Spatio-Temporal Grids for Daily Living Action Recognition

Srijan Das, Kaustubh Sakhalkar, Michal F Koperski, Francois Bremond

Spatio-Temporal Grids for Daily Living Action Recognition

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

Abstract

This paper address the recognition of short-term daily living actions from RGB-D videos. The existing approaches ignore spatio-temporal contextual relationships in the action videos. So, we propose to explore the spatial layout to better model the appearance. In order to encode temporal information, we divide the action sequence into temporal grids. We address the challenge of subject invariance by applying clustering on the appearance features and velocity features to partition the temporal grids. We validate our approach on four public datasets. The results show that our method is competitive with the state-of-the-art.

1. Introduction

Action recognition has been a popular subject in the field of computer vision because of its practical applications like video description, video surveillance, building smart homes, patient monitoring and so on. In this work, we focus on recognizing daily living activities which includes various challenges. Some of them includes intra class variation, for e.g. different subjects can perform the same action in different posture and time; occlusion; selecting the correct features to model an action and modeling spatio-temporal information, for e.g. if a person takes off his shoes and wear shoes, then spatial relationship (position of the shoes $w.r.t$ body) with respect to time is important to discriminate such actions.

Spatial granularity to focus on the relevant region of the image w.r.t time is an important aspect in Daily Living Action Recognition. Recent work on spatial attention [2, 31], attempted to focus on the important image patches which is learned over time. But such attention mechanisms are hard to train and mis-classify while providing wrong attention. Moreover, providing such spatio-temporal attention for daily living actions with similar environment, similar motion and color statistics is hard. So, we propose to use spatio-temporal grids to best describe the relevant image regions w.r.t time for RGB-D action recognition.

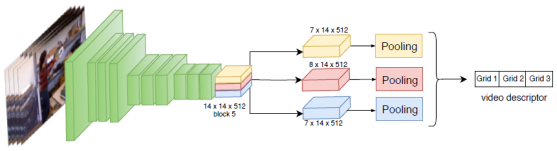

Figure 1. The Action Recognition Framework extracting CNN features from each body region followed by max-min pooling from each spatial grids.

Here, we address short-term actions. We use different cues to leverage the different modalities to model the actions. Our main contributions are

- introducing spatial grids on the body region of the subject to explore spatial granularity for recognizing actions.
- introducing temporal grids where the grid size is decided by the clustering results of appearances and velocity of the person performing action. This ensures division of homogeneous temporal grids.
- We validate our approach on 4 publicly available dataset achieving competitive performance on each of them.

We exploit the semantics of the person performing action by taking meaningful spatial grids keeping in mind of the actions generally occur in daily living actions. The purpose of this work is to show the discriminative power of using grids to model the appearance. The proposed framework with spatio-temporal grids has been depicted in fig. 1. In our framework, we exploit the different features from different networks by a classifier level fusion to show the combinational power of our proposed appearance based video descriptor.

2. Related Work

Earlier action recognition has been dominated by the use of motion based features like Dense Trajectories [35] and

its advancement IDT [36] followed by fisher vector encoding [28]. The emergence of deep learning networks changed the scenario. In [15], Karpathy et al. extended the image classification framework using VGG network [5] to a video classification framework. The frame-level features from the fully connected layer of VGG are aggregated over time using pooling operation. But such aggregation destroys the temporal information which is handled in [10]. In [10] the authors classify the actions using LSTMs fed upon frame-level CNN features. The elimination of simple video aggregation function by the use of LSTMs takes care of the temporal information. The authors in [32, 11] proposed to use appearance and motion information from optical flow images to recognize the actions. Most of the recent techniques in this field follow this similar strategy to take into account both appearance and motion [6, 4, 9].

The recent availability of easy and large scale depth data motivated authors to use different structures of LSTMs to recognize actions [39, 29]. In [39], the authors have presented a clear statistics of using different geometric features on RNNs to recognize actions. But most of these techniques fail to discriminate the daily living activities because of its challenges. Authors in [6, 8] extract spatial features from different parts of the body by first cropping out the parts from RGB frames using skeleton joint information. But, cropping different parts of the body and resizing them into 224×224 so as to feed the network is not a robust technique. This affects the image resolution and such cropping techniques may fail due to occlusion or noisy skeleton captured. Thus we propose to use the person localized technique to focus on the different parts of the body by employing grids in the convolutional feature maps to model the appearance information. Inspired from [19], we use $m \times n$ grids in the last convolutional feature map of VGG16 [5] to model the appearance information from different body region to enhance spatial granularity.

Recently, the authors in [40] have used different types of features extracted from C3D [34] and combined them with hidden Markov chaining mechanism. This inspires us to use score level fusion of different nature of features to model actions. The current advancement of 3D- CNN lead to a steep hike in recognition score of actions using [4]. But these video classification frameworks are not robust and hard to finetune on smaller datasets. And in practice, real-time data are often smaller in size and it incurs expensive costing, annotating them.

So, we propose a robust framework using geometric, appearance and motion based features to model daily living activities. We focus on modeling the static actions by the use of spatio-temporal grids from the appearance information.

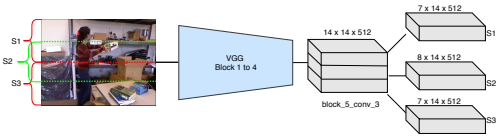

Figure 2. The action sequence on the subject bounding box is divided into spatial grids (from `block_5_conv_3`) of VGG-16. The grids focus on the upper (S1), middle (S2) and lower part (S3) of the body region of the subject performing action respectively.

3. Proposed Method

3.1. Spatio-temporal grids on CNN feature maps

Similar motion and less inter class variance in daily living activities are challenges to their recognition. We address these challenges by focusing on different parts of the body region.

Let us consider a video composed of T sequences represented by $\{X_1, X_2, X_3, ...X_T\}$. The features from the convolutional feature maps, retains spatial information of the frames. A feature map at timeframe t has a dimension of $N \times N \times D$, where N is the number of regions in an image and D is the feature vector for each region. These feature maps are divided into $m \times n$ spatial grids as in fig. 2. In our case, the grid size is set $m = 3$ and $n = 1$ since most of the daily living actions are performed either on top, middle or bottom spatial location of the human localized patches. This also ensures to encode the semantics of the action categories present in daily living activities. The dimension of the feature map reduces as we go deeper into the network, so we use overlapping grids to represent the actions within a grid as shown in fig. 2.

Pooling the frame level convolutional features over time destroys the temporal sequence of the action. Thus we propose to employ temporal grids within a video to embed the spatio-temporal relationship which results in 3D representation of an action. The temporal grids are equally spaced segments over the video sequence. In real world scenario, different subjects perform the actions with different velocity depending on their nature. We address such intra-class challenge by proposing dynamic temporal grids where the number of sequences in a temporal segment is not fixed and depends on the motion and appearance of the subject. This is done by employing $k - means$ clustering on the appearance and velocity of the human poses as shown in fig. 3.

Actions vary greatly in their appearance and motion characteristics over time. By explicitly modeling the temporal evolution of an action, we can take advantage of this inherent temporal structure especially in short-term actions. Action sequence in the middle of an action tend to have different motion and appearance than the starting and

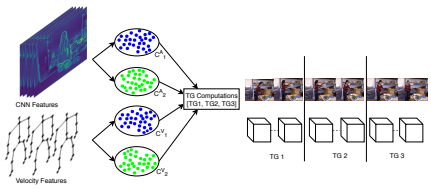

Figure 3. The *conv* features and three dimensional data are used to cluster the frames to be assigned to temporal segment TG_1, TG_2, TG_3. C_r^A and C_r^V are the clusters from appearance and velocity features respectively (for $r = 1, 2$).

ending frames as they perform complex body movements. Consider a video sequence with T sequences and x_{n1} and x_{n2} dimensional appearance and velocity features respectively. The appearance features are extracted from last convolutional layer (`block_5_conv_3`) and the velocity features are $[V_x, V_y, V_z]$, where $V_{r_j} = \frac{|r_j(t+\Delta t) - r_j(t=1)|}{\Delta t}$, for $j = \{1..\text{number of joints}\}$, t being the time sequence and $r \in \{x, y, z\}$. We employ k-means clustering on the $T \times x_{n1}$ appearance and $T \times x_{n2}$ velocity features independently with two clusters. Let C_1^A and C_2^A be the clusters from appearance features and C_1^V and C_2^V be the clusters from velocity features. One of the cluster from each feature consist of the starting and ending frames of the video sequence due to their similar motion and appearance in the first and last phase of the action. Assuming that C_1^A and C_1^V are the clusters with the staring and ending frames we divide the temporal grids into TG_1, TG_2, TG_3 by the following:

$$TG_1 \in \{1... \max(e_{TG_{C_1^A}}, e_{TG_{C_1^V}})\} \quad (1)$$

$$TG_2 \in \{(N(TG_1) + 1)...(S(TG_3) - 1)\} \quad (2)$$

$$TG_3 \in \{\max(f_{TG_{C_1^A}}, f_{TG_{C_1^V}})...T\} \quad (3)$$

where $e_{TG_{C_1^A}}$ is the latest sequence in $C_1^A \leq$ first sequence in C_2^A, $e_{TG_{C_1^V}}$ is the latest sequence in $C_1^V \leq$ first sequence in C_2^V, $f_{TG_{C_1^A}}$ is (1+ latest sequence in C_2^A) and $f_{TG_{C_1^V}}$ is (1+ latest sequence in C_2^V). The max operator to compute the sequences in the temporal grid is to ensure larger number of sequences in first and middle grids. Temporal grids TG_1 and TG_2 are defined first based on the clusters with 1 and then the pending sequences (computed from ending frames of TG_1 using N(.) and starting frames of TG_3 using S(.)) are assigned into temporal grid TS_2.

Thus the video descriptor V' of a spatial grid is defined by a concatenation of the static V_s and dynamic V_d video descriptor over the time frames in the different temporal segments of the video. Static descriptor ($V_s^{TG_r}$) for temporal grid TG_r is defined as

$$V_s^{TG_r} = [\max\{f_1, f_2, ..f_{TG_r}\}, \min\{f_1, f_2, ..f_{TG_r}\}] \quad (4)$$

where f_t is the $N' \times N \times D$ feature map ($N' = 7, 8$ depending on the grid) at time sequence t and dynamic descriptor ($V_d^{TG_r}$) for temporal grid TG_r is defined as

$$V_d^{TG_r} = [\max\{\Delta f_1, \Delta f_2, ..\Delta f_{TG_r}\}, \min\{\Delta f_1, \Delta_2, ..\Delta f_{TG_r}\}] \quad (5)$$

where $\Delta f_t = f(t + \Delta t) - f(t)$, for Δt=4. Thus the video descriptor is defined as

$$V' = [V_s^{TG_1}, V_d^{TG_1}, V_s^{TG_2}, V_d^{TG_2}, V_s^{TG_3}, V_d^{TG_3}] \quad (6)$$

These video descriptors from each grid are concatenated with $l_2 normalization$ to form the global video descriptor. This video descriptor is input to the SVM classifier to classify the actions based on their local appearances.

3.2. Motion features from Dense Trajectories

Most of the action recognition frameworks use motion information as an additional feature to model the actions [32, 11, 6]. This is done because motion is an important information for modeling the actions. So, in our framework, we use dense trajectories [35] followed by fisher vector encoding to model the motion information. In order to remove noise from the background which is similar and redundant in daily living activities, we focus on the subject performing actions spatially while computing dense trajectories. The pose information from the depth information allows us to focus on the subject spatially and encode the HoG, HoF, MBH features along the trajectories. In order to have a fixed size video descriptor for input to the linear SVM classifier, we use fisher vector encoding on these trajectory features as in [18].

3.3. Geometric features from LSTM

The availability of large scale 3-D data is leveraged for classifying actions with large dynamics like walking, standing up and so on. We use a 2-layer LSTM [12] which take 3D skeleton sequences as input. For daily living actions having similar motion, 3D spatial locations can discriminate the actions as in [9]. The latent representation of the skeleton sequences are extracted from the trained LSTM. The latent vector is a concatenated feature vector of the output hidden states of the LSTM from each time step.

3.4. Score Level Fusion of different features

Action Recognition problems are widely dispersed due to various inter-class variation. It is important to understand the different features required to model a particular action. Actions with low motion like *typing keyboard* and *reading* can be modeled using appearance features, actions with motion like *brushing teeth* and *shaking hands* can be modeled using motion based features and actions with high variance

or dynamics like *standing up* and *walking* can be modeled using geometric dynamics of the human poses.
So, we propose to use a score level fusion of geometric, appearance and motion based features to model the actions. A two level stacked SVM is employed to fuse all the features. The classifiers in the first level, classify the input features independently and the fusion is performed by a linear SVM classifier in the second level. The classifier in the second level takes the concatenated scores and weighted sum of the scores as input, to classify the actions.

4. Experiments

4.1. Dataset Description

To validate our proposed framework, we have used four public dataset based on daily living activities.
CAD-60 [33] - contains 60 RGB-D videos with 14 actions performed by 4 subjects. Small number of training samples in this dataset makes it a challenging dataset, specially for finetuning deep networks.
CAD-120 [33] - contains 120 RGB-D videos with 10 high level activities performed by 4 subjects. High inter class variation like stacking, unstacking objects and so on makes recognition of action harder in this dataset.
MSRDailyActivity3D 8a - contains 320 RGB-D videos with 16 actions performed by 16 subjects. Each action is repeated twice in sitting and standing position.
NTURGB+D [29] - contains 56880 RGB-D videos with 40 subjects performing 60 different actions. Samples are captured from 17 setups of camera and for each setup three cameras were located at the same height but from different horizontal angles: $-45\,\mathrm{deg}$, $0\,\mathrm{deg}$ and $+45\,\mathrm{deg}$. Each action is performed twice, once facing the left camera and once towards the right camera. The standard evaluations on these datasets include Cross-Subject evaluation where the training and testing split is made either by leave-one-person out schema or split mentioned in the dataset (as in NTURGB+D). We are not focusing on Cross-View problem and so, we have not evaluated cross-view accuracy on NTURGB+D dataset.

4.2. Implementation Details

For LSTM, we build a 2-layered stacked LSTM on the platform of keras toolbox [7] with TensorFlow [1]. Parameters like Dropout, gradient clipping and number of neurons in each LSTM lsyer for each dataset are used as in [9]. Adam optimizer [16] initialized with learning rate 0.005 is used to train the LSTM. The pose information is extracted from the pixel coordinates detected by middleware (like kinect sensor). These pose information are the concatenated pixel coordinates of the body joints of the subject performing the action.

4.3. Analysis of using spatio-temporal grids

The grids spatially focus on the different parts of the body region of the subject performing action. In table 1, we compare the performance of each grids and their combination with only using full body and parts based CNN features(P-CNN [6]). The features for full body and parts based CNN features are extracted from the last fully connected layer of VGG-16 whereas the features for spatial grids are extracted from the last convolutional layer. The statistics in table 1 clearly shows that the use of overlapping grids on the full body patches of the subject performing the action clearly outperforms the features extracted globally from the parts of the subject. This shows that spatial granularity is required to recognize certain action especially, having lower action dynamics.
Fig. 4 represents a bar plot showing how different spatio-temporal grids contribute to the actions indicating a contextual relationship. For actions like *relaxing on couch* and *using laptop*, the classifier trained on the middle overlapping grid is able to recognize such actions, mostly because of better object encoding and static appearance representation. For actions like *talking on couch* and *eating*, where the action occurs on the upper region of the subject, the grid specialized for upper spatial location is able to represent the action dynamics better than the other grids specialized for other body region. Based on our visualization, actions like *laying down on sofa*, *stacking and unstacking objects*, the motion of the action occurs in the lower region of the cropped body region (due to occlusion). Hence, it is recognized by the classifier trained on the grid focusing on the lower region of the person centric bounding box. Similarly, for actions like *using vacuum cleaner* and *still (doing nothing)*, either all the body region or none is taking part in performing the action, so it is captured by all the grids (not the bottom grid for *still* action).
Fig. 5 shows a histogram of average recognition accuracy on CAD-60, CAD-120 and MSRDailyActivity3D with variation of the action present in these datasets. With availability of 3D skeleton sequences of these actions, we computed a metric to measure how dynamic action is, (VAR) which is the average joint displacement per frame and is given by

$$r_{VAR} = \frac{\Sigma_{i=1}^{n_f} \frac{\Sigma_{j=1}^{n_{joints}} |r_j - r_{mean}|}{n_{joints}}}{n_f}, r \in \{X, Y, Z\}; \qquad VAR = \sqrt[2]{X_{VAR}^2 + Y_{VAR}^2 + Z_{VAR}^2} \tag{7}$$

Figure 5 clearly shows that the static actions with lower action dynamics (where the sample density is more) are well recognized by the fusion of all the spatio-temporal grids as compared to the CNN features from full body and all the parts of the subject body. Thus representing the actions with less motion which is a challenge for daily living activities is

Table 1. Performance Comparison of the spatio-temporal grids with full body and parts based CNN features. Here, Grid1, Grid2 and Grid3 signifies the top, overlapped middle and bottom spatio-temporal grids respectively.

	Full Body	P-CNN	Grid1	Grid2	Grid3	Grid (All)
CAD-60	73.53	85.29	83.82	83.82	66.17	86.76
CAD-120	56.45	63.70	65.32	78.22	78.22	78.22
MSRDailyActivity3D	71.875	78.12	76.87	76.56	67.5	78.75
NTU-RGB+D	44.56	48.71	49.45	44.43	31.54	65.25

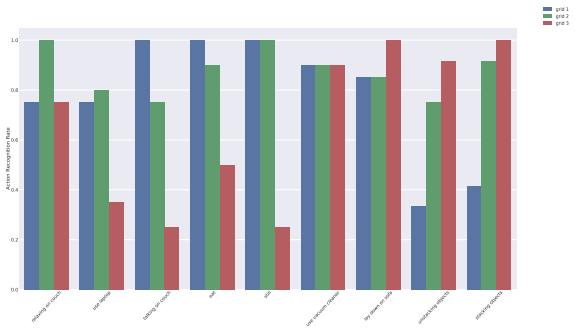

Figure 4. Performance comparison of the grids for selected actions. The top, overlapped middle and bottom grid are represented by blue, green and red color respectively.

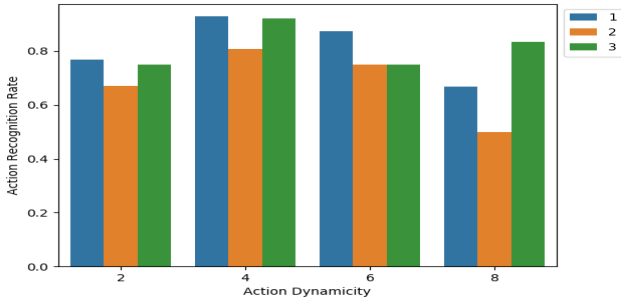

Figure 5. Plot of recognition accuracy vs variation of actions on CAD60, CAD120 and MSRDailyActivity3D for spatio-temporal grids(1), full body(2) and P-CNN features(3).

achieved by using the spatio-temporal grids.

The concept of having dynamic sequences in temporal grids is to maintain the consistency of the part of action assigned to a temporal grid for all the subject. This is to eliminate the challenge of a subject performing the action faster or slower than general. On CAD120 [33], we have validated our approach of assigning dynamic number of frames to each temporal grids. Firstly, the use of temporal segments instead of applying max-min operation on the whole video sequence boosted the performance from 73.38% to 76.61% which is because now the information from each part of the action sequence is provided to the classifier instead of destroying the whole temporal information by employing the pooling operation on the whole sequence. Secondly, we also observed that the recognition performance on a particular split is boosted from 64.51% to 70.96% on using dynamic number of sequences for each temporal grids instead of dividing the temporal grids into three equal halves. Such boosting in recognition can be observed for subjects performing the action with a trend deviated from the general trend.

4.4. Comparison with state-of-the-art

For datasets like CAD-60 and CAD-120 where most of the actions have less variation with respect to posture, appearance features are important. So, in these datasets spatio-temporal grids have contributed the most in achieving a competitive global recognition rate as compared to the state-of-the-art (results are in table 2 and 3). The performance of [8] similar to our performance on CAD-60, is not consistent on other datasets. Method [22] though outperforms our framework on CAD-120, is time expensive and the brute force enumeration over all settings of the latent variables cause extra computational cost. In MSRDailyActivity3D and NTURGB+D, the variation of some actions like *walking*, *standing up*, *shaking hands* and so on is more. Most of the actions have temporal evolution of spatial locations of the person performing the action. This is well captured by LSTM which is very data sensitive [9]. For MSRDailyAcitivity3D, classifier trained on features from last layer of LSTM outperforms the classifiers trained on motion and appearance features resulting a boost in the overall classification score as shown in table 4. The method in [30] reports same recognition rate as us on MSRDailyActivity3D but it is a kernel based method and evaluating such methods on large training samples are not tractable. The classification score of NTU [29] with respect to the state-of-the-art results are shown in table 5. In this dataset, the subject scale with respect to the global image scale is relatively small due to large subject to source camera distance. This makes the dataset unfit for using appearance based features whereas suitable for using 3D skeletons (as evident from state-of-the-art results). Due to a large variety of actions present in this dataset, the different features are able to capture actions based on their nature. Thus the fusion of these complementary features achieve competitive recognition score as in table 5. Attention mechanism based method [3] which outperforms our method on NTU are dif-

ficult to train and especially, focus of attention using spatial transformer network makes it data dependent. The consistent performance of our proposed framework on these dynamic dataset (smaller to larger in terms of size) proves the robustness of the framework.

5. Conclusion

Our contribution includes the use of spatio-temporal grids of convolutional feature maps to encode spatial granularity. We highlight the importance of using combination of different types of features namely, motion, geometry and appearance to model daily living activities. Our experimental analysis shows that the proposed spatio-temporal grids outperforms the existing 2D-CNN based features from parts of the subject body. The non-requirement of finetuning of RGB images in this framework and the score level fusion of independent classifiers trained on different features makes it robust on all categories of datasets. A future direction in this research can be to have a focus of attention in the network using these spatio-temporal grids. The recent evolution of 3D-CNNs is another area required 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] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] 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.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

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

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

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

[9] S. Das, M. Koperski, F. Bremond, and G. Francesca. A Fusion of Appearance based CNNs and Temporal evolution of Skeleton with LSTM for Daily Living Action Recognition. *ArXiv e-prints*, Feb. 2018.

[10] 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.

[11] 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.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

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

[14] 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.

[15] 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.

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

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

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

[19] M. Koperski and F. Bremond. Modeling spatial layout of features for real world scenario rgb-d action recognition. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 44–50. IEEE, 2016.

[20] 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.

[21] H. S. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–792–III–800. JMLR.org, 2013.

[22] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang. A deep structured model with radius–margin bound for 3d human activity recognition. *International Journal of Computer Vision*, 118(2):256–273, Jun 2016.

[23] 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.

[24] T. Liu, X. Wang, X. Dai, and J. Luo. Deep recursive and hierarchical conditional random fields for human action recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[25] C. Lu, J. Jia, and C. K. Tang. Range-sample depth feature for action recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 772–779, June 2014.

[26] B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recog-

Table 2. State-of-the-art performance on CAD-60 dataset. The performances of baseline methods are obtained from [18].

Method	Accuracy [%]
Order Sparse Coding [15]	65.30
Object Affordance [20]	71.40
HON4D [27]	72.70
Actionlet Ensemble [38]	74.70
MSLF [19]	80.36
JOULE-SVM [13]	84.10
P-CNN + kinect + Pose machines [8]	**95.58**
Dense Trajectories	73.53
LSTM (3D skeletons)	67.64
Proposed Method	**95.58**

Table 3. State-of-the-art performance on CAD-120 dataset. The performances of baseline methods are obtained from [18].

Method	Accuracy [%]
TDD [37]	80.38
SVM + CNN [22]	78.30
STS [21]	84.20
Object Affordance [20]	84.70
MSLF [19]	85.48
R-HCRF [24]	**89.80**
RSVM + LCNN [22]	**90.10**
Dense Trajectories	79.84
LSTM (3D skeletons)	60.48
Proposed Method	**89.23**

Table 4. State-of-the-art performance on MSRDailyActivity3D dataset. The performances of baseline methods are obtained from [18].

Method	Accuracy [%]
P-CNN + kinect + Pose machines [8]	84.37
Actionlet Ensemble [38]	85.80
MSLF [19]	85.95
BHIM [17]	86.88
JOULE-SVM [14]	95.00
Range Sample [25]	95.60
DSSCA-SSLM [30]	**97.50**
Dense Trajectories	83.44
LSTM (3D skeletons)	89.37
Proposed Method	**97.50**

Table 5. Cross-Subject recognition Accuracy comparison for NTURGB+D dataset.

Method	Accuracy [%]
P-LSTM [29]	62.93
ST-LSTM [23]	69.2
Geo-features [39] Ensemble	70.26
TS-LSTM [26]	74.60
DSSCA-SSLM [30]	74.86
CMN [40]	80.8
STA-Hands [2]	82.5
Glimpse Cloud [3]	**86.6**
Dense Trajectories	72.85
LSTM (3D skeletons)	66.53
Proposed Method	**84.22**

nition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3054–3062, 2016.

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

[28] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[29] 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.

[30] 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.

[31] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[32] 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.

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

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society.

[35] 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.

[36] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[37] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

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

[39] 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.

[40] 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.