Benchmarking a Multimodal and Multiview and Interactive Dataset for Human Action Recognition

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

***Abstract*—Human action recognition is an active research area in both computer vision and machine learning communities. In the past decades, the machine learning problem has evolved from conventional single-view learning problem, to cross-view learning, cross-domain learning and multitask learning, where a large number of algorithms have been proposed in the literature. Despite having large number of action recognition datasets, most of them are designed for a subset of the four learning problems, where the comparisons between algorithms can further limited by variances within datasets, experimental configurations, and other factors. To the best of our knowledge, there exists no dataset that allows concurrent analysis on the four learning problems. In this paper, we introduce a novel multimodal and multiview and interactive (M^{2}I) dataset, which is designed for the evaluation of human action recognition methods under all four scenarios. This dataset consists of 1760 action samples from 22 action categories, including nine person–person interactive actions and 13 person–object interactive actions. We systematically benchmark state-of-the-art approaches on M^{2}I dataset on all four learning problems. Overall, we evaluated 13 approaches with nine popular feature and descriptor combinations. Our comprehensive analysis demonstrates that M^{2}I dataset is challenging due to significant intraclass and view variations, and multiple similar action categories, as well as provides solid foundation for the evaluation of existing state-of-the-art algorithms.**

***Index Terms*—Action recognition, cross-domain learning, cross-view learning, multitask learning.**

I. Introduction

HUMAN action recognition has received increasing attention due to its rich real-world applications, such as human computer interaction, intelligent video surveillance, and multimedia content understanding and management [1]–[4]. Over more than a decade of active research, it is still a challenging research problem due to the following reasons. First, there exist large intraclass variations caused by the variation in speed and motion pattern, viewpoints, and background clutter. Second, action recognition is highly related with complex contextual information, such as scene characteristics and the interaction between objects and persons. Third, the diversity and the dynamic nature within an action category makes it difficult to model the salient action units, as well as the transition between consecutive actions. Finally, existing datasets contain limited samples for each action category. Hence, comprehensive analysis across various algorithms is not feasible, especially for the various machine learning problems discussed in the following sections.

Existing machine learning problem in action recognition can be generalized into four learning problems, namely single-view learning [5]–[10], cross-view learning [11]–[13], cross-domain learning [14]–[18], and multitask learning [19]–[26]. Single-view learning is the conventional approach and it is widely focused in [5]–[9]. Given the training data from a static camera view, it first learns a discriminative model for each action category and classify the unseen test data as one of the learn class. One of the most successful and widely applied approach is to first encode action sample with bag-of-word (BoW) descriptor, followed by using support vector machine (SVM) classifier to predict the genuine class [5]. Among the studied local feature detectors, space time interest points (STIPs) [27] and improved dense trajectories (iDTs) [28] has achieved state-of-the-art performance in the literature. Recently, Fisher vector encoding [29] has further improved the performance. Different from BoW and Fisher vector-based descriptor [30], [31] explore the efficacy of action attributes for this problem. Shao *et al.* [32] proposed the spatio-temporal Laplacian pyramid coding method, which is effective for single view action classification due to the discriminative visual feature. The core problem of single-view learning is that it does not consider the difference in appearance and motion characteristics from different camera views.

Manuscript received December 30, 2015; revised April 6, 2016 and June 17, 2016; accepted June 19, 2016. Date of publication July 18, 2016; date of current version June 14, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61472275, Grant 61572356, Grant 61502337, and Grant 61100124, in part by the Tianjin Research Program of Application Foundation and Advanced Technology under Grant 15JCYBJC16200, in part by the China Scholarship Council under Grant 201506255073, in part by the Elite Scholar Program of Tianjin University under Grant 2014XRG-0046, and in part by the National Research Foundation, Prime Ministers Office, Singapore, under its International Research Centre in Singapore Funding Initiative. This paper was recommended by Associate Editor L. Shao. *(Corresponding authors: An-An Liu and Yu-Ting Su.)*

A.-A. Liu, N. Xu, W.-Z. Nie, and Y.-T. Su are with the School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: liuanan@tju.edu.cn; ytsu@tju.edu.cn).

Y. Wong is with the Interactive and Digital Media Institute, National University of Singapore, Singapore.

M. Kankanhalli is with the School of Computing, National University of Singapore, Singapore.

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/TCYB.2016.2582918

TABLE I
KEY PROPERTIES OF EXISTING DATASETS AND THE PROPOSED M^{2}I DATASET

Category	Dataset	Modality			View	Interactive action		Category	Samples
		RGB	Depth	Skeleton		Person-person	Person-object		
Constrained	KTH [33]	√			1			6	2,391
	Weizmann [34]	√			1			10	90
	MSR Action3D [35]		√	√	1		√	20	402
Realistic	Hollywood [36]	√			1	√	√	8	233
	Hollywood2 [37]	√			1	√	√	12	3,669
	UCF Sports [38]	√			1		√	10	184
	UCF YouTube [39]	√			1		√	11	3,040
	UCF50 [40]	√			1	√	√	50	6,676
	UCF101 [41]	√			1	√	√	101	13,320
	HMDB51 [42]	√			1	√	√	51	6,849
Multi-View	IXMAS [43]	√			5			13	2,340
Multi-Modal	MSR DailyActivity3D [44]	√	√	√	1		√	16	320
	Indoor Activity [45]	√	√		1		√	12	-
	RGBD-HuDaAct [46]	√	√		1		√	12	1,189
Multi-View & Multi-Modal	Northwestern-UCLA [47]	√	√	√	3		√	10	1,475
	M^{2}I	√	√	√	2	√	√	22	1,784

Differing from single-view learning problem, cross-view learning aims to map features obtained from multiple views into a common feature space to handle the variations in visual appearance. In the case where a new action category is given, it can utilize the feature mapping model to perform action recognition between two different camera views. It is assumed that the action data are obtained from static views. Cross-view matrix factorization [48], [49] models the correlated latent semantic information of different views. Meanwhile, Bian *et al.* [50] took advantage of the complementary information from different viewpoints to understand the complicated human behaviors. It is noted that the multimodal and multiview data not only exists in human action recognition task, but also minifests in other computer vision problems [51]–[54]. The multimodal and multiview description naturally contains certain degree of semantic gap in object representation, leading to the challenges in retrieval, recognition, and other applications [55]. To bridge the semantic gap, a probabilistic semantic mapping approach [56] was introduced under the Hausdorff distance framework, addressing the semantic gap issue from the perspective of pairwise object matching, instead of object representation, and has shown satisfactory performance for the task of object retrieval.

Existing action dataset often contains limited sample for each action category (see Table I). To address this issue, cross-domain learning problem aims to leverage the small-scale data from target domain together with a large-scale data from an auxiliary domain to augment the generalization ability for model learning [57]. Adaptive multiple kernel learning proposed by Duan *et al.* [15] is a representative method. Last but not least, when there are multiple related tasks and the training dataset of individual task only contains limited annotated samples, multitask learning can implicitly augment the dataset to benefit model learning by jointly learning multiple related tasks to discover the latent shared information across tasks [19], [58]. Several representative methods for multitask learning were proposed by Zhou *et al.* [24] and Chen *et al.* [25].

A. Motivations

There are two key motivations for this paper. First, a recent benchmark analysis of human action recognition with the state-of-the-art BoW framework was conducted with various fusion methods [59]. It provided useful insight to understand the efficacy BoW framework on various datasets. This has stimulated the demand for similar benchmark projects for other learning problems, such as single-view, cross-view, cross-domain, and multitask learning problem. Despite the availability of many human action datasets (see Section II), there exists no dataset that allows concurrent evaluation of state-of-the-art methods for the aforementioned problems. Second, there exist few human action datasets which simultaneously cover the person–person and person–object interactive actions with multiple views and modalities, or cover limited action categories with relatively small number of videos [60]. Existing datasets usually consists of two key categories of action (i.e., atomic actions and interactive actions). Conventional action datasets mainly focus on atomic actions [33], [34], [43], which limits the generalization of relevant methods for real applications [61]. In recent years, there is an increasing interest in person–person and person–object interactive actions [62]. With the recent advancement of the cost-effective visual and depth sensors, more attention has been paid on multimodal data for various computer vision tasks [63]–[66]. Multiple human action recognition datasets have been constructed by leveraging both RGB and depth data [44], [67]. Since the reported performances are gradually saturating on most of public datasets [5], there exists a requirement of a new dataset with multiple camera views and modalities, as well as interactive actions.

B. Contributions

The contributions of this paper are as follows.

1) We introduce a new human action recognition dataset, namely multimodal and multiview and interactive (M^{2}I) dataset, which consists of person–person and person–object interaction captured using Kinect sensors. The proposed dataset has three key properties.
   a) It provides multimodal information (i.e., color images, depth information, and 3-D body joints) for joint-model analysis.
   b) Each action is simultaneously captured by two views (i.e., frontal view and side view) for cross-view analysis.
   c) All action samples are performed with unconstrained orientation and body movement, which results in high intraclass variability and interclass similarity.
2) We systematically benchmark the state-of-the-art approaches on the proposed dataset. Overall, we have evaluated one single-view approach [5], two cross-view approaches [68], five cross-domain approaches [15]–[18], and five multitask approaches [22]–[26]. All experiments are evaluated with the commonly used interest point detectors and descriptors.

Parts of this paper, appeared previously [69]. The remaining of this paper is organized as follows. Section II provides an overview of existing datasets for human action recognition. The details of the proposed M^{2}I dataset are elaborated in Section III. The details, evaluations, and discussion of the four recognition scenarios are shown in Sections IV–VII. Section VIII concludes this paper.

II. Overview of Human Action Datasets

In this section, we review the representative human action recognition datasets from five categories, namely constrained dataset, realistic dataset, multiview dataset, multimodal dataset, and multiview and multimodal dataset. The key properties of these datasets is summarized in Table I.

A. Constrained Dataset

Constrained dataset is the dominant type of dataset in the early research stage. Typically, these datasets are recorded using the visible color cameras, where actions are usually performed by few human subjects with static background and fixed viewpoint. The earliest datasets are KTH [33] and Weizmann [34] datasets. Generally, these datasets only consist of atomic action such as "running" and "waving." The simplicity of the action categories as well as the controlled recording environment simplifies the recognition task. Although KTH dataset consists of variations in lighting conditions and multiple view points, it is still far removed from real world scenarios. Furthermore, the reported performances on both datasets are close to perfect [34], [70]. Meanwhile, MSR Action3D [35] dataset was captured with the depth sensor for the recognition of instantaneous human activities. This single modality dataset contains small number of action samples.

B. Realistic Dataset

Realistic dataset aims to recognize natural human actions in diverse and realistic video settings. Action recognition in realistic environments is driven by the requirement of real applications on human action recognition facing the variations of expression, posture, motion, clothing, camera view, illumination, occlusion, and scene surroundings. Specifically, the Hollywood [36] and Hollywood2 [37] datasets are widely used for action recognition in movies. They are quite challenging since there are obvious variations of visual contents and camera movements. Although there exist action category overlaps between both Hollywood/Hollywood2 and M^{2}I datasets, they focus on different applications. The former deals with movie actions under uncontrolled environments with moving cameras while the latter is for interactive activity capturing under multimodal and fixed-camera settings. Meanwhile, UCF sports [38], UCF YouTube [39], UCF50 [40], UCF101 [41], and HMDB51 [42] datasets consist of a set of realistic action videos, collected from the Internet. They only contain the RGB data for human activity recognition. More challenging work has been done to recognize the abnormal crowd events via a spatio-temporal viscous fluid field by exploring both appearance of crowd behaviors and interaction among pedestrians [71], [72].

C. Multiview Dataset

Since individual action can generate different visual characteristics from different views, multiview datasets have been constructed by setting multiple synchronized cameras for action capturing. Specifically, IXMAS [43] dataset provides action samples observed from five views. However, it lacks realistic human activities and the dark environment makes it not suited for complex action recognition.

D. Multimodal Dataset

With the wide spread of Microsoft Kinect, multimodal datasets (Indoor Activity [45] dataset, RGBD-HuDaAct [46] dataset, etc.) have been built over the past few years. The multimodal data has several advantages over visible color data. First, the depth data provides 3-D structural information of the scene, which offers more information for posture recovery. Second, depth sensor can provide reliable data capture under low light scenario. Third, the multimodal data (RGB, depth, and skeleton data) can be integrated to benefit human action recognition. MSR DailyActivity3D [44] dataset recorded 16 daily human activities (e.g., drink, eat, read book, call cellphone, etc.) with a Kinect sensor. Each action is performed by one actor in the static background with a fixed viewpoint scenario. Similarly, Sung *et al.* [45] constructed an indoor activity dataset for human activity detection. This dataset includes indoor environment, such as office, kitchen, bedroom, bathroom, and living room, and 12 activity categories performed by four subjects. This dataset contains RGB, depth, and skeleton data. HuDaAct [46] dataset aims to encourage application of assisted living in health-care, where the proposed 12 categories of human daily activities are defined by health-care professionals.

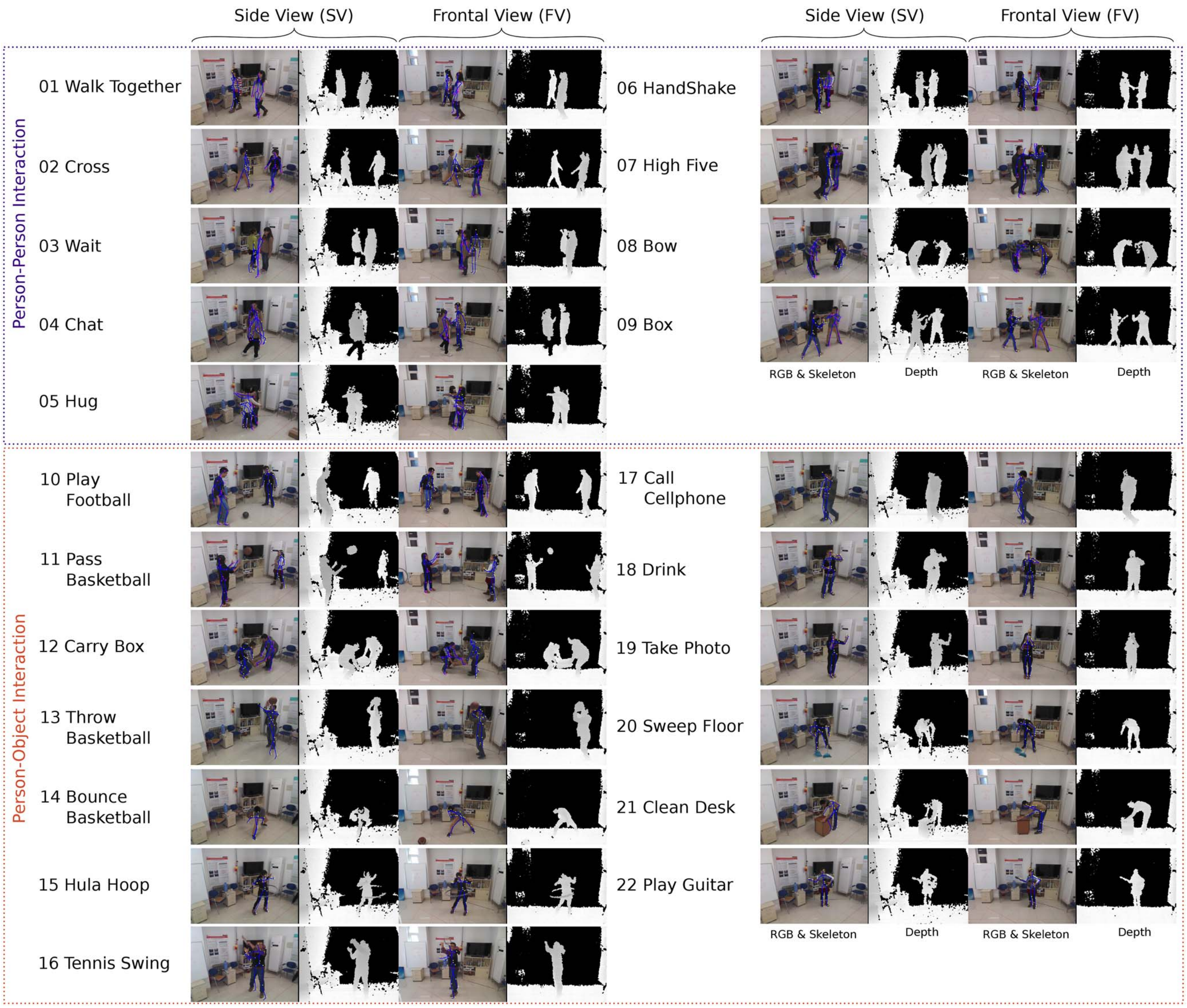


Fig. 1. Image samples and action categories of the proposed M^{2}I dataset.

E. Multiview and Multimodal Dataset

This category consists of RGB, depth, and human skeleton data captured simultaneously by multiple Kinect depth sensors. It aims to encourage studies on multimodality-based human action recognition and sensor fusion. Specifically, Northwestern-UCLA [47] dataset includes ten atomic and interactive action classes taken from three viewpoints, such as pick up with one hand (two hands), drop trash, etc.

III. M^{2}I Human Action Dataset

The proposed M^{2}I dataset[1] provides atomic actions, person–person interactive actions, and person–object interactive actions. In this dataset, two static Kinect depth sensors (frontal view and side view) were used to simultaneously capture the RGB image (320×240 pixels), depth image (320×240 pixels), and skeleton data (3-D coordinates of 20 joints). The dataset was recorded with 30 frames/s. The angle between the primary optical axes of two Kinect sensors was set at 60° to augment the view differences, which can further induce difficulty for shared knowledge discovery from multiple views. To increase the challenge, the indoor environment was set with the cluttered background and the illumination variation. The indoor environment for data capture is shown in Fig. 2.

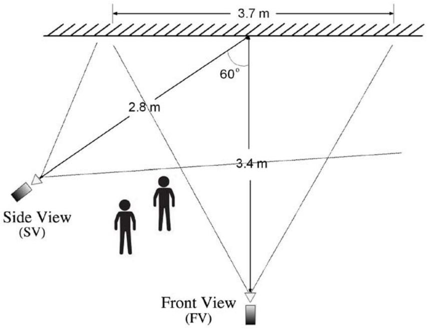


Fig. 2. Sensors configuration of proposed M^{2}I dataset.

[1] http://media.tju.edu.cn/m2i.html

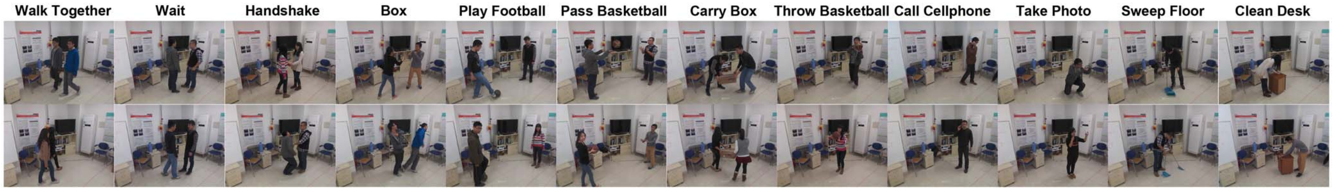


Fig. 3. Image samples of the intraclass variations of the proposed M^{2}I dataset.

Compared to the existing datasets, M^{2}I dataset contains most of common person–person and person–object interactive actions and consequently has richer action diversity. It consists of 22 action categories and 22 persons. The action in each action categories is performed twice by 20 person–person (or person–object) pairs. In total, M^{2}I dataset contains 1760 samples (22 actions × 20 pairs × 2 views × 2 runs). All the RGB, depth, and skeleton data are preprocessed to remove noise. Furthermore, we implemented background modeling and foreground extraction to provide masks for individual frames. Totally, M^{2}I dataset contains the following information: RGB data (6.79G), depth data (49.4G), mask (613M), and 3-D skeleton data (53.9M). The action samples in M^{2}I dataset are shown in Fig. 1 and the intraclass variances samples are shown in Fig. 3.

For evaluation, all samples are divided with respect to the pairs into training set (8 pairs), validation set (6 pairs), and test set (6 pairs). The action models are trained on training set and the validation set is used to optimize the parameters. The performance of action recognition are reported on test set.

IV. Single-View Learning

A. Learning Method

For single-view learning, the BoW+SVM framework [5] is, respectively, evaluated on the frontal view and the side view of M^{2}I dataset in both RGB and depth modalities, respectively. We extracted a set of local spatio-temporal features [27] and k-means algorithm was adopted to learn dictionary for individual views. Each action sample was represented as a bag-of-visual-words with respect to the corresponding dictionary with vector quantization encoding scheme. SVM with the χ^2-kernel was used for model learning. In addition, we also evaluate Fisher vector encoding [29], [73], [74] in this paper. Each action sample is represented by a $2DK$-dimensional descriptor, where D is the dimensionality of the descriptor and K is the number of Gaussians [75]. Linear SVM is employed for classification. SVM can be trained by optimizing its dual problem with the sequential minimal optimization algorithm.

B. Evaluation Protocol

In our experiment, two popular local salient features, STIP [27] [i.e., Harris3D detector with histogram of oriented gradients (HOGs) and histogram of oriented flow (HOF) descriptors] and iDT [28] [i.e., iDT detector with HOG, HOF, trajectory (Tra), and motion boundary histograms (MBH) descriptors] were evaluated. Specifically, for the BoW framework, we extracted a set of local spatio-temporal features, including iDT-Tra, iDT-HOG, iDT-HOF, iDT-MBH, and Harris3D-HOG+HOF. As for [28], iDT-HOG+HOF, iDT-HOF+MBH, and iDT-COM (concatenation of all descriptors) were utilized in the experiments. For the Fisher vector, we only use the iDT-COM feature for the evaluation due to higher computing cost. We used the original implementation and parameter settings provided by the respective authors. All combination of detectors and descriptors were evaluated for all learning problems in this paper.

For BoW descriptor, we clustered a subset of 100 000 randomly selected training features for dictionary learning. The size of the dictionary was empirically set with 1000. For Fisher vector encoding, we first reduce the dimensionality of the descriptor to 64 using principal component analysis [73]. The number of Gaussians is empirically set to $K = 256$ and randomly sample 256 000 features from the training set to train the Gaussian mixture model (GMM). For the multiclass classification, we applied the one-against-rest strategy and selected the optimal parameters by cross-validation.

C. Results

The experimental results are presented in Table IV for different combinations of detectors and descriptors on the frontal/side views in the RGB/depth modalities, respectively.

First, we compare the performance of Harris3D and iDT detectors. With the same HOG+HOF descriptors, iDT can significantly outperform Harris3D. iDT-HOG+HOF can achieve the absolute gains of 13.2%, 11.7%, 21%, 15.3% with respect to RGB-SV, RGB-FV, Depth-SV, Depth-FV, respectively. It indicates that iDT can detect more reliable spatial-temporal interest points with dense trajectories and camera motion correction compared against Harris3D, which can only extract sparse interest points with high motion salience. Moreover, it is observed that iDT performs worse in the depth modality than it does in the RGB modality. Specifically, the largest gap, 13.2%, can be obtained when iDT-Tra was implemented on the side view in RGB and depth modalities. It is quite understandable since it is difficult to track dense optical flow with the poor video quality in the depth modality.

Second, from the upper part of Table IV, iDT+Tra is the lowest performing descriptor when compared to the others because it is simply a concatenation of normalized displacement vectors, while the others are computed in the space-time volume aligned with the trajectory. iDT+HOG descriptor is better than iDT+Tra because it is computed with the orientation of image gradients and captures the static appearance characteristics. Consequently, iDT+HOG is around 3% and 5% better than iDT+Tra on RGB-FV and RGB-SV,

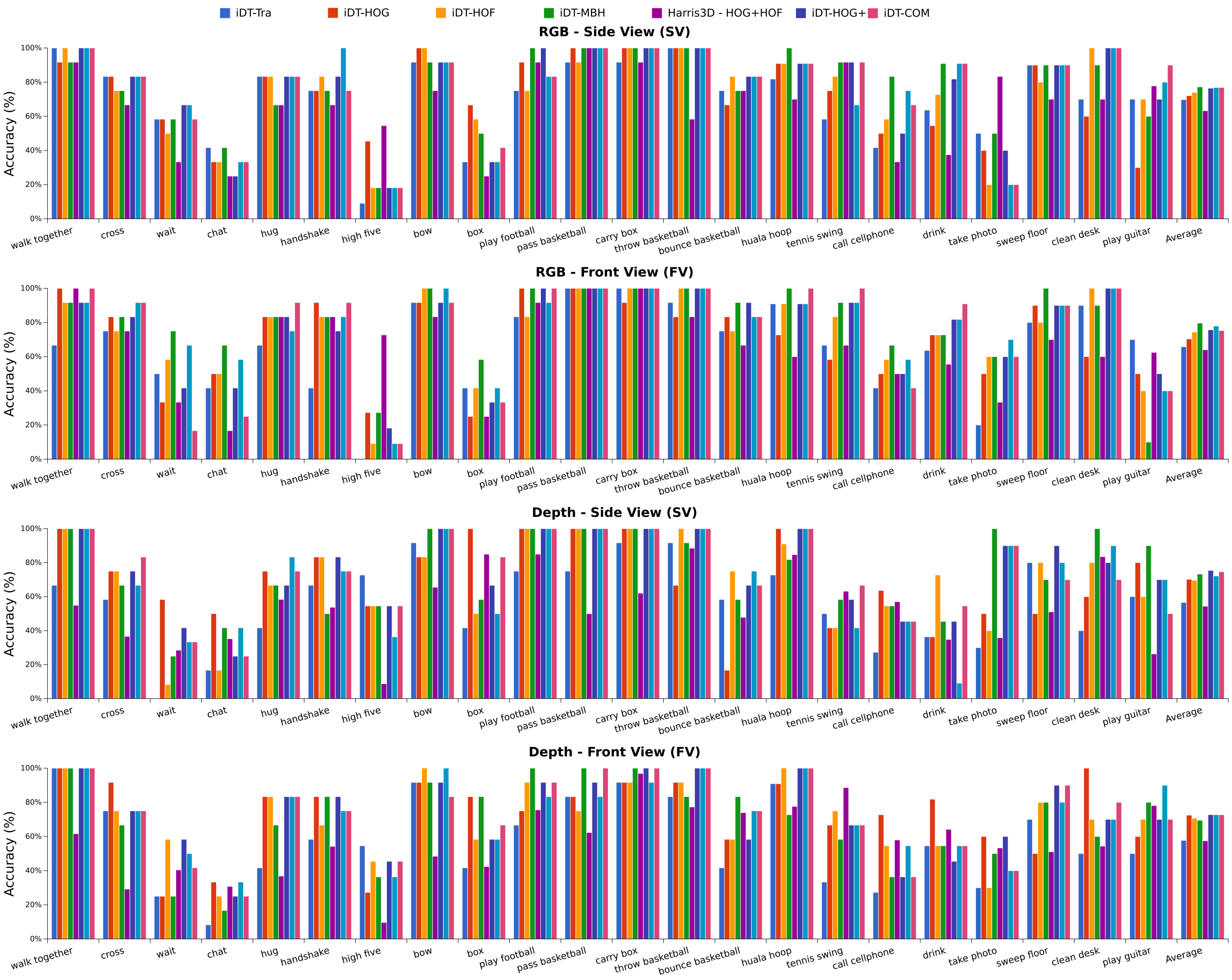


Fig. 4. Category-wise performance comparison under the single-view scenario. The horizontal axis indicates the action indexes and the vertical axis shows the category-wise accuracy.

respectively, and is around 13% and 15% better than iDT+Tra with the respective depth data. Comparing against Tra and HOG, HOF, and MBH obtained better performance by incorporating motion information from optical flow. HOF quantizes the orientation of flow vectors and MBH splits the optical flow into the horizontal and vertical components and quantizes the derivatives of each component, respectively. However, in the depth modality, MBH only shows slightly better results than HOG in the side view and under all the other settings (i.e., iDT-HOF-SV, iDT-HOF-FV, and iDT-MBH-FV), HOF and MBH perform worse than HOG because tracking optical flow is challenging and can even fail with more noise in the depth modality.

Third, as HOF represents zero-order motion information and MBH focuses on first-order derivatives, it is expected that combining HOF and MBH can improve the performance since they are complementary to each other [76]. As shown in Table IV, HOF+MBH is achieved 2% improvement from HOF in both RGB and depth modalities. However, the performance of HOF+MBH is slightly worse than MBH, which may be due to information redundancy and the existence of noise. Based on the iDT-COM feature, we also observed that Fisher vector descriptor always achieves better performance than BoW, which is benefit from the encoding of both first and second order statistics between the action descriptors and GMM [75]. Specifically, the average improvement on frontal view and side view is about 2% and 4%, respectively.

Last, we compare the category-wise performances on M^{2}I dataset (see Fig. 4). It is obvious that iDT-MBH performs better than the other descriptors. However, in the side view, the performances of 5/22 categories (i.e., *Chat, High Five, Box, Take Photo, and Play Guitar*) are below 60% which indicates that this dataset is very challenging with multiple similar actions (e.g., *Wait and Chat, Handshake and High Five*, etc.). Specifically, the performances of *High Five and Play Guitar* are below 20% with respect to SV and FV. Meanwhile, there is almost 50% gap between SV and FV for *Play Guitar*, which illustrates the significant intraclass variation of M^{2}I dataset.

TABLE II

PERFORMANCE OF DIFFERENT DETECTOR/DESCRIPTOR COMBINATIONS UNDER THE SUPERVISED/UNSUPERVISED SETTING FOR THE CROSS-VIEW SCENARIO (SV→FV: LEARNING IN THE SIDE VIEW AND TEST IN THE FRONTAL VIEW; FV→SV: LEARNING IN THE FRONTAL VIEW AND TEST IN THE SIDE VIEW)

		Supervised				Unsupervised			
Detector	Descriptor	RGB		Depth		RGB		Depth	
		SV→FV	FV→SV	SV→FV	FV→SV	SV→FV	FV→SV	SV→FV	FV→SV
iDT	Tra	43.3%	39.2%	30.1%	37.0%	48.8%	47.6%	34.2%	38.7%
iDT	HOG	60.7%	56.7%	34.5%	36.3%	72.8%	68.2%	41.3%	47.3%
iDT	HOF	74.1%	70.2%	40.2%	44.9%	**83.5%**	**79.5%**	51.2%	56.0%
iDT	MBH	16.9%	23.3%	37.4%	47.0%	15.7%	23.6%	48.9%	57.7%
Harris3D	HOG+HOF	57.7%	61.0%	**43.3%**	**57.7%**	67.4%	69.2%	37.5%	53.9%
iDT	HOG+HOF	**78.2%**	**72.1%**	39.7%	47.5%	82.7%	77.6%	51.6%	58.3%
iDT	HOF+MBH	75.8%	71.8%	41.6%	50.6%	82.3%	79.2%	**53.6%**	61.0%
iDT	COM	70.2%	67.7%	39.1%	50.2%	81.5%	77.5%	51.6%	**63.2%**

V. CROSS-VIEW LEARNING

A. Learning Method

In this paper, we evaluated the transferable dictionary pair learning method [68] for cross-view action recognition in both RGB and depth modalities, respectively. Specifically, we implemented the transferable dictionary pair learning method [68] in both supervised (shared actions in both views are labeled) and unsupervised (shared actions in both views are not labeled) settings to transferring sparse feature representations of videos from the source view to the target view.

1) Unsupervised Setting: Let $Y_s, Y_t \in \mathbb{R}^{n\times N}$ denote the feature representations of N videos of shared actions in the source and target views; $D_s, D_t \in \mathbb{R}^{n\times K}$ build a transferable dictionary pair for feature transformation; $X \in \mathbb{R}^{K\times N}$ is the common sparse representations of Y_s, Y_t; $\|x_i\|_0 \leq s$ is the sparsity constraint. The objective function for transferable dictionary pair learning can be formulated as

$$\min_{D_s,D_t,X} \quad \|Y_s - D_sX\|_2^2 + \|Y_t - D_tX\|_2^2 \quad \text{subject to} \quad \forall i,\ \|x_i\|_0 \leq s. \tag{1}$$

2) Supervised Setting: Let λ control the tradeoff between the reconstruction error and label consistent regularization. The matrix A denotes a linear transformation matrix which transforms the original sparse code X to be the most discriminative sparse feature space $\mathbb{R}^K$. The elements of matrix $Q = [q_1, \ldots, q_N] \in \mathbb{R}^{K\times N}$ consist of the ideal discriminative sparse codes of shared action videos in both views. The objective function for dictionary pair construction is given by

$$\min_{D_s,D_t,A,X} \quad \|Y_s - D_sX\|_2^2 + \|Y_t - D_tX\|_2^2 + \lambda\|Q - AX\|_2^2 \quad \text{subject to} \quad \forall i,\ \|x_i\|_0 \leq s. \tag{2}$$

The transferable dictionary pair $\{D_s, D_t\}$ in (1) and (2) can be efficiently learned with K-SVD [77]. Given the learned dictionary pairs $\{D_s, D_t\}$, we obtain sparse feature representations of the train and test videos in the source and target views. It can be solved by the following optimization problem:

$$\min_X \|Y - DX\|_2^2 \quad \text{subject to} \quad \forall i,\ \|x_i\|_0 \leq s. \tag{3}$$

The orthogonal matching pursuit algorithm [78] can be used to solve (3). At last, K-nearest neighbor algorithm was applied for action recognition.

B. Evaluation Protocol

We varied the dictionary dimension and sparsity coefficient within [50, 100, 200, 300] and [10, 20, 30, 40, 50] for optimal parameter selection, respectively. The K-nearest neighbor ($K = 1$) classifier was used to recognize unlabeled test videos. The leave-one-action-out strategy was used for evaluation as [68]. Different from the previous section, we exclude the evaluate of iDT-COM descriptor with Fisher vector encoding because the dimensionality of the resulting feature is too high for this learning problem.

C. Results

We evaluated the unsupervised and supervised transferable dictionary learning methods under the cross-view scenario. The results are shown in Table II. First, unsupervised learning approach can outperform the supervised method as there exists label bias for the supervised method. Although each pair of samples that captured from different views belong to the same action category, they exhibit diverse appearance and motion characteristics due to the change of viewpoints. Therefore, imposing label consistency regularization might have negative influenced the performance. Second, performance of iDT+MBH with RGB data is significantly lower than the others. MBH is the first-order derivative of optical flow and is much more discriminative with respect to specific views comparing to other descriptors. Consequently, it is extraordinarily difficult to learn a common feature space for multiviews when MBH is utilized. As shown in Fig. 5, the performance of most of actions are below 30% in both supervised and unsupervised methods in SV→FV case. The FV→SV case improves the results slightly but the category-wise accuracy are still below 55%. Specifically, the performances of *Bounce Basketball* and *Call Cellphone* in SV→FV case under the supervised method are only zero. These results show that MBH is not suitable for cross-view learning in the RGB modality. Moreover, the low performances of the cross-view experiments demonstrate

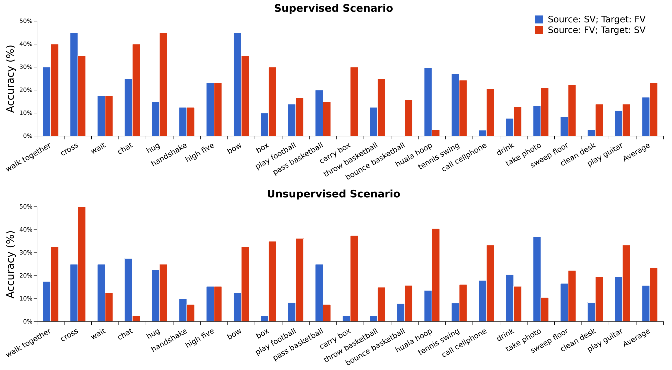


Fig. 5. Category-wise accuracy and average accuracy by iDT-MBH in the RGB modality under the cross-view scenario.

that M^{2}I dataset is very challenging with significant view difference. Third, as shown in Fig. 5, both views can convey complementary information. The performance of *Hug* in the FV→SV case is 30% better than the one in the SV→FV case since the frontal view can show more rich visual information while there exist serious occlusion when observed from the side view.

VI. Cross-Domain Learning

A. Learning Method

For the cross-domain scenario, one view is selected as the auxiliary domain D^A and the other view is regarded as the target domain $D^T = D_l^T \bigcup D_u^T$, where D_l^T and D_u^T represent the labeled and unlabeled data in the target domain, respectively; $D_l^T = (x_i, y_i)_{i=1}^N$, where x_i denotes the feature of the ith data and $y_i \in \{+1, -1\}$ is its binary label. Five representative cross-domain learning methods are evaluated, including adaptive transfer SVM (SVM-AT) [15], feature replication (FR) [16], multiple kernel learning (MKL) [17], domain transfer MKL (DT-MKL) [18], and adaptive MKL (A-MKL) [15]. We further compare these methods against the single-view learning method stated in Section IV.

1) SVM-AT: SVM-AT jointly utilizes the labeled samples from D^A and D^T to train a multikernel SVM classifier. In this paper, we utilized four types of kernel, including Gaussian, Laplacian, inverse square distance, and inverse distance kernels. The objective function can be formulated as

$$\begin{aligned}
&\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\
&\text{subject to} \ \ y_i\left(\frac{1}{M}\sum_{m=1}^{M} w_m^T\phi_m(x) + b\right) \geq 1 - \xi_i \\
&\qquad\qquad \xi_i \geq 0 \ \forall (x_i, y_i) \in D_l^T \cup D^A
\end{aligned} \tag{4}$$

where $\sum_i \xi_i$ measures the total classification error; $\|w\|^2$ is a regularization term that is inversely related to margin between training examples of two classes; M denotes the kernel number; and b is an unregularized bias term. This objective function seeks a decision boundary that achieves a small classification error and creates large margin.

2) FR: FR supposes that the data points x are drawn from a reproducing kernel Hilbert space H (RKHS) with positive semi-definite kernel $K : H \times H \to \mathbb{R}$. Then, K can be written as the dot product (in H) of two (perhaps infinite-dimensional) vectors: $K(x, x') = \langle \phi(x), \phi(x')\rangle_H$. Define ϕ^s and ϕ^t in terms of ϕ to map the auxiliary and target data to the expanded RKHS, as

$$\begin{aligned}
\phi^s(x) &= \langle \phi(x), \phi(x), 0\rangle \\
\phi^t(x) &= \langle \phi(x), 0, \phi(x)\rangle.
\end{aligned} \tag{5}$$

It can compute the kernel product between ϕ^s and ϕ^t by making use of the original kernel K and denotes the expanded kernel by $K'(x, x')$

$$K'\left(x, x'\right) = \begin{cases} 2K\left(x, x'\right) & \text{same domain} \\ K\left(x, x'\right) & \text{different domain.} \end{cases} \tag{6}$$

Equation (6) means that data points from the target domain have twice as much influence as auxiliary points for prediction. Then, K' is utilized to recompute the multiple kernels introduced in SVM-AT for modeling.

3) MKL: This approach considers the combination of kernel matrices for the SVM and shows that the optimization of the coefficient $d = [d_1, \ldots, d_M] \in \mathbb{R}^M$ for such combination can be formulated as

$$\begin{aligned}
&\min_{w_m, b, \xi_i} \ \frac{1}{2}\sum_{m=1}^{M} d_m\|w_m\|^2 + C\sum_{i=1}^{N}\xi_i \\
&\text{subject to} \ \ y_i\left(\sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b\right) \geq 1 - \xi_i \\
&\qquad\qquad \xi_i \geq 0, d_m \geq 0, \ \sum_{m=1}^{M} d_m = 1 \ \forall \ (x_i, y_i) \in D^A \cup D_l^T.
\end{aligned} \tag{7}$$

The SMO-based algorithm [17] is proposed to solve the combination of multiple kernels learning problem.

4) DT-MKL: This method aims to reduce the difference of data distribution between the auxiliary and target domains. The mismatch can be measured by maximum mean discrepancy (MMD) [79] based on the distance (DISK) between the means of samples from D^A and D^T in the RKHS

$$\begin{aligned}
DISK\left(D^A, D^T\right) &= \left\| \frac{1}{n_A}\Sigma_{i=1}^{n_A}\varphi\left(x_i^A\right) - \frac{1}{n_T}\Sigma_{i=1}^{n_T}\varphi\left(x_i^T\right)\right\|_H \\
&= \Omega(d) \quad \text{s.t} \quad d \in D
\end{aligned} \tag{8}$$

where x_i^A and x_i^T are the samples from D^A and D^T, respectively; a simplex $D = \{d \in \mathbb{R}^M | d \geq 0, \sum_m d = 1\}$ is the feasible set of linear combination coefficient d. Define $J(d)$ as the optimum value in (7) and θ as the balanced factor. The optimization problem in DT-MKL is then formulated as

$$\min_{d \in D} G(d) = \frac{1}{2}\Omega^2(d) + \theta J(d) \tag{9}$$

which can be learned by solving a semi-definite programming problem.

TABLE III
PERFORMANCES OF CROSS-DOMAIN SCENARIO WITH EIGHT FEATURE POINT DETECTORS AND DESCRIPTOR COMBINATIONS BY AUGMENTING THE TRAINING+VALIDATION DATA IN THE AUXILIARY DOMAIN (FV/SV) AND THE TARGET DOMAIN (SV/FV) TOGETHER FOR MODEL LEARNING AND THEN TESTING ON THE TARGET DOMAIN (SV/FV)

	Detector	Descriptor	**SVM** (SV / FV)	**SVM-AT** (SV / FV)	**FR** (SV / FV)	**MKL** (SV / FV)	**DT-MKL** (SV / FV)	**A-MKL** (SV / FV)
RGB	iDT	Tra	**69.8 / 65.8**%	59.9 / 61.0%	61.7 / 63.3%	64.1 / 64.2%	62.2 / 63.7%	62.9 / 63.6%
	iDT	HOG	72.1 / 70.4%	69.9 / 69.2%	71.1 / 71.7%	72.4 / 71.9%	71.6 / 71.8%	**72.7 / 72.0**%
	iDT	HOF	74.0 / 74.4%	71.4 / 72.9%	72.9 / 73.2%	73.9 / 73.9%	74.1 / 73.8%	**74.3 / 74.5**%
	iDT	MBH	**77.2 / 79.6**%	52.9 / 57.7%	53.6 / 58.9%	52.1 / 58.4%	52.3 / 57.2%	54.0 / 59.2%
	Harris3D	HOG+HOF	63.3 / 64.0%	62.9 / 62.9%	64.4 / 65.0%	63.7 / 64.5%	63.9 / 64.4%	**64.9 / 65.1**%
	iDT	HOG+HOF	76.5 / 75.7%	74.8 / 75.3%	76.2 / 76.2%	76.1 / 75.7%	76.1 / 75.8%	**77.0 / 76.9**%
	iDT	HOF+MBH	**76.8** / 78.0%	74.5 / 76.8%	75.0 / 78.2%	75.7 / 79.4%	75.9 / 79.0%	76.4 / **79.6**%
	iDT	COM	76.9 / 75.3%	76.8 / 72.8%	78.8 / 75.0%	78.1 / 75.1%	78.1 / 74.5%	**78.9 / 75.4**%
	iDT-COM (Fisher Vector)		78.0 / 79.5%	75.3 / 84.2%	75.5 / 80.3%	75.6 / 84.9%	77.1 / 83.6%	**78.1 / 85.0**%
Depth	iDT	Tra	56.6 / 57.7%	55.8 / 57.5%	57.6 / 58.5%	56.8 / 58.7%	57.5 / 58.2%	**58.2 / 59.2**%
	iDT	HOG	**70.2 / 72.5**%	64.3 / 64.3%	66.0 / 64.2%	64.2 / 63.0%	63.6 / 63.6%	66.2 / 65.3%
	iDT	HOF	**69.7 / 70.7**%	67.7 / 68.4%	68.9 / 69.9%	68.9 / 69.1%	68.4 / 68.9%	69.0 / 70.0%
	iDT	MBH	**73.3 / 69.5**%	63.9 / 66.9%	65.1 / 66.4%	65.0 / 66.8%	63.9 / 66.5%	66.6 / 67.1%
	Harris3D	HOG+HOF	54.4 / **57.5**%	55.2 / 55.2%	54.9 / 54.8%	55.4 / 53.9%	**56.0** / 54.1%	55.7 / 55.2%
	iDT	HOG+HOF	75.4 / 72.8%	72.7 / 72.7%	74.6 / 74.7%	73.4 / 73.7%	73.1 / 73.4%	**75.8 / 75.7**%
	iDT	HOF+MBH	**72.2 / 72.8**%	67.7 / 68.4%	68.4 / 68.4%	68.9 / 69.1%	68.4 / 68.9%	69.0 / 70.0%
	iDT	COM	74.7 / **72.7**%	73.2 / 68.8%	74.8 / 70.8%	73.9 / 69.8%	72.8 / 69.2%	**75.2** / 70.9%
	iDT-COM (Fisher Vector)		80.7 / 78.9%	73.7 / 79.3%	81.8 / 85.5%	82.9 / 84.8%	82.3 / 85.6%	**84.2 / 86.2**%

5) A-MKL: This method adapts the target classifier $f^T(x)$ from an auxiliary classifier $f^A(x)$ trained based on the samples from both domains. Specifically, the target decision function is defined as $f^T(x) = f^A(x) + \Delta f(x)$, where $\Delta f(x)$ is the so-called perturbation function. It can also employ multiple auxiliary classifiers by equally fusing them to obtain $f^A(x)$. Thus, $f^T(x)$ can be rewritten as follows:

$$f^T(x) = \sum_{p=1}^{P} \beta_p f_p^a(x) + \sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b \tag{10}$$

where $\sum_{p=1}^{P} \beta_p f_p^a(x)$ denotes the prelearned classifiers trained with the labeled data from both domains; $\beta = [\beta_1, \ldots, \beta_P] \in \mathbb{R}^P$ is the coefficient for linear combination; and $\Delta f(x) = \sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b$ is the perturbation function. The optimization problem of A-MKL can be formulated as

$$\begin{aligned} \min_{w_m, \beta, b, \xi_i} \quad & \frac{1}{2}\left(\sum_{m=1}^{M} d_m \|w_m\|^2 + \lambda \parallel \beta \parallel^2\right) + C \sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & y_i \left(\sum_{p=1}^{P} \beta_p f_p^a(x_i) + \sum_{m=1}^{M} d_m w_m^T \phi_m(x_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \ \forall (x_i, y_i) \in D^A \cup D_l^T \sum_{p=1}^{P} \beta_p = 1 \sum_{m=1}^{M} d_m = 1 \end{aligned} \tag{11}$$

where λ is the balanced factor. Equation (11) can be solved by existing SVM solvers as stated in [80].

B. Evaluation Protocol

In order to evaluate the impact by number of training data (both the target and auxiliary domains) on the performance of classification, we designed two cases of training data split for model learning.

Case 1: All data are used in the auxiliary domain, where N pairs are selected in the target domain.

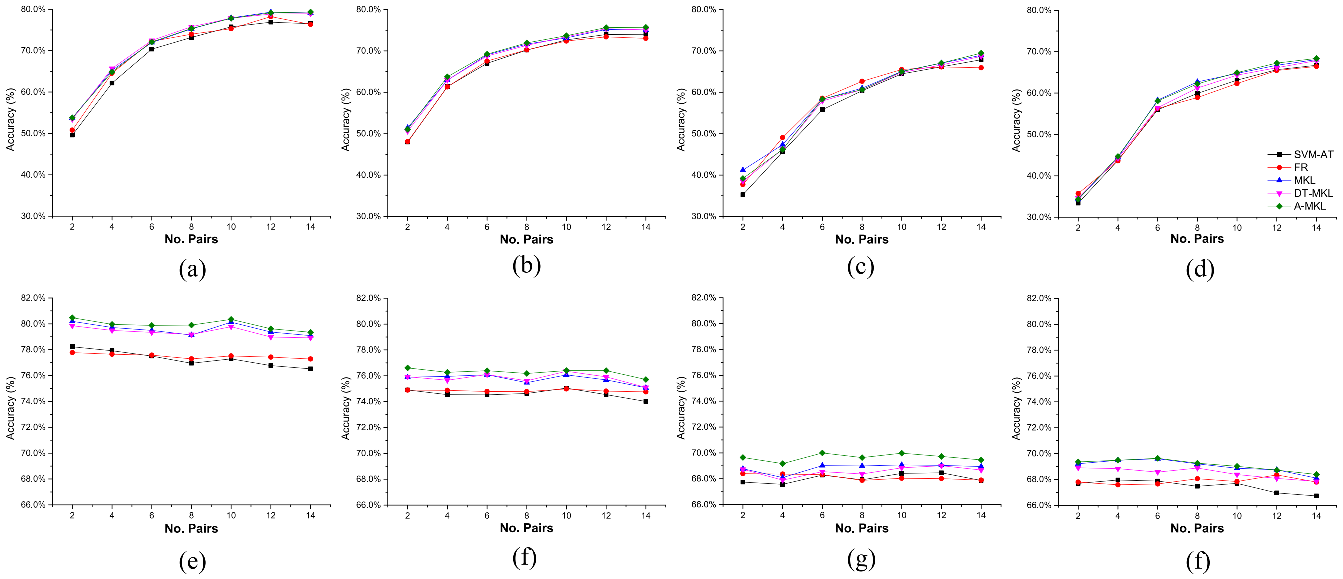
Case 2: N pairs are used in the auxiliary domain, where all data are selected in the target domain.

The value of N varied from 2 to 14. Note that the training data refers to training and validation set. The frontal view and the side view were varied for the target domain and the auxiliary domain.

C. Results

Table III reports the performance of five representative cross-domain learning methods and the single-view learning method with all detector and descriptor combinations in RGB/depth modalities by leveraging the training+validation data in the auxiliary domain (FV/SV) and the target domain (SV/FV) together for model learning and then testing on the target domain (SV/FV). From Table III, we observed that A-MKL consistently outperform SVM-AT, FR, MKL, and DT-MKL. This suggests that by adapting classifier with the multiple base kernels is effective. In addition, A-MKL prelearned average classifiers by minimizing both structural risk functional and mismatch between data distributions from two domains.

We also observed inconsistent performance of cross-domain learning methods when compared to the single-view learning method. Specifically, A-MKL outperforms the single-view method in the RGB modality but always fails in the depth modality. We believe this is due to two factors. First, for A-MKL and DT-MKL methods, it is illustrated that MMD criterion may not be utilized for all situations to capture the mismatch in the distributions between two domains. Second, for the entire cross-domain learning methods, the significant



(a) (b) (c) (d)

(e) (f) (g) (f)

Fig. 6. Classification accuracy of cross-domain scenario with RGB and depth modalities. Top row [i.e., (a)–(d)] varies in the number of data in the target domain (**T**), where the bottom row [i.e., (e)–(h)] augment the data in the auxiliary domain (**A**).

view difference augments the challenge for model learning and prediction. Therefore, it is important to develop robust feature representations and design proper objective function to take advantage of the learned classifiers from both domains for cross-domain prediction. In particular, the performances by iDT-MBH are worse than those by the single-view learning method since MBH is over-discriminative with respect to specific view. Therefore, MBH is not suited for cross-domain learning.

From Table III, we observed that cross-domain methods benefits mostly from Fisher vector, which illustrates the robust feature representations further improve the results. Specifically, RGB-FV, depth-SV, and depth-FV are significantly better than the other three. However, the performances of RGB-SV by Fisher vector is slightly worse than BoW, since both discriminative viewpoint and modality augment the challenge for cross-domain learning.

As iDT-HOF+MBH achieves the optimal performance in the BoW setting, we reviewed its performance with the variation of the person number in the training+validation set from the target/auxiliary domains (see Fig. 6). In case 1, refer to Fig. 6(a)–(d), the performances of all methods can be monotonically increased by augmenting the data in the target domain. This demonstrates that the information from other videos is helpful for improving the performance of robust target classifier for an individual class. Specifically, A-MKL can consistently outperform the other methods when the person number in the target domain increases, which shows the utilization of multiple base kernels as well as prelearned average classifiers can well cope with auxiliary videos. In case 1, refer to Fig. 6(e)–(h), the performances is relatively stable when the data in the auxiliary domain is increased and decrease when many data are augmented. One straightforward reason for the unsatisfactory performances is that the visual features from the auxiliary and target domains might have different distributions and thus it is more likely that the data from the auxiliary domain may degrade the action recognition performances in the target domain for case 2. Overall, A-MKL achieved the best performance in all scenario in case 2.

VII. Multitask Learning

A. Learning Method

We evaluated five representative multitask learning methods, including $\ell_{2,1}$-norm regularization multitask learning (denoted as L_{21}) [22], incoherent sparse and low-rank components multitask learning (denoted as SparseTrace) [23], clustered multitask learning (CMTL) [24], robust multitask learning (RMTL) [25], and robust multitask feature learning (rMTFL) [26]. We further compared these methods against the baseline of the single-view learning method introduced in Section IV.

1) L_{21}: This paper aims to constrain all models to share a common set of features. It is motivated by the penalty of group sparsity

$$\min_{W} \sum_{i=1}^{t} \left\| W_i^T X_i - Y_i \right\|_F^2 + \rho_1 \|W\|_{2,1} + \rho_2 \|W\|_F^2 \qquad (12)$$

where X_i denotes the input matrix of the ith task; Y_i denotes the corresponding label of X_i; W_i is the model for the ith task; $\|\cdot\|_{2,1}$ denotes $\ell_{2,1}$-norm of the matrix where $\|\cdot\|_F$ is the Frobenius norm of the matrix. The regularization parameter ρ_1 and ρ_2 controls group sparsity and ℓ_2-norm penalty, respectively. Equation (12) can be optimized by the iterative algorithm [22].

2) SparseTrace: This approach captures the task relationship by constraining the models of different tasks to share a low-dimensional subspace. The key idea is to decompose the task models W into two components, a sparse part P and a low-rank part Q. It solves the following multitask least squares problem:

$$\min_{W} \sum_{i=1}^{t} \left\| W_i^T X_i - Y_i \right\|_F^2 + \rho_1 \|P\|_1$$
$$\text{subject to} \quad W = P + Q,\ \|Q\|_* \leq \rho_2 \qquad (13)$$

TABLE IV

COMPARISON BETWEEN THE SINGLE-TASK LEARNING AND THE MULTITASK LEARNING FOR VARIOUS DETECTOR/DESCRIPTOR COMBINATIONS IN THE RGB/DEPTH MODALITIES

		RGB				Depth			
Detector	Descriptor	Single-task		Multi-task		Single-task		Multi-task	
		SV	FV	SV	FV	SV	FV	SV	FV
iDT	Tra	69.8%	65.8%	**72.2%**	**82.7%**	56.6%	57.7%	**85.8%**	**78.9%**
iDT	HOG	72.1%	70.4%	**78.4%**	**77.4%**	70.2%	72.5%	**74.3%**	**72.5%**
iDT	HOF	74.0%	74.4%	**82.1%**	**82.1%**	69.7%	70.7%	**77.8%**	**74.7%**
iDT	MBH	77.2%	**79.6%**	**80.8%**	68.6%	73.3%	69.5%	**75.5%**	**72.7%**
Harris3D	HOG+HOF	63.3%	64.0%	**78.7%**	**77.3%**	54.4%	57.5%	**68.0%**	**74.0%**
iDT	HOG+HOF	76.5%	75.7%	**80.6%**	**79.7%**	75.4%	72.8%	**81.5%**	**80.8%**
iDT	HOF+MBH	76.8%	78.0%	**77.0%**	**85.9%**	72.2%	72.8%	**81.4%**	**73.8%**
iDT	COM	76.9%	75.3%	**77.7%**	**81.4%**	74.7%	72.7%	**76.6%**	**75.0%**
iDT-COM (Fisher Vector)		78.0%	79.5%	**88.1%**	**88.7%**	80.7%	78.9%	**86.4%**	**88.0%**

where $\|Q\|_*$ is the trace norm of Q. ρ_1 and ρ_2 penalize the sparsity of P and the rank of Q, respectively. Equation (13) can be solved via accelerated gradient algorithm [23].

3) CMTL: This approach assumes that the models of tasks from the same group are closer to each other than those from different group. Based on the relaxed k-means clustering, the objective function can be formulated as

$$\begin{aligned} \min_{W} \quad & \sum_{i=1}^{t} \left\| W_i^T X_i - Y_i \right\|_F^2 \\ & + \rho_1 \eta (1+\eta) tr\Big(W(\eta I + M)^{-1} W^T\Big) \\ \text{subject to} \quad & tr(M) = k, M \preceq I, M \in S_+^t, \eta = \frac{\rho_2}{\rho_1} \end{aligned} \tag{14}$$

where β and α are the regularization parameters for task relatedness; $\eta = \beta/\alpha > 0$; $M = FF^T$; $F \in R^{t \times k}$ is an orthogonal cluster indicator matrix; $tr(M)$ is the trace of M; k is the number of clusters; I denotes the identity matrix of a proper size; $M \preceq I$ means that $I - M$ is positive semi-definite; S_+^t denotes the set of symmetric positive semi-definite matrices of size t by t; ρ_1 and ρ_2 are the regularization parameters. The accelerated projected gradient method [81] can be used for optimization.

4) RMTL: This approach assumes that all tasks are relevant and aims at identifying irrelevant tasks. The model W can be decomposed into two components, a low-rank structure L that captures task-relatedness and a group-sparse structure S that detects outliers. The following objective function is formulated by considering the incoherent group-sparse and low-rank multitask least square problem:

$$\begin{aligned} \min_{W} \quad & \sum_{i=1}^{t} \left\| W_i^T X_i - Y_i \right\|_F^2 + \rho_1 \|L\|_* + \rho_2 \|S\|_{1,2} \\ \text{subject to} \quad & W = L + S \end{aligned} \tag{15}$$

where ρ_1 penalizes the low-rank regularization on L; ρ_2 penalizes the $\ell_{2,1}$-norm penalty on S. The accelerated proximal method is used to solve this nonsmooth convex optimization [25].

5) rMTFL: This approach assumes that there is no outlier task. The model W can be decomposed into two components, a shared feature structure P that captures task relatedness and a group-sparse structure Q that detects outliers. The following objective function is formulated by considering the robust multitask feature learning with least squares loss:

$$\begin{aligned} \min_{W} \quad & \sum_{i=1}^{t} \left\| W_i^T X_i - Y_i \right\|_F^2 + \rho_1 \|P\|_{2,1} + \rho_2 \left\| Q^T \right\|_{2,1} \\ \text{subject to} \quad & W = P + Q \end{aligned} \tag{16}$$

where ρ_1 penalizes the joint feature learning on P and ρ_2 penalizes the column-wise group sparsity on Q. The accelerated gradient descent can be employed for solution [26].

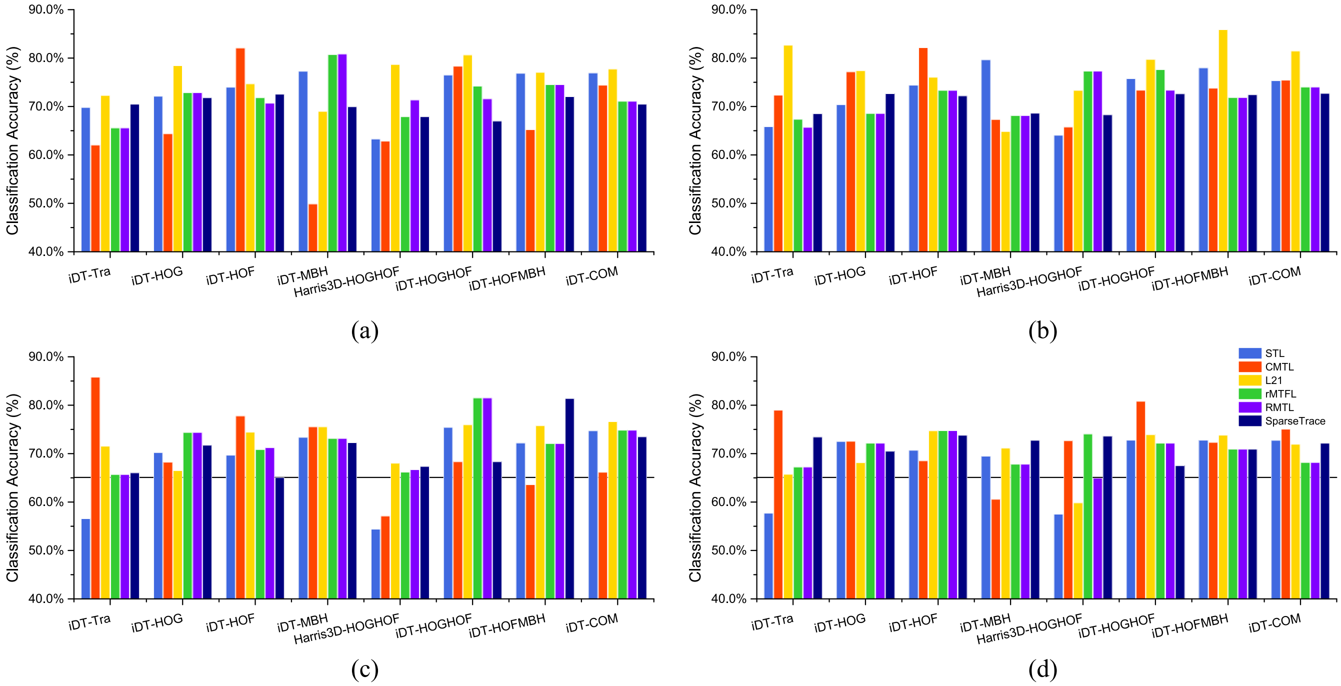
B. Evaluation Protocol

In this section, we evaluate 22 tasks, selected from 22 individual actions, on the frontal view and the side view in both RGB and depth modalities. We empirically varied ρ_1 and ρ_2 for various multitask methods to select the optimal parameters. For CMTL, RMTL, and rMTFL, ρ_1 and ρ_2 varied from 10^{-4} to 1. For SparseTrace, ρ_1 varied from 10^{-4} to 1 and ρ_2 varied from 6 to 22. For L_{21}, ρ_1 varied from 10 to 2000 and ρ_2 varied from 10^{-4} to 1.

C. Results

Table IV shows the results from the baseline single-view learning, also known as single-task learning (STL), and the optimal performance of all five multitask learning (MTL) methods. Performance for various detector and descriptor combinations are shown. It is clear that MTL consistently outperform the STL counterpart in most of the comparison, with the exception of iDT-MBH features with frontal view sensors and RGB modality. Furthermore, the improvement of multitask learning from STL is consistence on the iDT-COM feature with Fisher vector encoding, where an average of 8.5 percentage points improvement is observed across all experiments.

Fig. 7 compares the performances of STL and MTL methods including L_{21}, CMTL, rMTFL, RMTL, and SparseTrace. Meanwhile, we explore various detector/descriptor combinations under SV/FV in the RGB and depth modalities. Also, L_{21} usually outperforms the STL method under four circumstances. Meanwhile, with the same detector and descriptor combinations, L_{21} usually outperforms the other competing



(a)

(b)

(c)

(d)

Fig. 7. Classification accuracy of various multitask learning algorithms. (a) RGB: side view. (b) RGB: frontal view. (c) Depth: side view. (d) Depth: frontal view.

methods. Specifically, L_{21} obtains six best performances in Fig. 7(a). Moreover, it is obvious that not all MTL methods can outperform the STL method under our experimental protocols. Specifically, in the iDT-HOFMBH case of Fig. 7(a), only L_{21} can consistently outperform the STL method.

The individual analysis of the evaluated methods are as follows. First, L_{21} constrains all models to share a common set of features for all tasks by assuming there is no irrelevant task (i.e., outlier) and regarding all categories of M^{2}I dataset as the relevant tasks (inlier). SparseTrace focuses on a common low-dimensional subspace and decomposes the task models into sparse part P and low-rank part Q. The fact that L_{21} can outperform SparseTrace in most cases shows that a common set of features is more important for MTL than a common low-dimension subspace. Second, RMTL, rMTFL, and CMTL have similar sophisticated group structure. Fig. 7 shows that rMTFL usually performs better than RMTL, which illustrates that the joint feature learning and group discovery by rMTFL can benefit improving the multitask performance comparing against RMTL and CMTL. RMTL and CMTL mainly focus on task relatedness discovery.

VIII. Conclusion

In this paper, we introduce and benchmark an M^{2}I dataset with four representative learning problems (i.e., single-view learning, cross-view learning, cross-domain learning, and multitask learning) for human action recognition problem. A thorough evaluation with spatio-temporal salient point detectors (i.e., iDT and Harris3D) and descriptors (i.e., Tra, HOG, HOF, and MBH) are presented. The benchmark performance demonstrates that the proposed M^{2}I dataset is extremely challenging due to large intraclass variations, significant view variations, and multiple similar action categories. Furthermore, the benchmark results provides strong foundation for potential future work in the following directions. First, in order to enhance the robustness of visual representations, deep learning approaches should be leveraged to learn discriminative feature based on convolutional maps. This approach is expected to share the merits of spatio-temporal feature points and simultaneously discover the relationship between diverse modalities. Second, fusion of heterogeneous characteristics from multiple views (i.e., cross-view/cross-domain learning) remains a challenging computer vision task. As shown in the literature of other fields, fusion has the potential to cope with the mismatch in data distribution of various types of local feature. Last, but not least, we consider the appropriate structure for multitask learning to effectively improve the efficacy of exploiting the intrinsic relationships among tasks.

References

[1] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: A genetic programming approach," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 158–170, Jan. 2016.

[2] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 3164–3172.

[3] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[4] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, "Action2activity: Recognizing complex activities from sensor data," in *Proc. Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 1617–1623.

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

[6] M. Devanne *et al.*, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1340–1352, Jul. 2015.

[7] 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, Jul. 2015.

[8] 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. 2016.

[9] A. Mansur, Y. Makihara, and Y. Yagi, "Inverse dynamics for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1226–1236, Aug. 2013.

[10] A. Sen, M. M. Islam, K. Murase, and X. Yao, "Binarization with boosting and oversampling for multiclass classification," *IEEE Trans. Cybern.*, vol. 46, no. 5, pp. 1078–1091, May 2016.

[11] Y. Jiang *et al.*, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 688–701, Apr. 2015.

[12] Z. Gao, W. Nie, A. Liu, and H. Zhang, "Evaluation of local spatial–temporal features for cross-view action recognition," *Neurocomputing*, vol. 173, pp. 110–117, Jan. 2016.

[13] A. Farhadi and M. K. Tabrizi, *Learning to Recognize Activities From the Wrong View Point* (LNCS 5302). Heidelberg, Germany: Springer, 2008, pp. 154–166.

[14] Z. Cui *et al.*, "Flowing on Riemannian manifold: Domain adaptation by shifting covariance," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2264–2273, Dec. 2014.

[15] 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.

[16] H. Daumé, III, "Frustratingly easy domain adaptation," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, Prague, Czech Republic, 2007, pp. 256–263.

[17] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int. Conf. Mach. Learn.*, Banff, AB, Canada, 2004, p. 6.

[18] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.

[19] A.-A. Liu *et al.*, "Multiple/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, Jun. 2015.

[20] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and Z.-X. Yang, "Jointly learning multiple sequential dynamics for human action recognition," *PLoS One*, vol. 10, no. 7, pp. 1–21, 2015.

[21] A.-A. Liu *et al.*, "Single/multi-view human action recognition via regularized multi-task learning," *Neurocomputing*, vol. 151, pp. 544–553, Mar. 2015.

[22] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 41–48.

[23] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 457–464.

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

[25] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, San Diego, CA, USA, 2011, pp. 42–50.

[26] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Beijing, China, 2012, pp. 895–903.

[27] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[28] 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.

[29] X. Peng, C. Zou, Y. Qiao, and Q. Peng, *Action Recognition With Stacked Fisher Vectors* (LNCS 8693). Cham, Switzerland: Springer, 2014, pp. 581–595.

[30] B. Yao *et al.*, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 1331–1338.

[31] J. Zhang *et al.*, "Human action recognition bases on local action attributes," *J. Elect. Eng. Technol.*, vol. 10, no. 3, 2015, pp. 1264–1274.

[32] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[33] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Pattern Recognit.*, Cambridge, U.K., 2004, pp. 32–36.

[34] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[35] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, 2010, pp. 9–14.

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

[37] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 2929–2936.

[38] 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 Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.

[39] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1996–2003.

[40] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of Web videos," *Mach. Vis. Appl.*, vol. 24, no. 5, pp. 971–981, 2013.

[41] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," Center Res. Comput. Vis., Univ. Central Florida, Orlando, FL, USA, Tech. Rep. CRCV-TR-12-01, Nov. 2012.

[42] 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.*, Barcelona, Spain, 2011, pp. 2556–2563.

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

[44] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 1290–1297.

[45] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. AAAI Workshop Pattern Activity Intent Recognit.*, Stanford, CA, USA, 2011, pp. 47–55.

[46] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Consumer Depth Cameras for Computer Vision, Research Topics and Applications*. London, U.K.: Springer, 2013, pp. 193–208.

[47] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. Zhu, "Cross-view action modeling, learning, and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2649–2656.

[48] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2083–2090.

[49] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Gold Coast, QLD, Australia, 2014, pp. 415–424.

[50] W. Bian, D. Tao, and Y. Rui, "Cross-domain human action recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 298–307, Apr. 2012.

[51] A.-A. Liu, W. Z. Nie, Y. Gao, and Y. Su, "Multi-modal clique-graph matching for view-based 3D model retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2103–2116, May 2016.

[52] M. Liu, D. Zhang, D. Shen, and the Alzheimer's Disease Neuroimaging Initiative, "View-centralized multi-atlas classification for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 36, no. 5, pp. 1847–1865, 2015.

[53] L. Nie *et al.*, "Beyond doctors: Future health prediction from multimedia and multimodal observations," in *Proc. ACM Multimedia*, Brisbane, QLD, Australia, 2015, pp. 591–600.

[54] M. Liu, D. Zhang, E. Adeli-Mosabbeb, and D. Shen, "Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1473–1482, Jul. 2016.

[55] A.-A. Liu, Z. Wang, W. Nie, and Y. Su, "Graph-based characteristic view set extraction and matching for 3D model retrieval," *Inf. Sci.*, vol. 320, pp. 429–442, Nov. 2015.

[56] Y. Gao, M. Wang, R. Ji, X. Wu, and Q. Dai, "3-D object retrieval with Hausdorff distance learning," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 2088–2098, Apr. 2014.

[57] G. Wang, F. Wang, T. Chen, D.-Y. Yeung, and F. H. Lochovsky, "Solution path for manifold regularized semisupervised classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 308–319, Apr. 2012.

[58] 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.*, no. 99, p. 1, 2016.

[59] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understand.*, Mar. 2016.

[60] W. Choi and S. Savarese, "Understanding collective activities of people from videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1242–1257, Jun. 2012.

[61] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, 2015, pp. 115–129.

[62] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 835–848, Apr. 2013.

[63] D. Wu *et al.*, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.

[64] L. Sun *et al.*, "DL-SFA: Deeply-learned slow feature analysis for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2625–2632.

[65] A.-A. Liu *et al.*, "Coupled hidden conditional random fields for RGB-D human action recognition," *Signal Process.*, vol. 112, pp. 74–82, Jul. 2015.

[66] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, Mar. 2015.

[67] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, 2012, pp. 20–27.

[68] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," in *Proc. Brit. Mach. Vis. Conf.*, Surrey, U.K., 2012, pp. 1–11.

[69] N. Xu *et al.*, "Multi-modal & multi-view & interactive benchmark dataset for human action recognition," in *Proc. ACM Multimedia*, Brisbane, QLD, Australia, 2015, pp. 1195–1198.

[70] B. Wu, C. Yuan, and W. Hu, "Human action recognition based on context-dependent graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2609–2616.

[71] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei, "The large-scale crowd behavior perception based on spatio-temporal viscous fluid field," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1575–1589, Oct. 2013.

[72] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei, "Crowd event perception based on spatio-temporal viscous fluid field," in *Proc. AVSS*, Beijing, China, 2012, pp. 458–463.

[73] F. Perronnin, J. Sánchez, and T. Mensink, *Improving the Fisher Kernel for Large-Scale Image Classification* (LNCS 6314). Heidelberg, Germany: Springer, 2010, pp. 143–156.

[74] J. Wu, Y. Zhang, and W. Lin, "Good practices for learning to recognize actions using FV and VLAD," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2015.2493538.

[75] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[76] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 3551–3558.

[77] 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.

[78] 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.

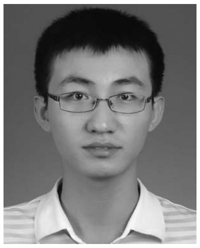
[79] K. M. Borgwardt *et al.*, "Integrating structured biological data by kernel maximum mean discrepancy," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, Fortaleza, Brazil, 2006, pp. 49–57.

[80] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[81] I. Daubechies, M. Fornasier, and I. Loris, "Accelerated projected gradient method for linear inverse problems with sparsity constraints," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 764–792, 2008.

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

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



Ning Xu is currently pursuing the doctor's degree with the School of Electronic 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, Tianjin, China.

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

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

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

Yongkang Wong (M'08) received the B.Eng. degree from the University of Adelaide, Adelaide, SA, USA, and the Ph.D. degree from the University of Queensland, Brisbane, QLD, Australia.

He is a Research Fellow and an Assistant Director of the SeSaMe Centre in the Interactive and Digital Media Institute, National University of Singapore, Singapore. He was a Graduate Researcher with NICTA's Queensland Laboratory, from 2008 to 2012. His current research interests include computer vision, machine learning, multi-camera analysis, and video surveillance.

Mohan Kankanhalli (F'14) received the B.Tech. degree from Indian Institute of Technology Kharagpur, Kharagpur, India, and the M.S. and Ph.D. degrees from Rensselaer Polytechnic Institute, Troy, NY, USA.

He is the Provost's Chair Professor with the Department of Computer Science, National University of Singapore (NUS), Singapore. He is also the Vice Provost for Graduate Education at NUS. His current research interests include multimedia computing, multimedia security, image/video processing, and social media analysis.

Prof. Kankanhalli was a recipient of the Singapore's National Research Foundation to set up the SeSaMe Centre. He is very 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, including *ACM Transactions on Multimedia Computing, Communications, and Applications*, *Springer Multimedia Systems*, and *Multimedia Tools and Applications*.