Action Recognition based on a mixture of RGB and Depth based skeleton

Srijan Das, Michal Koperski, François Bremond, Gianpiero Francesca

Action Recognition based on a mixture of RGB and Depth based skeleton

Srijan Das, Michal Koperski, Francois Bremond
INRIA, Sophia Antipolis
2004 Rte des Lucioles, 06902, Valbonne, France
name.surname@inria.fr

Gianpiero Francesca
Toyota Motor Europe
Hoge Wei 33, B - 1930 Zaventem
gianpiero.francesca@toyota-europe.com

Abstract

In this paper, we study how different skeleton extraction methods affect the performance of action recognition. As shown in previous work skeleton information can be exploited for action recognition. Nevertheless, skeleton detection problem is already hard and very often it is difficult to obtain reliable skeleton information from videos. In this paper, we compare two skeleton detection methods: the depth-map based method used with Kinect camera and RGB based method that uses Deep Convolutional Neural Networks. In order to balance the pros and cons of mentioned skeleton detection methods w.r.t. action recognition task, we propose a fusion of classifiers trained based on each skeleton detection method. Such fusion lead to performance improvement. We validate our approach on CAD-60 and MSRDailyActivity3D, achieving state-of-the-art results.

1. Introduction

Human Action Recognition is an important task in understanding the dynamic scenes and, it still remains a challenging task due to problems such as background clutter, partial occlusion, change in scale, viewpoint, lighting and appearance.

In this paper, we focus on comparing two skeleton detection methods and fuse them for developing a framework for human action recognition from the RGB-D videos. In RGB-D videos, most of the methods which achieve top results use skeleton detection. Skeleton based methods have become very popular on RGB-D due to the introduction of low-cost depth sensors such as Kinect. This made detection task much easier because segmentation on depth map is less challenging than on RGB. But RGB-D skeleton detection methods have problems when the subject covers too big distance, the depth map is noisy and it cannot work outdoors. Recent advancements in Convolutional Neural Networks (CNNs) has made it possible to detect the skeleton from RGB videos itself. Most of such approaches use a top down approach by first detecting the person. But such methods fail when person detection fails as it is prone to do when people are in close proximity. Moreover, in case of multi person frames, the computational cost increases. So, here we select a bottom up approach to detect the human skeleton using confidence maps for parts detection and Parts Affinity Fields (PAFs) as discussed in [4]. The pose machines works well in diverse scenarios of multi-person poses that contain many real world challenges like scale variation, dense crowd, occlusions. On the other hand it does not give good results in low lighting condition.

In this paper, we extract skeleton using both depth based method and RGB based method using CNNs and discuss their impact on Action Recognition task. To classify the actions, we follow the approach proposed by [5]. The input to the pose based CNN are the detected skeletons along with their corresponding RGB videos. The recent success of Convolutional Neural Networks (CNNs) have motivated us to find the CNN features for each body parts separately in each frame. Inspired from [5], we use flow based and appearance based CNN features computed from each part of the body which is aggregated further using max pooling to obtain the video descriptors.

The accuracy differs depending on selected skeleton extraction method. Based on fact above we propose to fuse the RGB-D skeletons and pose machines skeletons by fusing their classifiers scores (distances). We show our experimental results on two popular datasets CAD-60 and MSRDaily-Action3D datasets.

2. Related Work

Many authors in the past focused on methods based on local features [16, 23, 31]. Laptev *et al*. [15] have proposed Harris3D point detector. Some authors focused on depth point cloud methods [36] and they are robust to noise and occlusions. Currently, the Dense Trajectories [31] combined with Fisher Vector (FV) aggregation have shown good results.

Many authors [12, 11, 2, 10] have proposed a method to merge both RGB and depth information for action recognition. Kong *et al*. [10] used a projection function which is

learned based on both RGB and depth features.

There are many approaches [34, 1, 19] that use detected human skeleton for modeling actions, which became easier after the introduction of affordable depth sensors. Vemulapalli *et al*. [30] represented each skeleton using the relative 3D rotations between various body parts. Their skeletal representation becomes a point in a Riemannian manifold. Then, using this representation, they model human actions as curves in this manifold and perform classification in the Lie algebra. Wu *et al*. [33] proposed a hierarchical dynamic framework that first extracts high level skeletal features and then uses the learned representation for estimating emission probability to infer action sequences.

Recently, deep learning methods show some promising results in action recognition [14]. Deep learning methods require huge amount of annotated data for training. Some authors use pre-trained CNNs for action recognition [9, 32]. But application of CNNs in action recognition has shown little improvement so far [27, 37]. Mahasseni *et al*. [18] have proposed that action recognition in video can be improved by providing an additional modality in training data-namely, 3D human skeleton sequences. For recognition, they used Long Short Term Memory (LSTM) grounded via a deep CNN onto the video. They regularized the training of LSTM using the output of another encoder LSTM grounded on 3D human skeleton training data. Most of the recent action recognition works focus on using global aggregation of local descriptors. Some methods have used human joints and their temporal evolution to recognize actions. But human pose estimation is still a challenging task. Most approaches [22, 6, 28], for multi-person pose estimation have used a top down strategy where the person is detected first and then on the detected regions, poses are estimated. There are some approaches which uses bottom up approach as in [21] that jointly labels part detection candidates and associated them to individual people. This approach does not rely on person detection but involves solving an integer linear programming over fully connected graph which is an NP-hard problem. Thus the average processing time for a single image is in the order of hours. So, we have selected an approach [4] which uses bottom up approach to extract the skeletons using confidence maps for parts detection, Parts Affinity Fields (PAFs) for detecting the parts associations and greedy parse algorithm to quantify the correct detections.

We use the estimated poses from pose machines as well as skeletons from RGB-D to compute the CNN features. We selected the approach discussed in [5] which uses positions, appearance and motion of human body parts to compute the CNN features.

3. Proposed Method

The proposed method consists of two steps: skeleton detection described in section 3.1 and 3.2 and feature extraction for action recognition in section 3.3 and 3.4. Please note that we use two different skeleton inputs, compare them and fuse them to obtain better final action recognition accuracy.

3.1. Skeleton from RGB-D

One of the skeleton detection method that we use is using Kinect as discussed in [26]. It infers the body in a two stage process: first computes a depth map and then infer body position. The body parts are detected using a randomized decision forest, learned from over 1 million training examples. Inferring the body position is a two-stage process. First a depth map is computed and then the body position is inferred. It begins with 100,000 depth images with known skeletons from a motion capture system and then computer graphics is used to render all sequences for 15 different body types. Thus a million training examples are produced which are used to learn a randomized decision forest for mapping the depth images to body parts. Then, the mean shift algorithm is used to robustly compute the modes of probability distributions to transform the body image into a skeleton.

3.2. Skeleton from Pose Machines

Second method that we use for skeleton detection from RGB videos is [4]. The realtime multi person pose estimation algorithm is used to detect the 2D pose of multiple people in images. They present an explicit nonparametric representation of the keypoints that considers both position and orientation of human limbs. They also designed a CNN architecture for jointly learning the parts and parts association. They also use Part Affinity Fields (PAFs), a set of vector fields each of which encodes the location and orientation of a particular limb at each position in the image domain. Then, they use a greedy parsing algorithm to detect the correct candidates of the parts association using the PAFs and form the full body pose of all people in the image.

Detection using confidence Maps - The confidence Maps are obtained from the input images for detecting the parts. If $\mathrm{x}_{j,k} \in \mathbb{R}^2$ be the ground truth position of body part j for person k then, the value at location $\mathrm{p} \in \mathbb{R}^2$ in the confidence map $S_{j,k}$ for person k is given by

$$S^*_{j,k}(p) = e^{\frac{-||p-x_{j,k}||_2^2}{\sigma}} \quad (1)$$

The confidence map $S^*_j \in \mathbb{R}^{w \times h}$ with $w \times h$ being the dimension of the image and σ being chosen empirically. Ideal confidence map is an aggregation of peaks of all people in

a single map via a max operator

$$S_j^*(p) = \max_k S_{j,k}^*(p) \quad (2)$$

Association using PAFs - A measurement of the confidence for each pair of two part detections that they are associated from the same person is required. The part affinity field is a 2D vector that encodes the direction that points from one point to the other. Each type of limb has an associated field joining its two associated body parts. Ideal part affinity field to be predicted by the network combines the limbs of type c of all people into a single map.

During testing, the confidence score of each limb candidate by measuring the alignment of the predicted PAF with the candidate limb that would be formed by connecting the detected body parts.

Greedy Parsing algorithm - A set of body part detection candidates D_j for multiple people using non maxima supression on each predicted confidence map, where $D_j = \{d_j^m : j \in \{1, 2, ..., J\}, m \in \{1, 2..., N_j\}\}$ with N_j being the number of candidates of part type j and $d_j^m \in \mathbb{R}^2$ the location of the m-th detection candidate of body part type j. The detected body parts are required to be associated with other parts from the same person. A variable $Z_{mnj_1j_2} \in \{0, 1\}$ is defined to indicate whether two detection candidates $d_{j_1}^m$ and $d_{j_2}^m$ are connected and the goal is to find the optimal assignment for the set of all possible connections $Z = \{Z_{mnj_1j_2} : j_1, j_2 \in \{1, 2..., J\}, m \in \{1, 2..., N_j\}, n \in \{1, 2..., N_{j_2}\}\}$.

Thus a bipartite graph matching problem is to be solved in which nodes of the graph are D_{j_1} and D_{j_2} and edges are all possible connections between pair of detection candidates. Each edge is weighted with the part affinity aggregates. The goal is to find a matching with maximum weight for the chosen edges for which the Hungarian algorithm is used.

3.3. Pose based CNN

In pose based CNN, the images are cropped around the joints as discussed in [5] to get the different body part patches. These part patches are taken as input in CNN to get the CNN features. The body regions are represented with both motion based and appearance based CNN descriptors. These descriptors are extracted per frame and aggregated over time.

In order to construct the CNN features, we first compute the optical flow for each pair of frames using [3]. Then we, crop RGB image patches and flow patches right hand, left hand, upper body, full body and full image. Each patches obtained are resized to 224×224 in order to match the CNN input layer. We use two different architecture to obtain the appearance and flow based frame features. Each networks having 5 convolutional and 3 fully connected layers. The output of the last layer consists of 4096 values which is considered as the frame descriptor. For the RGB patches, we use VGG-f network that has been pre-trained on the ImageNet ILSVRC-2012 challenge dataset. For the flow patches, we use the motion network provided by [7] that has been pre-trained for action recognition task on the UCF-101 dataset.

From each descriptors f_t^p for each part p and each frame t of the video, we perform a max pooling over all the frames to obtain a fixed-length video descriptor. Finally, video descriptors for motion and appearance for all parts are normalized by dividing the video descriptors by the average l_2-norm of the f_t^p from the training set and concatenated to get the final CNN features.

We compute a χ^2-kernel from these CNN features which is the input to the SVM classifier.

3.4. Fusing RGB-D and Pose machines skeleton

Fusing the RGB-D and pose machines skeleton is the key idea of this work. This is done because there are instances which are discussed in section 4, where pose machines can detect the skeletons better than RGB-D and vice versa. RGB-D does work well when the subject is in front of the camera and sometimes pose machines fails in the skeleton detection when the subject gets mixed up with the background color. The pose based CNN features computes the features from the upper body, right hand, left hand, full image and full body patches. So, we put more importance to the patches on the upper body and the pose machines in such situations works well in detecting the skeleton on the upper side of the subject as discussed in section 4.1. Thus, it is very important to take the advantages of both the skeletons from pose machines and RGB-D which is done by fusing the classifier scores (distances). We report the accuracy of our approach using RGB-D skeletons, using RGB skeletons and using both the skeletons.

For classification based on either RGB-D skeleton or pose machine skeleton, we use SVM classifier with χ^2-kernel. From each skeleton, the $\chi^2(x, y)$-kernel is computed using equation 3 from the pose based CNN features.

$$\chi^2(x, y) = 1 - \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\frac{1}{2}(x_i + y_i)} \quad (3)$$

Let's define d as the distance of test example to SVM decision plane, then d_k is the distance of test example to decision plane of SVM trained on input from RGB-D skeleton and d_r is the distance of test example to decision plane of SVM trained on input from pose machine skeleton. To fuse both the classifiers, we propose to use the following weighted sum

$$d_f = \alpha d_k + (1 - \alpha) d_r \quad (4)$$

The value for α is found using cross-validation.

4. Experiments

We evaluate our framework on 2 public data-sets: CAD-60 [29] - contains 60 RGB-D videos with 4 subjects and 14 actions. The dataset contains the RGB-D frames which are used to find the skeletons using pose machines and depth information. MSRDailyActivity3D [34] - contains 320 RGB-D videos with 10 subjects and 16 actions. The dataset contains the RGB frames which are used to detect the skeletons using pose machines and world coordinates which are extracted in order to get the RGB-D skeleton information. This dataset has been captured in a living room and consists of all daily living activities.

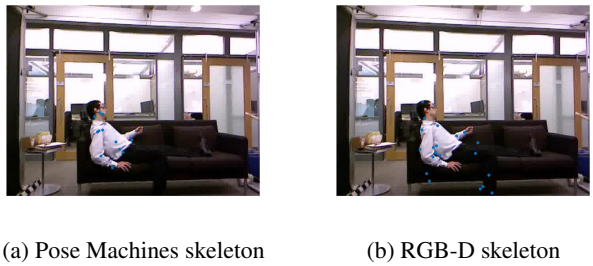
From both the above datasets, we extract the skeleton from RGB frames and their depth information. We evaluate these dataset by setting up a cross actor training/testing setup. We left out each actor from the training set and repeated the experiment for each of them.

4.1. Comparison of RGB and Depth based Skeleton

In this section, we provide qualitative result by visualizing the extracted skeleton and report the accuracy of the actions. For the qualitative results, we detect the poses of the person using pose machines and then use them to recognize the actions. We repeat the experiment to recognize the actions again with skeleton information from depth data and RGB data. From the confusion matrix, we can see that depth data recognize activities like drinking water and cooking(chopping) in CAD60 dataset with higher accuracy. On other hand, all other activities are recognized with similar accuracy using both the skeleton extraction methods. So, this difference in action recognition accuracy is because of the difference in the estimated poses. For activities such as drinking water RGB data does not work well to detect the skeleton, because in such activities the hand of the subject overlaps with the body. While actions like still, opening pill container, working on computer and so on, RGB data works similar to depth data to detect the skeleton because the hands are away from the body in the 3-D space and hence can be distinguished accurately from the other parts of the body.

In MSRDailyActivity3D dataset, the person performing the action stands close to a sofa which makes the skeleton extraction using pose machines difficult. In most of the cases, the depth based skeleton works better because of the frontal view of the subjects. For actions like vacuum cleaning, laying, walking and gaming, pose machines works a bit better than the depth based skeleton. The depth based skeleton works better for actions standing up and all other actions as compared to pose machines skeleton. This is because, in most of these actions, the subject has a frontal view and less motion of the body parts.

In Figure 1, the person lying on the sofa in MSRDaily-Action3D dataset is detected well by depth based method. We have only considered patches from the upper body parts and the full body for obtaining the CNN features, so pose machines gives better results in this case since it detects the skeleton more accurately on the upper side of the body.



(a) Pose Machines skeleton (b) RGB-D skeleton

Figure 1: Difference in skeleton detection on MSRDaily-Action3D

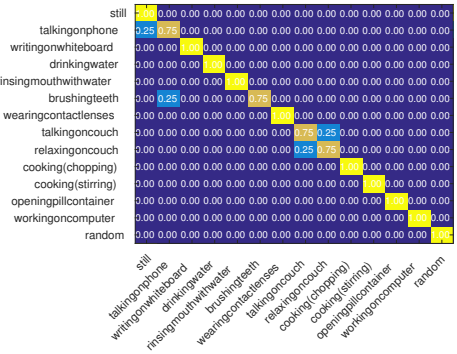


Figure 2: Confusion Matrix for CAD60 with depth based skeleton detection

For the quantitative results, we provide a brief analysis of the performance of our framework w.r.t. action recognition task based on the ground truth. As discussed earlier, in some cases depth based skeleton detection works better for action recognition task and in some case pose machines skeletons. Figure 2 and 3 represents the confusion matrix of action recognition using pose machine and depth based skeleton respectively in CAD60 dataset. We can see that depth data is working way better for actions such as drinking water and cooking(chopping). For other actions, both the extracted skeletons work similar in recognizing actions. For MSRDailyActivity3D, we can see that the pose machines is working better for actions such as playing game, laying and walking. For actions such as using laptop and cheering, both the methods show similar accuracy and for all other actions depth based skeleton works better.

4.2. Merging both the skeletons and comparison with the state-of-the-art

In this section, we report the overall accuracy of the proposed framework when applied on CAD-60 and MSRDai-

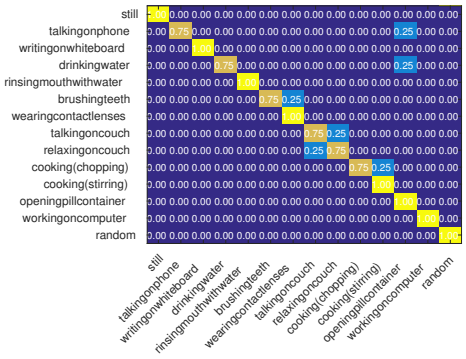


Figure 3: Confusion Matrix for CAD60 with pose machines detection

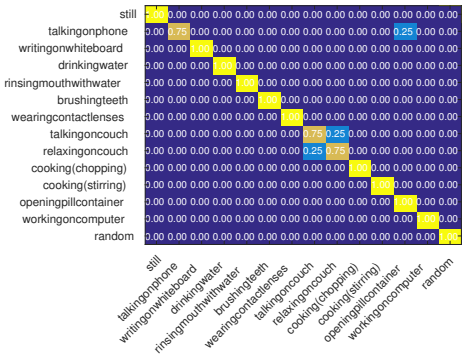


Figure 4: Confusion Matrix of fusion of depht based and pose machines skeleton detection. Actions like drinkingwater, brushingteeth and cooking(chopping) are improved.

lyAction3D dataset. The proposed approach outperforms the state-of-the-art on CAD-60 dataset as reported in Table 1 when we take any of the skeletons either from pose machines or depth data or from fusing the RGB and depth based skeleton. The accuracy for each actions on CAD-60 dataset are different when the skeleton input taken are different, and the accuracy improves on fusing the skeletons. Figure 4 is the confusion matrix of action recognition using the fusion of depth based and pose machines skeleton on CAD60. We can see that the fusion improves the recognition accuracy for activities such as drinking water, brushing teeth and cooking(chopping).

Our proposed framework works considerably well on MSRDailyAction3D dataset reported in Table 2. The accuracy is better when we take the depth based skeletons since in most of the actions, depth based skeleton detections are very good. So, the overall accuracy improves when we take the skeleton from both depth data and RGB information in this case since it exploits the advantages of both the detected skeletons.

Method	Accuracy [%]
STIP [38]	62.50
Order Sparse Coding [9]	65.30
Object Affordance [13]	71.40
HON4D [20]	72.70
Actionlet Ensemble [34]	74.70
JOULE-SVM [8]	84.10
MSLF [12]	80.36
Our Approach with Pose Machines	**91.17**
Our Approach with Kinect	**94.11**
Our Approach with kinect + Pose machine	**95.58**

Table 1: Recognition Accuracy comparison for CAD-60 dataset

Method	Accuracy [%]
NBNN [24]	70.00
HON4D [20]	80.00
STIP + skeleton [38]	80.00
SSFF [25]	81.90
DSCF [35]	83.60
Actionlet Ensemble [34]	85.80
RGGP + fusion [17]	85.60
Super Normal [36]	86.26
BHIM [10]	86.88
DCSF + joint [35]	88.20
MSLF [12]	85.95
Our Approach with Pose Machines	**80.63**
Our Approach with Kinect	**83.75**
Our Approach with kinect + Pose machine	**84.37**

Table 2: Recognition Accuracy comparison for MSRDailyActivity3D dataset

5. Conclusions

In this work we propose a framework to recognize actions from RGB-D videos. We use the skeleton detections from depth map as well as skeletons detected from RGB. We analyze the situations for different skeleton input on the action recognition task. We use pose based CNN architecture to extract CNN features from the part patches obtained from the input skeleton information and the input videos. We use χ^2 kernel from these CNN features to classify the actions. We show that both the skeleton detection methods carry complementary information as fusion improves the results. An interesting direction for future work is to model temporal evolution of frames using LSTM.

References

[1] B. Amor, J. Su, and A. Srivastava. Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories. *PAMI*, 38(1):1–13, Jan. 2016.

[2] P. Bilinski, M. Koperski, S. Bak, and F. Bremond. Representing Visual Appearance by Video Brownian Covariance Descriptor for Human Action Recognition. In *AVSS*, 2014.

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[5] G. Chron, I. Laptev, and C. Schmid. Ap-cnn: Pose-based cnn features for action recognition. In *ICCV*, 2015.

[6] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, 2014.

[7] G. Gkloxari and J. Malik. Finding action tubes. In *CVPR*, 2015.

[8] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*, 2014.

[10] Y. Kong and Y. Fu. Bilinear heterogeneous information machine for RGB-D action recognition. In *CVPR*, 2015.

[11] M. Koperski, P. Bilinski, and F. Bremond. 3D Trajectories for Action Recognition. In *ICIP*, 2014.

[12] M. Koperski and F. Bremond. Modeling spatial layout of features for real world scenario rgb-d action recognition. In *AVSS*, 2016.

[13] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *Int. J. Rob. Res.*, 32(8):951–970, July 2013.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[15] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[16] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[17] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, 2013.

[18] B. Mahasseni and S. Todorovic. Regularizing lstm with 3d human-skeleton sequences for action recognition. In *CVPR*, 2016.

[19] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, and A. Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In *ICIAR*, 2013.

[20] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013.

[21] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

[22] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele. Articulated people detection and pose estimation:reshaping the future. In *CVPR*, 2012.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[24] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *CVPRW*, 2013.

[25] A. Shahroudy, G. Wang, and T.-T. Ng. Multi-modal feature fusion for action recognition in rgb-d sequences. In *ISCCSP*, 2014.

[26] 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.

[27] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[28] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.

[29] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *ICRA*, 2012.

[30] R. Vemulapalli and R. Chellappa. Rolling rotations for recognizing human actions from 3dskeletal data. In *CVPR*, 2016.

[31] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[32] L. Wang, Y. Qiao, and X. Tang. Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors. In *CVPR*, 2015.

[33] D. Wu and L. Shao. Leveraging hierarchial parametric networks for skeletal joints based action segmentation and recognition. In *CVPR*, 2014.

[34] Y. Wu. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[35] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, 2013.

[36] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014.

[37] J. Yue-Hei, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

[38] Y. Zhu, W. Chen, and G. Guo. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing*, 32(8):453 – 464, 2014.