A Multi-modal & Multi-view & Interactive Benchmark Dataset for Human Action Recognition

Ning Xu[1], Anan Liu[1*], Weizhi Nie[1], Yongkang Wong[2], Fuwu Li[1], Yuting Su[1]
[1]School of Electronic Information Engineering, Tianjin University, China
[2] Interactive & Digital Media Institute, National University of Singapore, Singapore
*Corresponding author: anan0422@gmail.com

ABSTRACT

Human action recognition is one of the most active research areas in both computer vision and machine learning communities. Several methods for human action recognition have been proposed in the literature and promising results have been achieved on the popular datasets. However, the comparison of existing methods is often limited given the different datasets, experimental settings, feature representations, and so on. In particularly, there are no human action dataset that allow concurrent analysis on three popular scenarios, namely single view, cross view, and cross domain. In this paper, we introduce a Multi-modal & Multi-view & Interactive (M^2I) dataset, which is designed for the evaluation of the performances of human action recognition under multi-view scenario. This dataset consists of 1760 action samples, including 9 person-person interaction actions and 13 person-object interaction actions. Moreover, we respectively evaluate three representative methods for the single-view, cross-view, and cross-domain human action recognition on this dataset with the proposed evaluation protocol. It is experimentally demonstrated that this dataset is extremely challenging due to large intraclass variation, multiple similar actions, significant view difference. This benchmark can provide solid basis for the evaluation of this task and will benefit advancing related computer vision and machine learning research topics.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-*Indexing methods*

General Terms

Experimentation

Keywords

human action recognition, multi-modal, multi-view

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806315.

1. INTRODUCTION

Human action recognition has received increasing attention owing to its rich real-world applications, such as multimedia retrieval, human computer interaction, and video surveillance [1]. Depending on the environments, human actions may have different forms including atomic actions and interactions composed by person-object and person-to-person interactions. Existing action recognition dataset mainly focus on atomic actions (e.g., KTH [7], UCF50 [6], IXMAS [10], *etc.*), which limits the performance of video-based human activity recognition. Recently, researchers are engaging in person-person interaction actions and person-object interaction actions [2]. However, the current person-person interaction action dataset [3] and person-object interaction action dataset [5] only include limited action categories and video samples.

With the recent advancement of the cost-effective depth sensor (e.g., Kinect depth sensor), there is a significant attention of using depth data for various computer vision task in the research community. Multiple methods have been developed by leveraging RGB images and/or depth data on several popular RGB-D datasets (e.g., MSR Daily Activity [9], UTKinect [11], *etc.*). Until now the state-of-the-art approaches have achieved satisfactory performances on most of the public datasets [8].

Based on the evolution of human action datasets, it is necessary to prepare a challenging multi-modal & multi-view dataset with both person-person interaction and person-object interaction. To bridge the gap, we proposed a new video dataset which enumerate three major challenges to vision based on human action recognition. The first challenge is variation in modality, which contains RGB, depth and 3D body joints data captured simultaneously by Kinect sensors. The second challenge is variation in camera view, where the same actions can generate a different appearance from different perspectives. Moreover, double views is appropriate which is similar to the human physical features. The third challenge is the intra-class variability and inter-class similarity of actions. Individuals can perform an action in different directions with different characteristics of body part movements, and two actions may be only distinguished by very subtle spatio-temporal differences.

For human action recognition, there are three key tasks:

- *Single-view task*: The model learning and evaluation are independently conducted on two non-overlapping parts from the same dataset. One representative method was done by Wang *et al.*, which evaluated the bag of

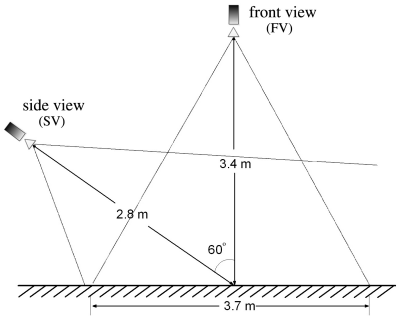

Figure 1: Camera configuration of the M^{2}I dataset.

visual word descriptor and SVM classifier on 9 popular datasets [11].

- *Cross-view task*: The model learning and evaluation are conducted on datasets with two overlapping views on the same observation. Transferable dictionary pair learning proposed by Zheng *et al.* [12] is a representative method. The basic philosophy is mapping the features from different views into the same feature space by the sparse representation framework to handle the cross-view action representation.

- *Cross-domain task*: It leverages a portion of the target dataset together with the auxiliary dataset for model learning and then evaluates on the remaining target dataset. Adaptive multiple kernel learning proposed by Xu *et al.* [4] is a representative method. It leverages fewer labeled data from the target domain and large-scale labeled data from the auxiliary domain together to augment the generalization ability of model learning.

However, there is lack of systematic evaluation of the representative methods under three scenarios on one benchmark dataset.

To tackle these problems, we benchmark a novel database for human action recognition. The major contributions are two-fold: (1) We contribute a Multi-modal & Multi-view & Interactive (M^{2}I) dataset[1] for human action recognition, which can be utilized for the evaluation of the three scenarios above. (2) We systematically evaluate three aforementioned representative methods (i.e., single-view, cross-view, and cross-domain scenarios) with the designed experimental protocol. Extensive experiments demonstrate that M^{2}I dataset is extremely challenging due to large intraclass variation, multiple similar actions, significant view difference. The details of the proposed new dataset is elaborated in Section 2, and the evaluation of the three experiment scenarios are shown in Section 3. Section 4 concludes the paper.

2. M^{2}I HUMAN ACTION DATASET

The proposed Multi-modal & Multi-view & Interactive (M^{2}I) dataset provide person-person interaction actions and person-object interaction actions. In this dataset, two static Kinect depth sensors were used to simultaneously capture the RGB image (320 × 240), depth image (320 × 240) and skeleton data (3D coordinates of 20 joints per frame) from both the front and side views. The dataset was recorded with 30 frames per second. The angle between the primary optical axes of two Kinects depth sensor was set with 60° to augment the view differences, which can further induce difficulty for shared knowledge discovery from multiple views. To increase the challenges, the layout of the indoor environment was set with cluttered background and illumination variation. The physical configurations of the data recording environment (indoor) is shown in Fig. 1.

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

Table 1: List of action categories of M^{2}I dataset.

Person-person Interaction		
1 Walk Together	2 Cross	3 Wait
4 Chat	5 Hug	6 Handshake
7 High Five	8 Bow	9 Box
Person-object Interaction		
10 Play Football	11 Pass Basketball	12 Carry Box
13 Throw Basketball	14 Bounce Basketball	15 Hula Hoop
16 Tennis Swing	17 Call Cellphone	18 Drink
19 Take Photo	20 Sweep Floor	21 Clean Desk
22 Play Guitar		

Compared against existing datasets, M^{2}I dataset contains most common person-person/person-object interaction actions and consequently has richer action diversity. It consists of 22 action categories (see Table 1) and a total of 22 unique individuals. Each action is performed twice by 20 groups (two persons in a group). In total, M^{2}I dataset contains 1760 samples (22 actions × 20 groups × 2 views × 2 run). All the RGB image, depth data, 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 (image sequence sample: 6.79G; video sample: 19.2G); Depth data (image sequence sample: 49.4G); mask (image sequence sample: 613M); 3D Skeleton data (53.9M). Image samples of selected action in the M^{2}I dataset is shown in Fig. 2.

For evaluation, all samples were divided with respect to the groups into a training set (8 groups), a validation set (6 groups) and a test set (6 groups). The classifiers are trained on the training set while the validation set is used to optimize the parameters. The final action recognition results are obtained with the test set.

3. EXPERIMENTS

We evaluated three representative methods under single-view, cross-view, and cross-domain scenarios on M^{2}I dataset.

3.1 Single-View Task

For single-view task, we respectively evaluated the BoW+SVM framework [8] on the front view and the side view of M^{2}I dataset in both RGB and depth. The popular dense trajectory+HOG&HOF spatiotemporal features were extracted. Then K-means was used to quantize them into visual words and a video is represented as the frequency histogram over the visual words. We empirically set the number of visual words with 1000. To limit the complexity, we cluster a subset of 100,000 randomly selected training features. Support vector machine (χ^2-kernel) was used for model learning. For multi-class classification, we applied the one-against-rest approach and selected the optimal parameters by cross validation. The split strategy was used for evaluation as [8] and

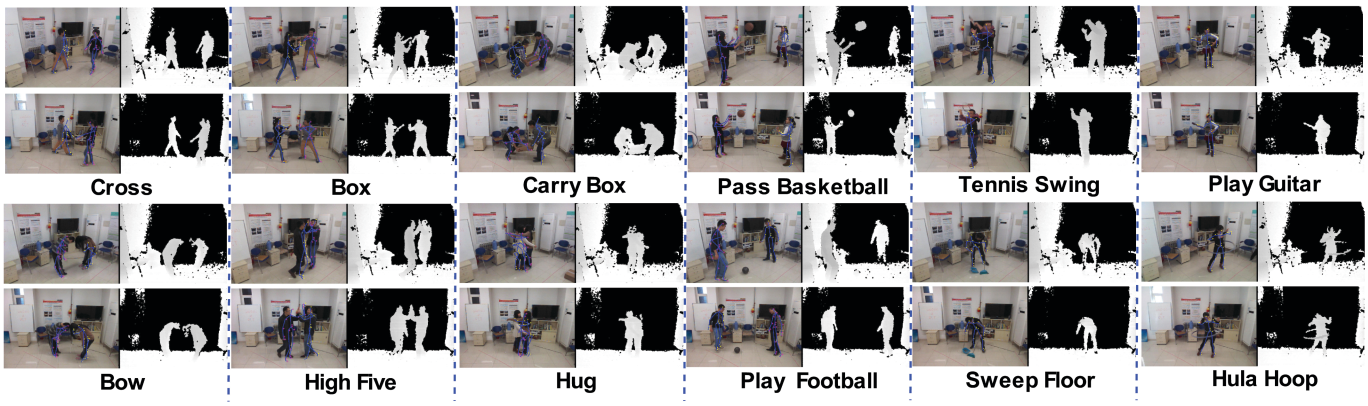

Figure 2: Image samples of selected action in the M^{2}I dataset. Each cell shows the front-view and side-view samples (RGB, depth, skeleton) of actions.

the model was trained on the training+validation data and tested on the test data.

We can achieve 75.7%/72.8% and 76.5%/75.4% on the front view and the side view of M^{2}I dataset in RGB/depth modality, respectively. With the category-wise comparison in Fig. 3, the performances in RGB modality can outperform most of those in depth modality with richer visual information and both modalities are complementary with each other. The low overall performances and multiple category-wise accuracy below 80% indicate that this dataset is very challenging with multiple similar actions (e.g., Wait & Chat, Handshake & High Five, *etc.*) and significant intraclass variation required when data capturing.

3.2 Cross-View Task

For feature representation, we respectively constructed the BoW features for individual views and individual modalities as introduced in Section 3.1. Then we implemented the transferable dictionary pair learning method [12] 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 to target view on the RGB and depth data of M^{2}I dataset, respectively. We varied the dictionary dimension and sparsity coefficient within [50, 100, 200, 300] and [10, 20, 30, 40, 50], respectively for the optimal performances. Then K-Nearest Neighbor (K=1) was utilized for action recognition. The leave-one-action-out strategy was used for evaluation as [12].

Fig. 4 shows: 1) In the RGB modality, learning in the side view to test in the front view can usually outperform learning in the from view to test in the side view; 2) In the depth modality, the conclusion is just the opposite since the front view usually has more significant depth variation; 3) The unsupervised learning case can outperform the supervised learning case by avoiding the confusion caused by the similar actions with different labels. The cross-view experiment also shows M^{2}I dataset is very challenging with significant view difference.

3.3 Cross-Domain Task

For cross-domain case, we selected one view as the target domain and the other as the auxiliary domain and mainly evaluated the Adaptive Multiple Kernel Learning (AMKL) method. As [4], we compared AMTL against the baseline by the single view-based method (SVM-T) introduced

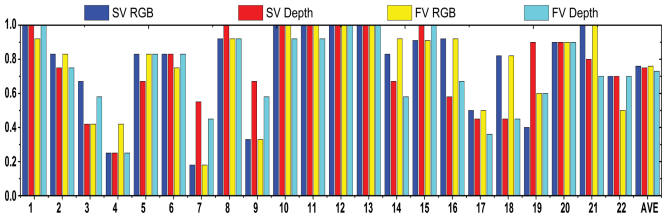

Figure 3: Category-wise accuracy and average accuracy on the RGB and Depth data under the single-view scenario. (SV: side view; FV: front view)

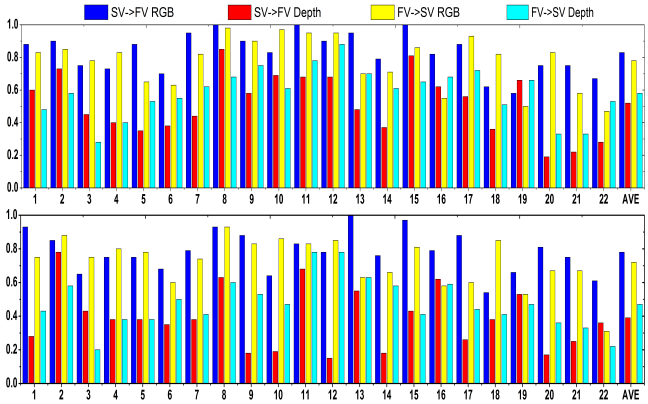

Figure 4: Category-wise accuracy and average accuracy on the RGB and Depth data under the cross-view scenario. The top bar figure corresponds to unsupervised approach and the bottom bar figure corresponds to supervised approach.

in Section 3.1 and four representative cross-domain learning methods, including DTSVM, MKL, FR, SVM-AT as [4]. Two different experiments for model learning were set: 1) Case 1: all the training+validation data of 14 persons in the auxiliary domain plus the data of N persons belonging to the training+validation part in the target domain; 2) Case 2: all the training+validation data of 14 persons in the target domain plus the data of N persons belonging to the training+validation part in the auxiliary domain. N was varied from 2 to 14. The front view and the side view were varied for the target domain and the auxiliary domain.

Fig. 5 shows that: 1) For Case 1, the performances of all methods can be monotonically increased by augmenting the

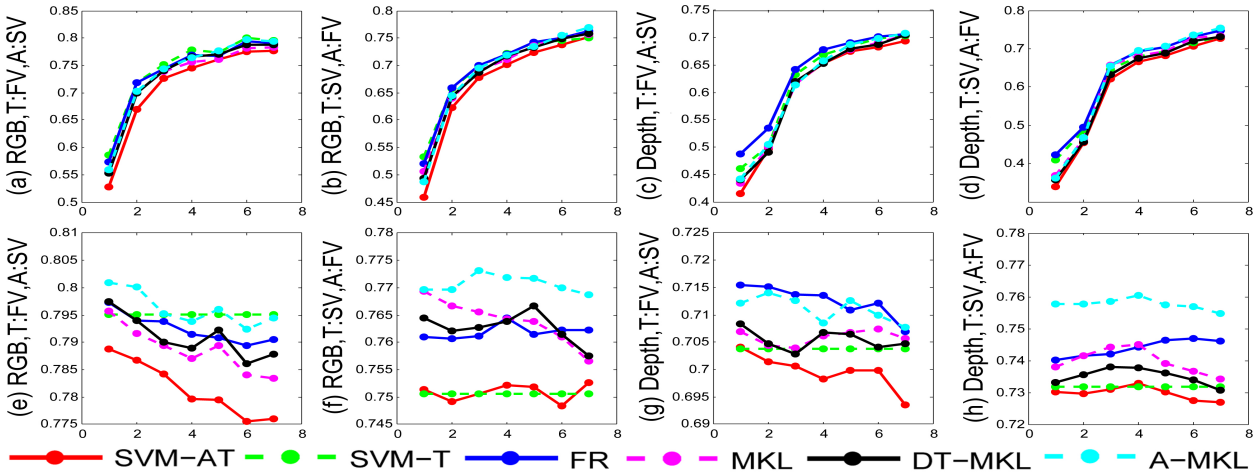

Figure 5: Overall accuracy on the RGB and Depth data under the cross-domain scenario. (a-d) Case 1; (e-h) Case 2. (T: target domain; A: auxiliary domain)

data from the target domain and AMKL can usually outperform the others by well leveraging the discovered knowledge from the auxiliary domain. 2) For Case 2, the performances can not be monotonically increased by augmenting the data from the auxiliary domain and it can even decrease since more samples in different feature space might have negative influence on useful knowledge transferring inbetween. This experiment further demonstrated that the dataset is extremely challenging with significant view differences.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce the Multi-modal & Multi-view & Interactive (M^2I) database for human action recognition. Moreover, we respectively evaluate three representative methods for the single-view, cross-view, and cross-domain human action recognition on this dataset. It is experimentally demonstrated that this dataset is extremely challenging due to large intraclass variation, multiple similar actions, and significant view difference. As the future work, we will evaluate more popular spatiotemporal features on this dataset and focus on multi-modal & multi-view information fusion to boost the performance.

5. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61472275, 61303208), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046).

6. REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16, 2011.

[2] X. Chang, W. Zheng, and J. Zhang. Learning person-person interaction in collective activity recognition. *IEEE Transactions on Image Processing*, 24(6):1905–1918, 2015.

[3] W. Choi and S. Savarese. Understanding collective activities of people from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1242–1257, 2012.

[4] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Anal. Mach. Intell*, 34(9):1667–1680, 2012.

[5] J. Hu, W. Zheng, J. Lai, S. Gong, and T. Xiang. Exemplar-based recognition of human-object interactions. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[6] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.*, 24(5):971–981, 2013.

[7] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004.*, pages 32–36, 2004.

[8] H. Wang, A. Kläser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[9] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1290–1297, 2012.

[10] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–7, 2007.

[11] L. Xia, C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, pages 20–27, 2012.

[12] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa. Cross-view action recognition via a transferable dictionary pair. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–11, 2012.