification.

$$f(W) = \|WF - Y\|_F^2 = \sum_{t=1}^{T} \|w_t F_t - y_t\|_F^2 \tag{3}$$

where Y consists of binary label vectors $\{y_i\}_{i=1}^{T}$ and $\|\cdot\|_F$ denotes the ℓ_2-norm (Frobenius norm) of the matrix.

3) Regularization & Constraint: For MDMTL, we utilize the Frobenius-norm regularization, $\|W\|_F^2$, to enhance the robustness of the models. The regularization term can be formulated as

$$Reg(W) = \rho\|W\|_F^2 \tag{4}$$

where the regularization parameter, ρ, controls the importance of the penalty term.

To discover the latent relatedness among multiple action categories, we design a constraint term that can help ensure that a specific number of rows in W will be non-zero; i.e., we control the number of sparse features in the model. The constraint term can be formulated as

$$Cons(W) = \sum_{g} I(\|w^g\| > 0) \le u \tag{5}$$

where w^g is the g-th row of W and $I(\cdot)$ is the indicator function. The constraint term indicates that the number of non-zero rows of W is no larger than u.

To this end, the objective function of MDMTL can be formulated as follows to jointly realize domain-invariant feature learning and multi-task modeling:

$$\begin{aligned} (W^*, P^*, R^*) = \arg\min_{W,P,R} &\sum_{t=1}^{T} \|w_t\{\sum_{i=1}^{V} r_{ij}(P_i x_{ij})\}_t - y_t\|_F^2 \\ &+ \rho\|W\|_F^2 \\ &s.t.\ \sum_{g} I(\|w^g\| > 0) \le u \end{aligned} \tag{6}$$

Algorithm 1 Solution of MDMTL

Input:
modality1: $X^{m1} := \{X_1^{m1}, X_2^{m1}, \cdots, X_{V^{m1}}^{m1}\}$
modality2: $X^{m2} := \{X_1^{m2}, X_2^{m2}, \cdots, X_{V^{m2}}^{m2}\}$
Output:
Action Models $W_{\mathrm{MD}\ (Multi-Domain)}$
Initialization:
Divide X_i based on the training-validation split
$X_i^{m1} := \{X_i^{[train](m1)}, X_i^{[val](m1)}\}$
$X_i^{m2} := \{X_i^{[train](m2)}, X_i^{[val](m2)}\}$
$Y, V := V^{m1} + V^{m2}$
• **Co-Embedding.**
Calculate $(P, S) = \arg\min_{P,S} \sum_{i=1}^{V} \|X_i - P_i S\|_2^2$
$P := \{P^{m1}, P^{m2}\}$
$P^{m1} := \{P_1^{m1}, \cdots, P_{V^{m1}}^{m1}\}$
$P^{m2} := \{P_1^{m2}, \cdots, P_{V^{m2}}^{m2}\}$
Compute $S^{m1} :=$ OMP (P^{m1}, X^{m1})
Compute $S^{m2} :=$ OMP (P^{m2}, X^{m2})
$S := \{S^{m1}, S^{m2}\}$
• **Fusion.**
$R :=$ KNN (S)
Compute $F = \{F^{[train]}, F^{[val]}\}$ by Eq. 8
• **Model Learning.**
Solve $\min_{W} \|WF^{[train]} - Y\|_F^2 + \rho\|W\|_F^2$
$s.t.\ \sum_g I(\|w^g\| > 0) \le u$
by *Algorithm 2*.
Output $W_{\mathrm{MD}} := W^*$

B. Optimization

To solve Eq. 6, we decompose the objective function into two consecutive steps: domain-invariant feature learning and multi-task learning. The procedure is summarized in Algorithm 1.

1) Domain-Invariant Feature Learning: Suppose that the multi-domain dataset consists of two modalities, X^{m1} and X^{m2}, with respect to the V^{m1} and V^{m2} camera views, respectively. This step consists of the following two key components.

a) Co-Embedding: This step aims to learn a group of co-embedding matrixes P to transfer the instance-level features into the common space $\mathbb{O}$. The objective function for learning P is given as follows:

$$\arg\min_{P,S} \sum_{i=1}^{V} \|X_i - P_i S\|_2^2 \quad s.t.\ \forall i, \|s_i\|_0 \leq z \tag{7}$$

where $\|s_i\|_0 \leq z$ is the sparsity constraint, which requires each s_i to have z or fewer non-zero terms. Here, both $\{P_i\}_{i=1}^{V}$ and S are unknown. We adopt the K-SVD algorithm [64] for the solution of both P & S since K-SVD is a highly effective method to learn an overcomplete co-embedding matrix for sparse signal representation.

Given the learned P_i, we can obtain the co-embedded features of the training and validation instances in the i-th domain using the OMP algorithm [65].

b) Fusion: A group of training data $S^{[train]} = \{S_i^{[train]}\}_{i=1}^{V}$ is used to learn a set of base classifiers. In this paper, we employ KNN classifiers to score the instances in space $\mathbb{O}$. We regard the output of the KNN classifier as the weight of the instance s_{ij}. The ensemble-learning strategy can be adopted to fuse instance-level sparse features to obtain domain-invariant feature representation. It can be formulated as follows:

$$f_j = \sum_{i=1}^{V} r_{ij} \times s_{ij} \tag{8}$$

where r_{ij} represents the weight of instance s_{ij} and f_j is the generated domain-invariant feature and belongs to the set-level matrix F.

2) Multi-Task Learning: Given the domain-invariant feature matrix F, we treat an individual human action category as one task and employ the MTL theory to explore action relatedness in between. In particular, the objective function can be converted into the following formulation after the first step:

$$\begin{aligned} &\min_{W} \sum_{t=1}^{T} \|w_t F_t - y_t\|_F^2 + \rho \|W\|_F^2 \\ &s.t. \sum_{g} I(\|w^g\| > 0) \leq u \end{aligned} \tag{9}$$

The optimization problem in Eq. 9 is non-convex. We adopt the iterative group hard thresholding (IGHT) framework [66] to optimize Eq. 9. The details are summarized in Algorithm 2.

The key idea of IGHT is to use the gradient information at the current iterate to provide the first-order approximation of the objective function. In particular, we use the combination of the linear approximation of the function $f(W)$ at a given point W_0 and a quadratic penalty term. Thus, Eq. 10 can be formulated as:

$$\begin{aligned} &\min_{W} \frac{1}{2}\|W - A\|_2^2 \\ &s.t. \sum_{g} I(\|w^g\| > 0) \leq u \end{aligned} \tag{10}$$

where $A = W^0 - \frac{1}{q}\nabla f(W^0)$. Eq. 10 can be regarded as an Euclidean projection problem, $proj(\cdot)$, that aims to find the optimal point to satisfy the constraint set that is closet to a fixed point A.

Algorithm 2 Solution of Model Learning

Input: $F^{[train]}, Y, \rho, \eta > 1$
Output: Action Models W_{MD}

```
Initialize: W^0, α^0 ← 1, L.
for t ← 1, 2, ... T do
    repeat
        A^t ← W^t − (1/L)∇f(W^t)
        W^t ← proj(A^t)
        L ← ηL
    until line search criterion is satisfied
    if the objective stop criterion satisfied then
        return W^t
    end if
end for
```

The solution of Eq. 10 admits a closed form given below:

$$w^g = \begin{cases} A^g, & if\ g \in \Omega_G \\ 0, & otherwise \end{cases} \tag{11}$$

where A^g is the g-th row of A and Ω_G is the index subset of $\{1, 2, \ldots, d\}$ of size u, including all rows of A that are among the top u rows of A in term of the length of the row vector.

To estimate the step size, which decides the distance of movement along the given search direction, we adopt the popular Lipschitz criterion.

When the objective function converges, we can obtain the action models W^*, which includes the latent relatedness among the action categories and shares domain-invariant information for action recognition.

IV. Experiments

We establish three experiments to evaluate the proposed method. 1) We compare the MDMTL framework against several representative methods, including single-view methods, multi-instance learning (MIL) methods, cross-view methods, cross-domain methods, and deep learning methods. 2) We execute a comprehensive study on domain-invariant feature learning in MDMTL. In particular, this study consists of three subexperiments: a. without the co-embedding setting (No-CE), b. co-embedding setting (CE), and c. the MDMTL framework. 3) We explore the effect of MDMTL on model learning. In particular, this exploration consists of three sub-experiments: a. multi-modal setting, b. multi-modal-late-fusion setting, and c. the MDMTL framework.

A. Dataset Description

We evaluate the proposed method on the INRIA Xmas Motion Acquisition Sequences (IXMAS), MSR Daily Activity 3D (DailyActivity3D), and Multi-modal & Multi-view & Interactive (M^2I) datasets. In IXMAS [30], which is the dataset with the largest number of camera views, we can explore the multi-domain action recognition induced by view-variant changes. In DailyActivity3D [31], which captures

Fig. 4. Exemplary frames from IXMAS. Each row shows one action captured from different views.

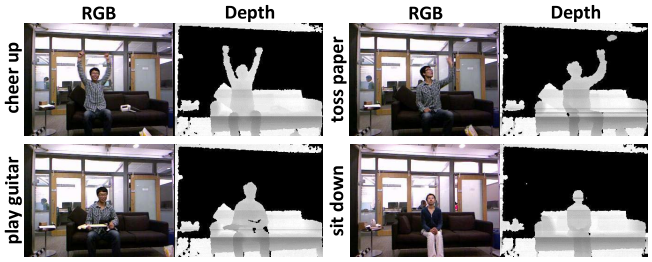


Fig. 5. Exemplary frames of DailyActivity3D.

daily person-object interaction actions in the RGB and depth modalities, we can explore the multi-domain action recognition induced by modality-variant changes. M^2I [32], the current largest multi-modal & multi-view dataset, contains many interactive actions. Further, its modal-variant and view-variant changes provide an evaluation environment for multi-domain action recognition.

1) IXMAS: The IXMAS view-invariant action recognition dataset [30] contains four side views and one top view of 11 daily-life actions performed 3 times by 10 actors. The actors were allowed to freely choose their positions and orientations. See Fig. 4 for exemplary frames from the IXMAS dataset. Because an action usually looks very different from different viewpoints, we group the videos of the same action captured from all five views into one set.

2) DailyActivity3D: The DailyActivity3D [31] records a human's daily actions in the living room with one Microsoft Kinect device. Some examples are shown in Fig. 5. It consists of 16 action categories involving person-object interactions. Each individual performs an action in two different poses, "sitting on sofa" and "standing". The total number of samples is 320. In this paper, we denote the examples of the same action in the RGB and depth modalities as one set.

3) M^2I: The M^2I [32] multi-domain dataset contains many common actions involving person-person and person-object interactions and consequently possesses rich action diversity. See Fig. 6 for exemplary frames from M^2I. It consists of 22 action categories performed by 22 unique individuals. Each action is performed twice by 20 pairs (where a pair consists of a person and an object or a person and another person). In total, the M^2I dataset contains 1760 samples

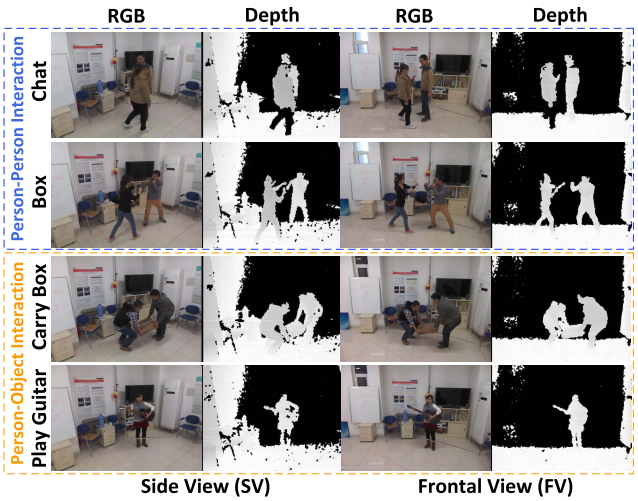


Fig. 6. Exemplary frames of M^2I are shown in the form of RGB and depth data with respect to Side View (SV) and Frontal View (FV), respectively.

(22 actions × 20 pairs × 2 views × 2 runs). M^2I contains the RGB and depth modalities. We group all videos of the same action in both the RGB and depth modalities into one set.

B. Experiment Setup

We first extracted the improved dense trajectories (iDT) detector [17] and chose HoGHoF descriptors [67] for the evaluation. Then, we used these points of interest to learn a group of codebooks with 1,000 codewords by K-means for each domain. Each action video was represented as a BoW [68] vector (a 1,000-dimensional histogram) using the corresponding domain-specific codebook. For the feature co-embedding component, the dimensionality and sparsity coefficient of the co-embedding matrix were varied within [50, 100, 200, 300] and [20, 30, 40, 50], respectively, to seek the optimal parameters. For the feature fusion component, we varied the parameter k from 1 to 10 in the learning procedure to obtain 10 KNN classifiers. The outputs of these KNN classifiers were averaged to obtain the final score for each instance. We empirically varied ρ and u in Eq. 9 for the multi-task modeling to select the optimal parameters. ρ was varied from 10^{-4} to 100, and u was varied from 10 to 50.

To enable a fair comparison with state-of-the-art approaches, we followed the same split as [7] for M^2I and [31] for DailyActivity3D. For IXMAS, we used the leave-one-action-out strategy as [9].

C. Comparison Against the State of the Arts

In this section, we compare MDMTL with several representative single-view methods, MIL methods, cross-view methods, and cross-domain methods. In particular, to ensure a fair comparison, we perform experiments in the single-modal scenario on M^2I.

We first evaluated MDMTL on the IXMAS dataset. Table II shows the average accuracy for each view pair between the state-of-the-art method [9] and MDMTL. Compared to [9], MDMTL can further improve the performances. We observe

TABLE II

PERFORMANCES ON IXMAS. $\{C_i\}_{i=0,1,\cdots,4}$ DENOTES THE FIVE DIFFERENT CAMERA VIEWS. THE FORMER AND LATTER NUMBERS IN THE BRACKET ARE THE AVERAGE RECOGNITION ACCURACIES BY CROSS-VIEW METHOD [9] AND OUR MDMTL FRAMEWORK, RESPECTIVELY. PARTICULARLY, EACH ROW CORRESPONDS TO A SOURCE (TRAINING) VIEW AND EACH COLUMN A TARGET (TEST) VIEW FOR [9]

Acc. (%)	C0	C1	C2	C3	C4
C0	-	(98.8,100)	(99.1,100)	(99.4,100)	(92.7,99.4)
C1	(98.8,100)	-	(99.7,99.7)	(92.7,100)	(90.6,99.7)
C2	(99.4,100)	(96.4,99.7)	-	(97.3,100)	(95.5,99.4)
C3	(98.2,100)	(97.6,100)	(99.7,100)	-	(90.0,100)
C4	(85.8,99.4)	(81.5,99.7)	(93.3,99.4)	(83.9,100)	-

TABLE III

PERFORMANCE COMPARISON ON DAILYACTIVITY3D

Algorithm	Accuracy
Holistic Dense Trajectories [68]	71.7%
Moving Pose [69]	73.8%
HON4D [70]	80.0%
Histograms of Depth Gradients and RDF [71]	81.3%
IPM [72]	83.3%
3D Key-Pose-Motifs [73]	83.5%
Actionlet Ensemble [31]	85.8%
MDMTL	**86.1%**

that the results are close to 100% because the environment and actions in the IXMAS dataset are simpler than those in the M^2I dataset. We notice that the recognition accuracy is slightly lower when camera 4 (top view) is used. The reason is that camera 4 was set above the actors and that all of the different actions look the same from the top view. Therefore, it is challenging to extract view-invariant information from the top view. Furthermore, dense trajectories [68] achieved an average accuracy of 93.5% on IXMAS. Comparatively, MDMTL can achieve 99.8% accuracy, which shows the advantages of multi-domain feature fusion.

Moreover, we evaluated MDMTL on the DailyActivity3D dataset. As shown in Table III, MDMTL can outperform all competing methods. In particular, our method outperforms actionlet ensemble [31] and 3D key-pose-motifs [73], which highly depend on the skeleton information to localize interaction parts and remove background noise.

IXMAS is a multi-view dataset, while DailyActivity3D is a multi-modal one. The proposed MDMTL focuses on the multi-domain (i.e., multi-view and multi-modal) scenario. Thus, we evaluate MDMTL in detail on M^2I, which contains the RGB and depth modalities with respect to the frontal view and the side view, respectively.

1) MDMTL Versus Single-View Methods: To evaluate the improvement achieved by MDMTL, we compare this framework against single-view methods, which solve instance-level recognition problems in the single-view scenario. In particular, two different experimental protocols were designed: 1) We used a non-linear support vector machine (SVM) with the χ^2-kernel [74]. 2) To investigate the impact of MTL, we replaced the SVM classifier with the popular ℓ_{21}-based MTL method described in [75].

To test MDMTL, we replaced the domain-invariant feature with the instance-level feature from the corresponding view. Thus, MDMTL was evaluated on both the side view (SV-MDMTL) and frontal view (FV-MDMTL) scenarios.

The experimental results are showed in Fig. 7. FV/SV-MDMTL can significantly outperform the other methods. In the FV case, the improvement is greater than 12% compared with the SVM method and approximately 7% compared with the ℓ_{21}-based MTL method. In the SV case, more than 7% improvement can be achieved over the SVM method, and approximately 3% improvement can be achieved over the ℓ_{21}-based MTL method. Two important observations can be made: (1) The ℓ_{21}-based MTL method is more than 4% better than the SVM method in both the SV and FV scenarios, which demonstrates that joint learning of multiple tasks using the knowledge shared among them can effectively improve the prediction performance compared with that for the independent learning of single tasks. (2) SV-MDMTL and FV-MDMTL perform approximately 3% and 8% better, respectively, than the ℓ_{21}-based MTL method, which illustrates that a single instance feature cannot appropriately represent its action category in the presence of multi-domain information. Consequently, domain-invariant features are more robust for human action recognition.

2) MDMTL Versus MIL Methods: In this section, we compare MDMTL with two representative MIL methods, Citation-kNN [76] and MI-SVM [77]. The major difference between MDMTL and these two MIL methods is how the domain-invariant feature of a single set is generated. MI-SVM chooses the instance vector with the highest score learned by the SVM classifier as the representative feature vector of a set. Citation-kNN uses a non-vectorial learning technique (i.e., kernel technology) to describe the distances between matrixes of sets.

We perform the MDMTL framework for multi-domain fusion and jointly compare the best performances of the Citation-kNN, MI-SVM and MDMTL methods in Fig. 8. In Citation-kNN [76], the notion of *citation* is adopted to consider not only the neighbors of a set B (we fixed the number of rank neighbors to 5 as [76]) but also the sets that count B as a neighbor (we also fixed the number of rank sets to 5 as [76]). In particular, the Hausdorff distance is employed, which provides a metric function measuring the distance between subsets of a metric space. By definition, two sets B_1 and B_2 are within a Hausdorff distance d of each other iff every point of B_1 is within the distance d from at least one point of B_2 and every point of B_2 is within the distance d from at least one point of B_1 [78]. However, the Hausdorff distance is very sensitive to even a single outlying point in B_1 or B_2. Therefore, any noisy instances will greatly influence the performance of Citation-kNN. For example, the performance for the action *Play Guitar* is near 0.

MI-SVM [77] used a non-linear sigmoid kernel. The functional margin of a set with respect to a hyperplane is given by $Y_I \max_{i \in I}(\langle w, x_i \rangle + b)$, where Y_I is the set category, x_i is the i-th instance feature, and w and b are the SVM parameters. For a set, the positive margin is defined as the

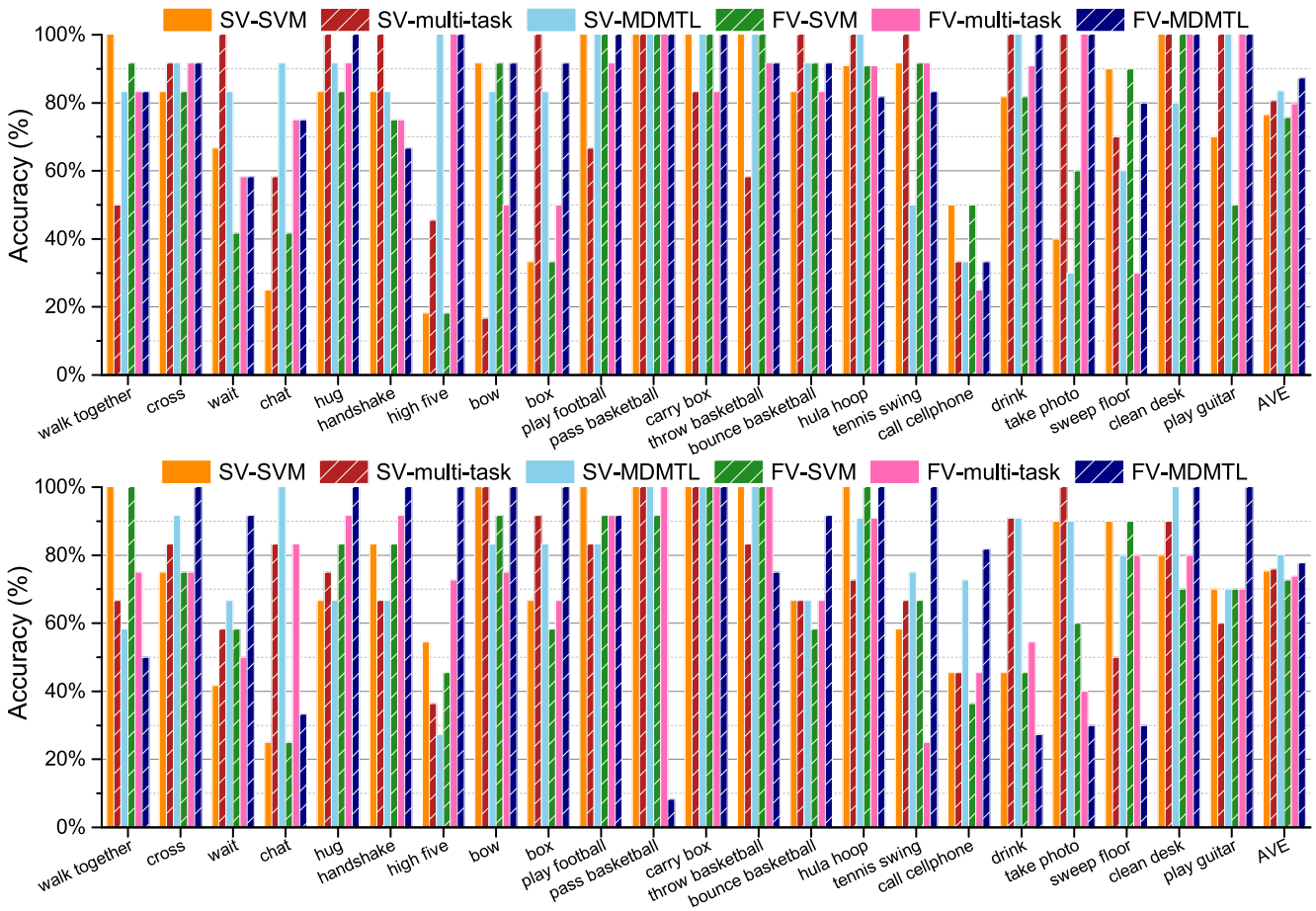


Fig. 7. Category-wise accuracy and average accuracy by the STL approaches and the MDMTL framework on M^2I. SV-MDMTL and FV-MDMTL denote the application of the MDMTL framework in side view and frontal view scenarios, respectively. Top: RGB modality; Bottom: depth modality.

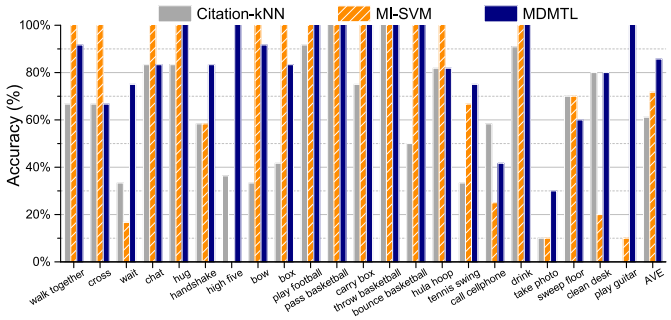


Fig. 8. Category-wise accuracy and average accuracy in the comparison of Citation-kNN [76], MI-SVM [77] and R-MDMTL on M^2I.

TABLE IV
COMPARISON AMONG THE CROSS-VIEW METHOD, CROSS-DOMAIN METHODS, AND MDMTL ON M^2I

Algorithm	RGB		Depth	
	SV	FV	SV	FV
Cross-view [9]	63.2%	65.6%	54.4%	55.6%
SVM-AT [79]	74.8%	75.3%	72.7%	72.7%
FR [80]	76.2%	76.2%	74.6%	74.7%
MKL [81]	76.1%	75.7%	73.4%	73.7%
DT-MKL [82]	76.1%	75.8%	73.1%	73.4%
A-MKL [79]	77.0%	76.9%	75.8%	75.7%
MDMTL	**83.5%**	**87.3%**	**80.2%**	**77.8%**

margin of the "most positive" instance, whereas the negative margin is defined with respect to the "least negative" instance. Because the set margin is completely defined by a single instance feature, it is also sensitive to noisy instances. For example, the performance for the action *High Five* is near 0.

As seen from the results above, the MDMTL framework incorporates domain-invariant information and outperforms Citation-kNN with 24% gain and MI-SVM with 14% gain.

3) MDMTL Versus Cross-View and Cross-Domain Methods: We compare the MDMTL framework with representative cross-view [9] methods and cross-domain [79]–[82] methods, which have been evaluated on M^2I by [7]. As shown in Table IV, the MDMTL framework achieves the best performance among all investigated methods. The reasons are explained as follows.

- Zheng *et al.* [9] proposed a cross-view method that generates a dictionary pair to encode all videos in a sparse space. However, this method exploits only the instance-level and view-invariant correspondences based on the shared dictionary pair. MDMTL generates domain-invariant features, which incorporate not only

TABLE V
PERFORMANCE COMPARISON WITH DEEP LEARNING-BASED METHODS ON M^2I

Algorithm	SV	FV
SFAM-D [83]	71.2%	83.0%
SFAM-S [83]	70.1%	75.0%
SFAM-RP [83]	79.9%	81.8%
SFAM-AMRP [83]	82.2%	78.0%
SFAM-LABRP [83]	72.0%	83.7%
Max-Score Fusion All [83]	87.6%	88.8%
Average-Score Fusion All [83]	88.2%	89.1%
Multiply-Score Fusion All [83]	89.4%	91.2%
I3D [84]	79.1%	74.5%
MDMTL	**89.5%**	**93.3%**

view-invariant information but also modality-invariant information.

- The five cross-domain methods attempt to leverage a large number of loosely labeled instances. However, they ignore the contributions of individual instances to the entire set. Therefore, they cannot effectively avoid the negative influence of noisy instances. MDMTL adopts the ensemble-learning strategy to fuse one action set, which scores each instance and represents its weight for the action pattern. Consequently, it can effectively suppress noisy instances by assigning low weights.

4) MDMTL Versus Deep Learning-Based Methods: We compare MDMTL with deep learning-based methods [83], [84]. Scene flow to action map (SFAM) [83] encodes the RGB-D video sequences as dynamic images and then uses ConvNets models to extract score vectors. In Table V, we present several variants of SFAM (i.e., SFAM-D, S, RP, AMRP, LABRP) that encode the spatiotemporal information in different aspects and the score-fusion SFAMs (i.e., Max, Average, Multiply-Score Fusion All) that can further improve the performance. In particular, SFAM [83] has been evaluated on M^2I with SV and FV scenarios. At the same time, we evaluate the I3D [84] model, which is a representative deep learning-based method, on M^2I. To ensure a fair comparison, we denote the examples of the same action in the RGB and depth modalities as one set with respect to the SV and FV scenarios, respectively. Moreover, we extract the features from the final average-pooling layer of the I3D model as domain-specific features in our framework. As shown in Table V, MDMTL can achieve competitive performance against I3D and variants of SFAM.

Similarly, we compared MDMTL with deep learning-based methods [85], [86] on DailyActivity3D. WHDMMs+ ConvNets [85] leverage the capability of ConvNets to mine discriminative features from different viewpoints of a depth sequence. DDI [86] (i.e., structured body/part/joint DDI) represents a depth sequence by skeleton guidance of three pairs of structured dynamic images at the body, part and joint levels, respectively. In particular, the skeleton data are the fine-grained and localized information for action recognition. As shown in Table VI, MDMTL can achieve competitive performance against these methods.

TABLE VI
PERFORMANCE COMPARISON WITH DEEP LEARNING-BASED METHODS ON DAILYACTIVITY3D

Algorithm	Accuracy
WHDMMs+ConvNets [85]	85.0%
Structured body DDI [86]	61.0%
Structured part DDI [86]	81.9%
Structured joint DDI [86]	93.1%
I3D [84]	82.8%
MDMTL	**93.8%**

D. Evaluation of Feature Learning

In this section, we evaluate the effectiveness of different components in domain-invariant feature learning in MDMTL. In particular, we designed two experimental protocols. To clearly evaluate the effectiveness of feature learning, we performed separate experiments on the RGB and depth modalities on M^2I.

1) Evaluation of Feature Co-Embedding: We designed two comparative experiments: 1) without feature co-embedding (No-CE setting) and 2) with feature co-embedding (CE setting), where the feature fusion is performed by mean pooling over the instance features of one set. These experimental settings allowed us to measure the improvements due to the co-embedding of multi-domain features. The objective function (Eq. 6) was directly used to handle the No-CE/CE settings.

The comparative results are presented in Fig. 9. It can be observed that compared with the No-CE setting, the CE setting achieves increased performance. The improvement was greater than 8% and 7% with respect to the RGB and depth modalities, respectively, which demonstrates the advantage achieved by multi-domain feature co-embedding. Particularly, in the RGB modality, for some actions that show significant viewpoint differences, such as *Walk Together*, *Chat*, and *Bounce Basketball*, CE can achieve significantly better performances than No-CE due to the discriminative features in the co-embedding space. Similarly, in the depth modality, the performances can be significantly improved by the CE setting for some actions, such as *Wait*, *Pass Basketball* and *Hula Hoop*.

2) Evaluation of Feature Fusion: We compare two experimental settings. One is the CE setting (with feature co-embedding and feature fusion by mean pooling) mentioned above. The other is the proposed MDMTL framework (with feature co-embedding and feature fusion by the ensemble-learning strategy). Based on the co-embedded features, the weight of each instance was generated according to the KNN classifiers as stated in Section III-B-(1). In particular, we varied the parameter k to obtain multiple KNN classifiers. The outputs of KNN classifiers were averaged to obtain the final weight for each instance. Then, we fused the co-embedded features of one set with the corresponding weights using Eq. 8.

Fig. 9 shows the comparison between the CE setting and the MDMTL framework. We observe improvements of more than 4% in the depth modality, which demonstrates the effectiveness of the fusion method. In particular, for some actions, such as *Drink*, and *Sweep Floor*, the instances from the same

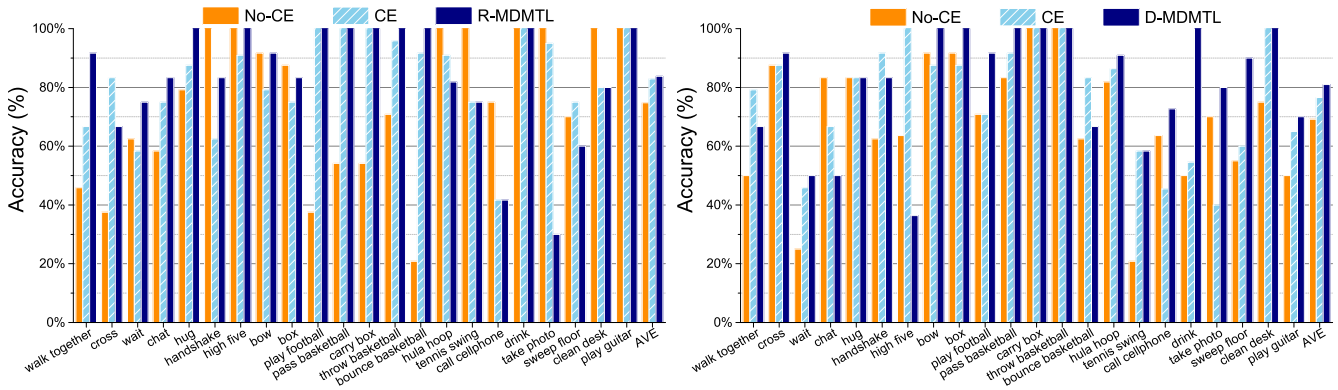


Fig. 9. Comparison of category-wise accuracy and average accuracy. No-CE denotes the method without both feature co-embedding and feature fusion components of MDMTL. CE denotes the method only with feature co-embedding but without feature fusion components of MDMTL. R-MDMTL and D-MDMTL denote the application of the MDMTL framework in the RGB and depth modalities, respectively. Left: RGB modality; Right: depth modality.

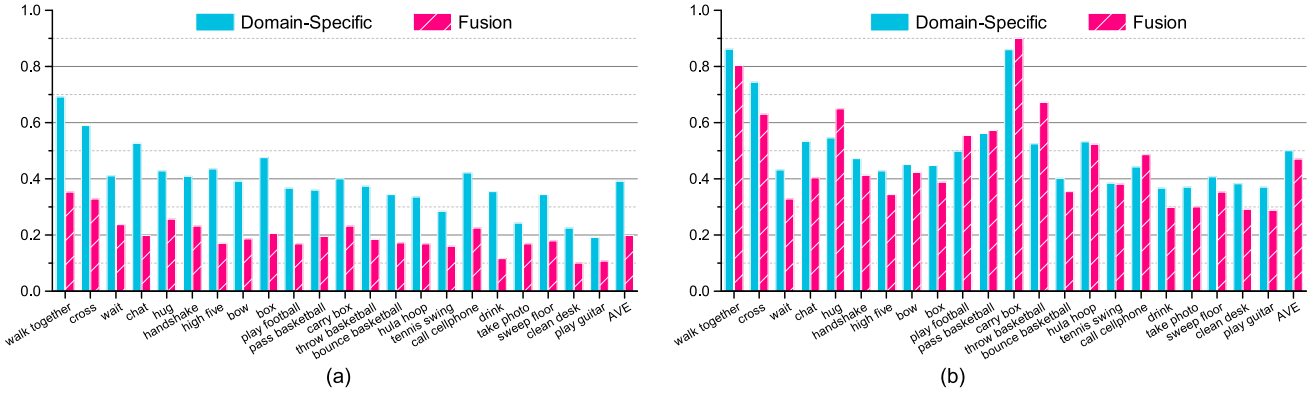


Fig. 10. Category-wise variation of domain-specific features and fusion features on M^2I. (a) Intra-class variation. (b) Inter-class variation.

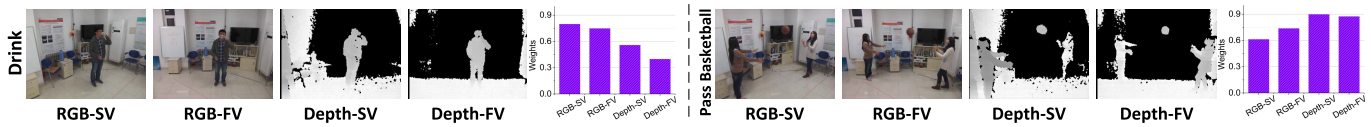


Fig. 11. Exemplary sets of action *Drink* (left) and *Pass Basketball* (right). The bars denote the weight of each instance with respect to RGB-SV, RGB-FV, Depth-SV, and Depth-FV, respectively.

set are fused based on the more discriminative weights than mean pooling. Therefore, the performances of MDMTL are better than those of the CE setting. Similarly, in the RGB modality, there are the significant improvements by MDMTL for some actions, such as *Wait*, *Hug*, and *Chat*.

To measure the degree of invariance, we compute the intra-class variation in both the domain-specific features and the fusion features on M^2I. In particular, we average the euclidean distances of pair-wise instances from each category as the variance measure. For the domain-specific process, we average the variations in all domains. As shown in Fig. 10 (a), the fusion process can significantly reduce the variation by learning the discriminative weights for each instance.

Furthermore, we compute the inter-class variation on M^2I. First, we represent each category by the mean feature of all category-wise instances. Then, we adopt the one-against-all strategy, where we fix one category and compute the euclidean distances with other categories. Finally, the obtained distances are averaged as the inter-class variation for the fixed category. As shown in Fig. 10 (b), we find that the inter-class variation is still significant after the fusion process.

In addition, Fig. 11 presents two sets from M^2I. We visualize the generated weights to provide intuition as to which instances should be focused on to predict the action category. For the action *Drink*, the RGB weights are higher than the depth ones since the local appearance changes in RGB are more significant than those in depth. However, for the action *Pass Basketball*, the inverse observation is made since the depth data can capture more significant moving information than the RGB data. These cases intuitively show that the proposed ensemble-learning strategy can quantize the contribution of each instance in one set.

E. Evaluation of Model Learning

In this section, we evaluated the model learning component in the MDMTL framework. To explore whether MDMTL

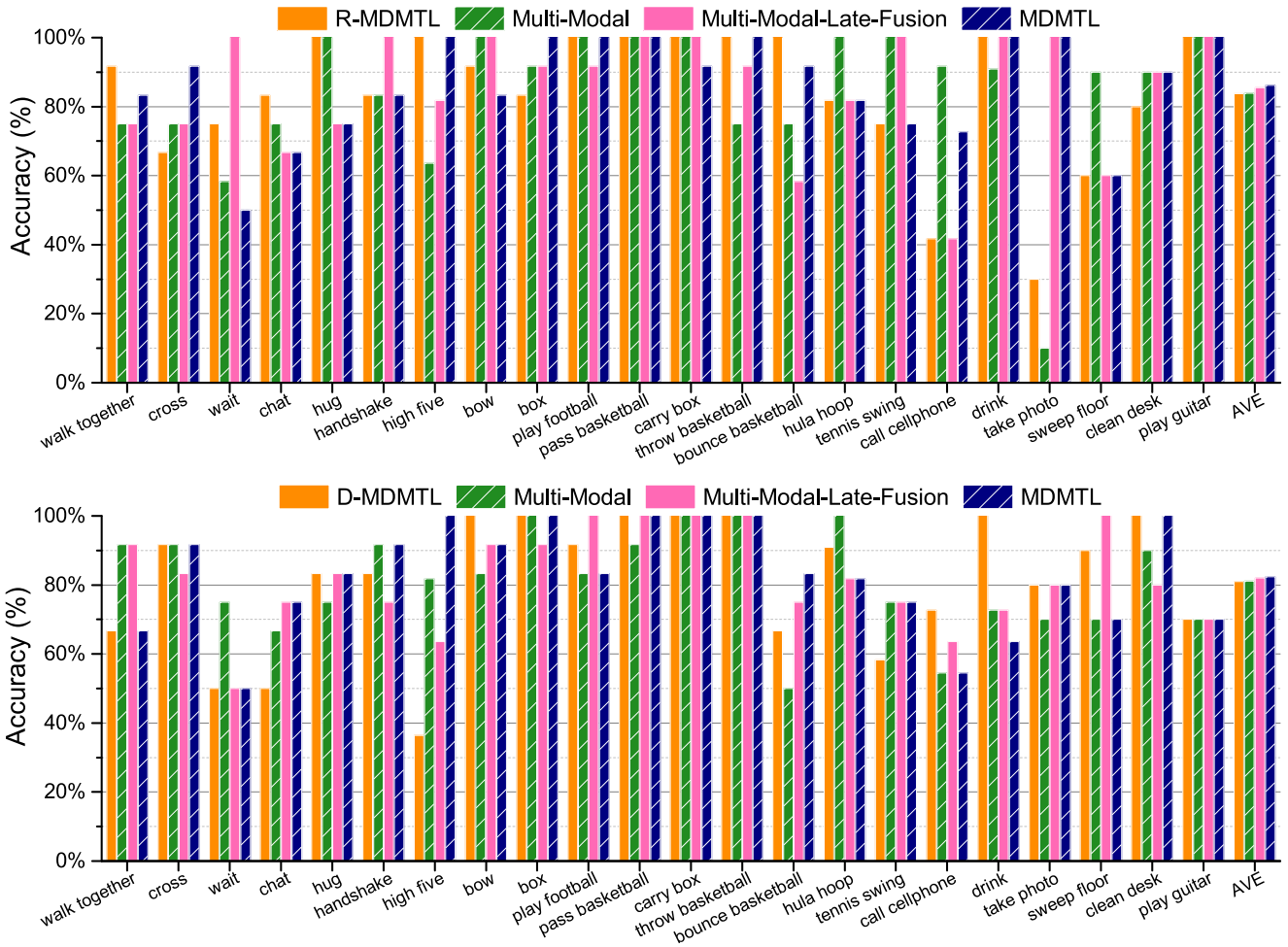


Fig. 12. Category-wise accuracy and average accuracy in the comparison of the R/D-MDMTL, Multi-Modal, Multi-Modal-Late-Fusion, and MDMTL methods on M^2I. Multi-Modal constructs modality-invariant tasks based on the MDMTL framework. Multi-Modal-Late-Fusion adds the function of late fusion into the Multi-Modal setting. Top: RGB modality; Bottom: depth modality.

can fuse multi-modal information to enhance performance, we proposed fusing two instances from both views with respect to the RGB and depth modalities, respectively, which led to 44 tasks for M^2I (22 action categories$\times$2 modalities). We set two comparison experiments, 1) multi-modal and 2) multi-modal-late-fusion, which will be detailed below. Meanwhile, we performed 3) the MDMTL framework, which jointly fuses four instances from 2 views and 2 modalities, leading to 22 tasks for M^2I.

1) Multi-Modal Setting: Multi-modal classifiers, W_{MM}, consist of W_{RGB} and W_{depth}. We employ W_{RGB} and W_{depth} to evaluate the view-invariant fused features with respect to the RGB and depth modalities, respectively. The results are presented in Fig. 12. The multi-modal setting performs better than R/D-MDMTL, which only use a single modality for model learning, as stated in Section IV-D. The multi-modal classifiers benefit from utilizing multi-modal information during MTL. In particular, the performances of the multi-modal setting are superior to those of the R/D-MDMTL setting for 14 and 12 of the total of 22 actions (e.g., *Hula Hoop* and *Tennis Swing*) with respect to the RGB and depth modalities, respectively. Specifically, the improvements for *Hula Hoop* are greater than 9%.

2) Multi-Modal-Late-Fusion Setting: To take full advantage of multi-modal classifiers W_{MM}, W_{RGB} and W_{depth} are combined to evaluate the view-invariant fused features in the RGB and depth modalities. In particular, we design the *multi-modal-late-fusion* setting. Each fused feature with respect to an individual modality is scored by the corresponding classifiers, and the higher score is taken as the final score for this action sample with respect to both modalities.

The multi-modal-late-fusion setting improves the results, as shown in Fig. 12. This approach can be regarded as a type of decision-level fusion, which combines the outputs of the RGB and depth classifiers to obtain the final prediction for an action sample. Improvements over R/D-MDMTL are observed for 12/13 of the total of 22 actions, which also demonstrates the complementarity of the multi-modal information.

3) MDMTL: The performances of MDMTL are shown in Fig. 12, which reveals that MDMTL can outperform the R/D-MDMTL, multi-modal, and multi-modal-late-fusion settings. In particular, the performances of MDMTL are superior to the competing methods for 10 and 13 of the total of 22 actions with respect to the RGB and depth modalities, respectively. MDMTL thoroughly integrates both multi-view

and multi-modal information to generate domain-invariant features for robust action model learning.

V. Conclusion

In this paper, we aim to solve the multi-domain action recognition problem by exploring and fusing view-invariant and modality-invariant information. We proposed the multi-domain & multi-task learning (MDMTL) framework, which jointly addresses the problems of domain-invariant feature learning and multi-task modeling. In particular, the framework learns a group of co-embedding matrixes from various domains and forces multi-domain instances of the same action to have similar embedded representations. The multi-task method is used to model the multi-domain recognition problem with the Frobenius-norm regularization term and the proposed sparse constraint term. The MDMTL framework was extensively evaluated on three public action datasets and was found to outperform the state-of-the-art approaches. In addition, we explored the effectiveness of the feature learning and model learning components in the MDMTL framework.

References

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[3] M. Ramezani and F. Yaghmaee, "A review on human action analysis in videos for retrieval applications," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 485–514, 2016.

[4] B. Su, J. Zhou, X. Ding, and Y. Wu, "Unsupervised hierarchical dynamic parsing and encoding for action recognition," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5784–5799, Dec. 2017.

[5] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *IEEE Multimedia*, vol. 21, no. 2, pp. 42–70, Apr./Jun. 2014.

[6] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5479–5490, Nov. 2016.

[7] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 7, pp. 1781–1794, Jul. 2017.

[8] Z. Qin and C. R. Shelton, "Event detection in continuous video: An inference in point process approach," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5680–5691, Dec. 2017.

[9] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," in *Proc. BMVC*, 2012, pp. 1–11.

[10] Y. Kong and Y. Fu, "Bilinear heterogeneous information machine for RGB-D action recognition," in *Proc. CVPR*, 2015, pp. 1054–1062.

[11] C. Jia and Y. Fu, "Low-rank tensor subspace learning for RGB-D action recognition," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4641–4652, Oct. 2016.

[12] B. Liang and L. Zheng, "Specificity and latent correlation learning for action recognition using synthetic multi-view data from depth maps," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5560–5574, Dec. 2017.

[13] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. ICCV*, 2003, pp. 432–439.

[14] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, 2011, pp. 3169–3176.

[15] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. CVPR*, 2011, pp. 3361–3368.

[16] Z. Jiang, Z. Lin, and L. S. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 533–547, Mar. 2012.

[17] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. ICCV*, 2013, pp. 3551–3558.

[18] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, Oct. 2014.

[19] H. Zhu, J.-B. Weibel, and S. Lu, "Discriminative multi-modal feature fusion for RGBD indoor scene recognition," in *Proc. CVPR*, 2016, pp. 2969–2976.

[20] Y. Kong and Y. Fu, "Discriminative relational representation learning for RGB-D action recognition," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2856–2865, Jun. 2016.

[21] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2430–2443, Dec. 2016.

[22] L. Zhang and D. Zhang, "Robust visual knowledge transfer via extreme learning machine-based domain adaptation," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4959–4973, Oct. 2016.

[23] G. Zhang, H. Sun, F. Porikli, Y. Liu, and Q. Sun, "Optimal couple projections for domain adaptive sparse representation-based classification," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5922–5935, Dec. 2017.

[24] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.

[25] X. Xu, T. M. Hospedales, and S. Gong, "Multi-task zero-shot action recognition with prioritised data augmentation," in *Proc. ECCV*, 2016, pp. 343–359.

[26] J. Zhou, J. Chen, and J. Ye. (2012). *MALSAR: Multi-Task Learning Via Structural Regularization*. [Online]. Available: http://jiayuzhou.github.io/MALSAR/

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

[28] Z. Fu, A. Robles-Kelly, and J. Zhou, "MILIS: Multiple instance learning with instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 958–977, May 2010.

[29] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. NIPS*, 2010, pp. 1813–1821.

[30] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. ICCV*, 2007, pp. 1–7.

[31] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. CVPR*, 2012, pp. 1290–1297.

[32] N. Xu, A. Liu, W. Nie, Y. Wong, F. Li, and Y. Su, "Multi-modal & multi-view & interactive benchmark dataset for human action recognition," in *Proc. ACM MM*, 2015, pp. 1195–1198.

[33] M. Ahmad and S.-W. Lee, "HMM-based human action recognition using multiview image sequences," in *Proc. ICPR*, 2006, pp. 263–266.

[34] F. Nie, J. Li, and X. Li, "Convex multiview semi-supervised classification," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5718–5729, Dec. 2017.

[35] X. Wu and Y. Jia, "View-invariant action recognition using latent kernelized structural SVM," in *Proc. ECCV*, 2012, pp. 411–424.

[36] Y. Song, L.-P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *Proc. CVPR*, 2012, pp. 2120–2127.

[37] F. Zhu, L. Shao, and M. Lin, "Multi-view action recognition using local similarity random forests and sensor fusion," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 20–24, 2013.

[38] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 412–424, Mar. 2012.

[39] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Process.*, vol. 93, no. 6, pp. 1445–1457, 2013.

[40] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. CVPR*, 2011, pp. 3209–3216.

[41] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Proc. CVPR*, 2012, pp. 1362–1369.

[42] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. CVPR*, 2008, pp. 1–8.

[43] Anwaar-ul-Haq, I. Gondal, and M. Murshed, "On dynamic scene geometry for view-invariant action matching," in *Proc. CVPR*, 2011, pp. 3305–3312.

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

[45] I. N. 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.

[46] M. Lewandowski, D. Makris, and J.-C. Nebel, "View and style-independent action manifolds for human activity recognition," in *Proc. ECCV*, 2010, pp. 547–560.

[47] Y. Jiang, F.-L. Chung, S. Wang, Z. Deng, J. Wang, and P. Qian, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 688–701, Apr. 2015.

[48] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 450–461, Feb. 2015.

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

[50] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 83–101, 2006.

[51] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. CVPR*, 2012, pp. 2855–2862.

[52] L. Cruz, D. Lucio, and L. Velho, "Kinect and RGBD images: Challenges and applications," in *Proc. 25th SIBGRAPI Conf. Graph., Patterns Images Tuts.*, 2012, pp. 36–49.

[53] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 1147–1153.

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

[55] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Learning Theory and Kernel Machines* (Lecture Notes in Computer Science), vol. 2777, B. Schölkopf and M. K. Warmuth, Eds. Berlin, Germany: Springer, 2003, pp. 567–580.

[56] T. Evgeniou, M. Pontil, and O. Toubia, "A convex optimization approach to modeling consumer heterogeneity in conjoint estimation," *Marketing Sci.*, vol. 26, no. 6, pp. 805–818, 2007.

[57] R. K. Ando, "Applying alternating structure optimization to word sense disambiguation," in *Proc. 10th Conf. Comput. Natural Lang. Learn.*, 2006, pp. 77–84.

[58] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, pp. 615–637, Apr. 2005.

[59] A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying, "A spectral regularization framework for multi-task structure learning," in *Proc. NIPS*, 2007, pp. 25–32.

[60] S. Thrun and J. O'Sullivan, "Clustering learning tasks and the selective cross-task transfer of knowledge," in *Learning to Learn*, S. Thrun and L. Pratt Eds. Boston, MA, USA: Springer, 1998, pp. 235–257.

[61] B. Bakker and T. Heskes, "Task clustering and gating for Bayesian multitask learning," *J. Mach. Learn. Res.*, vol. 4, pp. 83–99, May 2003.

[62] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. KDD*, 2011, pp. 42–50.

[63] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. ACM KDD*, 2012, pp. 895–903.

[64] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[65] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[66] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, Nov. 2009.

[67] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008, pp. 1–8.

[68] H. Wang, A. Kläser, 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.

[69] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. ICCV*, 2013, pp. 2752–2759.

[70] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. CVPR*, 2013, pp. 716–723.

[71] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. WACV*, 2014, pp. 626–633.

[72] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian, "Interaction part mining: A mid-level approach for fine-grained action recognition," in *Proc. CVPR*, 2015, pp. 3323–3331.

[73] C. Wang, Y. Wang, and A. L. Yuille, "Mining 3D key-pose-motifs for action recognition," in *Proc. CVPR*, 2016, pp. 2639–2647.

[74] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[75] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Multi-task learning for spatio-temporal event forecasting," in *Proc. ACM KDD*, 2015, pp. 1503–1512.

[76] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *Proc. ICML*, 2000, pp. 1119–1126.

[77] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. NIPS*, 2002, pp. 561–568.

[78] R. L. Devaney, "Measure, topology, and fractal geometry (Gerald A. Edgar)," *SIAM Rev.*, vol. 33, no. 4, pp. 668–669, 1991.

[79] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from Web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1667–1680, Sep. 2012.

[80] H. Daumé, III, "Frustratingly easy domain adaptation," in *Proc. ACL*, 2007, pp. 256–263.

[81] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. ICML*, 2004, p. 6.

[82] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank, "Domain transfer SVM for video concept detection," in *Proc. CVPR*, 2009, pp. 1375–1381.

[83] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *Proc. CVPR*, 2017, pp. 416–425.

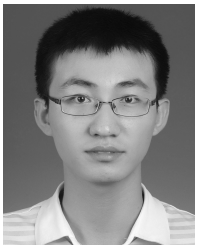
[84] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. CVPR*, 2017, pp. 4724–4733.

[85] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human Mach. Syst.*, vol. 46, no. 4, pp. 498–509, Aug. 2016.

[86] P. Wang, S. Wang, Z. Gao, Y. Hou, and W. Li, "Structured images for RGB-D action recognition," in *Proc. ICCV*, 2017, pp. 1005–1014.

An-An Liu (M'10) received the Ph.D. degree in electronic engineering from Tianjin University, China.

He was a Visiting Professor with the SeSaMe Centre, National University of Singapore, and a Visiting Scholar with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. His current research interests include computer vision and machine learning.



Ning Xu is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His current research interests include computer vision and machine learning.

Wei-Zhi Nie received the Ph.D. degree in electronic engineering from Tianjin University, China.

He was a Visiting Scholar with the National University of Singapore. He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include computer vision and machine learning.

Yu-Ting Su received the Ph.D. degree in electronic engineering from Tianjin University, China.

He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include computer vision and machine learning.

Yong-Dong Zhang (M'08–SM'13) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China. His current research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology.

He has authored over 100 refereed journal and conference papers. He was a recipient of the Best Paper Awards in PCM 2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011. He serves as an Editorial Board Member of the *Multimedia Systems* Journal and the IEEE TRANSACTIONS ON MULTIMEDIA.