This CVPR2014 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Cross-view Action Modeling, Learning and Recognition

Jiang Wang[1] Xiaohan Nie[2] Yin Xia[1] Ying Wu[1] Song-Chun Zhu[2]
[1]Northwestern University [2]University of California at Los Angles

Abstract

Existing methods on video-based action recognition are generally view-dependent, i.e., performing recognition from the same views seen in the training data. We present a novel multiview spatio-temporal AND-OR graph (MST-AOG) representation for cross-view action recognition, i.e., the recognition is performed on the video from an unknown and unseen view. As a compositional model, MST-AOG compactly represents the hierarchical combinatorial structures of cross-view actions by explicitly modeling the geometry, appearance and motion variations. This paper proposes effective methods to learn the structure and parameters of MST-AOG. The inference based on MST-AOG enables action recognition from novel views. The training of MST-AOG takes advantage of the 3D human skeleton data obtained from Kinect cameras to avoid annotating enormous multi-view video frames, which is error-prone and time-consuming, but the recognition does not need 3D information and is based on 2D video input. A new Multiview Action3D dataset has been created and will be released. Extensive experiments have demonstrated that this new action representation significantly improves the accuracy and robustness for cross-view action recognition on 2D videos.

1. Introduction

In the literature of video-based action recognition, most existing methods recognize actions from the view that is more or less the same as the training videos [6]. Their general limitation is the unpredictable performance in the situation where the actions need to be recognized from a novel view. As the visual appearances are very different from different views, and it is very difficult to find view-invariant features. Therefore, it is desirable to build models for cross-view action recognition, i.e., recognizing video actions from views that are unseen in the training videos. Despite some recent attempts [13, 7], this problem has not been well explored.

One possible approach is to *enumerate* a sufficiently large number of views and build dedicated feature and classifier for each view. This approach is too time consuming, because it requires annotating a large number of videos for all views multiplied by all action categories. Another possible approach is to *interpolate* across views via transfer learning [13]. This method learns a classifier from one view, and adapts the classifiers to new views. The performance of this approach is largely limited by the discrimination power of the local spatio-temporal features in practice.

In this paper, we approach this problem from a new perspective: creating a generative cross-view video action representation by exploiting the compositional structure in spatio-temporal patterns and geometrical relations among views. We call this model multiview spatio-temporal AND-OR graph model (MST-AOG), inspired by the expressive power of AND-OR graphs in object modeling [18]. This model includes multiple layers of nodes, creating a hierarchy of composition at various semantic levels, including *actions*, *poses*, *views*, *body parts* and *features*. Each node represents a conjunctive or disjunctive composition of its children nodes. The leaf nodes are appearance and motion features that ground the model. An important feature of the MST-AOG model is that the grounding does not have to be at the lowest layer (as in conventional generative models), but can be made at upper layers to capture low resolution spatial and temporal features. This compositional representation models geometry, appearance, and motion properties for actions. Once the model is learned, the inference process facilitates cross-view pose detection and action classification.

The AND/OR structure of this MST-AOG model is simple, but the major challenges lie in the learning of geometrical relations among different views. This paper proposes novel solutions to address this difficult issue. To learn the multiple-view structure, we take advantage of the 3D human skeleton produced by Kinect sensors as the 3D pose annotation. This 3D skeleton information is only available in training, but not used for cross-view action recognition. The projection of the 3D poses enables explicit modeling of the 2D views. Our model uses a set of discrete views in training to interpolate arbitrary novel views in testing. The appearances and motion are learned from the multiview training video and the 3D pose skeletons.

To learn the multiple-pose structure, we design a new

discriminative data mining method to automatically discover the frequent and discriminative poses. This data-driven method provides a very effective way to learn the structure for the action nodes. Since this hierarchical structure enables information sharing (e.g., different *view* nodes share certain body *part* nodes), MST-AOG largely reduces the enormous demands on data annotation, while improving the accuracy and robustness of cross-view action recognition, as demonstrated in our extensive experiments.

2. Related Work and Our Contributions

The literature on action recognition can be roughly divided into the following categories:

Local feature-based methods. Action recognition methods can be based on the bag-of-words representation of local features, such as HOG [1] or HOF [9] around spatio-temporal interest points [8]. Transfer learning-bsaed cross-view action recognition methods [3, 11, 25] are based on local appearance features. Hankelet [10] represents actions with the dynamics of short tracklets, and achieves cross-view action recognition by finding the Hankelets that are relative invariant to viewpoint changes. Self temporal similarity [7] characterizes actions with temporal self-similarities for cross-view action recognition. These methods work well on simple action classification, but they usually lack discriminative power to deal with more complex actions.

2D Pose-based methods. Recently, human pose estimation from a single image has make great progresses [20]. There is emerging interest in exploiting human pose for action recognition. Yao et al. [22] estimates the 2D poses from the images, and matches the estimated poses with a set of representative poses. Yao et al. [23, 24] developed spatio-temporal AND-OR graph to model the spatio-temporal structure of the poses in an actions. Desai et al. [2] learns a deformable part model (DPM) [4] that estimates both human poses and object locations. Maji et al. [14] uses the activations of *poselets*, which is is a set of pose detectors. Ikizler-Cinbis et al. [6] learns the pose classifier from web images. [21] proposes a coupled action recognition and pose estimation method by formulating pose estimation as an optimization over a set of action-specific manifold. In general, these methods were not specifically designed to handle cross-view actions. In contrast, this paper presents a new multi-view video action recognition approach.

3D skeleton-based methods. Pose-based action recognition generally needs a large amount of annotated poses from images. Recently, the development of depth cameras offers a cost-effective method to track 3D human poses [17]. Although the tracked 3D skeletons are noisy, it has been shown that they are useful to achieve good results in recognizing fine-grained actions [19]. In addition, the 2D DPM model can be extended to 3D [5, 12] to facilitate multi-view object detection. Parameswaran [15] proposes view-invariant canonical body poses and trajectories in 2D invariant space. In this paper, our proposed method uses the tracked 3D skeleton as supervision in training, but it stands out from other skeleton-based method because it does not need 3D skeletons inputs for action recognition in testing.

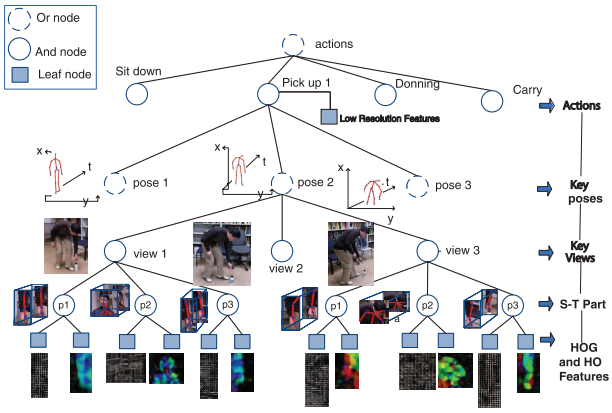


Figure 1. The MST-AOG action representation. The geometrical relationship of the parts in different views are modeled jointly by projecting the 3D poses into the given view, see Fig. 2. The parts are discriminately mined and shared for all the actions.

In comparison with the literature, this paper makes the following contributions:

- The proposed MST-AOG model is a compact but expressive multi-view action representation that unifies the modeling of geometry, appearance and motion.
- Once trained, this MST-AOG model only needs 2D video input to recognize actions from novel views.
- To train this MST-AOG model, we provide new and effective methods to learn its parameters, as well as mining its structure to enable effective part sharing.

3. Multi-view Spatio-Temporal AOG

3.1. Overview

Being a multi-layer hierarchical compositional model, the proposed multiview spatio-temporal AND-OR graph (MST-AOG) action representation is able to compactly accommodate the combinatorial configurations for cross-view action modeling. It consists of AND, OR and leaf nodes at various layers, and each node is associated with a score computed from its children. An AND node models the conjunctive relationship of its children nodes, and its score takes the summation over those of its children. An OR node captures the disjunctive relationship or the mixture of possibilities of its children node, and its score takes the maximum over its children. A leaf node is observable and is associated with evidence, and thus grounds the model.

The structure of the proposed MST-AOG model is shown in Fig. 1. The *root* node is an OR node representing the mixture of the set of all actions. We regard an action as a

sequence of discriminative 3D poses. A 3D pose exhibits a mixture of its projections on a set of 2D views. A 2D view includes a set of spatio-temporal parts, and each part is associated with its appearance and motion features. Thus, the *action* nodes, *view* nodes and *part* nodes are AND nodes, and *pose* nodes are OR nodes. We will discuss the scores and parameters for these nodes in the following subsections.

The strong expressive power of an AOG [18] lies in the structure of layered conjunctive and disjunctive compositions. Moreover, MST-AOG shares the part nodes across different views via interpolation. An example will be given when discussing the *action* node in Sec. 3.4.

3.2. Pose/View Nodes and 3D Geometry

To handle multi-view modeling, we introduce *pose* and *view* nodes. A *pose* node is an OR node that models the association of spatio-temporal patterns to a 3D pose projected to various views (each of which is a *view* node). For each *view* node, it captures the AND relationship of a number of parts (i.e., the limb of the human). Each *part* node captures its visual appearance and motion features under a specific view θ. Specifically, we use a star-shaped model for the dependencies among body parts, inspired by DPM [4], as Fig. 2 shows. Their 2D locations are denoted by $\mathcal{V} = \{\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$, where $\boldsymbol{v}_0$ is for the root part (the whole pose). Denote by I the image frame. We define the score associated with the i-th *part* node to be $S_{\mathcal{R}}(\boldsymbol{v}_i, I, \theta)$ (details will be provided in Sec. 3.3).

Two factors contribute to the score of a *view* node: the score of its children *part* nodes $S_{\mathcal{R}}(\boldsymbol{v}_i, I, \theta)$ and the spatial regularization among them $S_i(\boldsymbol{v}_0, \boldsymbol{v}_i, \theta)$ that specifies the spatial relationship between the root part and each child part. Such spatial regularization measures the compatibility among the parts from view θ (we only consider the rotation angle, details will follow). In view of this, the total compatibility score of a *view* node is written as:

$$S_{\mathcal{V}}(\boldsymbol{v}_0, \theta) = \sum_{i=0}^{N} S_{\mathcal{R}}(\boldsymbol{v}_i, I, \theta) + \sum_{i=1}^{N} S_i(\boldsymbol{v}_0, \boldsymbol{v}_i, \theta) \quad (1)$$

where $\boldsymbol{v}_i$ is the location of the part i, and θ is the view.

The 2D global location of a 2D pose is set to be the location of the root part, i.e.,$\boldsymbol{v}_0$. As the *pose* node is an OR node, the score for a *pose* node is computed by maximizing the scores from its children *view* nodes:

$$S_{\mathcal{P}}(\boldsymbol{v}_0) = \max_{\theta} S_{\mathcal{V}}(\boldsymbol{v}_0, \theta) \quad (2)$$

The evaluation of the spatial regularization of the parts needs a special treatment, because a *pose* node represents a 3D pose and it can be projected to different views to lead to different part relationships explicitly, as illustrated in Fig. 2.

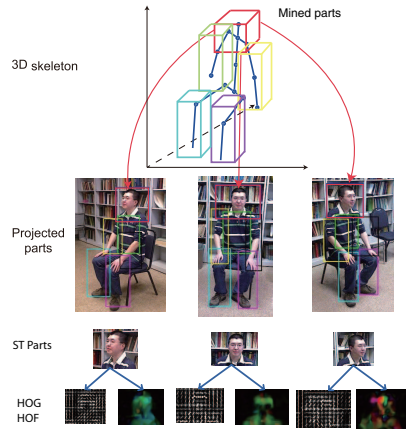


Figure 2. 3D parts and projected parts in different views.

The 3D geometrical relationship of the parts can be modeled as the 3D offsets of the i-th part with respect to the root part. Each offset can be modeled as a 3D Gaussian distribution with the mean $\boldsymbol{\mu}_i$ as well as diagonal covariance matrix $\boldsymbol{\Sigma}_i$.

$$\log P(\Delta \boldsymbol{p}_i) \propto -\frac{1}{2} (\Delta \boldsymbol{p}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\Delta \boldsymbol{p}_i - \boldsymbol{\mu}_i) \quad (3)$$

where $\Delta \boldsymbol{p}_i = (\Delta x_i, \Delta y_i, \Delta z_i)$ is the 3D offset between the part i and the root part. Here $\boldsymbol{\mu}_i$ can be estimated using the 3D skeleton data, and $\boldsymbol{\Sigma}_i$ will be learned (in Sec. 5).

The distribution of the 3D part offsets is projected to 2D for a given view. Here we assume scaled orthographic projections: $\boldsymbol{Q}_i^{\theta}$

$$\boldsymbol{Q}_i^{\theta} = \begin{bmatrix} k_1 \cos\theta & 0 & -k_1 \sin\theta \\ 0 & k_2 & 0 \end{bmatrix} \quad (4)$$

where θ is a rotation angle of the view, and k_1 and k_2 are the scale factors for two image axes. In training, we take advantage of the 3D skeleton data from Kinect cameras. Since we have the ground truth 3D (from 3D skeleton data) and 2D (from multiview videos) locations in our training data, these parameters can be easily estimated. The orthographic projection approximation works well in practice because the actors are sufficiently far away from the camera when performing actions. Since $\boldsymbol{Q}_i^{\theta}$ is a linear transform, the resulting projected 2D offset distribution is also a Gaussian distribution, with mean $\boldsymbol{\mu}_i^{\theta} = \boldsymbol{Q}_i^{\theta} \boldsymbol{v}_i$ and covariance matrix $\boldsymbol{\Sigma}_i^{\theta} = \boldsymbol{Q}_i^{\theta} \boldsymbol{\Sigma} (\boldsymbol{Q}_i^{\theta})^T$. Thus the 2D spatial pairwise relationship score $S_i(\boldsymbol{v}_0, \boldsymbol{v}_i, \theta)$ can be written as follows:

$$\begin{aligned} S_i(\boldsymbol{v}_0, \boldsymbol{v}_i, \theta) = ((\boldsymbol{\Sigma}_i^{\theta})_{11}^{-1}, (\boldsymbol{\Sigma}_i^{\theta})_{22}^{-1}, (\boldsymbol{\Sigma}_i^{\theta})_{12}^{-1})^T \cdot \\ (-\Delta u_i^2, -\Delta v_i^2, -2\Delta u_i \Delta v_i) \end{aligned} \quad (5)$$

where $(\Delta u_i, \Delta v_i) = \boldsymbol{v}_i - \boldsymbol{v}_0 - \boldsymbol{\mu}_i^{\theta}$ is the 2D deformation between the i-th part and the root part.

This 3D geometrical relationship is shared and learned across different views. The 2D geometrical relationship of

the novel views can be obtained by projecting the 3D geometrical relationship to the novel views.

3.3. Part Node and Motion/Appearance

The spatio-temporal patterns of a part under a view are modeled as its motion and appearance features. Each *part* has an *appearance* node with score $A_i(\boldsymbol{v}_i, I, \theta)$, and a *motion* node with score $M_i(\boldsymbol{v}_i, I, \theta)$. They capture the likelihood (or compatibility) of the appearance and motion of part i located at $\boldsymbol{v}_i$ under view θ, respectively. The score associated with a *part* node is thus written as:

$$S_{\mathcal{R}}(\boldsymbol{v}_i, I, \theta) = A_i(\boldsymbol{v}_i, I, \theta) + M_i(\boldsymbol{v}_i, I, \theta). \quad (6)$$

We exploit commonly used HOG [1] and HOF [9] features to represent the appearance and motion of a given part, respectively. In order to model the difference and correlation of the appearance and motion for one part in different view, we discretize the view angle θ into M discrete bins (each bin corresponds to a view node), and use exponential interpolation to obtain the appearance and motion features in the view bins. The appearance score function $A_i(\boldsymbol{v}_i, I, \theta)$ and motion score function is defined as

$$A_i(\boldsymbol{v}_i, I, \theta) = \frac{\sum_{m=1}^{M} e^{-d^2(\theta,\theta_m)} \boldsymbol{\phi}_{i,m}^T \phi(I, \boldsymbol{v}_i, \theta)}{\sum_{m=1}^{M} e^{-d^2(\theta,\theta_m)}} \quad (7)$$

where $e^{-d^2(\theta,\theta_m)}$ is the exponential of angular distance between the view θ and the view of bin m, $\boldsymbol{\phi}(I, \boldsymbol{v}_i, \theta)$ is the HOG features at the location $\boldsymbol{v}_i$ in image I under the view θ. $\boldsymbol{\phi}_{i,m}$ is the HOG templates of view bin m, and need to be learned from the training data (see Sec. 5). The motion score function $M_i(\boldsymbol{v}_i, I, \theta)$ is defined and learned from HOF features in a similar way.

Thus, the part node of different nodes are shared across different views via interpolation. We can learn the appearance/motion of the part nodes for the novel views via interpolation.

3.4. Action Node

Basically an action consists of a number of $N_{\mathcal{P}}$ 3D discriminative poses, but it is insufficient for an *action* node to include only a set of *pose* nodes for two reasons. First, when the image resolution of the human subject is low, further decomposing the human into body parts is not plausible, as detecting and localizing such tiny body parts will not be reliable. Instead, low resolution visual features may allow the direct detection of rough poses. Suppose we have $N_{\mathcal{L}}$ low resolution features, denoted by $\boldsymbol{\varphi}_i, i = 1, 2, \cdots, N_{\mathcal{L}}$. We simply use a linear prediction function $\sum_i^{N_{\mathcal{L}}} \boldsymbol{w}_i^T \boldsymbol{\varphi}_i$ to evaluate low-resolution-feature action prediction score. The weights $\boldsymbol{w}_i$ can be learnt for each low-resolution features. We use two low-resolution features: intensity histogram and size of the bounding boxes of the foreground.

Therefore, an *action* node consists of two kinds of children nodes: a $N_{\mathcal{P}}$ number of *pose* nodes and a $N_{\mathcal{L}}$ number of leaf nodes for low-resolution grounding. The score of an *action* node evaluates:

$$S_{\mathcal{A}}(l) = \sum_i^{N_{\mathcal{P}}} S_{\mathcal{P}}^i(\boldsymbol{v}_0) + \sum_i^{N_{\mathcal{L}}} \boldsymbol{w}_i^T \boldsymbol{\varphi}_i \quad (8)$$

where $S_{\mathcal{P}}^i(\boldsymbol{v}_0)$ is the score of the i-th *pose* node, and $\boldsymbol{w}_i$ is weights of the low-resolution features to be learned (as discussed in Sec. 5).

4. Inference

Given an input video from a novel view, the inference of MST-AOG calculates the scores of all the nodes so as to achieve cross-view action classification. Since this MST-AOG model is tree-structured, inference can be done via dynamic programming. The general dynamic programming process contains bottom-up phase and top-down phase, which is similar to sum-product and max-product algorithm in graphical model.

4.1. Cross-view Pose Detection

The states of the *pose* nodes, *view* nodes, and *part* nodes are their locations and scales. The score for a *view* node is defined in Eq. (1), and the score for a *pose* node is defined in Eq. (2). The inference of a *pose* node is simply comparing the scores of all the child *view* nodes at each location and scale, and finding the maximum score.

For a *view* node, since the score function (1) is convex, we can maximize the score in terms of the locations of the parts $\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N$ very efficiently using distance transform [4]. The inference step can be achieved by convolving the input frame and its optical flow with the appearance and motion templates of all the parts from different views and obtain the response maps. Then for each view bin, we can compute its projected part offset relationship. Using the distance transform, we can efficiently calculate the response map for the poses under this view bin. This also enables the estimation of the novel view by finding the view bin that has the largest view score.

4.2. Action Classification

We apply the spatio-temporal pyramid to represent the spatio-temporal relationship of poses and low-resolution features for action recognition. The scores of the *pose* nodes and the *low-resolution feature* nodes at different locations and frames constitute a sequence of response maps. We apply the max-pooling over a spatio-temporal pyramid. The response of a cell in the pyramid is the maximum among all responses in this cell.

We divide one whole video into 3-level pyramid in the spatio-temporal dimensions. This yields $1 + 8 + 64 = 73$

dimensional vector for each response map. Then, we can use the linear prediction function defined in Eq. (8) to compute the score of an action. The *action* node with the maximum score corresponds to the predicted action. Although this representation only acts as a rough description of the spatial-temporal relationships between the poses, we find it achieves very good results on our experiments.

5. Learning

The learning process has two tasks. The first is to learn the MST-AOG parameters, e.g., the appearance and motion patterns of each part in the *part* nodes, 3D geometrical relationship in the *view* and *pose* nodes, and the classification weights in the *action* nodes. The second task is to discover a dictionary of discriminative 3D poses to determine the structure of the MST-AOG model.

5.1. Learning MST-AOG Parameters

Learning MST-AOG parameters for the *part* and *view* nodes can be formulated as a latent structural SVM problem. The parameters of the latent SVM include: the variance $\mathbf{\Sigma}_i$ in Eq. (3), the appearance and motion templates $\boldsymbol{\beta}_{i,m}$ and $\boldsymbol{\gamma}_{i,m}$ in Eq. (7).

Although we have the non-root part locations and the view available in the training data, since we are more interested in predicting the pose rather than the precise location of each part and the view, we treat the locations of the parts $\boldsymbol{v}_j$ and the view θ as latent variables. And we apply a latent SVM to learn the our model using the labeled location of the parts and the view angle as initialization. This treatment is more robust to the noise in the training data.

For each example $\boldsymbol{x}_n$, we have its class label $y_n \in \{-1, +1\}$, $n \in \{1, 2, \cdots, N\}$. The objective function is:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{\Sigma}_i} \frac{1}{2}\|[\beta, \boldsymbol{\gamma}, \mathbf{\Sigma}_i]\|_2^2 + C\sum_{n=1}^{N} \max\left(0, 1 - y_n S_{\mathcal{P}}(\boldsymbol{v}_0 : \boldsymbol{x}_n)\right) \tag{9}$$

where $S_{\mathcal{P}}(\boldsymbol{v}_0 : \boldsymbol{x}_i)$ is defined in Eq. (2), which is the total score for example $\boldsymbol{x}_i$.

The learning is done by iterating between optimizing $\beta, \gamma, \mathbf{\Sigma}_i$, and calculating the part locations and the views of the positive training data.

For each pose, we use the samples whose distances are less than η to this pose in the positive videos as positive examples, and randomly sample 5000 negative training examples from negative videos. We apply two bootstrapping mining of hard negatives during the learning process. As the action score Eq. (8) is a linear function, the parameter $\boldsymbol{w}_i$ can be easily learned via a linear SVM solver.

5.2. Mining 3D Pose Dictionary

To learn the structure of the MST-AOG, we propose an effective data mining method to discover the discriminative 3D poses, which are specific spatial configurations of a subset body parts.

5.2.1 Part Representation

The 3D joint positions are employed to characterize the 3D pose of the human body. For a human subject, 21 joint positions are tracked by the skeleton tracker [17] and each joint i has 3 location coordinates $\boldsymbol{p}_j(t) = (x_j(t), y_j(t), z_j(t))$, a motion vector $\boldsymbol{m}_j(t) = (\Delta x_j(t), \Delta y_j(t), \Delta z_j(t))$ as well the visibility label $h_j(t)$ at a frame t. $h_j(t) = 1$ indicates that the j-th joint is visible in frame t and $h_j(t) = 0$ otherwise. The location coordinates are normalized so that they are invariant to the absolute body position, the initial body orientation and the body size. We manually group the joints into multiple parts.

5.2.2 Part Clustering

Since the poses in one action are highly redundant, we cluster the examples of each part to reduce the size of the search space, and to enable part sharing. Let part k be one of the K parts of the person and $\mathcal{J}_k$ be the set of the joints of this part. For each joint $j \in \mathcal{J}_k$ in this part, we have $\boldsymbol{p}_j = (x_j, y_j, z_j)$, $\boldsymbol{m}(j) = (\Delta x, \Delta y, \Delta z)$, and $\boldsymbol{h}_i \in \{0, 1\}$ as its 3D position, 3D motion and visibility map, respectively. For a certain part, given the 3D joint positions of the two examples s and r, we can define their distance:

$$\begin{aligned} D_k(s, r) = \sum_{j \in \mathcal{J}_k} & (\|\boldsymbol{p}_j^s(t) - \hat{\boldsymbol{S}}\boldsymbol{p}_j^r(t)\|_2^2 \\ & + \|\boldsymbol{m}_j^s(t) - \hat{\boldsymbol{S}}\boldsymbol{m}_j^r(t)\|_2^2)\,(1 + h_{s,r}(t)) \end{aligned} \tag{10}$$

where $\hat{\boldsymbol{S}}$ is a similarity transformation matrix that minimizes the distance between the part k of the example s and the example t. The term $h_{s,r}$ is a penalty term based on the visibility of the joint j in the two examples: $h_{s,r}(j) = a$ if $v_s(j) = v_r(j)$ and is 0 otherwise. Since this distance is non-symmetric, we use a symmetric distance as the distance metric: $\bar{D}(s, r) = (D(s, r) + D(r, s))/2$.

Spectral clustering is performed on the distance matrix. We remove the clusters that have too few examples, and use the rest of the clusters as the candidate part configurations for mining. We denote the set of all candidates part configurations for the part k as: $\mathcal{T}_k = \{t_{1k}, t_{2k}, \cdots, t_{N_k k}\}$, where each t_{ik} is called a *part item* represented by the average joint positions and motions in the cluster.

5.2.3 Mining Representative and Discriminative Poses

The discriminative power of a single part is usually limited. We need to discover poses (the combinations of the parts) that are discriminative for action recognition.

For a pose $\mathcal{P}$ that contains a set of part items $\mathcal{T}(\mathcal{P})$, with each part item in this set belonging to different part. we define the spatial configuration of a poses as the 3D joint positions and motions of all the part items in this pose.

The activation of a pose $\mathcal{P}$ with configuration $\boldsymbol{p}_{\mathcal{P}}$ in a video v_i can be defined as: $a_p(i) = \min_t e^{-D(\boldsymbol{p}_{\mathcal{P}}, \boldsymbol{p}^t_{\mathcal{P}})}$, where $\boldsymbol{p}^t_{\mathcal{P}}$ is the 3D joint positions of the poses $\mathcal{P}$ in the t-th frame of video, and $D(.,.)$ is a distance function defined in Eq. (10). If very similar poses exist in this video, the activation is high. Otherwise, the activation is low. One discriminative pose should have large activation in the videos in a given category, while having low activation vector in other categories. We define the support of the pose $\mathcal{P}$ for category c as:$Supp_p(c) = \frac{\sum_{c_i=c} a_p(i)}{\sum_{c_i=c} 1}$, where c_i the category label of video v_i, and the discrimination of the poses p as: $\text{Disc}_p(c) = \frac{Supp_p(c)}{\sum_{c' \neq c} Supp_p(c')}$.

We would like to discover the poses with large support and discrimination. Since adding one part item into a pose always creates another pose with lower support, i.e., $Supp_{\mathcal{P}}(c) < Supp_{\mathcal{P}'}(c)$ if $\mathcal{T}(\mathcal{P}) \supset \mathcal{T}(\mathcal{P}')$. Thus we can use the Aprior-like algorithm to find the discriminative poses. In this algorithm, we remove the non-maximal poses from the discriminative pose pool. For a pose $\mathcal{P}$, if there exist a pose $\mathcal{P}'$ such that $\mathcal{T}(\mathcal{P}) \subset \mathcal{T}(\mathcal{P}')$ and both $\mathcal{P}$ and $\mathcal{P}'$ are in the set of discriminative and representative poses, then $\mathcal{P}$ is a non-maximal pose.

This algorithm usually produces an excessive large number of poses, we prune the sets of discriminative poses with the following criteria. Firstly, we remove poses that are similar to each other. This can be modeled as a set-covering problem, and can be solved with a greedy algorithm. We choose a pose $\mathcal{P}$ with highest discrimination, and remove the poses whose distance is less than a given threshold. Secondly, we remove the poses with small validation scores for the detectors trained for these poses.

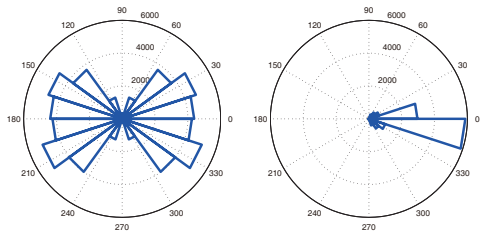
6. Experiments

We evaluate the proposed method on two datasets: the Multiview Action3D Dataset, collected by ourselves and the MSR-DailyActivity3D dataset [19].

In all our experiments, we only use the videos from a single unknown view for testing, and do not use the skeleton information or the videos from multiple views.

6.1. Northwestern-UCLA Multiview Action3D Dataset

Northwestern-UCLA Multiview 3D event dataset [1] contains RGB, depth and human skeleton data captured simultaneously by three Kinect cameras. This dataset include 10 action categories: *pick up with one hand*, *pick up with two hands*, *drop trash*, *walk around*, *sit down*, *stand up*, *donning*, *doffing*, *throw*, *carry*. Each action is performed by 10 actors. Fig. 4 shows some example frames of this dataset. The view distribution is shown in Fig. 3. This dataset contains data taken from a variety of viewpoints.



(a) (b)

Figure 3. The view distributions of the Multiview-Action3D dataset (left) and MSR-DailyActivity3D dataset (right).

The comparison of the recognition accuracy of the proposed algorithm with the baseline algorithms is shown in Table 1. We compare with virtual views [11], Hankelet [10], Action Bank [16] and Poselet [14]. For Action Bank, we use the actions provided by [16] as well as a portion of the videos in our dataset as action banks. For Poselet, we use the Poselets provided by [14]. We also compare our model with training one dedicated model for each view, which is essentially a mixture of deformable part models (DPM), to compare the robustness of the proposed method under different viewpoints with DPM model. We have 50 *pose* nodes for all the actions and 10 child *view* nodes for one *pose* node for both mixture of DPM and MST-AOG. The number of the *part* nodes in DPM and MST-AOG is both 1320 (different poses can have different number of parts). MST-AOG also has 2 child *low-resolution feature* nodes for each *action* node. These parameters are chosen via cross-validation. In MST-AOG, the appearance/motion and geometrical relationship of the *part* nodes are shared and learned across different *view* nodes, but the mixture of DPM treats them independently.

We perform recognition experiments under three settings.

- *cross-subject* setting: We use the samples from 9 subjects as training data, and leave out the samples from 1 subject as testing data.
- *cross-view* setting: We use the samples from 2 cameras as training data, and use the samples from 1 camera as testing data.
- *cross-environment* setting: We apply the learned model to the same action but captured in a different environment. Some of the examples of the cross environment testing data are shown in Fig. 4.

These settings can evaluate the robustness to the variations in different subjects, from different views, and in different environments.

[1] http://users.eecs.northwestern.edu/ jwa368/my_data.html

Figure 4. Sample frames of Multiview Action3D dataset, cross environment test data, and MSR-DailyActivity3D dataset [19].

Method	C-Subject	C-View	C-Env
Virtual View [11]	0.507	0.478	0.274
Hankelet [10]	0.542	0.452	0.286
Action Bank [16]	0.246	0.176	N/A
Poselet [14]	0.549	0.245	0.485
Mixture of DPM	0.748	0.461	0.688
MST-AOG w/o Low-S	0.789	0.653	0.719
MST-AOG w Low-S	**0.816**	**0.733**	**0.793**

Table 1. Recognition accuracy on Multiview-3D dataset.

The proposed algorithm achieves the best performance under all three settings. Moreover, the proposed method is rather robust under the cross-view setting. In contrast, although the state-of-the-art local-feature-based cross-view action recognition methods [10, 11] are relatively robust to viewpoint changes, the overall accuracy of these methods is not very high, because the local features are not enough to discriminate the subtle differences of the actions in this dataset. Moreover, these methods are sensitive to the changes of the environment. The Poselet method is robust to environment changes, but it is sensitive to viewpoint changes. Since the mixture of DPM does not model the relations across different view, its performance degrades significantly under cross-view setting. The comparison of the recognition accuracy of the different methods under cross-view setting is shown in Fig. 6. We also observe that utilizing low-resolution features can increase the recognition accuracy, and the proposed method is also robust under cross environment setting.

The confusion matrix of the proposed methods with low-resolution features under cross-view setting is shown in Fig. 5. The actions that cause most confusion are "pick up with one hand" versus "pick up with two hands", because the motion and appearance of these two actions are very similar. Another action that causes a lot of confusion is "drop trash", because the movement of dropping trash can be extremely subtle for some subjects.

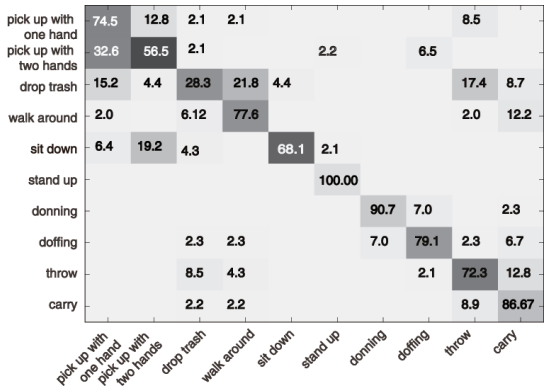


Figure 5. The confusion matrix of MST-AOG on multiview data under cross-view setting (with low-resolution features).

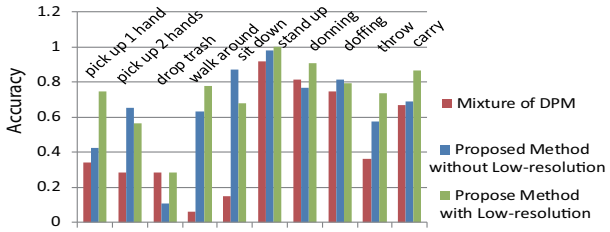


Figure 6. The recognition accuracy under cross-view setting.

6.2. MSR-DailyActivity3D Dataset

The MSR-DailyActivity3D dataset is a daily activity dataset captured by a Kinect device. It is a widely used as a Kinect action recognition benchmark. 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". Some example frames are shown in Fig. 4. The view distribution of this dataset can be found in Fig. 3. Although this dataset is not a multiview dataset, we compare the performance of the proposed method with the baseline methods to validate its performance on single view action recognition.

We use the same experimental setting as [19], using the samples of half of the subjects as training data, and the samples of the rest half as testing data. This dataset is very challenging if the 3D skeleton is not used. The Poselet method [14] achieves 23.75% accuracy, because many of the actions in this dataset should be distinguished with motion information, which is ignored in the Poselet method. STIP [8] and Action Bank [16] do not perform well on this dataset, either. The proposed MST-AOG method achieves a recognition accuracy of 73.5%, which is much better than the baseline methods.

Notice that the accuracy of Actionlet Ensemble method in [19] achieves 85.5% accuracy. However, the proposed method only needs one RGB video as input during testing,

Method	Accuracy
STIP [8]	0.545
Action Bank [16]	0.23
Poselet [14]	0.2375
Actionlet Ensemble [19]	0.835[a]
MST-AOG	**0.731**

Table 2. Recognition accuracy for DailyActivity3D dataset.

[a]This result is not directly comparable with MST-AOG, because it uses 3D skeleton.

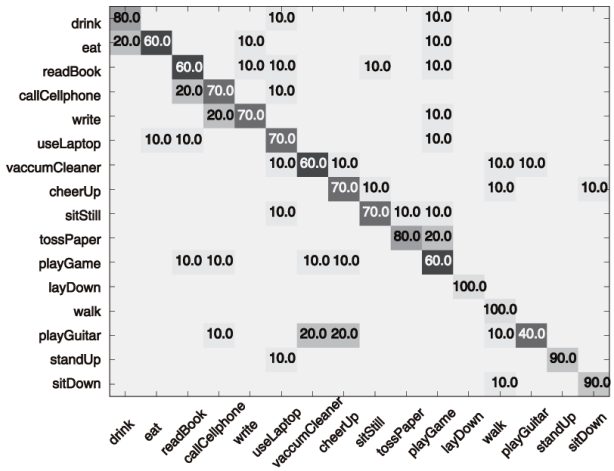


Figure 7. The confusion matrix of MST-AOG on MSR-DailyActivity3D dataset.

while Actionlet Ensemble method requires depth sequences and Kinect skeleton tracking during testing.

The confusion matrix of the proposed method on MSR-DailyActivity3D dataset is shown in Fig. 7. We can see that the proposed algorithm performs well on the actions that are mainly determined by poses or motion, such as "stand up", "sit down", "toss paper", "cheer up", "call cellphone". However, recognizing some actions requires us to recognize objects, such as "playing guitar" and "play games". Modeling the human-object interaction will improve the recognition accuracy for these actions.

7. Conclusion

We propose a new cross-view action representation, the MST-AOG model, that can effectively express the geometry, appearance and motion variations across multiple view points with a hierarchical compositional model. It takes advantage of 3D skeleton data to train, and achieves 2D video action recognition from unknown views. Our extensive experiments have demonstrated that MST-AOG significantly improves the accuracy and robustness for cross-view, cross-subject and cross-environment action recognition. The proposed MST-AOG can also be employed to detect the view and locations of the actions and poses. This will be our future work.

7.0.1 Acknowledgement

This work was supported in part by DARPA Award FA 8650-11-1-7149, National Science Foundation grant IIS-0916607, IIS-1217302, and MURI grant ONR N00014-10-1-0933. Part of the work was done when the first author was visiting VCLA lab in UCLA.

References

[1] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893. IEEE, 2005. 2, 4

[2] C. Desai and D. Ramanan. Detecting Actions , Poses , and Objects with Relational Phraselets. In *ECCV*, 2012. 2

[3] A. Farhadi and M. K. Tabrizi. Learning to recognize activities from the wrong view point. In *ECCV*, pages 154–166. Springer, 2008. 2

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–45, Sept. 2010. 2, 3, 4

[5] S. Fidler, S. Dickinson, and R. Urtasun. 3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model. In *NIPS*, pages 1–9, 2012. 2

[6] N. Ikizler-Cinbis and S. Sclaroff. Web-Based Classifiers for Human Action Recognition. *Multimedia, IEEE Transactions on*, 14(4):1031–1045, 2012. 1, 2

[7] I. N. Junejo, E. Dexter, I. Laptev, and P. Patrick. Cross-View Action Recognition from Temporal Self-Similarities. In *ECCV*, 2008. 1, 2

[8] I. Laptev. On Space-Time Interest Points. *IJCV*, 64(2-3):107–123, Sept. 2005. 2, 7, 8

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008. 2, 4

[10] B. Li, O. I. Camps, and M. Sznaier. Cross-view activity recognition using hankelets. In *CVPR*, pages 1362–1369. IEEE, 2012. 2, 6, 7

[11] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, pages 2855–2862. Ieee, June 2012. 2, 6, 7

[12] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. *CVPR*, pages 1688–1695, June 2010. 2

[13] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, pages 3209–3216. Ieee, June 2011. 1

[14] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*. IEEE, June 2011. 2, 6, 7, 8

[15] V. Parameswaran and R. Chellappa. View invariance for human action recognition. *IJCV*, 66(1):83–101, 2006. 2

[16] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, number May, 2012. 6, 7, 8

[17] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 2, 5

[18] Z. Si and S.-C. Zhu. Learning AND-OR Templates for Object Recognition and Detection. *PAMI*, 2013. 1, 3

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *CVPR*, 2012. 2, 6, 7, 8

[20] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts resenting shape. *CVPR*, 2011. 2

[21] A. Yao, J. Gall, and L. Van Gool. Coupled action recognition and pose estimation from multiple views. *IJCV*, 100(1):16–37, 2012. 2

[22] B. Yao and F.-F. Li. Action Recognition with Exemplar Based 2.5D Graph Matching. In *ECCV*, 2012. 2

[23] B. Yao, X. Nie, Z. Liu, and S.-C. Zhu. Animated pose templates for modelling and detecting human actions. *PAMI*, 36(3):436–452, 2014. 2

[24] B. Yao and S.-C. Zhu. Learning deformable action templates from cluttered videos. In *ICCV*, pages 1507–1514, 2009. 2

[25] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi. Cross-view action recognition via a continuous virtual path. In *CVPR*, pages 2690–2697. IEEE, 2013. 2