M&M: Recognizing Multiple Co-evolving Activities From Multi-source Videos

Lan Zhang, Mu Yuan, Daren Zheng, Xiang-Yang Li
University of Science and Technology of China
Hefei, China
zhanglan@ustc.edu.cn, ym0813@mail.ustc.edu.cn, zdr123@mail.ustc.edu.cn, xiangyangli@ustc.edu.cn

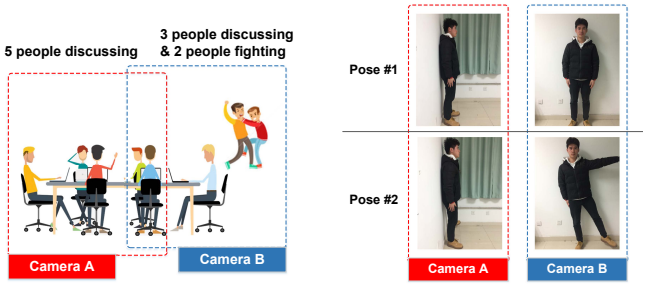
***Abstract*—The wide deployment of surveillance systems has shown the necessity to recognize human activities in videos. Existing work achieved high recognition accuracy for single-person activities and some group activities. In this paper, we identify a more challenging issue: recognizing multiple co-evolving but asynchronous activities from a set of videos captured by multiple cameras that share overlapped views. To address this issue, we design a system M&M to fuse objects and 2D skeletons from multi-source videos to reconstruct the complete 3D view-invariant model of multiple activity scenarios, which enables accurate recognition of cross-source activities. By embedding the 3D model with a concise graph representation, we propose an efficient recognition method in a bottom-up manner to achieve high accuracy and good scalability to changing complexity of the captured scenario. We collect and release a dataset containing correlated multi-source videos for multiple co-evolving activities, and evaluate our design on it. Experimental results show that M&M achieves 91.2% average accuracy for 10 types of co-evolving activities.**

***Index Terms*—activity recognition, cross-source videos, co-evolving activities**

I. Introduction

Recognizing activities in video streams has a wide range of applications, for example, intelligent surveillance systems, professional behavior analysis, and smart robotics. The remarkable achievement of deep learning is significantly contributing to the development of activity recognition. Over the past few years, a large amount of effort has been devoted to this task. Most works focus on recognizing single-person activities given pre-trimmed video clips, which train activity models base on labeled pre-trimmed video clips, e.g., [1]–[8]. Several works leverage features from multiple views to recognize a single activity [9], [10]. Some works recognize more complicated group activities with multiple people involved using models such as hierarchical random filed model [11], deep CNNs-based graphical model [12], and two-level LSTM [13].

Existing works achieved high recognition accuracy for single-person activities and some group activities. In our work, we consider a more realistic and more challenging situation, where multiple co-evolving but asynchronous activities are captured by multiple cameras that share overlapped views. In real-world applications, it is common to monitor a large area using multiple cameras from different angles. Views of these cameras have some overlaps, but each of which may only obtain incomplete information of one or more activities.



(a) Cross-source co-evolving activities.

(b) One-view indistinguishable activities.

Fig. 1. Examples of cross-source co-evolving activities and one-view indistinguishable activities.

As the scenario shown in Figure 1(a), in the view of Camera A, there are five people discussing; in the view of Camera B, there are three people discussing and two people fighting. However, the truth is that there are six people discussing and two people starting to fight. It is important to understand what's really going in many complicated real-life scenarios. To achieve this goal, we propose a novel system **M&M** to recognize multiple co-evolving activities from multi-source videos. Several challenges make it non-trivial to design such a system:

1. Complicated spatiotemporal features of multiple activities: those activities co-evolve with each other, but start or end asynchronously. Multiple people and objects are involved in those activities, and their locations may change continuously. All these features raise great challenges for video segmentation and activity localization, embedding as well as recognition.

2. Changing scenarios and limited training data: previous work focus on learning the global spatiotemporal features of one activity using models like Structural RNN, Markov Random Filed and so on. However, existing models cannot adapt to different scenarios with multiple co-evolving activities and more involved people and objects. It is also difficult to obtain a large set of labeled training data for various scenarios. It is challenging to design a model scalable to changing complexity and scale of the captured scenario with limited training data.

3. One-view indistinguishable activities: some activities cannot be distinguished given only one-view video. As the example in Figure 1(b), in the view of Camera A, Pose #1 and Pose #2 look the same. With the help of Camera B, we

This paper has been accepted for publication at DCOSS'21. This version is a preprint identical in content to the version to be published.

find that they are two totally different poses. This issue has been neglected by most existing work, which requires careful design to utilize multi-source videos.

4. Incomplete information of cross-source activity: considering a monitoring system with multiple cameras, it is common that information of an activity captured by each camera is incomplete due to the limited viewing field of the camera or occlusion, which significantly increases the difficulty for cross-source information fusion. Note that, here we distinguish our multi-source recognition problem from the existing multi-view recognition problem, which assumes every camera captures complete information of the target activity.

Facing the aforementioned challenges raised by multiple activities and multi-source videos, we design **M&M** to enable recognition of multiple one-view indistinguishable activities and cross-source activities. Our innovation is to take multi-source videos as an opportunity to reconstruct the fused 3D view-invariant model of the multiple activity scenario, embed the complete 3D model, and parse the model in an bottom-up manner to recognize multiple activities in complicated scenarios. To the best of our knowledge, **M&M** is the first step to recognize multiple co-evolving activities from multi-source videos, which shows great value in various real-world applications. The main contributions and results of this paper are summarized as follows:

1) To achieve a complete view-invariant description of multiple one-view indistinguishable activities and cross-source activities, we propose to reconstruct complete 3D scenario using multiple views of 2D human skeletons, meanwhile fuse information from multiple sources in the feature space.

2) To achieve concise embedding and scalable recognition of complicated activities, we propose to embed the complete 3D model into a graph structure with human poses, objects and their interactions. Based on the embedding, we design an efficient activity learning method requiring only small size training data, and a bottom-up parsing-based recognition method with good scalability. A Hidden Markov Model (HMM) based method is designed to recognize co-evolving activities in untrimmed videos. Our method is naturally able to detect multiple activity regions as well as segment video streams according to the recognized activities.

3) We collect a video dataset named **M&MD** including 10 types of co-evolving activities captured by 5 cameras and release the dataset with labels, which is the first dataset contains correlated multi-source videos for multiple activity recognition task. We implement our design and evaluate it on the **M&MD** dataset. Our evaluation results show that the 3D reconstruction error is less than 5 cm with even only two cameras, and **M&M** achieves 91.2% average accuracy for recognizing 10 types of activities.

II. Design Overview

A. Motivation and Basic Idea

The wide adoption of cameras and video surveillance systems makes strong demands for recognizing human activities in videos. Most previous work focused on improving the accuracy of single-person activity recognition, some work proposed methods to recognize group activity. In real applications, activities in videos could be much more complicated. It is very common that multiple activities happen in the same area. It is also common to monitor an area using several cameras, each of which only captures part of one or more activities. As a result, it is necessary to recognize multiple co-evolving activities using multi-source videos, however few work has been done to address this issue.

In this paper, we focus on multiple cross-source activities recognition, and propose a novel system **M&M**. As mentioned in Section I, critical challenges including complicated spatiotemporal features of multiple activities, highly changing scenarios, one-view indistinguishable activities and incomplete information of cross-source activity, make the design of **M&M** non-trivial.

The main ideas of our design are two-fold:

1) Given correlated multi-source videos, incomplete information of each video makes cross-source activity recognition very difficult. Instead of conducting recognition on single-source 2D videos, we only extract human skeletons and objects from 2D videos, and take multiple sources as an opportunity to reconstruct a complete view-invariant 3D multi-activity scenario consisting of people and objects. Taking advantage of the overlapped view of multiple cameras, we process multi-view based 3D reconstruction and multi-source information fusion simultaneously. Learning and recognition are conducted on the fused 3D model.

2) Facing complicated and continuously changing features of multiple activities and limited training data, we need a compact embedding mechanism and scalable learning and recognition methods to achieve high accuracy and robustness. We propose to cluster 3D skeletons into a set of basic poses in an unsupervised way, and design a concise graph descriptor to embed human poses, objects and their interactions. To achieve good scalability, instead of end-to-end model training, based on our embedding, we propose a reduction-based method, which recognizes activities in a bottom-up manner. The reduction rule can be learned from a small set of training data.

B. System Design

Figure 2 illustrates the overview of our design and system workflow. Taking multi-source videos, which share certain overlapped view, as input, **M&M** runs the following modules in a row to output the labels of recognized activities as well as their respective involved people.

1) Object detection. Human-object interaction is important for recognizing the performed activity. Our object detection module takes videos as input and outputs the class and location of objects in the video, which are important input of the sub-activity embedding module. Any sophisticated object detection method can be adopted for this component. One type of approaches [14] first generate potential bounding boxes for objects by region proposal method, then use a single-object classifier for classification. Some works [15] treat the detection problem as a single regression problem that utilizes a single

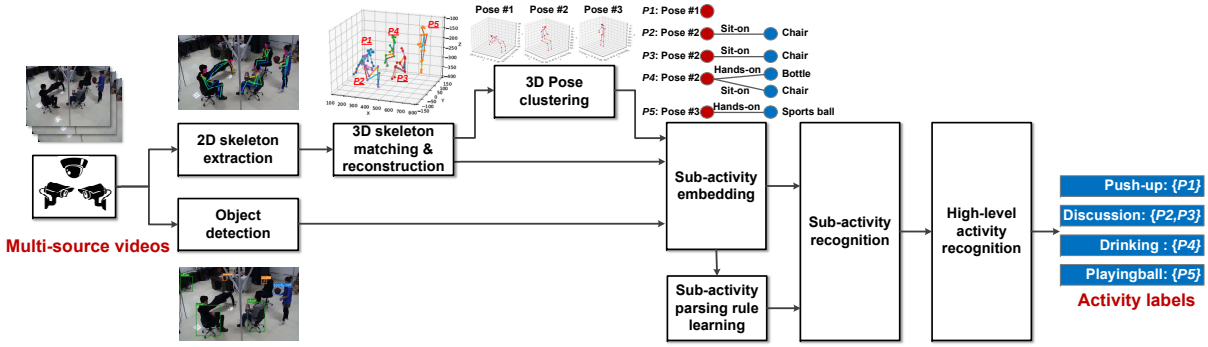


Fig. 2. Overview of system design.

CNN to produce probability for each grid cell in the image. In our implementation, we employ YOLO [16] for object detection because of its efficiency and fairly high accuracy.

2) 2D skeleton extraction. Human-object interaction and human-human interaction are important for activity recognition, and both require the information of human action. Our system aims to analyze the human action based on the view-invariant 3D skeleton model. 2D skeleton extraction is the first step for this goal. Previous work of estimating multi-person 2D skeletons can be divided into two main classes: the two-step approach and the part-based approach. Two-step framework [17], [18] detects human bounding boxes first, then extracts skeleton in each box by a single-person pose estimator. Part-based framework [19]–[21] detects body parts for multiple people then assembles them for individuals. In our implementation, we utilize OpenPose [21] for 2D skeleton extraction due to its robust performance and real-time efficiency. The 2D skeleton joint locations from multi-source videos are used to reconstruct the 3D activity model.

3) 3D skeleton matching and reconstruction. In order to solve the action ambiguity and activity information incompleteness in the one-view video, we build the view-invariant 3D activity scenario based on reconstructing 3D human skeleton using multi-view 2D skeletons. Simultaneously, we fuse information from multi-source videos by determining the optimal matching of 2D skeletons between different sources. The 3D skeleton matching and reconstruction module in our system is the key to recognize activities captured by multi-source videos. This step not only transfers the activity scenario from 2D into 3D, but also fuses multi-source incomplete information into a complete one. The output 3D skeletons and matching results are used to generate a complete description of the multi-source activity scenario.

4) 3D pose clustering. The 3D pose clustering module aims to decrease the feature dimension of the description of sub-activity. This module clusters 3D skeletons into a set of basic poses. Low-dimension description of sub-activity increases the robustness and efficiency of our system.

5) Sub-activity embedding, learning and recognition. Based on the output of object detection module, 3D skeleton matching module and 3D pose clustering module, we use a graph descriptor to embed the sub-activity. The learning step requires fairly few training samples to learn the reduction rules for the graph descriptor of sub-activity. And we treat the recognition step as a graph coarsening process, which can be done quite efficiently and achieve good scalability. The output sub-activity labels and involved people compose the sub-activity sequence, which is the input of the high-level activity recognition module.

Algorithm 1: Skeleton Matching Estimation

input: 2D Skeleton Lists L_1, L_2, Threshold t

repeat
  errorDict $\leftarrow \{\}$
  for s_i *in* L_1 **do**
    for s_j *in* L_2 **do**
      F $\leftarrow$ FindFundamentalMatrix(s_i, s_j)
      errorDict[(i,j)] $\leftarrow$ SUM(ABS($s_j^T F s_i$))
  $idx_1, idx_2 \leftarrow$ SortByValue(errorDict,'reverse')
  minerror $\leftarrow$ errorDict[(idx_1, idx_2)]
  if *minerror* $< t$ **then**
    matching.$append((idx_1, idx_2))$
    $L_1.pop(idx_1)$, $L_2.pop(idx_2)$
until *minerror* $\geq t$;

6) High-level activity recognition. High-level activity recognition model is the final step to achieve our goal. This module solves the problem of modeling and recognizes sub-activity sequences. Due to the low dimension of sub-activity sequence, the learning step also requires few training samples, which gives our system good scalability and efficiency.

In the following three sections, we present the detailed design of three core modules of our system.

III. 3D Skeleton Matching and Reconstruction

As shown in Figure 3, this module aims to determine the optimal matching of skeletons from multi-source videos via minimizing the 3D reconstruction error and reconstruct the fused 3D activity scenario. For the skeleton matching and 3D reconstruction problem, we make two assumptions: (1)intrinsic parameters of cameras are known; (2)any pair of 2 cameras

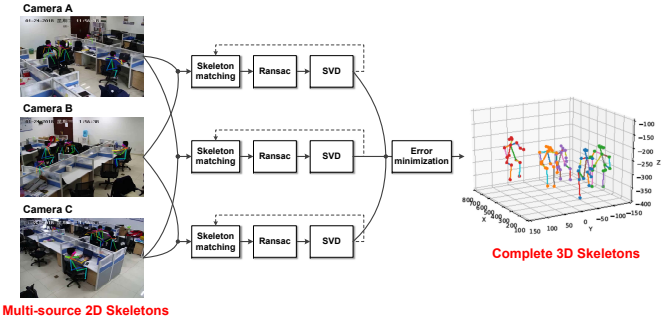


Fig. 3. Skeleton matching and 3D reconstruction from multi-source 2D skeletons.

share some overlapped view and at least one person has ever appeared in the overlapped area.

A. Greedy Algorithm For Matching Estimation

We design a greedy algorithm to determine the matching of 2D skeletons from different views. Algorithm 1 presents the matching estimation algorithm in detail. Given a candidate matching pair, that is a group of corresponding points in two views, the function $FindFoundamentalMatrix$ computes the fundamental matrix by solving a linear triangulation problem [22]. According to the property of the fundamental matrix, $x_2^T F x_1 = 0$ for perfect matching. For each candidate matching pair, we compute the sum of the absolute value of matrix $x_2^T F x_1$ as the error of found fundamental matrix. Iteratively, we find the candidate matching that has the minimal error, and the error is less than a threshold value. Then we move the matching pair from the candidate list to the final matching list. The threshold is an experiential value and determined based on the experiment.

B. 3D Skeleton Reconstruction

According to the matching result, this module outputs the 3D skeleton for each person. To reconstruct 3D skeletons, we compute the projection matrix by decomposing the fundamental matrix using SVD method. Given the projection matrix, 3D skeleton of every 2D skeleton can be obtained. When there are more than 2 cameras, we have a group of 3D skeletons computed from each pair of cameras. We need to fuse them into one complete model. Based on the confidence of 2D skeleton estimation in each view, we use a weighted average strategy to minimize error of the reconstructed 3D skeleton. For each pair of cameras Cam_i and Cam_j, the confidence c_{ij} of the reconstructed skeleton is computed as the average of 2D skeleton estimation confidences from each camera. The weighted averaging 3D skeleton $\mathcal{S}$ is computed by:

$$\mathcal{S} = \sum_{i \neq j} \frac{c_{ij} \mathcal{S}_{ij}}{\sum_{i \neq j} c_{ij}} \tag{1}$$

Now, we have constructed a complete view-invariant 3D activity scenario.

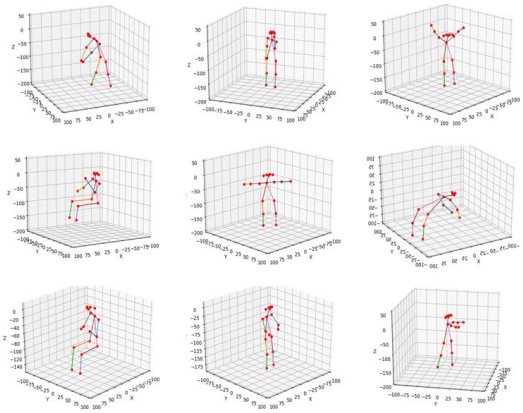


Fig. 4. Illustration of pose clustering results.

IV. Sub-activity Embedding and Recognition

Based on the object detection, 3D skeleton matching and reconstruction results, we embed sub-activity using a concise graph descriptor. We propose a novel method to learn reduction rules and recognize sub-activity based on the graph descriptor.

A. 3D Pose Clustering

3D skeleton is a view-invariant descriptor for human pose and the normalization process makes 3D skeleton scale-invariant. We use KMeans algorithm to cluster normalized 3D skeletons into a set of basic poses. Figure 4 shows some example center poses clustered from the dataset we collected. The 3D pose clustering module reduces the feature space from 3D coordinates of 18 joints to K clusters, which benefits the efficient and robust learning of reduction rules from small training samples.

B. Sub-Activity Recognition

Given synchronous frames from multi-source video streams, our goal is to determine the sub-activity and involved people.

1) Sub-activity embedding. We embed the spatial structure of the sub-activity as a graph $G(V, E)$. The input data contains object labels with bounding boxes and pose labels with skeleton keypoint coordinates. V consists of object nodes and pose nodes. E is computed according to the spatial relations among pose and object nodes by pre-defined rules. The spatial relations include human-human interactions and human-object interactions. For example, for a pose node, if the hand keypoint is inside the bounding box of an object node, then the edge "Hands-On" will be generated to connect these the corresponding pose node and object node.

2) Reduction rules learning. We propose to recognize sub-activity by reduction. The reduction rules for parsing the graph descriptor are learned from training samples. We treat the learning of reduction rules as training a function $\Phi : V \times E \times V \mapsto S$, where S is the set of sub-activity labels. The mapping function Φ aims to collapse a pair of nodes

(v_i, v_j) and the edge e_{ij} between them into a sub-activity label s_k. To learn the mapping fuction, we embed the node-edge-node sub-graph as a vector as shown in Figure 5.

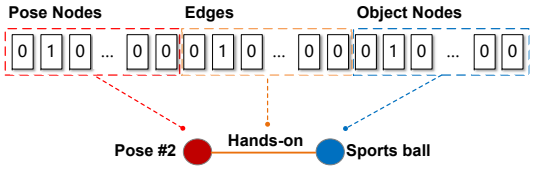


Fig. 5. Sub-graph embedding.

We use the nearest neighbor classifier to learn this mapping function. Attribute to robust and concise embedding, a small set of training samples is sufficient to learn a good mapping function. For example, in our experiments, we can learn the mapping function with only 5 samples for each sub-activity. Our method has good scalability to changing activity scenarios, that is, rules learned from two people's interaction or interaction between one person and one object could be applied to parse graph descriptor with multiple people and objects.

3) Reduction-based recognition. We design a hierarchical graph coarsening algorithm to recognize the sub-activity in a bottom-up manner based on the learned mapping function. Our 3-level graph coarsening algorithm considers human-object interactions, human-human interactions and human pose without interaction step by step. This hierarchical parsing approach coarsens the graph descriptor into several connected components, which represent the multiple co-evolving activities. The computation complexity is $O(|E|)$, so it can be conducted very efficiently and applied to parse graph descriptor of complicated scenario. Another property of the approach is that it naturally detects multiple activity regions, due to its bottom-up manner.

V. High-level Activity Recognition

Activities in video data contain both spatial and temporal structures. We learn the spatial structure as presented in Section IV. In this section, we learn the temporal structure based on Hidden Markov Models. Figure 6 shows two examples of the temporal structure of high-level activities. The sweeping activity consists of two main processes (sweeping floor follows fetching broom) and the gymnastics activity is composed of 4 successive gymnastics movements.

Hidden Markov Models (HMM) have been widely applied to modeling temporal sequence data in tasks like speech recognition, handwriting recognition, and so on. For each high-level activity, we use an HMM to model the learning and reduction process as an automata [23]. The visible states of HMM are the sub-activities of corresponding high-level activity. The training objective is to learn the parameters: state transition probability A, observation symbol probability matrix B and initial state probability matrix Π. After training steps, for multiple high-level activities, we obtain a group of HMMs, which generate the probability of observing each high-level activity according to the input sub-activity sequences. If the possibility exceeds an experimental threshold value, the HMMs output corresponding high-level activity label.

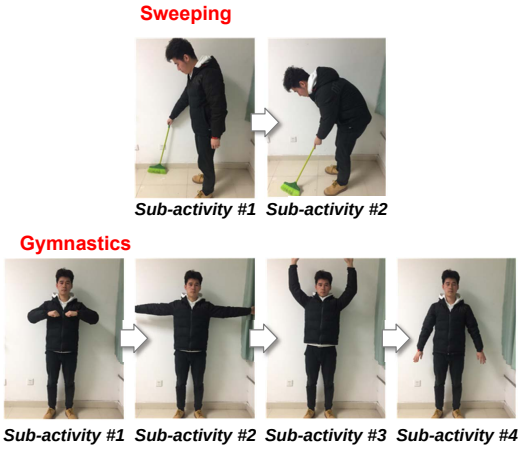


Fig. 6. Illustration of sub-activities of two high-level activities in our dataset.

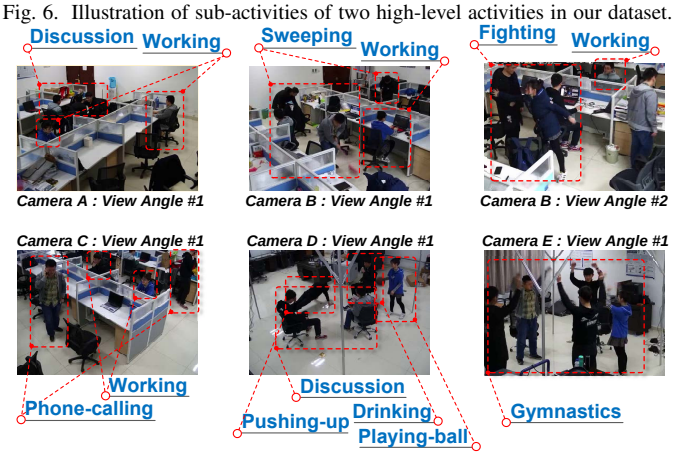


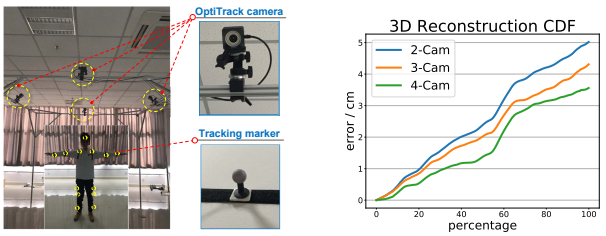
Fig. 7. Illustration of high level activities in our dataset.

We use Bayesian Model Merging algorithm to learn both the structure of the HMMs and the parameters from training samples [24]. Bayesian model merging procedure for HMM considers all possible state merging possibilities. Briefly, the state merging procedure chooses two states to merge such that the Bayesian posterior probability is maximized. It stops when no merging results increase the posterior probability of the current model. A good property of the Bayesian model merging algorithm is that it works for small training samples, while Baum-Welch method may not coverage. Due to this property, our experimental results show that learning from 5 samples of video clips for a high-level activity, the HMMs perform accurate recognition for 10 types of high-level activities.

VI. Experiments

A. Dataset and Experiment Configuration

We evaluate our method on one dataset we collected – Multi-source And Multiple activity Dataset (**M&MD**), which is composed of two sub-datasets captured in different scenarios. The first scene is a large office (with 25 desks) monitored by 3 cameras and the other scene is a laboratory monitored by 2 cameras. The video resolution is 704×576 and the fps is 24. We collected 5 sets of videos, 2 sets in the office scene and 3 sets in the lab scene. So in all we have $2 \times 3 + 3 \times 2 = 12$ untrimmed videos captured by 5 cameras and the total time is



(a) 3D skeleton ground-truth data collection.

(b) Cumulative distributions of 3D skeleton reconstruction errors.

Fig. 8. 3D skeleton reconstruction.

45.4 minutes. In the office scene, we captured 6 types of high-level activities: {1) discussion, 2) phone-calling, 3) working, 4) moving-chair, 5) fighting, 6) sweeping}. In the lab scene, we captured 4 types of high-level activities: {1) playing-ball, 2) pushing-up, 3) drinking, 4) gymnastics}. Note that, all these activities are co-evolving, which means at any moment there is more than one class of activities are in progress and captured by cameras. Figure 7 illustrates some snapshots of our dataset. As we can see, these scenarios are complicated and highly noisy. We manually annotated **M&MD** one frame per second with sub-activity labels, high-activity labels and the number of people involved, and release this dataset. Compared with existing datasets (see Table I), **M&MD** possesses two main characteristics: (1) Continuous co-evolving activities happen in untrimmed videos; (2) Cameras in the same scene have overlap views, but information from any single video source is incomplete.

a) Experiment setting.: We conduct experiments on a machine with Intel Core i7-5930K CPU3.50GHz×12, GeForce GTX 1080×1, 62.8GiB memory.

B. 3D Skeleton Reconstruction

To evaluate the accuracy of the 3D skeleton reconstruction approach, we used OptiTrack cameras and markers to obtain the ground truth of skeleton joints. OptiTrack camera captures infrared video and supports high-speed (240 FPS) and sub-mm 3D precision tracking of the marker. Figure 8(a) shows the cameras, markers and the experiment setting. We use 4 OptiTrack cameras to track 13 keypoints of the volunteer. We asked the volunteer to move freely, and captured his activities using 4 cameras while the OptiTrack system tracking the location of each keypoints. We conducted 3D skeleton reconstruction using videos from 2 cameras, 3 cameras and 4 cameras, respectively. We evaluated the reconstruction results for each frame compared to the ground truth. Figure 8(b) plots the CDF of reconstruction error for 600×4 frames. In this work, we use the weighted averaging strategy to minimize the error of reconstructed skeletons from each pair of cameras. As we can see, using more cameras leads to better reconstruction accuracy. Using 2 cameras, errors of more than 99% frames are less than 5 cm and errors of more than 99% frames are less than 4 cm when using all 4 cameras.

C. Activity Classification

We use **M&MD** to evaluate the performance of our method for recognizing multiple co-evolving activities from multi-source videos. Here we use two accuracy metrics: $Accuracy_1 = \frac{N_1}{N}, Accuracy_2 = \frac{N_2}{N}$, where N denotes the number of total frames, N_1 denotes the number of frames that recognized activity is correct but the number of people involved could be wrong, N_2 denotes the number of frames that both the recognized activity and the number of people involved are correct. Figure 9 shows the high-level activity recognition accuracy for 10 types of activities in **M&MD**. The average of $Accuracy_1$ is 91.2% and the average of $Accuracy_2$ is 70.5%, which means that in 91.2% cases **M&M** correctly recognized multiple activities from multiple untrimmed videos, and in 70.5% **M&M** outputs not only the correct activity labels but also the correct number of involved people. We find that 7 in 10 activities are recognized well with more than 94% $Accuracy_1$. But activities like drinking and phone-calling are recognized relatively poorly with 75% $Accuracy_1$. We analyzed the recognition results and found two main reasons: (1) related objects are small and hard to detect, e.g., bottle and phone; (2) related human poses are subtle and hard to discriminate. In the future, the development of object detection techniques and more accurate skeleton estimation can greatly benefit our system.

D. System Efficiency

We also evaluated the time cost of modules in our system. Figure 10(a) shows the runtime per frame for those frame-based modules. The results show that, the most costly module is 2D skeleton extraction (about 95 ms per frame). The 3D skeleton reconstruction costs about 60 ms per frame and object detection costs about 48 ms per frame. Due to our compact embedding and efficient reduction method, the sub-activity recognition costs less than 15 ms per frame. For pose clustering and high-level activity recognition modules, we measure the time cost of the 5 subsets of videos. Figure 10(b) shows their runtime against different video lengths. We can see that compared to the video length, the time cost of pose clustering and high-level activity recognition is very low. Besides, the time cost has a near-linear growth with the video duration. The results exhibit the efficiency and scalability of our system.

VII. Related Work

To the best of our knowledge, our work is the first to recognize multiple co-evolving activities captured by multiple cameras. **M&M** is related to previous works in the following aspects.

a) 3D pose estimation.: Previous works attempting to estimate 3D human pose from videos can be divided into two main classes according to the number of cameras used: monocular camera-based methods and multi-view video-based methods. Typically estimating 3D pose from monocular video needs the prediction of 2D heatmaps for joints and fitting of skeleton. [30] used deep feed-forward networks to infer 3D locations based on 2D joint locations. [31] designed a two-stream architecture to utilize both 2D and 3D image features

Dataset	Multi-view	Untrimmed	Interaction person-person	Interaction person-object	Co-evolving Activities
UCF101 [25]	✗	✗	✓	✓	✗
HMDB51 [26]	✗	✗	✓	✓	✗
IXMAS [27]	✓	✗	✗	✗	✗
MSR Daily Activity3D [28]	✗	✗	✓	✓	✗
ActivityNet [29]	✗	✓	✓	✓	✗
M&MD	✓	✓	✓	✓	✓

TABLE I
PROPERTIES OF EXISTING DATASETS AND OUR **M&MD**

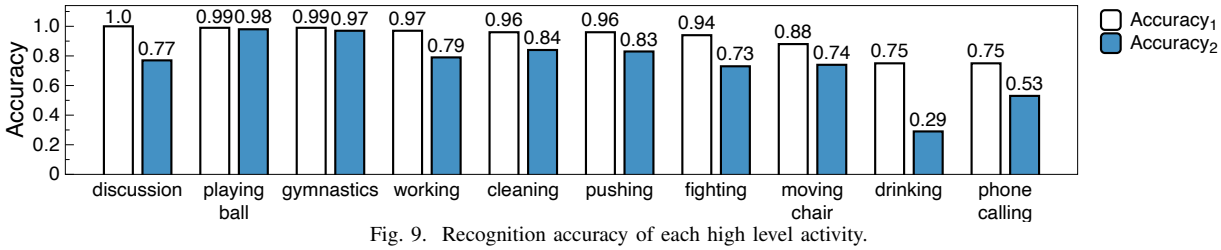
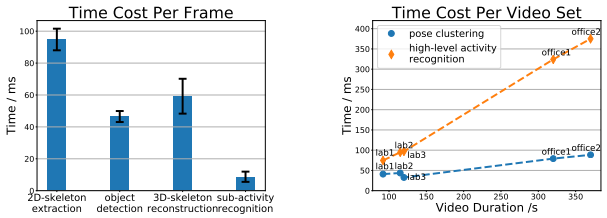


Fig. 9. Recognition accuracy of each high level activity.



(a) Runtime per frame. (b) Runtime per video set.

Fig. 10. Time cost of each component of our system.

to estimate 3D pose. The recent work [32] achieved real-time accurate estimation with a single RGB camera. Multi-view video-based methods utilizing the multiple projections of the human pose to compute the optimal spatial location of each camera [33], [34]. [35] used sums of spatial Gaussians (SoG) body model and achieved nearly real-time performance. Some works [36] can deal with 3D pose estimation of multiple people, while most of other works focused on estimating pose of a single person.

All those works assume every camera captures complete information of the target pose/activity. While, in our work, each camera captures only incomplete information due to limited field of view and occlusion. It is worth mentioning that we conduct 3D activity reconstruction and multi-source information fusion simultaneously.

b) Human activity recognition.: Works recognizing human activities in videos can be summarized as two main types: (1)single-view video-based method; (2)multi-view video-based method. For single-view video-based methods, modeling the spatiotemporal structure of activity in video is the key point. Some works use RGB videos as the input. [2] used dense trajectories feature with SVM classifier. [4]–[6] used a novel ConvNet-based model to extract the spatiotemporal structure. And some works consider more complicated group activity recognition tasks. [37] designed a unified framework to coherently learn and recognize collective activities which involve multiple people. [11] proposed a hierarchical random filed model to model the dependency of frames in the temporal dimension. [12] used a deep CNNs-based graphical model to recognize group activities. [13] utilized a two-level LSTM to model the high dynamics of both person and group in videos. Some works use RGB-D videos with depth information to increase the recognition accuracy [38], [39]. For multi-view video-based methods, it is important to discover the correlation of features from multiple views. [9] employed epipolar geometry to compute the corresponding points in multi-view videos. [10] designed a discriminative and structured dictionary learning model to fuse multiple views for human activity recognition.

Most of those works consider only single-person activity and some of them consider group activities. In this paper, we focus on a more challenging task, that is recognizing multiple co-evolving complicated activities in multi-source videos.

VIII. CONCLUSION

In this paper, we consider the task of recognizing multiple co-evolving activities from multi-source videos. We proposed a system **M&M** to reconstruct the complete 3D activity scenario, designed a concise and scalable embedding for multiple activities, as well as an efficient bottom-up method to learn and recognize co-evolving activities with high accuracy. We also collected and released a video dataset for this task. Our work is the first step, to our best knowledge, to solve the hitherto most complicated multiple activity recognition tasks using multi-source videos, and our experiments show the promising performance of **M&M**.

ACKNOWLEDGMENT

This research is supported by the National Key R&D Program of China 2017YFB1003003, NSF China under Grants No. 61822209, 61932016, China National Funds for Distinguished Young Scientists with No.61625205, and Key Research Program of Frontier Sciences, CAS. No. QYZDY-SSW-JSC002.

REFERENCES

[1] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[2] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2593–2600.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[6] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[7] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1617–1632, 2017.

[8] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 056–12 065.

[9] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 984–989.

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

[11] M. R. Amer, P. Lei, and S. Todorovic, "Hirf: Hierarchical random field for collective activity recognition in videos," in *European Conference on Computer Vision*. Springer, 2014, pp. 572–585.

[12] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," *arXiv preprint arXiv:1506.04191*, 2015.

[13] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 1971–1980.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[15] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[17] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3178–3185.

[18] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[19] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[22] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[23] R. Parekh and V. Honavar, "Grammar inference, automata induction, and language acquisition," *Handbook of natural language processing*, pp. 727–764, 2000.

[24] A. Stolcke and S. Omohundro, "Hidden markov model induction by bayesian model merging," in *Advances in neural information processing systems*, 1993, pp. 11–18.

[25] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2556–2563.

[27] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–7.

[28] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[29] H. Fabian Caba, E. Victor, G. Bernard, and N. Juan Carlos, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[30] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *IEEE International Conference on Computer Vision*, vol. 206, 2017, p. 3.

[31] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in *International Conference on Computer Vision (ICCV)*, no. EPFL-CONF-230311, 2017.

[32] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 44, 2017.

[33] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view pictorial structures for 3d human pose estimation." in *Bmvc*. Citeseer, 2013.

[34] M. Burenius, J. Sullivan, and S. Carlsson, "3d pictorial structures for multiple view articulated pose estimation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3618–3625.

[35] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, "Fast articulated motion tracking using a sums of gaussians body model," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 951–958.

[36] A. Elhayek, E. de Aguiar, A. Jain, J. Thompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Marconiconvnet-based marker-less motion capture in outdoor and indoor scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 3, pp. 501–514, 2017.

[37] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 215–230.

[38] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[39] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5308–5317.