Deep-Temporal LSTM for Daily Living Action Recognition

Srijan Das, Michal Koperski, Francois Bremond, Gianpiero Francesca

Deep-Temporal LSTM for Daily Living Action Recognition

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 propose to improve the traditional use of RNNs by employing a many to many model for video classification. We analyze the importance of modeling spatial layout and temporal encoding for daily living action recognition. Many RGB methods focus only on short term temporal information obtained from optical flow. Skeleton based methods on the other hand show that modeling long term skeleton evolution improves action recognition accuracy. In this work, we propose a deep-temporal LSTM architecture which extends standard LSTM and allows better encoding of temporal information. In addition, we propose to fuse 3D skeleton geometry with deep static appearance. We validate our approach on public available CAD60, MSRDailyActivity3D and NTU-RGB+D, achieving competitive performance as compared to the state-of-the art.

1. Introduction

In this work we focus on solving problem of daily living action recognition. This problem facilitates many applications such as: video surveillance, patient monitoring and robotics. The problem is challenging due to complicated nature of human actions such as: pose, motion, appearance variation or occlusions. Holistic RGB approaches focus on computing hand-crafted or deep features (eg. CNN). Such methods usually model short term temporal information using optical flow. Long term temporal information is either ignored or modeled using sequence classifiers such as HMM, CRF or most recently LSTM. Introduction of low-cost depth sensors and advancements in skeleton detection algorithms lead to increased research focus on skeleton based action recognition. 3D skeleton information allows to build action recognition methods based on high level features, which are robust to view-point and appearance changes [31]. RNNs are a variant of neural nets capable of handling sequential data, applied on this problem to model the dynamics of human motion. Existing work on RNN based action recognition model the long term contextual information in the temporal domain to represent the action dynamics and so on. The traditional many-to-one LSTMs used for video classification, takes decision based on the feature obtained at the last time stamp, failing to incorporate the prediction of video class over time in the loss backpropagated. This disables the LSTM to model long-term motion for action classification.

In this paper we propose a many-to-many model for video classification namely, deep-temporal LSTM in which the features extracted and the loss backpropagated is computed over time. We show that this latent representation of video from LSTM leads to better temporal encoding. We also propose to fuse RGB information and skeleton information. We claim that it is especially important in daily-living action recognition, where many actions have similar motion and pose footprint (eg. drinking and taking pills), thus it is very important to model appearance of objects involved in the action accurately. This paper shows (1) that LSTM can model temporal evolution of activities when the loss is computed over every temporal frames rather than relying on the last time step, (2) deep appearance and motion based features selected from appropriate image region for action classification. In our work, we propose to use late fusion of skeleton based LSTM classifier with appearance based CNN classifier. Both classifiers work independently and we fuse their classification scores to obtain final classification label. In this way, we take advantage of LSTM classifier which is able to capture long term temporal evolution of pose and CNN classifier which focuses on static appearance features.

We validate our work on 3 public daily-activity datasets: CAD60, MSRDailyActivity3D and NTU-RGB+D. Our experiments show that we obtain competitive results on all the datasets as compared to the state-of-the-art.

2. Related Work

Previous work on human action recognition was centric with the use of dense trajectories [26] combined with Fisher Vector (FV) aggregation. The introduction of low cost kinect sensors have made it possible to detect the skele-

ton poses of human body easily which can be exploited to recognize actions as in [27] or 3D trajectories [11].

The emergence of Deep Learning in computer vision has improved the results in terms of accuracy of action recognition as they show some promising results [14]. One of the key point to use deep learning for action recognition is that, it not only focuses on extracting the deep features from CNNs but for instance considers the temporal evolution of these features using the Recurrent Neural Networks (RNNs).

In [21], authors have used two stream networks for action recognition. One for appearance features from the RGB frames and one for flow based features from optical flow. They propose to fuse these features in the last convolutional layer rather than fusing them in the softmax layer which further improves their accuracy of the framework. Cheron *et al*. [4] have used different parts of the skeleton to extract the CNN features from each of them. These features are aggregated with max-min pooling to classify the actions. The authors claim to use the temporal information by taking the difference of these CNN features followed by the max-min aggregation. But this aggregation ignores the temporal modeling of the spatial features. The recent advancement in this field led to the use of 3D CNNs in C3D and I3D [3] for action classification reporting high accuracy. But these networks containing huge number of parameters are difficult to train on small datasets. They also do not address the long term dependencies of the actions.

The availability of informative three dimensional human skeleton data led to the use of RNNs which are capable of modeling the dynamics of human motion. Shahroudy *et al*. [19] proposed the use of stacked LSTMs namely, Deep-LSTM and also, p-LSTM where separate memory cells are dedicated for each body part of an individual. Another variant of LSTM is proposed in [15], where the authors introduce a new gating mechanism within LSTM to learn the reliability of the sequential data. Some studies also reports the use of different types of feature on RNNs as in [7, 31]. Authors in [7] modeled the spatio-temporal relationship by feeding the LSTM with CNN features from fc-6 of VGG network. But this strategy is valid for datasets with very dynamic actions and not applicable on similar based motion characterized actions. Authors in [31] have represented 3D skeletons using distance based features and feed them into 3 layer LSTM. They have also proposed joint line distance to be the most discriminative features for action classification. From the above discussion, it is clear that action recognition tasks focus on improving appearance and motion based features and temporal features through RNN modeling separately. But both spatial layout and temporal encoding is important to model daily living activities. This is because of the presence of low motion actions like *typing keyboard, relaxing on couch* and so on where spatial layout is important, and similar actions like *drinking water, brushing hair* and so on where temporal encoding is important. Thus we propose to combine features from convolutional network and recurrent network to encode appearance-motion and temporal information together in a model.

In this paper, we use body translated joint coordinates from the depth information to find the discriminative dynamics of the actions using 3-layer LSTM followed by a SVM classification for the temporal stream. We show that our LSTM based features can model the actions temporally better that the existing LSTM architectures with similar input sequences. For deep spatial features, we extended [4] to produce CNN features considering both the flow and appearance features from different image regions. We employ a feature selection mechanism to use the most informative image-region over the training data for classifying actions. Since temporal information modeling along with encoding spatial layout is an important dimension in action recognition, so we focus on using the fusion of deep spatial features along with temporal information to recognize actions using a late fusion to learn semantic concepts from unimodal features.

3. Proposed Method

3.1. Deep-Temporal LSTM

LSTMs being a special kind of RNNs can model the time information as in [31]. LSTM mitigates the vanishing gradient problem faced by RNNs by utilizing the gating mechanism over an internal memory cell. The gates enable the LSTM to determine what new information is going to be stored in the next cell state and what old information should be discarded. Such recurrent model receives inputs sequentially and models the information from the seen sequence with a componential hidden state h_t:

$$h_t = f_h(h_{t-1}, v_t; \theta_h) \tag{1}$$

where LSTM is our recurrent function f_h with parameters θ_h. We omit the gates from the equations so as to keep the notation simple. The input to the recurrent model is the context vector v_t which is described below.

The main focus of the existing methods includes using the RNNs to discover the dynamics and patterns for 3D human action recognition. The sequential nature of the 3D skeleton joints over the time makes the RNN learn the discriminative dynamics of the body. In this work, we use transformed body pose information on a 3-layer stacked LSTM so as to model the temporal information as shown in fig. 1. The main reason for stacking LSTM is to allow for greater model complexity, to perform hierarchal processing on large temporal tasks and naturally capture the structure of sequences. A pre-processing step is performed to normalize the 3D skeleton in camera coordinate

system as in [19]. The 3D skeleton joint is translated to the $hip - center$ followed by a rotation of the X axis parallel to the 3D vector from "right hip" to the "left hip", and Y axis towards the 3D vector from "spine base" to "spine". At the end, we scale all the 3D joints based on the distance between "spine base" and "spine" joints. Thus the transformed 3D skeleton v_t at time frame t which is represented as $[x_{r,t}, y_{r,t}, z_{r,t}]$ for $r \in$ joints (J) and (x, y, z) being the spatial location of r^{th} joint is input to the LSTM at time stamp t. We normalize the time steps in videos by padding with zeros. This is done to keep fixed time steps in LSTM to process a video sample.

Traditionally, authors in [7, 31, 19] solve action recognition problem as a many to one sequence classification problem. They compute the loss at the last time step of the video which is backpropagated through time. In this work, we compare the LSTM cell output with the true label of the video at each timestep. In this way we get time-step sources to correct errors in the model (via backpropagation) rather than just for each video (giving rise to the term *deep-temporal*). Thus the cost function of the LSTM for videos is computed by averaging the loss at each frame as follows

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} y_i \log(p_{it}) \tag{2}$$

where L is categorical cross-entropy computed for N video samples in a batch over T frames, y is the sample label and $p_{it} \in (0, 1) : \Sigma_t = 1 \forall i, t$ is the prediction for a video. This loss L is back-propagated through time. Here, the LSTM treats each temporal sequence independently as a sample, whose prediction is again determined by the current and previous gate states. This method provides better performance compared to the minimization of the loss at the last time step only due to better feedback backpropagated through time. So, we extract the latent vectors from every time step of the last layered LSTM using it as a feature extractor.

We train deep-temporal LSTM with parameters θ_h on the input sequence $V = \{v_t\}$ with loss L resulting in a hidden state representation $h = \{h_t\}$. Each element in h_t is again a latent vector represented as $h_t = \{h_{j,t}\}$, where j is the index over the hidden state dimension. This latent vector constituting $h'_i = h_r$ represent the action dynamics at time instant $r \in \{1 \cdots T\}$ for i^{th} sample, qualifying it as a representative vector over time. This latent vector h represents a better and more complex representation of the long-term dependencies among the input 3D sequential data. This temporal latent vector h is input to a linear SVM classifier for action classification. The 3-D matrix $H_n = \{h'_1, h'_2, \cdots h'_n\}$ for n training sample is input to the SVM to learn the mapping $\mathbb{X} \rightarrow \mathbb{Y}$, where $h'_n \in \mathbb{X}$ and $y \in \mathbb{Y}$ is a class label. The features $\mathbb{X}$, extracted from trained LSTM are used to learn a classifier:

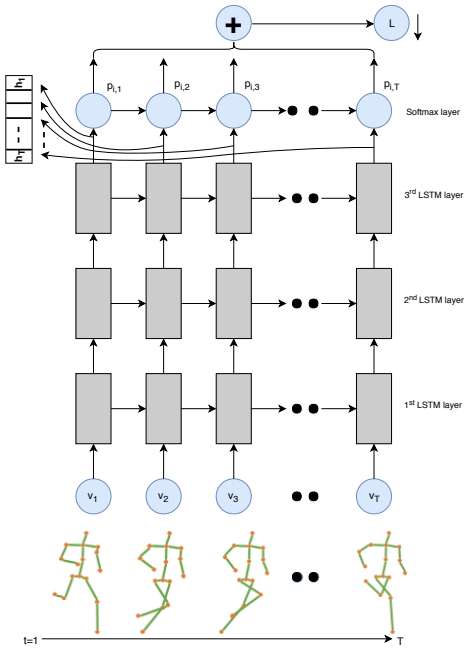

Figure 1. Three-layer stacked LSTM with $t = T$ time steps. The skeleton joint coordinates v_t are input at each time step. L is the loss computed over time and h is the latent vector from last layer LSTM which is input to the linear SVM classifier.

$y = f_{SVM}(h'_n, \alpha)$, where α, the parameters of the function f_{SVM}.

3.2. Pose based CNNs

In [4], the authors have used the concept of two streams for recognizing actions on the different parts of the subject extracted from their skeleton joint information. This inspires us to use the deep features from different body regions of the subject to represent their appearance and motion features. The main objective behind using these features is to model the static appearances along with encoding the object information carried while performing the actions. We extend [4] by invoking deeper networks for feature extraction and employing a feature selection technique to select the best image region involved in the action dataset.

In our pose based action network, CNN features from the left hand, right hand, upper body, full body and full images from each frame (cropped using their 2D joint information) are extracted to represent each body region for the classification task as illustrated in fig. 2. Our experimental studies show that this body region representation leads to a lot of redundancy. Sometimes, wrong patches extracted due to side view actions which mislead the classifier to select a wrong action. Thus we propose a technique to select the best representation of the appearance feature by focusing on the body region with the most discriminative information. The patch

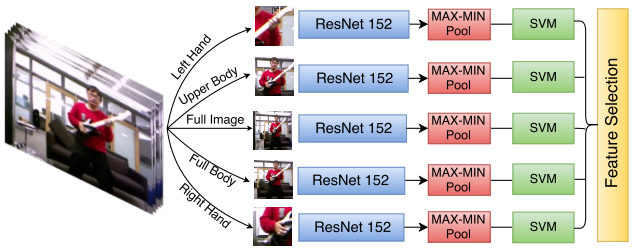

Figure 2. Each image frames are divided into five parts from their pose information which are input to ResNet-152 followed by max-min pooling. The classification from the SVM determines the part to be selected. The figure is only depicted for RGB stream and the Flow stream follows similar trend.

representation for a given image-region i is convolutional network f_g with parameters θ_g, taking as input a crop taken from image I_t at the position of the part patch i:

$$z_{t,i} = f_g(crop(I_t, patch_i); \theta_g) \qquad i = \{1, ..., 5\} \quad (3)$$

We use pre-trained Resnet-152 for f_g to extract the deep features from the last fully connected layer which yields 2048 values described as our frame descriptors $z_{t,i}$. These frame descriptors are aggregated over time using max and min pooling so as to focus the most salient values on the feature maps representing the video descriptors z_i.
The feature selection is done by feeding these CNN features z_i to linear SVM for classification separately for each patch i. These SVM produce classification scores on cross validation set separately for each patch i. We select the patch i of image-region with the best classification score on validation set. This allows us to select the best body region for characterizing the appearance feature. As per our observation, these selected appearance features not only represent the best static appearances but also have the best combinational power with the motion based information. Motion based information signifies the optical flow information which is computed similarly using the feature selection mechanism.

3.3. Fusing Geometric and Appearance-Motion Features for Action Modeling

In this work, we propose to combine the discriminative power of 3D skeleton sequences with RGB based appearance and motion. Authors in [29, 32, 21, 8] attempts to fuse the appearance from RGB information, poses and motion. Here, the novelty lies in an attempt to fuse features from convolutional network and recurrent network in order to encode spatial and temporal features together.
The body based CNN features model the salient features in the global video. But these features are not discriminative enough to model the difference between actions with less intra-class variance. On the other hand, the features from LSTM models the temporal evolution of the salient features over the entire video. It captures the geometric evolution of the activity performed by a subject. So, the idea in this paper is to fuse the appearance based CNN features with the temporal evolution of the body translated skeleton joints. This is done by fusing the classification scores from SVM trained on deep appearance features and SVM trained on latent LSTM features.

4. Experiments

4.1. Dataset Description

For evaluating our framework, we use **CAD-60** [23] containing 60 RGB-D videos with 14 actions performed by 4 subjects, **MSRDailyActivity3D** [27] containing 320 RGB-D videos with 16 actions performed by 16 subjects and **NTU-RGB+D** [19] containing 56880 RGB-D videos with 60 actions performed by 40 subjects.
We evaluate CAD-60 and MSRDailyActivity3D by setting up a cross subject training and testing validation set up. For NTU-RGB+D, we follow the training/testing protocol mentioned in [19]. Our view transformation on 3D skeleton is performed to handle the side view actions performed by the subject on fixed camera, we are not focusing on Cross-View problem. Hence, we have not evaluated cross-view accuracy on NTURGB+D dataset.

4.2. Implementation Details

We build our LSTM framework on the platform of keras toolbox [5] with TensorFlow [1]. The concept of Dropout [22] is used with a probability of 0.5 to eliminate the problem of overfitting. The concept of Gradient clipping [24] is used by restricting the norm of the gradient to not to exceed 1 in order to avoid the gradient explosion problem. Adam optimizer [10] initialized with learning rate 0.005 is used to train both networks.

4.3. Ablation Study

In this section, we present the performance and analysis of the two cues used independently for action classification. Table 1 shows the performance of different variants of LSTM with skeleton joints as input, used in the state-of-the-art. The performance of our proposed deep temporal LSTM feature extractor followed by linear SVM classifier outperforms the other LSTM variants. This is because of considering the predictions at each time step of the video sequence and classifying the latent representant of time for a video sample, which improves the temporal modeling of the classifier. This also opens a direction of using our approach with different input features and even with LSTM variants proposed in the state-of-the-art for LSTMs.
Table 2 shows the effectiveness of our feature selection

mechanism on different image regions for modeling the actions. It confirms the presence of irrelevant features (due to the combination of all body regions) which deviates the classifier decision boundary from the ideal one. This is also justified by the fact that some image regions may not be extracted correctly in some sequences and some may not have any significance in modeling the action. For instance, extracting the features from right hand of a person drinking water with left hand is of no significance.

Method	CAD-60	MSRDaily Activity3D	NTU-RGB+D
Traditional LSTM	64.65	80.90	60.69
Deep LSTM [19]	-	-	60.7
P-LSTM [19]	-	-	62.93
ST-LSTM (Joint-chain) [15]	-	-	61.7
Deep Temporal LSTM	**67.64**	**91.56**	**64.49**

Table 1. Comparison of different approaches with body translated skeleton coordinates on CAD-60, MSRDailyActivity3D and NTU-RGB+D. The numbers here denote the accuracy [%].

Method	CAD-60	MSRDaily Activity3D	NTU-RGB+D
P-CNN [4]	95.59	87.81	48.71
FS**(P-CNN)**	**97.06**	**89.06**	**58.69**

Table 2. Effectiveness of pose based CNN features with feature selection mechanism on CAD60, MSRDailyActivity3D and NTU-RGB+D. The numbers here denote the accuracy [%] and FS corresponds to the feature selection mechanism.

4.4. Comparison with the state-of-the-art

Table 3 presents the state-of-the-art comparison of our proposed fusion of depth based LSTM features and pose based CNN features to classify actions. The complementary nature of the LSTM and CNN based networks are evident from the boosted performance for MSRDailyActivity3D and NTU-RGB+D on fusion. The presence of static actions like *cooking, talking on phone, relaxing on couch* and so on in CAD-60 do not enable the LSTM to recognize the dynamicity of the actions. This explains the gainless accuracy reported for CAD-60 on fusing the features. We outperform state-of-the-art results on CAD-60 and MSRDailyActivity3D. [32] and [2] outperforms our proposed method using multi-stream 3D convolutions and attention mechanism respectively. Such mechanisms are hard to train and may not have consistent performance on smaller dataset. We observe that in NTU-RGB+D, short term motion is important which can be modeled using dense trajectory features [25](IDT-FV). Thus, we combine the IDT-FV features using a late fusion of individual classification score significantly boosting the performance over using individual features only and resulting in state-of-the-art performance on NTU-RGB+D.

Method	Accuracy [%]
Object Affordance [13]	71.40
HON4D [18]	72.70
Actionlet Ensemble [27]	74.70
MSLF [12]	80.36
JOULE-SVM [9]	84.10
P-CNN + kinect + Pose machines [6]	95.58
Proposed Method	**97.06**
P-CNN + kinect + Pose machine [6]	84.37
Actionlet Ensemble [27]	85.80
MSLF [12]	85.95
DCSF + joint [28]	88.20
JOULE-SVM [9]	95.00
DSSCA-SSLM [20]	97.50
Proposed Method	**98.44**
JOULE-SVM [9]	60.23
Geometric features [31]	70.26
Enhanced skeleton visualization [16]	75.97
Ensemble TS-LSTM [17]	74.60
DSSCA-SSLM [20]	74.86
VA-LSTM [30]	79.4
Chained Multistream Network [32]	80.8
STA-hands [2]	82.5
Proposed Method	**74.75**
Proposed Method + IDT-FV	**84.22**

Table 3. Recognition Accuracy comparison for CAD-60 (1^{st} section), MSRDailyActivity3D (2^{nd} section) and NTU-RGB+D (3^{rd} section) dataset. Proposed Method signifies Deep-Temporal LSTM + $FeatureSelection$ (P-CNN).

5. Conclusion

In this work, we propose a deep-temporal LSTM which models better temporal sequences as compared to the state-of-the-art architectures on the same input features and extended the pose based CNN action network by employing a feature selection mechanism. We also present the idea of fusing the pros of 3D skeleton based geometric features with appearance and motion based deep features to classify daily living activities.

A future direction lies in exploring different efficient features and variants of gating mechanism of LSTMs with our proposed approach. In the appearance-motion stream, the feature selection mechanism to select the appropriate image region is globally decided over the dataset. An attempt to select the appropriate feature for each sample and employing such a mechanism in the network itself is a direction to be explored.

References

[1] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] F. Baradel, C. Wolf, and J. Mille. Human action recognition: Pose-based attention draws focus to hands. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 604–613, Oct 2017.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[4] G. Cheron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV*, 2015.

[5] F. Chollet et al. Keras, 2015.

[6] S. Das, M. Koperski, F. Bremond, and G. Francesca. Action recognition based on a mixture of rgb and depth based skeleton. In *AVSS*, 2017.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 1933–1941. IEEE, 2016.

[9] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2186–2200, Nov 2017.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[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] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 816–833, Cham, 2016. Springer International Publishing.

[16] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[17] B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3054–3062, 2016.

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

[19] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] A. Shahroudy, T. T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[21] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

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

[24] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

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

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

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

[29] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond Short Snippets: Deep Networks for Video Classification. *ArXiv e-prints*, Mar. 2015.

[30] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157, March 2017.

[32] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2923–2932. IEEE, 2017.