Zineng Xu, Verónica Vilaplana, Josep Ramon Morros. (2018) Action tube extraction based 3D-CNN for RGB-D action recognition. *CBMI - 16th International Conference on Content-Based Multimedia Indexing*, *2018* : IEEE, 2018. Pp. 1-6 Doi: 10.1109/CBMI.2018.8516450.

Zineng Xu, Verónica Vilaplana, Josep Ramon Morros. (2018) Action tube extraction based 3D-CNN for RGB-D action recognition. *CBMI - 16th International Conference on Content-Based Multimedia Indexing*, *2018* : IEEE, 2018. Pp. 1-6 Doi: 10.1109/CBMI.2018.8516450.

Action Tube Extraction based 3D-CNN for RGB-D Action Recognition

Zineng Xu
Computer Science
Universitat Politècnica de Catalunya
Barcelona, Spain
zineng.xu@est.fib.upc.edu

Verónica Vilaplana
Signal Theory and Communications
Universitat Politècnica de Catalunya
Barcelona, Spain
veronica.vilaplana@upc.edu

Josep Ramon Morros
Signal Theory and Communications
Universitat Politècnica de Catalunya
Barcelona, Spain
ramon.morros@upc.edu

***Abstract*—In this paper we propose a novel action tube extractor for RGB-D action recognition in trimmed videos. The action tube extractor takes as input a video and outputs an action tube. The method consists of two parts: spatial tube extraction and temporal sampling. The first part is built upon MobileNet-SSD and its role is to define the spatial region where the action takes place. The second part is based on the structural similarity index (SSIM) and is designed to remove frames without obvious motion from the primary action tube. The final extracted action tube has two benefits: 1) a higher ratio of ROI (subjects of action) to background; 2) most frames contain obvious motion change. We propose to use a two-stream (RGB and Depth) I3D architecture as our 3D-CNN model. Our approach outperforms the state-of-the-art methods on the OA and NTU RGB-D datasets.**

***Index Terms*—action recognition, action tube extraction, 3D-CNN**

I. Introduction

RGB-D based action recognition has attracted much attention in recent years due to the advantages that depth information brings to the combined data modality. Depth is insensitive to illumination changes and has rich 3D structural information of the scene. Traditional studies on RGB-D action recognition use different kinds of methods [1], [2] to compute handcrafted features. With the recent development of deep learning, a few methods [3]–[6] have been developed based on Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN). These methods take as input either RGB, depth or both of them as independent streams and fuse the recognition scores of individual modalities. More recently, cooperative training a single CNN on both RGB visual features and depth features [7] has achieved state-of-the-art performance.

It is noteworthy that most works are focused on the architecture of deep learning model and very few works focus on the frame preprocessing. For frame extraction, the general method is uniform sampling of fixed number of frames [8]. However, this approach will miss some frames that contain an important amount of motion. This may affect the performance as motion is the most important clue for action recognition. Another common method [6] is to keep all frames, and to split them into several fixed-length clips. For video-based prediction, the model averages the predictions over all clips and provides the final prediction for the input video. This method has some weaknesses as it will break the completeness of action and may lead to a hard representation of action. For frame rescaling, the common approach is to crop the center area from original frames and resize to a fixed resolution [6], [8]. However, the subjects involved in the action are not always in the center of frames and their location could change through time. Furthermore, the subjects can appear very small in the video frame when they are far from the camera. In this case, fixed-size crop is not enough. In this paper, we propose a simple, yet effective, novel action tube extractor that takes as input a trimmed video containing one action, and outputs an spatio-temporal action tube (with fixed number of frames). Then, the action tube can be directly fed to the action recognition model. Our proposed action tube extractor can solve the problems mentioned above. For action recognition model, we propose to use I3D [9] as our 3D-CNN model. 3D ConvNets provide a natural approach to video modeling because they can learn motion features from RGB/Depth inputs directly, achieving excellent results. The approach is illustrated in Fig. 1.

The main contributions of this paper are:

- A method to extract spatio-temporal action tubes with a fixed number of frames
- The application of this method to improve an I3D model using RGB-D data for action recognition

Our approach has been evaluated on two challenging datasets, Office Activity (OA) [10] and NTU RGB-D [11] datasets. Experimental results achieved are state-of-the-art.

II. Proposed algorithm

The algorithm presented here consists of two parts, as illustrated in Fig. 1. The first part is our action tube extractor. It takes as input a trimmed video (a sequence of N frames containing a single action) and outputs an action tube. The second part is a RGB-D two-stream network. The inputs of the network are extracted action tubes using the method proposed in the first part. We propose to use the Inception 3D (I3D) architecture by Carreira and Zisserman [9] to model temporal context. It is designed based on the Inception architecture [12], but replaces 2D convolutions with 3D convolutions. Temporal

978-1-5386-7021-7/18/$31.00 © 2018 IEEE

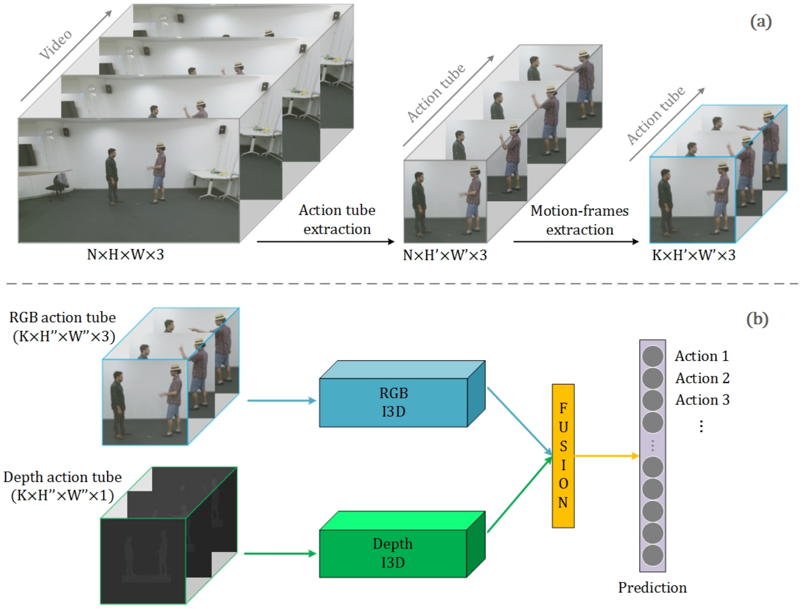

Fig. 1. Illustration of our approach for RGB-D based human action recognition. (a) Given a video, we extract the action tube by human detection (MobileNet-SSD). Then, the structural similarity index (SSIM) is used to extract motion-frames (contains obvious motion) to form final action tube. (b) Extracted RGB-Depth action tubes are classified with two-stream I3D model. Late fusion is performed to combine RGB-Depth information.

information is kept throughout the network. At test time, late fusion [13] is applied to combine RGB and depth information. In this section, we first describe our proposed action tube extractor (Sect. II-A), and then the two-stream I3D for action recognition (Sect. II-B).

A. Action tube extractor

Our action tube extractor involves two steps (see Fig. 1-(a)). The first step performs the action tube extraction. An action tube is defined as a spatio-temporal region following an actor performing an action, in this case a sequence of cropped frames through the video that contain the subjects of a given action. In order to achieve this goal, human detection is applied on each frame to generate the action tube. The action tube extraction has two benefits: 1) removing most useless background information; 2) increasing the area of region-of-interest (subjects). The second part is designed to perform a temporal sampling of the video sequence to remove frames without motion. This way, the video can be sampled using a fixed number of frames, as needed by the I3D model. Finally, we get an action tube with a fixed number (K) of frames. In the following we describe the action tube extraction and temporal sampling in detail.

Action tube extraction: The method is illustrated in Fig. 2. Considering efficiency, we propose to use MobileNet-SSD [14] as the human detection algorithm. This model is pre-trained on VOC0712 (2007+2012) [15] dataset. For each input frame, if there are more than one person, the network outputs more than one bounding boxes. Here, we make a slight modification to output only one bounding box for each frame consisting on selecting the rectangle enclosing all detected regions (see Fig. 2, second column). The final bounding box contains all the detected persons. Finally, we get N bounding boxes from N frames. As in some frames MobileNet-SSD can fail to detect humans, we use the detected bounding boxes of adjacent frames. Then, we generate a bounding box (Fig. 2, third column, black dashed box) that contains these N bounding boxes. Finally, the expanded bounding box (Fig. 2, third column, black solid box) is applied on each video frame to generate the action tube.

Motion-frames extraction: Motion is the most important information for action recognition. However, in most cases there are lots of similar frames (without motion change) in the extracted action tube. Thus, it is important to extract frames with obvious motion change. In order to extract those frames, we propose to use structural similarity index (SSIM) [16], that measures the similarity between two frames. We choose SSIM as it can be computed very efficiently and when applied to successive frames gives a good indication of the amount of motion. Fig. 3 shows some examples. We can see that frames without motion have higher SSIM value than frames with obvious motion. In other words, lower SSIM value indicates the frames with obvious motion. The motion-frames extraction is illustrated in Fig. 4. The first frame is always kept, and the rest $K-1$ frames are extracted according to the SSIM values. The SSIM is calculated from every two consecutive frames. We adopt two steps of extraction: local extraction and global extraction. For local extraction step, we extract one frame with lowest SSIM value every 16 frames. For global extraction step, we extract first $K-1-N_{local}$ frames with lowest SSIM values. For most simple actions (see Fig. 4. falling), global extraction is enough. However, for some complex actions (can be divided into several sub-actions, see Fig. 4. sleeping), local extraction is necessary because in some cases motion could occur only in one of the sub-actions so the remaining ones would not be represented in the final sampling. Our method combines a sort of uniform sampling with more detailed sampling where motion is present. At the end of this process, we obtain an action tube with only K frames.

B. Two-stream I3D

In [17], deep architectures used for action recognition are categorized in four groups: 2D models, motion-based input features, 3D models and temporal networks. In the first group, [18] uses a pre-trained model on one or more frames which are sampled from the whole video. Then, the entire video is labeled by averaging the result of the sampled frames. To consider temporal information, in the second group, [19] and [20] compute 2D motion features like optical flow. Afterwards, these features are exploited as different input streams of a 2D network. The third group introduces 3D filters in the convolutional and pooling layers to learn discriminative features along both spatial and temporal dimensions [21], [9]. The input data of these networks are a fixed length sequence of frames. Finally in the fourth category, Recurrent Neural Networks (RNN) and variations [5], [11] are utilized to process temporal information.

Among previous methods, two-stream (RGB frames and optical flow) 2D-CNN architecture achieved state-of-the-art results on RGB datasets. More recently, Carreira and Zisserman proposed Inception 3D (I3D) architecture [9], and this model achieves state-of-the-art performance on a wide range of video classification benchmarks. Therefore, I3D has been selected in this paper to be extended and analyzed for RGB-D data. In this paper, two modalities (RGB and depth) are used as the input data for I3D to form the two-stream I3D architecture (see Fig. 1-(b)). In trimmed activity recognition, the length of video is usually less than 10 seconds. As I3D needs a fixed number of frames as the input, we set the frame number K = 32. For late fusion, we average scores from the RGB and depth streams.

III. Experimental results

The proposed method has been evaluated on two RGB-D benchmarks, i.e. Office Activity (OA) [10] and NTU RGB-D [11] datasets. In the following, we proceed by briefly to describe the datasets (Sect. III-A), implementation details (Sect. III-B), and then we present the comparison to the state-of-the-art (Sect. III-C) and discussion (Sect. III-D).

A. Datasets

OA dataset. It covers the regular daily activities taken place in office. The dataset consists of 1,180 sequences, containing 20 classes of activities performed by 10 subjects. Specifically, it is divided into two subsets, each of which contains 10 classes of activities: OA1 (complex activities by a single subject) and OA2 (complex interactions by two subjects). For fair comparison and evaluation, we follow the same protocol, and thus 5-fold cross validation is adopted by ensuring that the subjects in training set are different with those in testing set.

NTU RGB-D dataset. To our best knowledge, it is currently the largest action recognition dataset in terms of training samples for each action. The dataset has 56,880 sequences, containing 60 actions performed by 40 subjects aged between 10 and 35. It consists of 80 different camera views. For fair comparison and evaluation, the same protocol as that in [11] was used. It has both cross-subject and cross-view evaluation. According to previous papers [8], [11], [22], [23], cross-subject is harder than cross-view. Therefore, we only focus on the cross-subject evaluation. In the cross-subject evaluation, samples of subjects 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35 and 38 were used as training and samples of the remaining subjects were reserved for testing.

B. Implementation details

The frame resolution of extracted action tube is resized to 300×300. For RGB stream, the I3D networks are initialized with Kinetics [24] pre-trained models. For OA dataset, we also applied random left-right flipping consistently for each video during training. For very short videos ($N < 32$), we looped the last frame $32 - N$ times without motion-frames extraction.

C. Comparison to the state-of-the-art

We compare our proposed approach to some state-of-the -art results on two challenging datasets.

OA dataset. On this dataset, we apply our method on the two OA subsets. As shown in Table I, our model performance is much better than the state-of-the-art method on both subsets, with improvements in accuracy larger than 20%. We see that using a combination of RGB and depth outperforms the individual modalities, as was expected. From the results, we can conclude that visual recognition of actions (interactions) by two subjects (OA2) is harder than recognition of actions by a single subject (OA1). In most cases, interactions by two subjects are more abstract/complex than actions performed by a single subject.

TABLE I
Comparison of the proposed method with state-of-the-art approach on OA1 and OA2 Datasets

OA1 Dataset			
Method	RGB	Depth	RGB+Depth
R-SVM-LCNN [10] (2016)	60.4%	65.2%	69.3%
Ours	**87.7%**	**84.8%**	**91.9%**
OA2 Dataset			
Method	RGB	Depth	RGB+Depth
R-SVM-LCNN [10] (2016)	46.3%	51.1%	54.5%
Ours	**77.5%**	**72.8%**	**82.2%**

NTU RGB-D dataset. Table II lists the performance of the proposed method and previous works. The proposed method has been compared with some state-of-the-art skeleton-based, depth-based and RGB+Depth based methods that were previously reported on this dataset. We can see that the proposed method outperforms all these previous approaches.

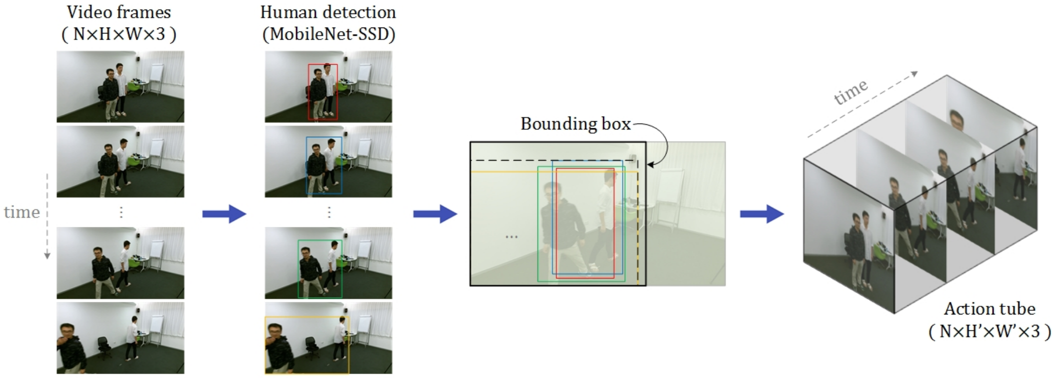

Fig. 2. Overview of action tube extraction. Pre-trained MobileNet-SSD is performed to detect subjects in each video frame. The final bounding box (black solid box) is applied on every frame to form the action tube.

TABLE II
COMPARATIVE ACCURACIES OF THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS ON NTU RGB-D DATASET (CROSS-SUBJECT EVALUATION)

Method	Skeleton	RGB	Depth	RGB+Depth
SSSCA-SSLM [7] (2017)	-	-	-	74.86%
c-ConvNet [8] (2018)	-	-	-	86.42%
D-CNN [22] (2018)	-	-	87.08%	-
HCN [23] (2018)	86.5%	-	-	-
Ours	-	91.95%	86.02%	**93.56%**

Detailed results, including the confusion matrices and per class accuracies can be found in the Additional Material document [25]

D. Discussion

To better analyze the performance of the proposed model, we take a closer look at actions that are highly confusing to the two-stream I3D structure (Fig. 6 shows the confusion matrix for the NTU RGB-D dataset). As presented in Fig. 5, such action pairs include reading vs. writing, nod head/bow vs. pickup, nausea or vomiting condition vs. nod head/bow, showing object vs. shaking hands, chatting vs. chatting and eating, and arranging files vs. looking for objects. From these samples, we can observe that these misclassified actions are inherently confusing. In order to deal with such actions, we may need to obtain fine-grained information. This will be our future work.

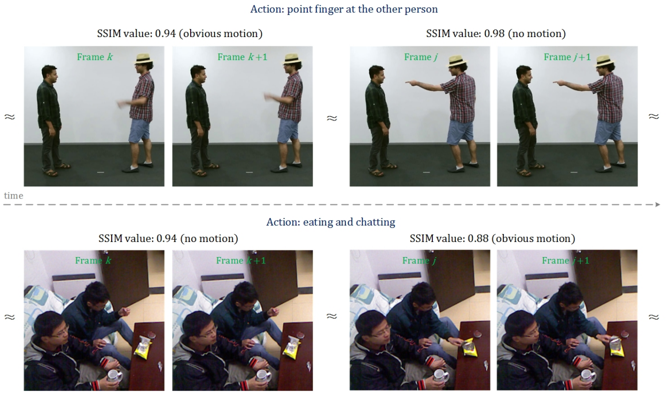
Action: point finger at the other person
SSIM value: 0.94 (obvious motion)
SSIM value: 0.98 (no motion)
Frame k
Frame k+1
Frame j
Frame j+1
≈
time
Action: eating and chatting
SSIM value: 0.94 (no motion)
SSIM value: 0.88 (obvious motion)
Frame k
Frame k+1
Frame j
Frame j+1

Fig. 3. Examples of SSIM value of consecutive frames. Frames with obvious motion have lower values than those with no motion

In order to further demonstrate the effectiveness of the action tube extractor, we compare the results of our method against a similar system where the action tube extractor has been replaced by a more traditional approach consisting in cropping the center region and using uniform sampling. A region of size $H \times H$ is cropped from the original frame, where H is the size of shorter side. The extracted frames are resized to 300×300 pixels. Finally, these resized frames are fed into I3D model. For this test, we used only the RGB modality for simplicity. The comparisons are shown in Table III. We can see that our proposed action tube extractor provides an improvement in accuracy around 3% on both OA and NTU RGB-D datasets. This is a strong demonstration of the effectiveness of our proposed action tube extractor.

TABLE III
COMPARISON OF PERFORMANCE WITH AND WITHOUT ACTION TUBE EXTRACTOR ON THE TWO DATASETS (RGB MODALITY)

Dataset	with ATE	w/o ATE
NTU RGB-D	**91.95%**	89.29%
OA1	**87.7%**	84.2%
OA2	**77.5%**	73.9%

CONCLUSIONS

In this paper, we propose a novel action tube extractor for 3D action recognition. It takes as input a trimmed video and outputs an action tube. The action tube contains much less background information, and has higher ratio of ROI (subjects) to background. Besides, every frame of the extracted action tube contains obvious motion change. Then the extracted RGB/Depth action tubes are directly fed into two-stream I3D model. An extensive experimental analysis shows the benefits

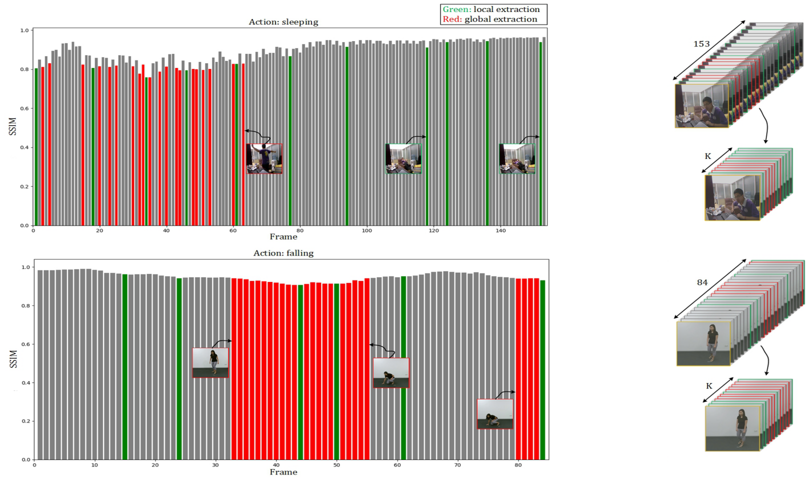

Fig. 4. Overview of motion-frames extraction. The bar plot represents SSIM values of every two consecutive frames. Green frames are locally selected frames, red frames are globally selected. The K extracted frames consists of green frames, red frames and the first frame. (here, K=32)

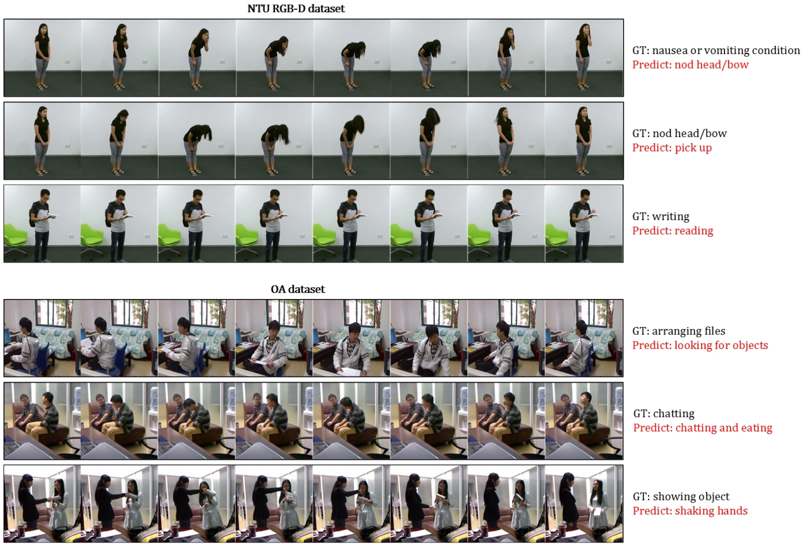

Fig. 5. Some incorrect action recognition results on the test set of OA and NTU RGB-D datasets.

of our proposed approach, which achieves state-of-the-art results on both OA and NTU RGB-D datasets.

IV. Acknowledgements

This work was supported by the project MALEGRA, TEC2016-75976-R, financed by the Spanish Ministerio de Economa, Industria y Competitividad and the European Regional Development Fund (ERDF)

References

[1] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal Quads: Human Action Recognition Using Joint Quadruples," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 4513–4518.

[2] C. van Gemeren, R. T. Tan, R. Poppe, and R. C. Veltkamp, "Dyadic interaction detection from pose and flow," in *Human Behavior Understanding*, H. S. Park, A. A. Salah, Y. J. Lee, L.-P. Morency, Y. Sheikh, and R. Cucchiara, Eds. Springer International Publishing, 2014, pp. 101–115.

[3] H. Li and C. Y. Suen, "Robust face recognition based on dynamic rank representation," *Pattern Recognition*, vol. 60, pp. 13–24, 2016.

[4] H. Li, W. Hu, W. Wang, and Z. Xie, "Automatic dictionary learning sparse representation for image denoising." *Journal of Grey System*, vol. 30, no. 2, 2018.

[5] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR 2015*, vol. 07-12-June, 2015, pp. 1110–1118.

[6] R. Zhao, H. Ali, and P. Van Der Smagt, "Two-stream RNN/CNN for action recognition in 3D videos," in *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017-September, 2017, pp. 4260–4267.

[7] A. Shahroudy, S. Member, and T.-t. Ng, "Deep Multimodal Feature Analysis for Action Recognition in RGB + D Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1045–1058, 2017.

[8] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative Training of Deep Aggregation Networks for RGB-D Action Recognition," in *AAAI*, 2018.

[9] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *The IEEE Conference on*

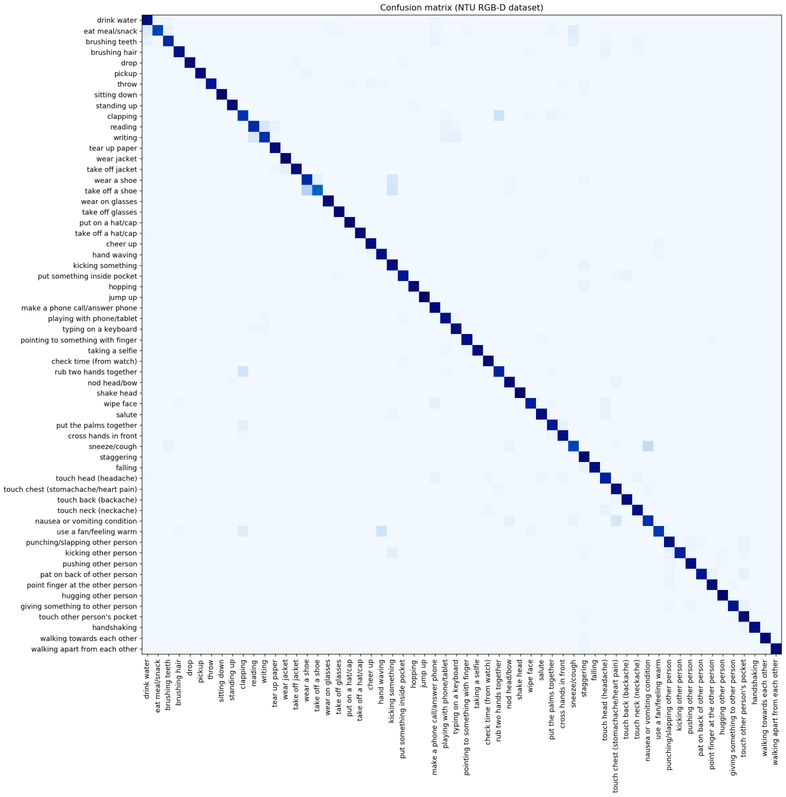

Fig. 6. Confusion matrix for the NTU dataset

Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: http://arxiv.org/abs/1705.07750

[10] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A Deep Structured Model with RadiusMargin Bound for 3D Human Activity Recognition," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 256–273, 2016.

[11] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in *CVPR 2016*, 2016.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *CVPR 2016*, 2016.

[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *ArXiv*, p. 9, 2017.

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc 2007) results (2007)," 2008.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "Deep learning for action and gesture recognition in image sequences: A survey," in *Gesture Recognition*. Springer, 2017, pp. 539–578.

[18] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4597–4605.

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

[20] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3164–3172.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[22] P. Wang, W. Li, Z. Gao, C. Tang, and P. Ogunbona, "Depth Pooling Based Large-scale 3D Action Recognition with Convolutional Neural Networks," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1051–1061, 2018.

[23] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *cvpr 2017*, 2017.

[25] Z. Xu, V. Vilaplana, and J. R. Morros, "Additional results for action tube extraction based 3d-cnn for rgb-d action recognition," Tech. Rep. [Online]. Available: https://imatge.upc.edu/web/resources/additional-results-action-tube-extraction-based-3d-cnn-rgb-d-action-recognition