Learning Actionlet Ensemble for 3D Human Action Recognition

Jiang Wang, *Student Member, IEEE,* Zicheng Liu, *Senior Member, IEEE,* Ying Wu, *Senior Member, IEEE,* and Junsong Yuan, *Member, IEEE,*

Abstract—Human action recognition is an important yet challenging task. Human actions usually involve human-object interactions, highly articulated motions, high intra-class variations, and complicated temporal structures. The recently developed commodity depth sensors open up new possibilities of dealing with this problem by providing 3D depth data of the scene. This information not only facilitates a rather powerful human motion capturing technique, but also makes it possible to efficiently model human-object interactions and intra-class variations. In this paper, we propose to characterize the human actions with a novel *actionlet ensemble* model, which represents the interaction of a subset of human joints. The proposed model is robust to noise, invariant to translational and temporal misalignment, and capable of characterizing both the human motion and the human-object interactions. We evaluate the proposed approach on three challenging action recognition datasets captured by Kinect devices, a multiview action recognition dataset captured with Kinect device, and a dataset captured by a motion capture system. The experimental evaluations show that the proposed approach achieves superior performance to the state-of-the-art algorithms.

Index Terms—Action recognition, Kinect, ensemble method, human pose, human-object interaction

1 Introduction

Recognizing human actions has many applications including video surveillance, human computer interfaces, sports video analysis and video retrieval. Despite remarkable research efforts and many encouraging advances in the past decade, accurate recognition of the human actions is still a quite challenging task. There are two major issues for human action recognition. One is the sensory input, and the other is the modeling of human actions that are dynamic, ambiguous and interactive with other objects.

Human motion is articulated in nature. Extracting such highly articulated motion from monocular video sensors is a very difficult task. This difficulty largely limits the performance of video-based human action recognition, as indicated in the studies in the past decade. The recent introduction of the cost-effective depth cameras may change the picture by providing 3D depth data of the scene, which largely eases the task of object segmentation. Moreover, it has facilitated a rather powerful human motion capturing technique [28] that outputs the 3D joint positions of the human skeleton. As we will show in this paper, although the estimated 3D skeleton alone is not sufficient to solve the human action recognition problem, it greatly alleviates some of the difficulties in developing such a system.

The depth cameras in general produce better quality 3D depth data than those estimated from monocular video sensors. Although depth information alone is very useful for human action recognition, how to effectively combine such 3D sensory data with estimated 3D skeletons is nontrivial. First, the 3D skeleton alone is not sufficient to distinguish the actions that involve human-object interactions. For example, "drinking" and "eating snacks" exhibit very similar skeleton motions. Additional information is needed to distinguish the two actions. Second, human actions may have specific temporal structure. For example, the action "washing a mug" may consist of the following steps: "arriving at the mug", "taking the mug", "arriving at the basin" and "dumping the water". The temporal relationship of these steps is crucial to model such actions. Finally, human actions may have strong intra-class variations. A person may use either his left hand or right hand to make a phone call, and different people have different ways of washing a plate. Modeling these variations is also challenging.

This paper proposes novel features to represent human actions in depth data. First of all, we propose a new 3D appearance feature called *local occupancy pattern* (LOP). Each LOP feature describes the "depth appearance" in the neighborhood of a 3D joint. Translational invariant and highly discriminative, this new feature is also able to capture the relations between the human body parts and the environmental objects that the person is interacting with. Secondly, to represent the temporal structure of an action, we propose a new temporal representation called

- *J. Wang and Y. Wu are with the EECS Department, Northwestern University, Evanston, IL 60208 USA. E-mail: wangjiangb@gmail.com; yingwu@eecs.northwestern.edu.*
- *Z. Liu is with Microsoft Research, Redmond, WA 98052 USA. E-mail: zliu@microsoft.com.*
- *J. Yuan is with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798. E-mail: jsyuan@ntu.edu.sg.*

Manuscript received 26 Dec. 2012; revised 20 Aug. 2013; accepted 17 Sep. 2013. Date of publication 08 Oct. 2013. Date of current version 29 Apr. 2014.
Recommended for acceptance by R. Vidal.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier 10.1109/TPAMI.2013.198

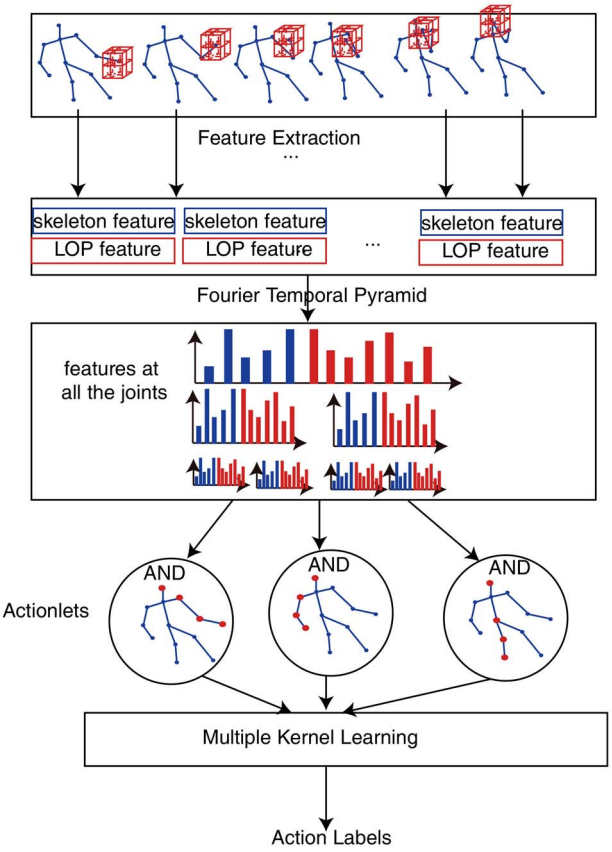

Fig. 1. General framework of the proposed approach.

Fourier Temporal Pyramid. This representation is insensitive to temporal sequence misalignment, robust to noise, and is discriminative for action recognition.

More importantly, we propose a new model called the *Actionlet Ensemble Model*, illustrated in Fig. 1. The articulated human body has a large number of kinematic joints, but a particular action may only involve a small subset of them. For example, for right-handed people, action "drinking" typically involves joints "right wrist" and "head". Thus the combinational feature of the two joints is a discriminative feature. For left-handed people, action "drinking" typically involves joints "left wrist" and "head". Therefore the combinational feature of joints "left wrist" and "head" is another discriminative feature for this action. Therefore, we introduce the concept of *actionlet*. An *actionlet* is a conjunction of the features for a subset of the joints. As the number of possible *actionlets* is enormous, we propose a novel data mining solution to discover *discriminative actionlets*. An action is then represented as an *actionlet ensemble*, which is a linear combination of the *actionlets* whose discriminative weights are learnt via a multiple kernel learning method. This new action model is more robust to the errors in the features, and it can better characterize the intra-class variations in the actions.

Our main contributions include the following three aspects. First, this paper proposes the *actionlet ensemble* model as a new way of characterizing human actions. Second, we propose a novel feature called local occupancy pattern, which is shown through our extensive experiments to be well suitable for the depth data-based action recognition task. Third, the proposed Fourier temporal pyramid is a new representation of temporal patterns, and it is shown to be robust to temporal misalignment and noise.

The proposed features and models are evaluated on five benchmark datasets: CMU MoCap dataset [1], MSR-Action3D dataset [19], MSR-DailyActivity3D dataset, Cornell Activity dataset (CAD-60) [29] and Multiview 3D Event dataset. The first dataset contains 3D joint positions captured by a multi-camera motion capturing system, and the other four datasets are captured with Kinect devices. Our extensive experimental results show that the proposed method is able to achieve significantly better recognition accuracy than the state-of-the-art methods. Moreover, we demonstrate that the proposed algorithm is insensitive to noise and translation and can handle view changes.

After a brief review of the related work in Section 2, the proposed LOP feature and the Fourier temporal pyramid are described in Section 3. Section 4 presents the *actionlet ensemble* model and its learning method. The empirical evaluations are given in Section 5. This paper is an extension of the conference paper [32].

2 Related Work

Actions are spatio-temporal patterns. There are two important issues in action recognition: the extraction and representation of suitable spatio-temporal features, and the modeling and learning of dynamical patterns.

Features can be sensor-dependent. In video-based methods, it is a common practice to locate spatio-temporal interest points like STIP [15], and then use the local distributions of the low-level features like gradients and optical flow (e.g., HOF [16] or HOG [9]) to represent the local spatio-temporal pattern. When we want to use depth data, however, because there is no texture in the depth map, these local features are not suitable.

It is generally agreed that knowing the 3D joint positions is helpful for action recognition. Multi-camera motion capture (MoCap) systems [4] can produce accurate 3D joint positions, but such special equipment is marker-based and expensive. It is still a challenging problem to develop a marker-free motion capturing system using regular video sensors. Cost-effective depth cameras have been used for motion capturing, and produced reasonable results, despite the noise when occlusion occurs. Because of the difference in the motion data quality, the action recognition methods designed for MoCap data might not be suitable for depth cameras.

In the literature, there have been many different temporal models for human action recognition. One way to model the human actions is to employ generative models, such as a Hidden Markov model (HMM) and Conditional Random Field (CRF). [20] used HMM over pre-defined relative positions obtained from the 3D joints. [13] used CRF over 3D joint positions. Similar approaches are also proposed to model human actions in normal videos [7], [24]. The 3D joint positions that are obtained via skeleton tracking from depth maps sequences are generally more noisy than the MoCap data. When the difference between the actions is small, without careful selection of the features,

determining the accurate states is usually difficult, which undermines the performance of such generative models. Moreover, with limited amount of training data, training a complex generative model is prone to overfitting.

Temporal patterns can also be modeled by a linear dynamical systems or a nonlinear Recurrent Neural Network [22]. Although these approaches are good models for time series data and are robust to temporal misalignment, it is generally difficult to learn these models from limited amount of training data.

Another method for modeling actions is the dynamic temporal warping (DTW) [23], which defines the distance of two time series as their edit distance. The action recognition can be done through nearest-neighbor classification. DTW's performance heavily depends on a good metric to measure the frame similarity. Moreover, for periodic actions (such as "waving"), DTW is likely to suffer from large temporal misalignment thus degrading classification performance [18].

Different from these approaches, we propose a *Fourier Temporal Pyramid* for temporal pattern representation. The Fourier temporal pyramid is a descriptive model. It does not require complicated learning as in the generative models (e.g., HMM, CRF and dynamical systems), and it is much more robust than DTW to noise and temporal misalignment.

In the actions with a complex articulated structure, the motions of the individual parts may be correlated. The relationship among these parts (or high-order features) is often more discriminative than the individual ones. Such combinatorial features can be represented by stochastic AND/OR structures. This idea has been pursued for face detection [8], human body parsing [41], object recognition [39], and human object interaction recognition [38]. This paper presents an initial attempt of using the AND/OR ensemble approach for action recognition. We propose a novel data mining solution to discover the discriminative conjunction rules based on [2], which is a branch-and-bound algorithm that guarantees to find all the frequent patterns efficiently, and apply multiple kernel learning framework to learn the ensemble. Other work that involves learning the interactions of human joints include poselet model [3] and phraselet model [10]. Poselet has been successfully applied in action recognition by mining discriminative appearance patterns to classify actions [21]. These models learn the relationship among human parts in annotated images.

Recently, a lot of efforts have been made to develop features for action recognition in depth data and skeletons. [19] represents each depth frame as a bag of 3D points along the human silhouette, and utilizes HMM to model the temporal dynamics. [31] learns semi-local features automatically from the data with an efficient random sampling approach. [30] also uses spatio-temporal occupancy patterns, but all the cells in the grid have the same size, and the number of cells is empirically set. [36] proposes a dimension-reduced skeleton feature, and [37] develops a histogram of gradient feature over depth motion maps. [25] selects most informative joints based on the discriminative measures of each joint. [40] utilizes distances between all pairs of joints as features and multiple instance learning for feature selection. [27] utilize Kinect cameras to recognizes

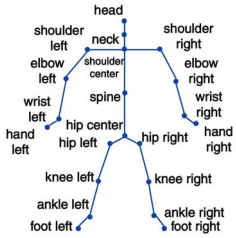

Fig. 2. Human joints tracked with the skeleton tracker [28].

dance actions. [6] uses linear dynamic systems to model the dynamic medial axis structures of human parts and proposes discriminative metrics for comparing sets of linear dynamics systems for action recognition, but it organizes skeleton joints into human parts manually rather than automatically learns from data. Our work is the first attempt to model the structure and relationship among the human parts and achieves state-of-the-art performance on multiple benchmark datasets.

3 Spatio-Temporal Features

This section gives a detailed description of two types of features that we utilize to represent the actions: the 3D joint position feature and the Local Occupancy Pattern (LOP). These features can characterize the human motions as well as the interactions between the objects and the human. In addition, the Fourier Temporal Pyramid is proposed to represent the temporal dynamics. The proposed features are invariant to the translation of the human body and robust to noise and temporal misalignment. The orientation normalization method, which can improve the proposed method's robustness to human orientation changes, is also discussed.

3.1 Invariant Features for 3D Joint Positions

The 3D joint positions are employed to characterize the motion of the human body. One key observation is that representing the human movement as the pairwise relative positions of the joints results in more discriminative features.

For a human subject, 21 joint positions are tracked by the skeleton tracker [28] and each joint i has 3 coordinates $\boldsymbol{p}_i(t) = (x_i(t), y_i(t), z_i(t))$ at a frame t. The illustration of the skeleton joints are shown in Fig. 2. The coordinates are normalized so that the motion is invariant to the initial body orientation and the body size. The details of the orientation normalization can be found in Section 3.4.

For each joint i, we extract the pairwise relative position features by taking the difference between the position of joint i and any other joint j:

$$\boldsymbol{p}_{ij} = \boldsymbol{p}_i - \boldsymbol{p}_j, \tag{1}$$

The 3D joint feature for joint i is defined as:

$$\boldsymbol{p}_i = \{\boldsymbol{p}_{ij} | i \neq j\}.$$

Although enumerating all the joint pairs introduces some information that may be irrelevant to our classification task, our system is capable of selecting the joints that are most relevant to our recognition task. The selection will

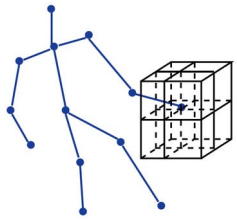

Fig. 3. Local occupancy pattern feature models the "depth appearance" around each joint. Note that there is a local occupancy pattern feature for every joint.

be handled by the *actionlet* mining algorithm as discussed in Section 4.

Relative joint position is actually a quite intuitive way to represent human motions. Consider, for example, the action "waving". It can be interpreted as "arms above the shoulder and moving left and right". This can be effectively characterized through the pairwise relative positions.

3.2 Local Occupancy Patterns

Using the 3D joint positions alone is insufficient to represent an action, especially when an action includes the interactions between human subject and other objects. Therefore, it is necessary to design a feature to describe the local "depth appearance" for the joints. In this paper, the interaction between the human subject and the environmental objects is characterized by the *Local Occupancy Patterns* or LOP at each joint. For example, suppose a person is drinking a cup of water. When the person fetches the cup, the space around his/her hand is occupied by the cup. Afterwards, when the person lifts the cup to his/her mouth, the space around both the hand and the head is occupied. The occupancy information can be useful to characterize this interaction and to differentiate the drinking action from other actions.

In each frame, as described below, an LOP feature computes the local occupancy information based on the 3D point cloud around a particular joint. The temporal dynamics of these occupancy patterns can discriminate different types of interactions. An illustration of the spatial-temporal occupancy pattern is shown in Fig. 3. Note that we only draw the LOP box for a single joint in Fig. 3, but in fact, a local occupancy pattern is computed for every joint.

At frame t, we have the point cloud generated from the depth map of this frame. For each joint j, its local region is partitioned into $N_x \times N_y \times N_z$ spatial grid. Each bin of the grid is of size (S_x, S_y, S_z) pixels. For example, if $(N_x, N_y, N_z) = (12, 12, 4)$ and $(S_x, S_y, S_z) = (6, 6, 80)$, the local $(72, 72, 320)$ region around a joint is partitioned into $12 \times 12 \times 4$ bins, and the size of each bin is $(6, 6, 80)$.

The number of points at the current frame that fall into each bin b_{xyz} of the grid is counted, and a sigmoid normalization function is applied to obtain the feature o_{xyz} for this bin. In this way, the local occupancy information of this bin is:

$$o_{xyz} = \delta\Big(\sum_{q \in \text{bin}_{xyz}} I_q\Big) \tag{2}$$

where $I_q = 1$ if the point cloud has a point in the location q and $I_q = 0$ otherwise. $\delta(.)$ is a sigmoid normalization function: $\delta(x) = \frac{1}{1+e^{-\beta x}}$. The LOP feature of a joint i is a vector consisting of the feature o_{xyz} of all the bins in the spatial grid around the joint, denoted by $\boldsymbol{o}_i$.

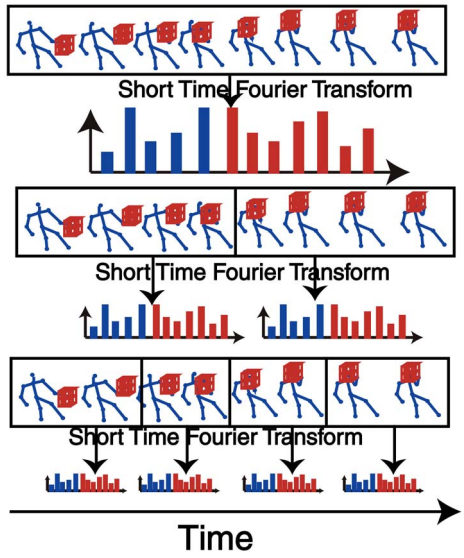

Fig. 4. Illustration of the Fourier Temporal Pyramid.

3.3 Fourier Temporal Pyramid

Two types of features are extracted from each frame t : the 3D joint position features $\boldsymbol{p}_j[t]$, and the LOP features $\boldsymbol{o}_j[t]$. In this subsection, we propose the Fourier temporal pyramid to represent the temporal patterns of these frame-level features.

When using the current cost-effective depth camera, we always experience noisy depth data and unreliable skeletons. Moreover, temporal misalignment is inevitable. We aim to design a temporal representation that is robust to both noisy data and the temporal misalignment. We also want such temporal features to be a good representation of the temporal structure of the actions. For example, one action may contain two consecutive sub-actions: "bend the body" and "pick up".

The proposed Fourier Temporal Pyramid is a descriptive representation that satisfies these properties. It is partly inspired by the Spatial Pyramid approach [17]. In order to capture the temporal structure of the actions, in addition to the global Fourier coefficients, we recursively partition the action into a pyramid, and use the short time Fourier transform for all the segments, as illustrated in Fig. 4. The final feature is the concatenation of the Fourier coefficients from all the segments.

For each joint j, let $\boldsymbol{g}_j = (\boldsymbol{p}_j, \boldsymbol{o}_j)$ denote its overall feature vector, where $\boldsymbol{p}_j$ is its 3D pairwise position vector and $\boldsymbol{o}_j$ is its LOP vector. Let N_j denote the dimension of $\boldsymbol{g}_j$, i.e., $\boldsymbol{g}_j = (g_1, \ldots, g_{N_j})$. Note that each element g_n is a function of time and we can write it as $g_n[t]$. For each time segment at each pyramid level, we apply Short Fourier Transform [26] to the element $g_n[t]$ and obtain its Fourier coefficients, and we utilize its low-frequency coefficients as features. The Fourier Temporal Pyramid feature at joint j is defined as the low-frequency coefficients at all levels of the pyramid, and is denoted as $\boldsymbol{G}_j$.

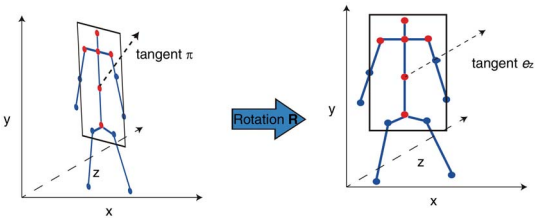

Fig. 5. Orientation alignment first fits a plane to the joints shown as red in the figure. Then, we compute a rotation matrix that rotates this plane to the x-y plane.

The proposed Fourier Temporal Pyramid feature has several benefits. First, by discarding the high-frequency Fourier coefficients, the proposed feature is robust to noise. Second, this feature is insensitive to temporal misalignment, because a temporally translated time series has the same Fourier coefficient magnitudes. Finally, the temporal structure of the actions is characterized by the pyramid structure.

3.4 Orientation Normalization

The proposed joint position features and local occupancy patterns are generally not invariant to the human orientation. In order to make the system more robust to human orientation changes, we perform the orientation normalization using the tracked skeleton positions. An illustration of the orientation normalization procedure is shown in Fig. 5.

In our experiment, we employ the up-right pose for orientation normalization. We find the frames where the human is approximately in an up-right pose, and use the pose of these frames for orientation alignment. If there is no up-right pose in a sequence, we do not perform orientation normalization for this sequence. For each frame where the human subject is in an up-right pose, we fit a plane to the joints "head", "neck", "hip", "left shoulder", and "right shoulder". The plane normal is used for orientation normalization.

Denote the 3D positions of the joints "head", "neck", "hip", "left shoulder", and "right shoulder" by $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_5$, respectively. The plane $f(\boldsymbol{p}) = \pi^T[\boldsymbol{p}; 1] = 0, \|\pi\|^2 = 1$, that best fits these joints can be found by minimizing the sum of the distances of the points $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_5$ to the plane:

$$\min_{\pi} \sum_{i=1}^{5} \|f(\boldsymbol{p}_i)\|^2 = \min_{\pi} \|\boldsymbol{P}\pi\|^2 \quad s.t. \|\pi\|^2 = 1 \tag{3}$$

where $\boldsymbol{P}$ is an constraint matrix defined as

$$\begin{bmatrix} \boldsymbol{p}_1 & \boldsymbol{p}_2 & \boldsymbol{p}_3 & \boldsymbol{p}_4 \\ 1 & 1 & 1 & 1 \end{bmatrix}^T \tag{4}$$

The plane parameters $\pi = [\pi_x; \pi_y; \pi_z; \pi_t]$ that minimize Eq. (3) are the right singular vector of $\boldsymbol{P}$ corresponding to the smallest singular value, which can be found by singular value decomposition.

In addition, we employ RANSAC procedure [11] to robustly estimate the plane. The RANSAC procedure iterates between the plane fitting step and the outlier detection step. The plane fitting step fits a plane to the non-outlier points by solving Eq. (3). The outlier detection step identifies the points that are too far from the plane as the outliers. The RANSAC procedure is more robust to the outliers of the 3D joint positions. When some joints are incorrectly tracked or the human pose we employ is not precisely upright, the RANSAC procedure can still robustly find the correct plane with small error.

To use the fitted plane for orientation normalization, we find a rotation matrix $\boldsymbol{R}$ that maps orientation of the plane $f(\boldsymbol{p}) = \pi^T[\boldsymbol{p}; 1] = 0$ to the x-y plane: $u(\boldsymbol{p}) = \boldsymbol{e}_z[\boldsymbol{p}; 1] = 0$, where $\boldsymbol{e}_z$ is the vector $[0; 0; 1; 0]$. Denote the normal of the plane $f(\boldsymbol{p}) = 0$ and $u(\boldsymbol{p}) = 0$ as

$$\pi' = \frac{[\pi_x; \pi_y; \pi_z]}{\|[\pi_x; \pi_y; \pi_z]\|_2} \tag{5}$$

$$\boldsymbol{e}'_z = [0; 0; 1] \tag{6}$$

This is equivalent as rotating the plane normal from π' to $\boldsymbol{e}'_z$, shown in Fig. 5. The rotation axis $\boldsymbol{x}$ and rotation angle θ of the rotation matrix $\boldsymbol{R}$ can be found as:

$$\boldsymbol{x} = [x_1; x_2; x_3] = \frac{\pi' \times \boldsymbol{e}'_z}{\|\pi' \times \boldsymbol{e}'_z\|} \tag{7}$$

$$\theta = \cos^{-1}\left(\frac{\pi'.\boldsymbol{e}'_z}{\|\pi'\|.\|\boldsymbol{e}'_z\|}\right) \tag{8}$$

Then the rotation matrix $\boldsymbol{R}$ can be defined according to exponential map:

$$\boldsymbol{R} = \boldsymbol{I}\cos\theta + \boldsymbol{A}\sin\theta + (1 - \cos\theta)\boldsymbol{x}\boldsymbol{x}^T \tag{9}$$

where $\boldsymbol{A}$ is a skew-symmetric matrix corresponding to $\boldsymbol{x}$

$$\boldsymbol{A} = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix} \tag{10}$$

When there are more than one frame with up-right pose, orientation normalization utilizes the average of the fitted plane normals of all the up-right poses in this sequence.

This rotation matrix can be applied to the 3D joint positions and 3D point cloud of all the frames for orientation normalization.

In addition to orientation normalization, we also perform scale normalization. The scale of the body can be estimated from the average pairwise distances of the skeleton joints "head", "neck", "hip", "left shoulder", and "right shoulder".

4 Actionlet Ensemble

To deal with the errors of the skeleton tracking and better characterize the intra-class variations, an *actionlet ensemble* approach is proposed in this section as a new representation of human actions.

An *actionlet* is defined as a conjunctive (AND) structure on the base features. One base feature is defined as the Fourier Temporal Pyramid features of an individual joint. A novel data mining algorithm is proposed to discover the *discriminative actionlets*, which are highly representative of one action and highly discriminative compared to the other actions.

Once we have mined a set of discriminative actionlets, a multiple kernel learning [5] approach is employed to learn an actionlet ensemble structure that combines these discriminative actionlets.

4.1 Mining Discriminative Actionlets

The human body consists of a large number of kinematic joints, but a particular action may only involve a small subset of them. For example, for right-handed people, action "calling cellphone" typically involves joints "right wrist" and "head". Therefore, the combinatorial feature of the two joints is a discriminative feature. Moreover, strong intra-class variation exists in some human actions. For left-handed people, action "calling cellphone" typically involves joints "left wrist" and "head". Therefore, the combinatorial feature of joint left wrist and head is another discriminative feature for this action. We propose the *actionlet ensemble* model to effectively characterize the combinatorial structure of human actions. An *actionlet* is a conjunction (AND) of the features for a subset of the joints. We denote an *actionlet* as its corresponding subset of joints $S \subseteq \{1, 2, \ldots, N\}$, where N is the total number of joints. Since one human action contains an exponential number of the possible actionlets, it is time consuming to construct an ensemble from all of the possible actionlets. In this section, We propose an effective data mining technique to discover the *discriminative actionlets*.

We employ the training data to determine whether an actionlet is discriminative. Suppose we have the training pairs $(\boldsymbol{x}^{(i)}, y^{(i)})$, where $\boldsymbol{x}^{(i)}$ is the features of i-th example and $y^{(i)}$ is the label of the i-th example. In order to determine how discriminative each individual joint is, a SVM model is trained on the feature $\boldsymbol{G}_j$ of each joint j. For each training example $\boldsymbol{x}^{(i)}$ and the SVM model on the joint j, the probability that its classification label $y^{(i)}$ is equal to an action class c is denoted as $P_j(y^{(i)} = c|\boldsymbol{x}^{(i)})$, which can be estimated from the pairwise probabilities by using pairwise coupling approach[34].

Since an actionlet takes a conjunctive operation, it predicts $y^{(i)} = c$ if and only if every joint $j \in S$ (the joint contained in this actionlet) predicts $y^{(i)} = c$. Thus, assuming the joints are independent, the probability that the predicted label $y^{(i)}$ is equal to an action class c given an example $\boldsymbol{x}^{(i)}$ for an actionlet S can be computed as:

$$P_S(y^{(i)} = c|\boldsymbol{x}^{(i)}) = \prod_{j \in S} P_j(y^{(i)} = c|\boldsymbol{x}^{(i)}) \tag{11}$$

Define $\mathcal{X}_c$ as the set of the training data with class label c: $\{i{:}t^{(i)} = c\}$. For a discriminative actionlet, the probability $P_S(y^{(i)} = c|\boldsymbol{x}^{(i)})$ should be large for some data in $\mathcal{X}_c$, and be small for all the data that does not belong to $\mathcal{X}_c$. Define the confidence score for actionlet S as

$$\text{Conf}_S = \max_{i \in \mathcal{X}_c} \log P_S(y^{(i)} = c|\boldsymbol{x}^{(i)}) \tag{12}$$

and the ambiguity score for actionlet S as

$$\text{Amb}_S = \frac{\sum_{i \notin \mathcal{X}_c} \log P_S(y^{(i)} = c|\boldsymbol{x}^{(i)})}{\sum_{i \notin \mathcal{X}_c} 1} \tag{13}$$

The discriminativeness of an actionlet S can be characterized by its confidence score Conf_S and ambiguity score Amb_S. A discriminative actionlet should exhibit large confidence score Conf_S and small ambiguity score Amb_S. Since one action contains an exponential number of actionlets, it is time consuming to enumerate all actionlets. We propose an Aprior-based data mining algorithm that can effectively discover the discriminative actionlets.

Algorithm 1: Discriminative Actionlet Mining

```
Take the set of joints, the feature G_j on each joint j, the
  number of the classes C, thresholds T_conf and T_amb.
Train the base classifier on the features G_j of each
  joint j.
for Class c = 1 to C do
    Set P_c, the discriminative actionlet pool for class c
      to be empty : P_c = {}. Set l = 1.
    repeat
        Generate the l-actionlets by adding one joint
          into each (l − 1)-actionlet in the discriminative
          actionlet pool P_c.
        Add the l-actionlets whose confidence scores
          are larger than T_conf to the pool P_c.
        l = l + 1
    until no discriminative actionlet is added to P_c in this
      iteration;
    remove the actionlets whose ambiguities scores
      are larger than T_amb in the pool P_c.
end
return the discriminative actionlet pool for all the
  classes.
```

An actionlet S is called an l-actionlet if its cardinality $|S| = l$. One important property of the actionlet is that if we add a joint $j \notin S$ to an $(l-1)$-actionlet S to generate an l-actionlet $S \cup \{j\}$, we have $\text{Conf}_{S \cup \{j\}} \leq \text{Conf}_S$, i.e., adding a new joint into one actionlet will always reduce the confidence score.

As a result, the Aprior mining process [2] can be applied to select the actionlets with large Conf_S and small Amb_S. The Aprior-based algorithm is essentially a branch and bound algorithm that effectively prunes the search space by eliminating the actionlets that do not have the confidence score larger than the threshold. If the confidence score Conf_S of an actionlet S is already less than the confidence threshold, we do not need to consider any actionlets S' with $S' \supset S$, because the confidence score of these actionlets $\text{Conf}_{S'} < \text{Conf}_S$ is less than the confidence threshold.

The outline of the mining process is shown in Algorithm 1. For each class c, the mining algorithm outputs a discriminative actionlet pool P_c which contains the actionlets that meet our criteria: $\text{Amb}_S \leq T_{\text{amb}}$ and $\text{Conf}_S \geq T_{\text{conf}}$.

The speed of the proposed data mining algorithm is a lot faster than naively enumerating all the candidate actionlets. We implement the proposed data mining algorithm with Python and run it on a Corei7-2600K machine with 8 GB memory. In our experiment on MSR-DailyActivity3D dataset, which contains 20 human joints and 320 sequences, we set the threshold for confidence score$T_{\text{conf}} = -1$, and the threshold for ambiguity score $T_{\text{amb}} = -2$. The data mining algorithm generates 180 actionlets in 5.23 seconds. In contrast, naively enumerating all the candidate actionlets takes 307 seconds under the same environment.

Since we do not impose the constraints that the discriminative actionlets are significantly different from each other,

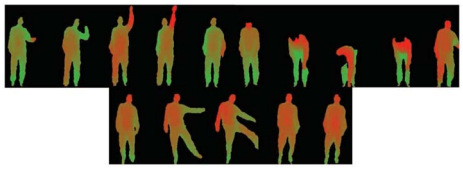

Fig. 6. Sample frames of the MSR-Action3D dataset.

there may be some redundancies among the discovered discriminative actionlets. We will employ multiple kernel learning algorithm to select the discriminative actionlets as described in the next subsection.

4.2 Learning Actionlet Ensemble

The discriminative power of a single actionlet is limited. In this subsection, we propose to learn an *actionlet ensemble* with multiple kernel learning approach.

An *actionlet ensemble* is a linear combination of the actionlet classifiers. For each actionlet S_k in the discriminative actionlet pool, we train an SVM model on it as an actionlet classifier, which defines a joint feature map $\Phi_k(\boldsymbol{x}, y)$ on data $\mathcal{X}$ and labels $\mathcal{Y}$ as a linear output function $f_k(\boldsymbol{x}, y)$ parameterized with the hyperplane normal $\boldsymbol{w}_k$ and bias b_k:

$$\begin{aligned} f_k(\boldsymbol{x}, y) &= \langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}, y) \rangle + b_k \\ &= \sum_i \alpha_{ik} K_k((\boldsymbol{x}_i, y_i), (\boldsymbol{x}, y)) + b_k \end{aligned} \tag{14}$$

Where each kernel $K_k(.,.)$ corresponds to the conjunctive features of an actionlet. The predicted class y for $\boldsymbol{x}$ is chosen to maximize the output $f_k(\boldsymbol{x}, y)$.

Multiclass-MKL considers a convex combination of p kernels, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \beta_k K_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Equivalently, we consider the following output function:

$$f_{\text{final}}(\boldsymbol{x}, y) = \sum_{k=1}^{p} \left[\beta_k \langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}, y) \rangle + b_k \right] \tag{15}$$

We aim at choosing $\boldsymbol{w} = (\boldsymbol{w}_k), \boldsymbol{b} = (\boldsymbol{b}_k), \beta = (\beta_k), k = 1, \ldots, p$, such that given any training data pair $(\boldsymbol{x}^{(i)}, y^{(i)})$, $f_{\text{final}}(\boldsymbol{x}^{(i)}, y^{(i)}) \geq f_{\text{final}}(\boldsymbol{x}^{(i)}, u)$ for all $u \in \mathcal{Y} - \{y^{(i)}\}$. The resulting optimization problem becomes:

$$\begin{aligned} &\min_{\beta, \boldsymbol{w}, \boldsymbol{b}, \xi} \frac{1}{2} \Omega(\beta, \boldsymbol{w}) + C \sum_{i=1}^{n} \xi_i \\ &\text{s.t. } \forall i: \xi_i = \max_{u \neq y_i} l(f_{\text{final}}(\boldsymbol{x}^{(i)}, y^{(i)}) - f_{\text{final}}(\boldsymbol{x}^{(i)}, u)) \end{aligned} \tag{16}$$

where C is the regularization parameter and l is a convex loss function, and $\Omega(\beta, \boldsymbol{w})$ is a regularization term on β and $\boldsymbol{w}$. Following the approach in [12], we choose $\Omega(\beta, \boldsymbol{w}) = \|\beta\|_1^2 + C_2 \|\boldsymbol{w}\|_2^2$. Since there exists redundancies among the discriminative actionlets discovered with the data mining algorithm, the l_1 regularization $\|\beta\|_1^2$ acts as a feature selection regularization by encouraging a sparse β , so that an ensemble of a small number of non-redundant actionlets is learned. The regularization $\|\boldsymbol{w}\|_2^2$ encourages the actionlet classifiers to have large margin.

This problem can be solved by iteratively optimizing β with fixed $\boldsymbol{w}$ and $\boldsymbol{b}$ through sparse solver, and optimizing $\boldsymbol{w}$ and $\boldsymbol{b}$ with fixed β through a generic SVM solver such as LIBSVM.

TABLE 1
Recognition Accuracy Comparison for MSR-Action3D Dataset

Method	Accuracy
Recurrent Neural Network [22]	0.425
Dynamic Temporal Warping [23]	0.540
Hidden Markov Model [20]	0.630
Action Graph on Bag of 3D Points [19]	0.747
Histogram of 3D Joints [35]	0.789
Random Occupancy Pattern [31]	0.862
Eigenjoints [36]	0.823
Sequence of Most Informative Joints [25]	0.471
Proposed Method with Absolute Joints Positions	0.685
Proposed Method	**0.882**

5 Experimental Results

We choose CMU MoCap dataset [1], MSR-Action3D dataset [19], MSR-DailyActivity3D dataset, Cornell Activity dataset [29], and Multiview 3D Event dataset to evaluate the proposed action recognition approach. In all the experiments, we use two-level Fourier temporal pyramid, with 1/4 length of each segment as low-frequency coefficients. The coefficients of all levels are concatenated sequentially. The empirical results show that the proposed framework outperforms the state-of-the-art methods.

5.1 MSR-Action3D Dataset

MSR-Action3D dataset [19] is an action dataset of depth sequences captured by a depth camera. This dataset contains twenty actions: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pick up & throw*. Every action was performed by ten subjects three times each. The frame rate is 15 frames per second and resolution 640×480. Altogether, the dataset has 402 action sequences with a total of 23797 frames of depth maps. Some examples of the depth sequences are shown in Fig. 6.

Those actions were chosen to cover a variety of movements of arms, legs, torso and their combinations. The subjects were advised to use their right arm or leg if an action is performed by a single arm or leg. Although the background of this dataset is clean, this dataset is challenging because many of the actions in the dataset are highly similar to each other.

The 3D joint positions are extracted from the depth sequence by using the real time skeleton tracking algorithm proposed in [28]. Since there is no human-object interaction in this dataset, we only extract the 3D joint position features in this experiment.

We compare our method with the state-of-the-art methods on the cross-subject test setting [19], where the examples of half of the subjects are used as training data, and the rest of the examples are used as testing data. The comparison of the recognition accuracy is shown in Table 1. The recognition accuracy of the dynamic temporal warping is only 54%, because some of actions in the dataset are very similar to each other, and there are typical large temporal misalignment in the dataset. The accuracy of

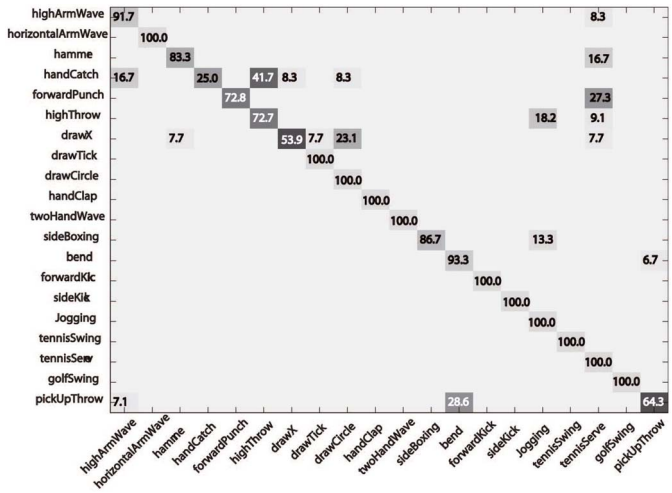

Fig. 7. Confusion matrix for MSR-Action3D dataset.

recurrent neural network is 42.5%, while he accuracy of Hidden Markov Model is 63%. The recently proposed joint-based action recognition methods, including Histogram of 3D joints, Eigenjoints and Sequence of most informative joints, achieve accuracy 78.9%, 82.3% and 47.06%, respectively. The proposed method achieves an accuracy of 88.2%. This is a very good performance considering that the skeleton tracker sometimes fails and the tracked joint positions are quite noisy. We also compare the proposed relative joint position features with the absolute joint position features. The proposed method using absolute joint positions achieves much worse accuracy than the proposed method using relative joint positions.

The confusion matrix is illustrated in Fig. 7. For most of the actions, our method works very well. The classification errors occur if two actions are too similar to each other, such as "hand catch" and "high throw", or if the occlusion

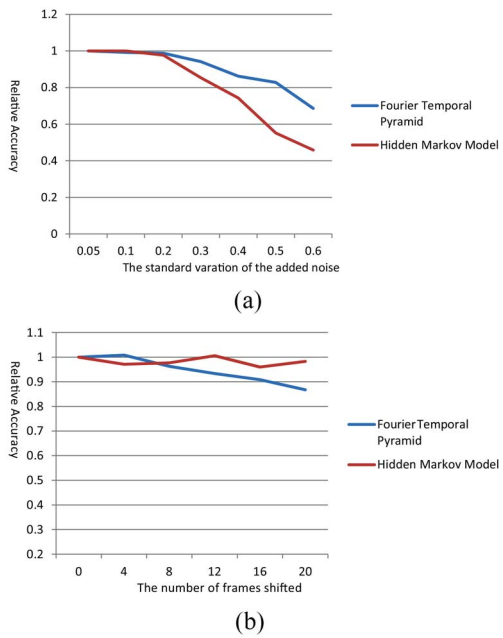

Fig. 8. Relationship between the relative accuracy and the variance of noise or temporal misalignment.

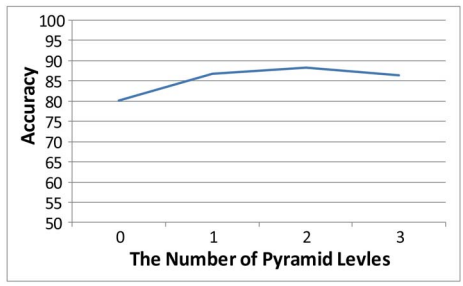

Fig. 9. Recognition accuracy of the proposed Actionlet Ensemble method using different levels of Fourier Pyramid.

is so large that the skeleton tracker fails frequently, such as the action "pick up and throw".

The proposed temporal representation Fourier Temporal Pyramid has two advantages: robustness to the noise and temporal misalignment, which are common in the action sequences captured with Kinect camera. In this experiment, we compare its robustness with a widely utilized temporal representation: Hidden Markov Model. The comparison of the noise robustness of the Fourier Temporal Pyramid features and that of Hidden Markov Model is shown in Fig. 8(a). In this experiment, we add white Gaussian noise to the 3D joint positions of the samples, and compare the relative accuracies of the two methods. For each method, its relative accuracy is defined as the accuracy under the noisy environment divided by the accuracy under the noiseless environment. We can see that the proposed Fourier Temporal Pyramid feature is much more robust to noise than Hidden Markov Model, because the clustering algorithm employed in Hidden Markov Model to obtain hidden states is relatively sensitive to noise, especially when the different actions are similar to each other.

The temporal shift robustness of the proposed method and the Hidden Markov model is also compared. In this experiment, we circularly shift all the training data, and

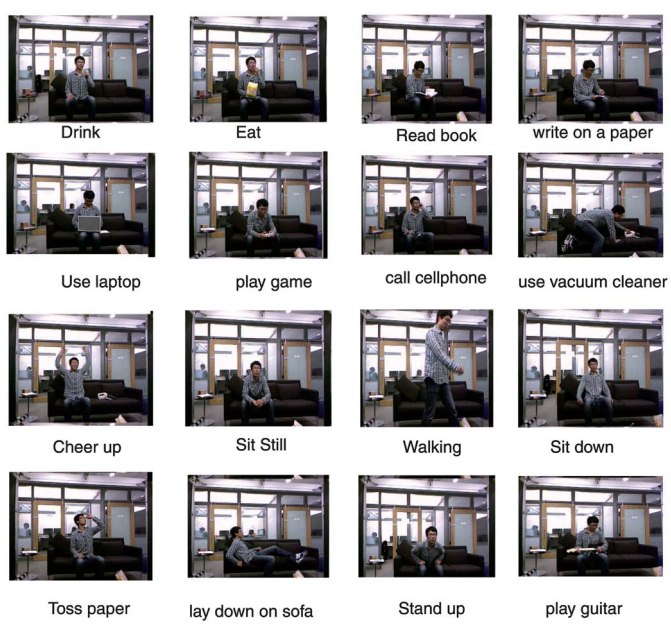

Fig. 10. Sample frames of the DailyActivity3D dataset.

TABLE 2
Recognition Accuracy Comparison for MSR-DailyActivity3D Dataset

Method	Accuracy
Dynamic Temporal Warping [23]	0.54
Random Occupancy Pattern [33]	0.64
Only LOP Features	0.43
Only Joint Position Features	0.68
SVM on Fourier Temporal Pyramid Features	0.78
Actionlet Ensemble on LOP Features	0.61
Actionlet Ensemble on Joint Features	0.74
MKL on All the Base Features	0.80
Actionlet Ensemble	**0.86**

keep the testing data unchanged. The relative accuracy is shown in Fig. 8(b). Hidden Markov Model is very robust to the temporal misalignment, because learning a Hidden Markov Model does not require the sequences to be temporally aligned. We find that the proposed approach is also robust to the temporal shift of the depth sequences, though the Fourier Temporal Pyramid is slightly more sensitive to temporal shift.

Thus, compared with widely applied Hidden Markov Model, the proposed Fourier Temporal Pyramid is a temporal representation that exhibits more robustness to noise while retaining the Hidden Markov Model's robustness to temporal misalignment. These properties are important for the action recognition with the depth maps and joint positions captured by Kinect devices, which can be very noisy and contain strong temporal misalignment.

Another advantage of the proposed Fourier Temporal Pyramid is its robustness to the number of action repetitions in the sequences. In order to evaluate the robustness of the proposed method to the number of action repetitions, we manually replicate all the action sequences two times and four times and apply the proposed algorithm to the new sequences. The recognition accuracy is 86.45% and 86.83% for two-times repetitions and four-times repetitions, respectively. If we repeat half of the action sequences two times, and the other half of the action sequences four times,

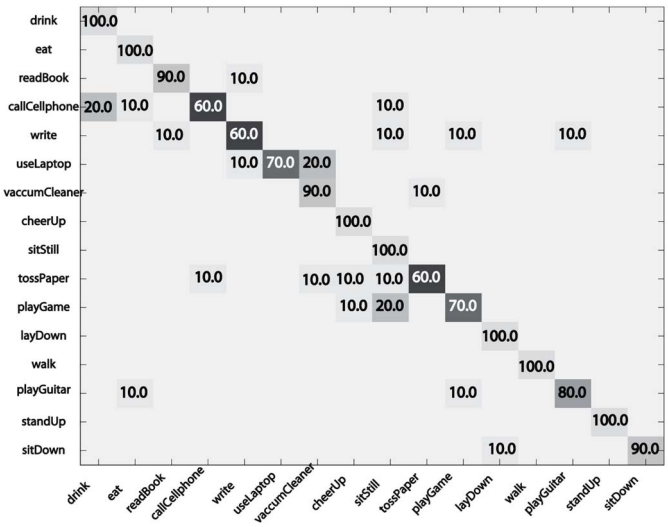

Fig. 11. Confusion matrix of the proposed method on DailyActivity3D dataset.

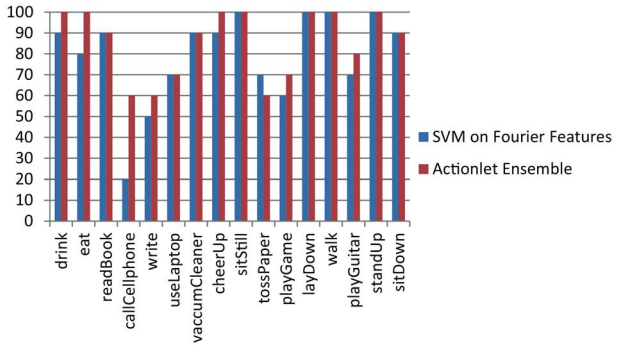

Fig. 12. Comparison between the accuracy of the proposed actionlet ensemble method and that of the support vector machine on the Fourier Temporal Pyramid features.

the recognition accuracy is 84.24%. This experiment shows that the proposed method is relatively insensitive to the number of action repetitions.

We also study the relationship between the levels of Fourier pyramid and the recognition accuracy of the proposed Actionlet Ensemble method, and the result is shown in Fig. 9. Each level of pyramid divides one temporal segment into two parts. For example, 2-level Fourier Temporal Pyramid contains 1, 2, 4 segments in level 0, 1 and 2, respectively. We can see that the proposed method achieves the best performance when the number of pyramid levels is 2, although the performance is quite close when the number of pyramid levels is 1 or 3.

5.2 DailyActivity3D Dataset

DailyActivity3D dataset is a daily activity dataset captured by a Kinect device. There are 16 activity types: *drink*, *eat*, *read book*, *call cellphone*, *write on a paper*, *use laptop*, *use vacuum cleaner*, *cheer up*, *sit still*, *toss paper*, *play game*, *lay down on sofa*, *walk*, *play guitar*, *stand up*, *sit down*. If possible, each subject performs an activity in two different poses: "sitting on sofa" and "standing". The total number of the activity sequences is 320. Some example activities are shown in Fig. 10.

This dataset is designed to cover daily activities in a living room. This dataset is more challenging than MSR-Action3D dataset. When the performer stands close to the sofa or sits on the sofa, the 3D joint positions extracted by the skeleton tracker are very noisy. Moreover, most of the activities involve the humans-object interactions.

We apply the cross-subject setting to evaluate the proposed algorithm on this dataset. Half of the subjects are used as training data, while the other half are used as testing data. Table 2 shows the accuracies of different methods. By employing an actionlet ensemble model, we obtain a

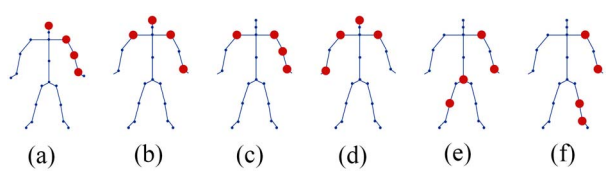

Fig. 13. Examples of the mined actionlets. The joints contained in each actionlet are marked as red. (a), (b) are actionlets for "drink" (c), (d) are actionlets for "call". (e), (f) are actionlets for "walk".

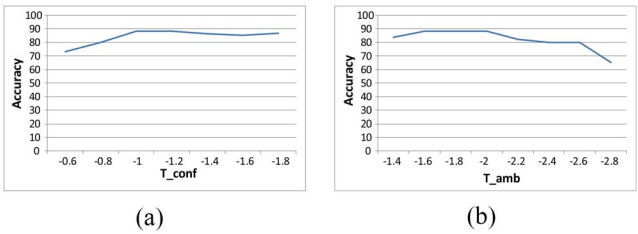

Fig. 14. Relationship between the recognition accuracy and the parameters T_{conf} and T_{amb}, which are the threshold of the confidence score and ambiguity score in the actionlet mining algorithm, respectively.

recognition accuracy of 85.75%. This is a decent result considering the challenging nature of the dataset. If we directly train an SVM on the Fourier Temporal Pyramid features, the accuracy is 78%. When only the LOP feature is employed, the recognition accuracy drops to 42.5%. If we only use 3D joint position features without using LOP, the recognition accuracy is 68%. If we train an Multiple Kernel Learning classifier on all the base features, the recognition accuracy is 80%. We also evaluate Random Occupancy Pattern (ROP) [33] method on MSR-DailyActivity3D dataset. Since, the ROP method requires a segmentation of the human, we manually crop out the human, and apply the ROP method on this dataset. The accuracy of ROP method is 64%.

Fig. 11 shows the confusion matrix of the proposed method. The proposed approach can successfully discriminate "eating" and "drinking" even though their motions are very similar, because the proposed LOP feature can capture the shape differences of the objects around the hand. Fig. 12 compares the accuracy of the actionlet ensemble method and that of the support vector machine on the Fourier Temporal Pyramid features. We can observe that for the activities where the hand gets too close to the body, the proposed actionlet ensemble method can significantly improve the accuracy. Fig. 13 illustrates some of the actionlets with large kernel weights discovered by our mining algorithm.

We also study the effect of the parameters of the actionlet mining algorithm on the recognition accuracy, shown in Fig. 14. In this experiment, we adjust T_{amb} while fixing $T_{conf} = -1$ and adjust T_{conf} while fixing $T_{amb} = -1.8$. We find that the proposed data mining algorithm is not sensitive to these two parameters as long as they are in a reasonable range. However, setting T_{conf} too high or setting T_{amb} too low may seriously undermine the recognition accuracy because the actionlet mining algorithm rejects discriminative actionlets in these cases. On the other hand, setting T_{conf} too low or setting T_{amb} too high may lead to a large number of the actionlets generated by the actionlet mining algorithm, which greatly slows down the actionlet mining and the actionlet ensemble learning procedures.

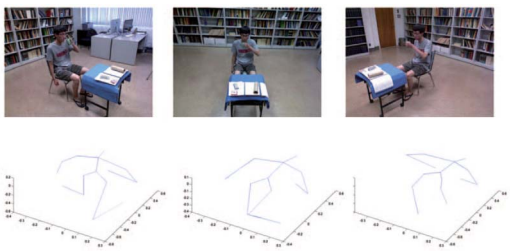

Fig. 15. Action captured from three views and their aligned skeletons.

Fig. 16. Sample frames of the Multiview 3D event dataset.

5.3 Multiview 3D Event Dataset

Multiview 3D event dataset[1] contains RGB, depth and human skeleton data captured simultaneously by three Kinect cameras. This dataset includes 8 event categories: *drink with mug*, *make a call*, *read book*, *use mouse*, *use keyboard*, *fetch water from dispenser*, *pour water from kettle*, *press button*. Each event is performed by 8 actors. Each actor repeats each event 20 times independently with different object instances and action styles. An example action captured from three view points is illustrated in Fig. 15. In total, there are 480 sequences per action class. Fig. 16 shows some examples of this dataset. The background of this dataset is relatively clean. The difficulty of this dataset is to generalize across different views. We apply our algorithm to this dataset to evaluate the robustness of our algorithm across different views.

Before applying the proposed approach to this dataset, we first perform a human orientation normalization described in Section 3.4. Although the human body orientations are aligned, action recognition across multiple views is still challenging due to the following two reasons. Firstly, the occlusions of different views are very different even for the same action, as shown in Fig. 15. Since the occluded joints are usually non-critical joints ("legs"), these occlusions can be handled effectively by the proposed actionlet ensemble model, because the actionlet ensemble model usually does not contain the non-critical joints. Secondly, the orientation normalization is usually not perfect due to skeleton tracking noise and errors. The proposed actionlet ensemble model is also robust to these noises thanks to the Fourier temporal pyramid representation. As a result, the proposed algorithm achieves very good performance on cross-view action recognition experiment.

1. This dataset will be released to public.

TABLE 3
Recognition Accuracy Comparison for Multiview 3D Event Dataset

Method	C-Subject	C-View
Dynamic Temporal Warping [23]	0.4712	0.4533
Hidden Markov Model [20]	0.848	0.6187
Actionlet Ensemble	**0.8834**	**0.8676**

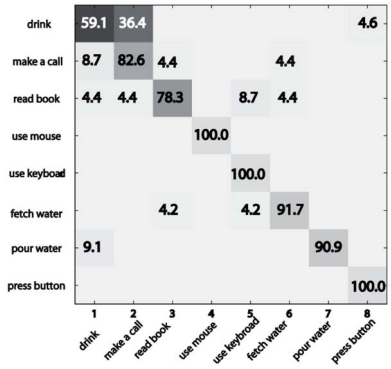

Fig. 17. Confusion matrix for Multiview 3D event dataset on cross-subject setting.

First, we perform cross-subject recognition experiment. In this setting, we use examples of 1/3 of the subjects as training data and the rest of the examples as testing data. We implement the dynamic temporal warping [23] and Hidden Markov Model [20] and compare the proposed model to these models. The classification accuracy comparison of these algorithms is shown in Table 3. The dynamic temporal warping achieves 47.12%, and the hidden Markov model achieves 84.8% accuracy on this setting, while the proposed algorithm achieves an accuracy of 88.34%. Most of the confusion occurs between the actions "drinking" and "make a phone call", because the movement of these two actions are very similar.

Then, we perform evaluation under the cross-view recognition setting. In this setting, the data are partitioned into three subsets each corresponding to a different camera. We use the subset from one camera as the testing data and use the data from the other two cameras for training. Three-fold cross-validation is applied to measure the overall accuracy. The results are listed in Table 3. One observation is that the proposed algorithm is quite robust across multiple views. The proposed algorithm achieves an accuracy of 86.75% on cross-view setting, which is only 1.4% lower than the accuracy on cross-subject setting. The confusion matrix of the proposed algorithm under cross-subject and cross-view settings are shown in Fig. 17 and Fig. 18, respectively.

The experimental results show that, with orientation normalization, the proposed algorithm can achieve good action recognition accuracy under cross-view action recognition setting.

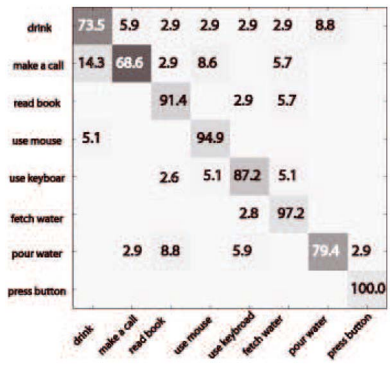

Fig. 18. Confusion matrix for Multiview 3D event dataset on cross-view setting.

TABLE 4
Recognition Accuracy Comparison for Cornell Daily Activity Dataset

Method	S-Person	C-Person
MEMM [29]	81.15	51.9
Object Offordances [14]	N/A	71.4
Actionlet Ensemble	**94.12**	**74.70**

5.4 Cornell Activity Dataset

Cornell Activity dataset (CAD-60)[29] contains the RGB frames, depth sequences and the tracked skeleton joint positions captured with Kinect cameras. The actions in this dataset can be categorized into 5 different environments: office, kitchen, bedroom, bathroom, and living room. Three or four common activities were identified for each environment, giving a total of twelve unique actions: "rinsing mouth", "brushing teeth", "wearing contact lens", "talking on the phone", "drinking water", "opening pill container", "cooking (chopping)", "cooking (stirring)", "drinking water", "talking on couch", "relaxing on couch", "talking on the phone", "writing on whiteboard", "drinking water", "working on computer"

The recognition accuracy is shown in Table 4. We employ the same experimental setup as [29]: The same-person experiment setup employs half of the data of the same person as training, and the other half is used as testing. The cross-person experiment setup uses leave-one-person-out cross-validation. The proposed method achieves an accuracy of 97.06% for the same-person setup and 74.70% for the cross-person setup. Both results are better than those of the state-of-the-art methods.

The confusion matrices of the proposed algorithm on Cornell Activity dataset under the same-person setting and the cross-subject setting are shown in Figs. 19 and 20, respectively. We can see that the proposed algorithm correctly classifies most of the actions under the same-person setting. The cross-person setting is more challenging, and we find that many actions are classified into "still" under this setting, because the motions of these actions are very

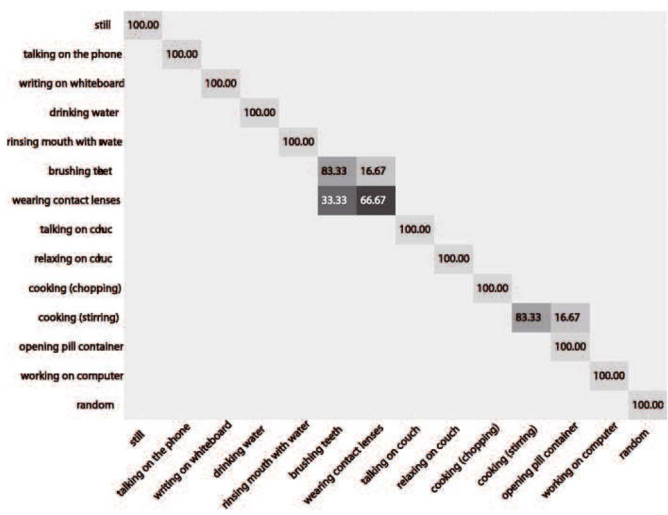

Fig. 19. Confusion matrix for Cornell Activity dataset on same-person setting.

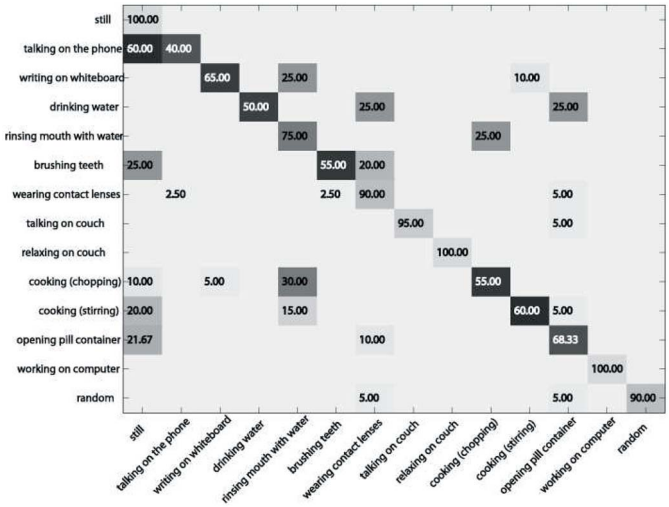

Fig. 20. Confusion matrix for Cornell Activity dataset on cross-person setting.

subtle and there are serious noises in Kinect skeleton tracking. Thus, it is difficult to distinguish the action "still" from those actions with subtle motions.

5.5 CMU MoCap Dataset

We also evaluate the proposed method on the 3D joint positions extracted by a motion capture system. The dataset we use is the CMU Motion Capture (MoCap) dataset.

Five subtle actions are chosen from CMU MoCap datasets following the configuration in [13]. The five actions differ from each other only in the motion of one or two limbs. The actions in this dataset include: *walking*, *marching*, *dribbling*, *walking with stiff arms*, *walking with wild legs*. The number of segments of each action is listed in Table 5. The 3D joint positions in CMU MoCap dataset are relatively clean because they are captured with high-precision camera array and markers. This dataset is employed to evaluate the performance of the proposed 3D joint position-based features on 3D joint positions captured by Motion Capture system.

The comparison of the performance is shown in Table 6. Since only the 3D joint positions are available, the proposed method only utilizes the 3D joint position features. It can be seen that the proposed method achieves comparable results with the state-of-the-art methods on the MoCap dataset.

6 Conclusion

In this paper, we propose a novel actionlet ensemble model that characterizes the conjunctive structure of 3D human actions by capturing the correlations of the joints that are representative of an action class. We also propose two novel features to represent 3D human actions with depth and skeleton data. Local occupancy pattern describes "depth appearance" in the neighborhood of a 3D joint. Fourier temporal pyramid describes the temporal structure of an action. The proposed features effectively discriminate human actions with subtle differences and human-object interactions and are robust to noise and temporal misalignment. Our extensive experiments demonstrated the superior performance of the proposed approach to the state-of-the-art methods. In the future, we aim to exploit the effectiveness of the proposed technique for the understanding of more complex activities.

TABLE 5
Description of the Subtle Action Dataset

Name	Number of segments
Walking	69
Marching	23
Dribbling	11
Walking with Stiff Arms	26
Walking with Wild Legs	28

TABLE 6
Recognition Accuracy Comparison for CMU MoCap Dataset

Method	Accuracy
Dynamic Temporal Warping [23]	0.6427
CRF with learned manifold space [13]	0.9827
Proposed Method	**0.9813**

Acknowledgment

This work was supported in part by National Science Foundation grant IIS-0347877, IIS-0916607, U.S. Army Research Laboratory and the U.S. Army Research Office under grant ARO W911NF-08-1-0504, and DARPA Award FA 8650-11-1-7149. Part of this work was done when J. Wang is doing an internship at Microsoft Research Redmond.

References

[1] CMU Graphics Lab Motion Capture Database [Online]. Available: http://mocap.cs.cmu.edu/
[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th VLDB*, vol. 1215. Santiago, Chile, 1994, pp. 487–499.
[3] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. CVPR*, Kyoto, Japan, 2009.
[4] L. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *Proc. 5th ICCV*, Cambridge, MA, USA, 1995.
[5] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 131–159, 2002.
[6] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3D discriminative skeletal features for human action recognition," in *Proc. HAU3D13*, Portland, OR, USA, 2013.
[7] H. Chen, H. Chen, Y. Chen, and S. Lee, "Human action recognition using star skeleton," in *Proc. 4th ACM Int. Workshop Video Surveillance Sensor Networks*, New York, NY, USA, 2006, pp. 171–178.
[8] S. Dai, M. Yang, Y. Wu, and A. Katsaggelos, "Detector ensemble," in *Proc. IEEE Conf. CVPR*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Computer Soc. Conf. CVPR*, San Diego, CA, USA, 2005, pp. 886–893.
[10] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. 12th ECCV*, Berlin, Germany, 2012.
[11] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 916–954, Sept. 2008.
[13] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia, "Discriminative human action recognition in the learned hierarchical manifold space," *Image Vis. Comput.*, vol. 28, no. 5, pp. 836–849, May 2010.
[14] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, 2013.
[15] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2–3, pp. 107–123, Sept. 2005.
[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. CVPR*, Anchorage, AK, USA, pp. 1–8, 2008.
[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. CVPR*, vol 2. Washington, DC, USA, 2006.
[18] L. Li and B. Prakash, "Time series clustering: Complex is simpler!" in *Proc. 28th ICML*, Bellevue, WA, USA, 2011.
[19] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. Human Communicative Behavior Analysis Workshop (in Conjunction with CVPR)*, 2010.
[20] F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost," in *Proc. 9th ECCV*, 2006, pp. 359–372.
[21] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. IEEE Conf. CVPR*, Providence, RI, USA, June 2011.
[22] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proc. 28th ICML*, Bellevue, WA, USA, 2011.
[23] M. Muller and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proc. 2006 ACM SIGGRAPH/Eurographics Symp. Computer Animation*, Aire-la-Ville, Switzerland, pp. 137–146.
[24] H. Ning, W. Xu, Y. Gong, and T. Huang, "Latent pose estimator for continuous action recognition," in *Proc. 10th ECCV*, Marseille, France, 2008, pp. 419–433.
[25] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 24–38, Jan. 2014.
[26] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete Time Signal Processing* (Prentice Hall Signal Processing Series). Upper Saddle River, NJ, USA: Prentice Hall, 1999.
[27] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proc. 2011 ACM SIGGRAPH/Eurographics SCA*, New York, NY, USA, p. 147.
[28] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Providence, RI, USA, 2011.
[29] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE ICRA*, Saint Paul, MN, USA, 2012.
[30] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. M. Campos, "STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences," in *Proc. 17th Iberoamerican Congr. Pattern Recognition Buenos Aires*, Buenos Aires, Argentina, 2012.
[31] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. 12th ECCV*, Florence, Italy, 2012, pp. 1–14.
[32] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. CVPR*, Providence, RI, USA, 2012.
[33] J. Wang, J. Yuan, Z. Chen, and Y. Wu, "Spatial locality-aware sparse coding and dictionary learning," in *Proc. ACML*, 2012.
[34] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, Aug. 2004.
[35] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints the University of Texas at Austin," in *Proc. CVPR 2012 HAU3D Workshop*.
[36] X. Yang and Y. Tian, "EigenJoints-based action recognition using naïve-bayes-nearest-neighbor," in *Proc. CVPR 2012 HAU3D Workshop*.
[37] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. 20th ACM Int. Conf. Multimedia*, Nara, Japan, 2012.
[38] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. IEEE Conf. CVPR*, San Francisco, CA, USA, 2010.
[39] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Proc. IEEE Conf. CVPR*, Providence, RI, USA, 2011.
[40] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, and S. Brook, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Computer Soc. Conf. CVPR 2012 HAU3D Workshop*, Providence, RI, USA.
[41] L. L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille, "Max margin and/or graph learning for parsing the human body," in *Proc. IEEE Conf. CVPR*, Anchorage, AK, USA, Jun. 2008.

Jiang Wang received the B.S. and M.S. degrees in electronic engineering from Fudan University, Shanghai, China, in 2007 and 2010, respectively. Currently, he is a Ph.D. candidate with Northwestern University, Evanston, IL, USA. His current research interests include human action recognition object recognition and deep learning.

Zicheng Liu is a Senior Researcher at Microsoft Research Redmond, WA, USA. He received the B.S. degree in mathematics from Huazhong Normal University, Wuhan, China, in 1984, the M.S. degree in operation research from the Institute of Applied Mathematics, Chinese Academy of Sciences, China, in 1989, and the Ph.D. degree in computer science from Princeton University, Princeton, NJ, USA, in 1996. Before joining Microsoft Research, he was at Silicon Graphics Inc. His current research interests include human activity recognition, 3D face modelling and animation, and multimedia signal processing. He is a senior member of IEEE.

Ying Wu received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, the M.S. degree in automation from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana-Champaign, IL, USA, in 1994, 1997, and 2001, respectively. He joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA, as an Assistant Professor, in 2001, where he is currently an Associate Professor of Electrical Engineering and Computer Science. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction. He was the recipient of the Robert T. Chien Award at UIUC in 2001 and the NSF CAREER Award in 2003. He is an Associate Editor of the *IEEE Transactions On Pattern Analysis And Machine Intelligence, the SPIE Journal of Electronic Imaging, and the IAPR Journal of Machine Vision and Application*.

Junsong Yuan received the B.Eng. degree in communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2002, the M.Eng. degree from the National University of Singapore, Singapore, and the Ph.D. degree from Northwestern University, Evanston, IL, USA. He joined Nanyang Technological University (NTU), Singapore, as a Nanyang Assistant Professor, in 2009, where he is currently the Program Director of Video Analytics with the Infocomm Center of Excellence, School of Electrical and Electronic Engineering. He was selected for the Special Class for the Gifted Young of the Huazhong University of Science and Technology in 1998. He holds three U.S. patents and two provisional U.S. patents. His current research interests include computer vision, video analytics and mining, multimedia searches, human computer interaction, and biomedical image analysis. He was a recipient of the Outstanding EECS Ph.D. Thesis Award from Northwestern University and the Doctoral Spotlight Award from the IEEE Conference on Computer Vision and Pattern Recognition in 2009. He is the Co-Chair of two workshops at the IEEE Conference on Computer Vision and Pattern Recognition in 2012.