© 2018 IJSRST | Volume 4 | Issue 8 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Neoteric Approach Based on Multi Task Learning Network for Skeletal 3D Action Recognition

T. Seshagiri[1], S. Varadarajan[2]
[1]Research Scholar, Rayalaseema University, Kurnool, Associate Professor, Shree Institute of Technical Education, Tirupati, India
[2]Professor, Department of Electronics & Communicationengineering, Svu Engineering College, Tirupati, India

ABSTRACT

This paper presents a new representation of skeleton sequences for 3D action recognition. Existing methods based on hand-crafted features or recurrent neural networks cannot adequately capture the complex spatial structures and the long term temporal dynamics of the skeleton sequences, which are very important to recognize the actions. In this paper, we propose to transform each channel of the 3D coordinates of a skeleton sequence into a clip. Each frame of the generated clip represents the temporal information of the entire skeleton sequence, and one particular spatial relationship between the skeleton joints. The entire clip incorporates multiple frames with different spatial relationships, which provide useful spatial structural information of the human skeleton. We also propose a Multi-task Learning Network (MTLN) to learn the generated clips for action recognition. The proposed MTLN processes all the frames of the generated clips in parallel to explore the spatial and temporal information of the skeleton sequences. The proposed method has been extensively tested on challenging benchmark datasets. Experimental results consistently demonstrate the superiority of the proposed learning method for 3D action recognition compared to existing techniques.

I. INTRODUCTION

Human action recognition has a wide range of applications, including video surveillance, human-machine interaction and robot control [1]. Nowadays due to the prevalence of highly-accurate and affordable depth devices, there are more and more works using depth videos for computer vision tasks [2], [3], [4], [5]. Action recognition based on 3D skeleton sequences has also been attracting increasing attention [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Compared to RGB videos, skeleton data is more robust to cluttered backgrounds and illumination changes [17], with human actions described as movements of the skeleton joints [18]. To recognize an action from a skeleton sequence, the temporal dynamics of the skeleton sequence and the spatial structure among the joints need to be exploited for a good understanding of the action class [10]. Hidden Markov Models

(HMMs) [19], [20], Conditional Random Fields (CRFs) [21] and Temporal Pyramids (TPs) [7], [22] have been used to model the temporal structure of a sequence. To exploit the spatial structure among the joints, various features have been investigated, such as histogram of joint positions [6], pairwise relative position [22] and 3D rotation and translation [7]. These traditional models are based on hand-crafted features, which cannot effectively capture the long-term temporal structure and the discriminant spatial information of the skeleton sequence [8]. Recently, recurrent neural networks (RNNs) with LongShort Term Memory (LSTM) neurons [23], [24] have also been used to model the spatial and temporal information of skeleton sequences for action

recognition [8], [25], [10], [9], [11]. LSTM networks operate the input sequentially and return an output at each timestep. Human actions are generally very complex with many timesteps. The earlier timesteps of an action sequence might contain ambiguous sub-actions and the context of the entire sequence needs to be learned to accurately recognize the action. Although LSTM networks are designed to explore the problem of long-term temporal dependency, they are incapable of memorizing the information of an entire sequence with many timesteps [26], [27]. Besides, it is also difficult to construct deep LSTM networks to learn high-level features [28], [29].

In order to learn the long-term temporal information and the complex spatial information of the skeleton sequences for action recognition, in this paper, we transform each skeleton sequence into three clips. Each clip consists of only a few frame images, as shown in Figure 1.

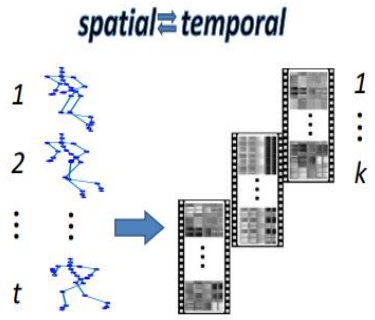

Figure 1. A skeleton sequence of an arbitrary length t is transformed into three clips

We propose to change every skeleton succession to another portrayal, i.e., three clips, to permit worldwide long haul fleeting demonstrating of the skeleton grouping by utilizing profound CNNs to take in progressive highlights from outline pictures. (31) We acquaint a MTLN among the procedure; the entire CNN highlights of the boundaries in conveyed cuts, during this manner take the spatial formation as well as the provisional information of the skeleton gathering.

The MTLN enhances the execution by using inherent connections among various frames of the created clips. Our test results come about show that MTLN performs superior to connecting or then again pooling the highlights of the housings. [32] The proposed approach accomplishes the best in class execution on three skeleton datasets.

II. RELATED WORKS

In this section, we briefly review relevant literature on skeleton-based action recognition methods using hand-crafted features and using deep learning networks. Hand-crafted Features Traditional methods extract handcrafted features and utilize sequential models to represent the spatial temporal information of the skeleton sequences. Xia et al. [6] computed histograms of 3D joint locations (HOJ3D) to represent each frame of the skeleton sequences, and used discrete hidden Markov models (HMMs) to model the temporal dynamics.

Wang et al. [45] represented actions with the histogram of spatial-part-sets and temporal-part-sets, which are constructed from a part pose dictionary. Chaudhry et al. [46] used a set of Linear Dynamical Systems (LDSs) to encode a hierarchy of spatial temporal information of the 3D skeleton data, and performed action recognition using discriminative metric learning. Vemulapalli et al. [7] used the 3D rotations and translations between various body parts as a representation, and modeled the skeleton sequence as a curve in the Lie group. Lv et al. [19] extracted a set of features corresponding to the motion of an individual joint or multiple joints, and used HMM to model the temporal dynamics. Wu et al. [20] concatenated the posture motion and the offset features as representation, and estimated the emission probability for action inference using a deep neural network.

Hussein et al. [47] computed the covariance matrices of the trajectories of the joint positions over hierarchical temporal levels to model the skeleton sequences. Wang et al. [22] computed the pairwise relative positions of each joint with other joints to represent each frame of the skeleton sequences, and used Fourier Temporal Pyramid (FTP) to model the temporal patterns. Yang et al also used the pairwise relative positions of the joints to characterize the posture features, the motion features, and the offset features of the skeleton sequences, and applied Principal Component Analysis (PCA) to the normalized features to compute EigenJoints as representations. These traditional methods are based on handcrafted features, which are not powerful enough to extract discriminant spatial temporal information from the skeleton sequences for action recognition.

Liu et al. [11] designed a spatial temporal LSTM with a Trust Gate to jointly learn both the spatial and temporal information of skeleton sequences and to automatically remove noisy joints. Although LSTM networks are designed to explore long-term temporal dependencies, it is still difficult for LSTM to memorize the information of the entire sequence with many timesteps [26], [27]. In addition, it is also difficult to construct deep LSTM to extract highlevel features [28], [29]. In contrast to LSTM, CNNs can extract such high-level features, but do not have the capacity to model the long-term temporal dependency of the entire video [50].

III. PROPOSED SYSTEM

To resolve this problem, we propose to transform the skeleton sequences into clips, which allows for the spatial and temporal information learning of the skeleton sequences based on CNNs. More specifically, we propose a novel MTLN to exploit the intrinsic relationships among different frames of the clips for action recognition. The classification of each frame is treated as a separate task.

MTLN jointly learns multiple tasks of classification, and then outputs multiple predictions. Each prediction corresponds to one task. The labels of all the tasks are the same as the action label of the skeleton sequence. During training, the loss value of each task is individually computed using its respective class scores. The network parameters are learned using the total loss that is defined by the sum of the loss values of all tasks.

During testing, the class scores of all tasks are averaged to form the final prediction of the action class. The proposed method captures both the temporal and the spatial structural information of the skeleton sequences and also makes the representation more robust to view variations. MTLN explores both the spatial and temporal information of the skeleton sequence from the generated clips for action recognition. In our experiments, we compare the proposed method with other methods. We also compare the multitask learning of the clips with the single-task learning of an individual frame, as well as feature concatenation and pooling methods of multiple frames, to show the advantages of the proposed clip representation and learning method.

IV. CLIP GENERATION

Instead of frame images, skeleton sequences only provide the 3D trajectories of the skeleton joints. In this section, we introduce two different methods of transforming the skeleton sequences to a set of clips. Each clip consists of several images to allow for spatial temporal feature learning based on deep CNNs. More specifically, for each skeleton sequence, both methods generate three clips. Each clip corresponds to one channel of the 3D coordinates of the skeleton joints. Each frame of the clips includes the information of one particular spatial relationship between the

skeleton joints and the temporal information of the entire sequence. Each clip aggregates multiple frames with different spatial relationships, which provides important information of the spatial structure of the joints

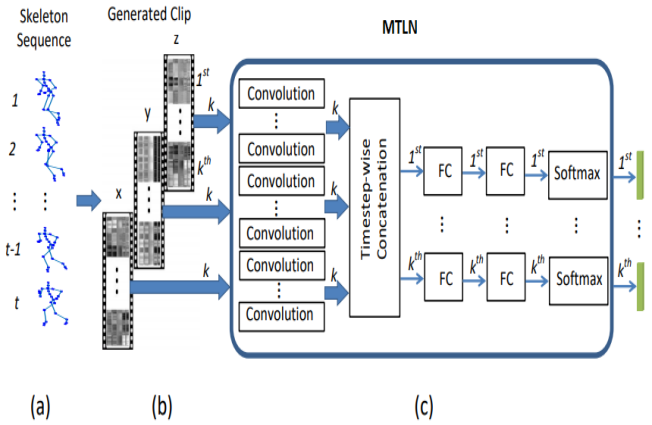

Figure 2. structural design of the future system. particular skeleton series (a), three clips (b) related toward the three channels of the cylindrical coordinates are generated.

A important approved CNN show (c) and a transient mean pooling layer (d) are utilized to separate a reduced interpretation from each edge of the clips. The amount produce CNN interpretations of the three clips in the meantime step are connect together, coming about four component vectors (e). Every part vector tests to the transient data of the skeleton association and a difficult spatial connection of the skeleton joints.

The proposed MTLN (f) which joins a totally linked (FC) layer, a redressed instantly unit (ReLU), a different FC layer and a Soft max layer normally shapes the four section vectors in equivalent and yields four blueprints of set scores (g), each identifying with one endeavor of portrayal using one component vector. Times of introduction, the difficulty estimations of the four assignments are demonstrated portray the misfortune estimation of the framework used to revive the formation parameters. For taxing the set scores of the four coursework are inwards at the average to express the previous calculation of the expansion class.

The robust and invariant transient data of the first skeleton arrangement could be caught with the capable CNN representations gained from each casing picture. The instance route of movement of each joint of a skeleton interest group can be tended to as three 1D bring to light regions appearing in a different way relation to the three channels of the 3D Cartesian headings (x, y,and z).

$$\Omega = \{q_i \in R^3 : i = 1, m\}(1)$$

wherever m is the measure of the skeleton joints, and qi = [xi; yi; zi] speaks to the 3D facilitate of the ith joint.

To change Where m is the quantity of the skeleton joints, and qi = [xi; yi; zi] speaks to the 3D facilitate of the ith joint. line measurement in a successive inquire. The four reference joints are perused four body parts, particularly, the left shoulder, the correct shoulder, the left hip and the correct hip.

For each reference joint, a course of action of vectors can be controlled by enlisting the qualification of bearings between the reference joint and alternate joints. Each arrangement of vectors mirrors specific spatial connections between the joints.

Let the reference joint be

$$q_0^k = [x_0^k y_0^k z_0^k] k = 1, \ldots .4 \text{ , and define}$$

$$V_k \triangleq \{q - q_0^k : q \epsilon \Omega , k = 1, \ldots ,4 \}(2)$$

Where V_k is the set of the vectors of the k^{th} reference joint in one frame.

V. CLIP LEARNING

The three CNN highlights of the three fastens in the mean time step are associated in a section vector, which verbalizes to the concise data of the skeleton movement and one specific spatial relationship between the skeleton joints in three channels tube

shaped directions. At that point the element vectors ever steps are together handled in parallel utilizing multi-undertaking adapting, in this way to use their inborn connections for activity response.

Give the enactment at the I th a chance to push and the j th fragment of the k th feature direct be x i,jk. After transient mean pooling, the yield of the kth highlight outline given by

$$y^k = [y_1^k \ldots, y_j^k \ldots, y_{14}^k]$$

OR

$$y_j^k = \frac{1}{1}\sum_{i=0}^{14} \max(0, x_{i,j}^k) \quad (3)$$

1.3.3 Multi-Task Learning Network:

Multi-Task Learning Network is then proposed to mutually process the four element vectors to use their natural connections for activity acknowledgment. The characterization of each component vector is dealt with as a different assignment with a similar grouping name of the skeleton succession.

i. During training, the class scores of each task are used to compute a loss value.

$$L_k(Z_k Y) = \sum_{i=1}^{m} y_i \left(-\log\left(\frac{exp_{z_{ki}}}{\sum_{j=1}^{m} exp_{z_{kj}}}\right)\right) (4)$$

$$= \sum_{i=1}^{m} y_i \left(\log \sum_{j=1}^{m} exp_{z_{kj}}\right) - zk_i$$

Where z_k is the vector fed to the Softmax layer generated from the k^{th} input feature, m is the amount of action classes and y_i is the ground-truth label for class i.

ii. Then the loss values of all tasks are summed up to generate the final loss of the network used to update the network parameters.

VI. RESULTS

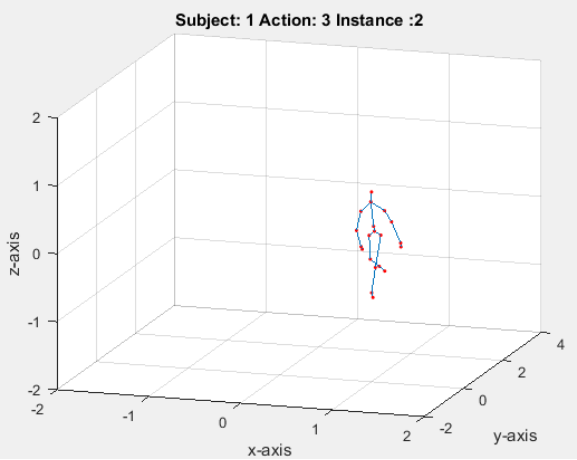

Figure 3. Input Image

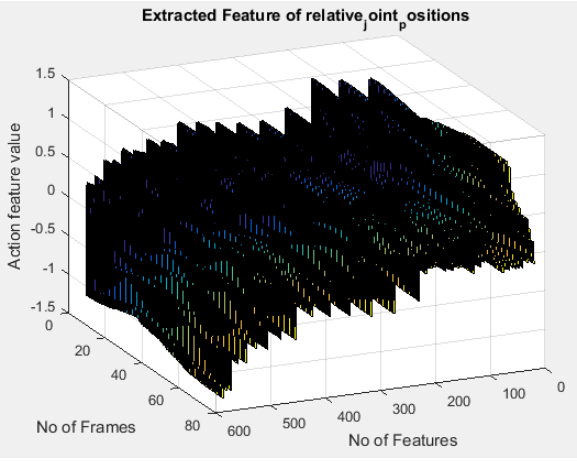

Figure 4. Extracted feature of relative joint positions

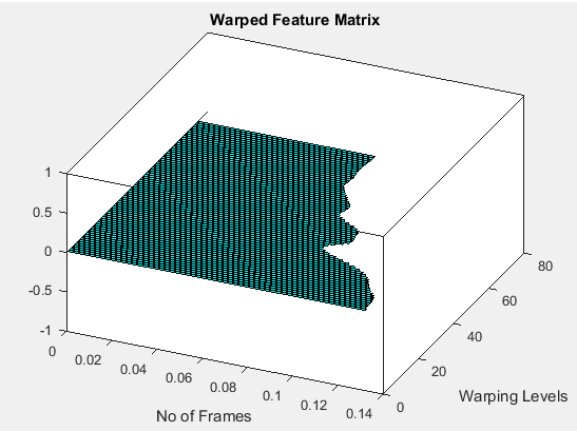

Figure 5. Warped Feature matrix

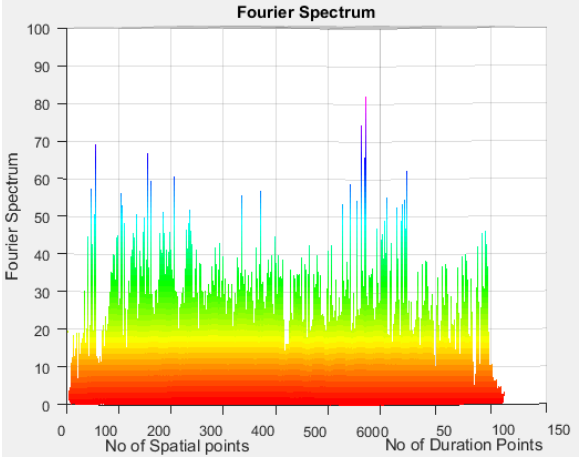

Figure 6. Fourier spectrum

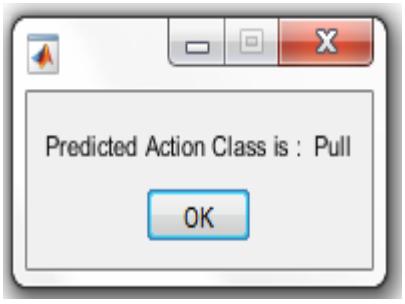

Figure 7. Predicted action

Table 1. Comparison table for different methods

Method	Accuracy	Sensitivity	Specificity
DRNN	73.9051	77.2302	78.2121
Uni-directional	85.7011	82.3212	80.4567
Bi-directional	86.9317	83.8361	83.9317
MTLN	93.4237	87.2398	85.6127

VII. CONCLUSION

In this paper, we have proposed to transform a skeleton sequence into three clips for robust feature learning and action recognition. Each frame of the generated clips depicts the temporal information of the skeleton sequence. The entire clips incorporate different spatial relationships between the joints and provide useful spatial structural information of the skeleton sequence. The generated clips are then processed with an MTLN to capture both the spatial and temporal information for action recognition. MTLN learns the clips in a multi-task learning manner in order to utilize the intrinsic relationships between the clip frames. This improves the performance (compared to the concatenation or the pooling methods). We have tested the proposed method on datasets and have compared it to previous state-of-the-art methods and several baselines. Experimental results have shown the effectiveness of the proposed new representation and feature learning method.

VIII. REFERENCES

[1]. X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 804-811.

[2]. G. Zhang, J. Liu, Y. Liu, J. Zhao, L. Tian, and Y. Q. Chen, "Physical blob detector and multi-channel color shape descriptor for human detection," Journal of Visual Communication and Image Representation, 2018.

[3]. G. Zhang, J. Liu, H. Li, Y. Q. Chen, and L. S. Davis, "Joint human detection and head pose estimation via multistream networks for rgb-d videos," IEEE Signal Processing Letters, vol. 24, no. 11, pp. 1666-1670, 2017.

[4]. P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for rgb-d action recognition," arXiv preprint arXiv:1801.01080, 2017.

[5]. H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," IEEE transactions on pattern analysis and machine intelligence, 2017.

[6]. L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 20-27.

[7]. R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 588-595.

[8]. Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110-1118.

[9]. A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 10W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in AAAI Conference on Artificial Intelligence (AAAI), 2016. 11J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in European Conference on Computer Vision (ECCV). Springer, 2016, pp. 816-833.

[10]. P. Koniusz, A. Cherian, and F. Porikli, "Tensor representations via kernel linearization for action recognition from 3d skeletons," arXiv preprint arXiv:1604.00239, 2016.

[11]. P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016, pp. 102-106.

[12]. Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," IEEE Signal Processing Letters, vol. 24, no. 6, pp. 731-735, 2017.

[13]. J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, and G. Wang, "Skeletonbased action recognition using spatio-temporal lstm network with trust gates," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[14]. J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeletonbased human action recognition with global context-aware attention lstm networks," IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 1586-1599, 2018.

[15]. F. Han, B. Reily, W. Hoff, and H. Zhang, "space-time representation of people based on 3d skeletal data: a review," arXiv preprint arXiv:1601.01006, 2016.

[16]. M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. Springer, 2013, pp. 149-187.

[17]. F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class adaboost," in European Conference on Computer Vision (ECCV), 2006, pp. 359-372.

[18]. D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 724-731.

[19]. C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," Computer Vision and Image Understanding, vol. 104, no. 2, pp. 210-220, 2006. 22J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1290- 1297.

[20]. A. Graves, "Neural networks," in Supervised Sequence Labelling with Recurrent Neural Networks. Springer, 2012, pp. 15-35.

[21]. A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in IEEE International Conference on

Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 6645-6649.

[22]. V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4041-4049.

[23]. J. Weston, S. Chopra, and A. Bordes, "Memory networks," arXiv preprint arXiv:1410.3916, 2014. 27J. Gu, G. Wang, and T. Chen, "Recurrent highway networks with language cnn for image captioning," arXiv preprint arXiv:1612.07086, 2016. 28T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 4580-4584.

[24]. R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," arXiv preprint arXiv:1312.6026, 2013. 30Y. LeCun, Y. Bengio et al., "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, vol. 3361, no. 10, p. 1995, 1995.

[25]. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," arXiv preprint arXiv:1405.3531, 2014.

[26]. D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 3642- 3649.

[27]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[28]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014. 35C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[29]. Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1600-1609.

[30]. Q. Ke and Y. Li, "Is rotation a nuisance in shape recognition?" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4146-4153.

[31]. Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, "Human interaction prediction using deep temporal features," in European Conference on Computer Vision Workshops. Springer, 2016, pp. 403-414.

[32]. Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Leveraging structural context models and ranking score fusion for human interaction prediction," IEEE Transactions on Multimedia, 2017. 40W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2016, pp. 4898-4906.

[33]. K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 28-35.

[34]. A. Yao, J. Gall, G. Fanelli, and L. Van Gool, "Does human action recognition benefit from pose estimation?"," in Proceedings of the 22nd British machine vision conference-BMVC 2011,

2011. 43R. Caruana, "Multitask learning," in Learning to learn. Springer, 1998, pp. 95-133.

[35]. Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[36]. C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 915-922.

[37]. R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bioinspired dynamic 3d discriminative skeletal features for human action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 471-478.

[38]. M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." in IJCAI, vol. 13, 2013, pp. 2466-2472.