Deep Learning For Human Activity Recognition

Muhammad Umer Siddique
UMJI-SJTU
Graduate(MSc)
Shanghai
umer_siddiqie@sjtu.edu.cn

Ailane Mohamed Toufik
SJTU
Graduate(PhD)
Shanghai
ailane8@sjtu.edu.cn

Delmwin Baeka
UMJI-SJTU
Undergrad(BSc)
Shanghai
dbaekajnr@sjtu.edu.cn

Abstract

In this paper, we introduce a simple Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) based method for Human Action Recognition (HAR) by incorporating the memory attention module in RNN for temporal and then CNN for spatial information to reduce the difficulty in complex action recognition task. Our RNN module is like a residual module that apply the attention learning to the rectangular cooardinates of skeleton data and the first Convolutional Neural Network CNN takes depth images directly and then another CNN used which takes the output of first CNN and RNN as images and then model the spatial information on these sequence of images. These two CNNs and RNN form a single architecture which the model accuracy. We validate our model on two different kinds of datasets, our first and main dataset is University of Texas at Dallas Multimodal Human Action Dataset(UTD-MHAD) which consist of 27 different actions from 4 females and 4 males subjects and second one is NTU RGB+D dataset which contains 60 classes and 17 subjects. Our proposed architecture can be trained directly with the skeleton instead of using different skeleton based descriptors which might loose some useful information and by using this simple approach along with the separate CNN for depth dataset dataset we achieve the better accuracy —- on th validation and —- on test data, which is significantly higher than the baseline architecture. For the purpose of fairness we also compare our model results with other state of art model for HAR.

1. Introduction

Human Action Recognition (HAR) is a research field that piqued the researchers interests since late 19s due to its wide range of applications in many different fields like Visual Surveillance, Video Retrieval [15], human-Robot Interaction, medicines and sociology etc.

Action recognition is considered relatively more challenging in the area of computer vision due to many reasons among which: Different actions which makes the representation complex to discriminate the actions. Secondly, it involves identification of different actions from video like medical images. Last and not list, the main reason is that it requires a huge computational cost.

Before the invention of kinects cameras [25] and high computing powers action recognition was considered as one of the most difficult tasks for computers. As the data in the form of waves (inertial), depth images (kinects, point clouds) or skeleton based like accelerometers and gyroscopes which makes it more difficult than the usual recognition task.

Also the information can be directly retrieved from the sensors, but in our research we will use the already collected state-of-art dataset collected by the research group of University of Texas at Dallas [1]. Instead of going for the traditional approaches, where feature extraction is applied independently in order to capture relevant features [22]. Then, machine learning based approaches are used for classification and/or recognition purpose. Alternatively, a deep learning architecture which consist of recurrent neural network (RNN) [13] and convolutional neural networks (CNNs) [11] is adopted. In this approach, the temporal information is directly extracted from the skeleton data by using the rectangular coordinates and applying the attention like residual learning style [20] and then pass the output to the CNN which is trained to capture the spatial information of the input. By adopting this approach we do not need to extract the features manually, instead, the model extracts the features automatically and relate spatial information with less error rate than a human. As we have depth image descriptors and skeleton data in rectangular coordinates form, so we can use the keras functional API to create multiple input channel for RNN and CNN and use seperate input for CNN and RNN and then combine the outputs of each module and pass it to the CNN which now have the more specific information which is helps CNN to accurately recognize the action and put into the right class from which it belongs. Our contributions in this work are:

- Reduce the time taken to generate descriptors from

skeleton dataset by working directly with 3D skeleton data

- Proposed End-to-End spatiotemporal Memory Attention + CNN Networks for Both Skeleton and Depth data

- Proposed End-to-End spatiotemporal Memory Attention + CNN Networks for Both Skeleton and Depth data

2. Related works

The article [3], [17] uses the single and double stream network for action recognition while in 2015 [8] uses the convolutional neural network. The paper [2] compares the results of CNNs based action recognition with other approaches: SVM (Support Vector Machine), KNN (classification based on the top k nearest neighbors), MV (Mean and Variance) and DBN (Deep Belief Betwork). Among them, the first two methods and the third method show the best results for the Opportunity Activity Recognition and Hand Gesture data sets, respectively, and got high attraction by significantly improved results without manual hand engineering. In their experiments they showed that the usage of the ultra-accurate networks surpasses other methods. Deep convolutional networks (DCN) are implemented in [9]. The authors of this article claim that when data from the sensors is extracted independently by hand correlation between different signals is ignored, so they transfer all signals from the accelerometer and gyroscope to the activity image, which carries more information: the hidden relationship between any pair of signals. The authors compare DCNN and DCNN + with SVM and the Feature Selection method. The accuracy of the DCNN + reaches 97% on the UCI data, while the SVM and the Feature selection method reach 96 and 91%, respectively. In the article by Morales & Roggen [14] these results are compared with the results of [23]. Basing on the recent success of recurrent neural networks for time series domains, they proposed a deep generic framework for activity recognition relying on convolutional and LSTM recurrent units, that are suitable for multimodal wearable sensors; can perform sensor fusion naturally; do not require expert knowledge in designing features; and explicitly model the temporal dynamics of feature activations.

3. Model

We got the idea of memory attention module from this paper [21] where they just use the skeleton data as input to the RNN and then use the CNN for spatial temporal. Second we got the idea of of fusing multiple channels for better feature extraction and results from this paper [10] where the author inputs multiple descriptors to multiple CNNs and then fuse them to get the higher accuracy. In the second paper where author first generate different depth and skeleton data descriptors as input to the multiple CNNs and mention several results by just using skeleton, depth and both of them. In their results they just got 50% accuracy by using the skeleton data descriptors because when we generate images of the skeleton data the joints angle, pitch and other useful information lost. So in our work we well take the direct input as x,y,z in 3 different channels of RNN and also inputs the depth images to the CNN then we will use the other CNN architecture for spatial temporal information. In short our model have two parts which we will briefly elaborate here as Temporal Attention Module and Spatial Temporal Module.

3.1. Temporal Attention Module

The input skeleton data is a sequence of multi-frame 3D joint coordinates forming an action. Let $O = X; Y; Z \in R^{T\times N\times 3}$, where $X \in R^{T\times N}, Y \in R^{T\times N}, Z \in R^{T\times N}$, denotes N joints along T frames with x, y, and z coordinates.

As shown in the Fig. 1, given a matrix X the Temporal Attention Module starts a specific attention in BiGRU which is a forward and backward memory cell. BiGRU capture the temporal information like residual style as

$$X^{J} = X + F(X) \tag{1}$$

and combines and input and re-calibrated features in a single framework. F(X) which is recalibrated features defined as

$$F(X) = F_M(X) \cdot F_A(X) \tag{2}$$

where $F_M(X)$ and $F_A(X)$ are the outputs of BiGRU and Sigmoid function of attentional module. Be note that both $F_M(X)$ and $F_A(X)$ are 2D vectors $\in R^{T\times N}$

In temporal attention module we also use a single CNN for depth based images and then fuse both the above 3 RNNs channels output and CNN output to pass through the spatial temporal module. For the CNN architecture we start from a state of the art base architecture, like Resnet [19], Densnet [7], and VGGnet [18] etc. It will not effect the model performance so any kind of CNN architecture can be used. For the UTD-MHAD dataset we use the Resnet-18 [4] while for the NTU RGB+D dataset we use the Resnet-101 [4]. There is no any special reason behind why to choose two different networks it just depends on the amount and type of dataset. But for the temporal attention module resent style architecture should be used which is shown in the Fig. 1 and already proved a better architecture [].

3.2. Spatio-Temporal Convolution Module

Conventional attention methods in skeleton action recognition are limited by the modeling capacity of RNNs [24]. This module is introduced based on CNNs to extract the enhanced spatio-temporal features from the output of RNNs

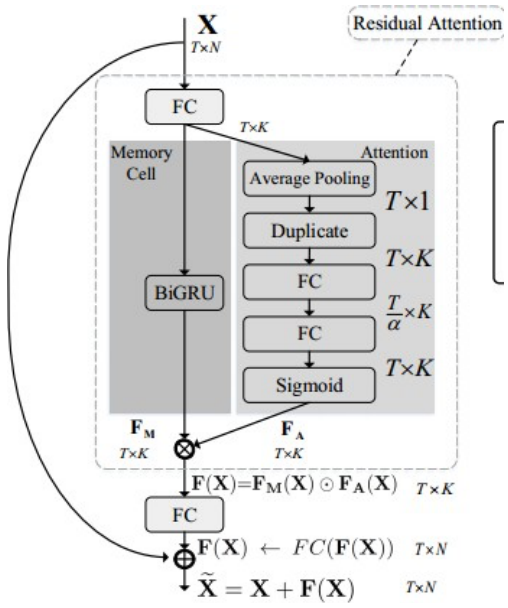

Figure 1. Temporal Attention Recalibration Module

channel named as X,Y,Z and then pass these outputs along with the output of depth data CNN to this module. By leveraging the robustness to deformation of CNNs, this module further extracts high-level feature representations and achieve better results than simply using the skeleton or depth data descriptors. There is no strict rules in choosing the right CNN architecture, any CNN should be work with this kind of model. But for simplicity we will only mention the Resnet-18 results here, although we have done experiments with Resnet-101, VGGNet etc. Fig. 2 depicts the whole network that we have used in our experiments.

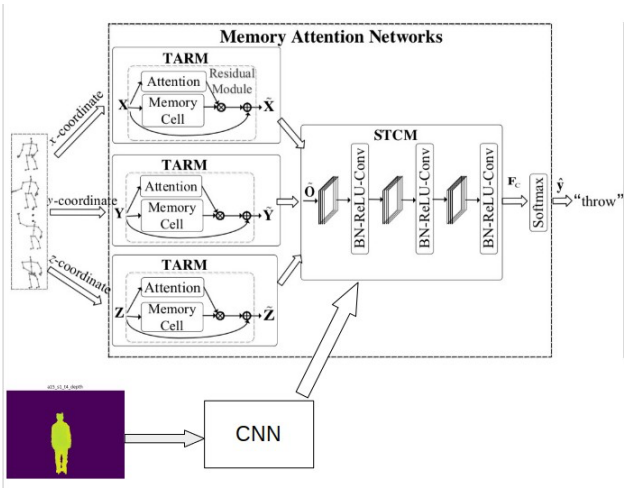

Figure 2. Our Proposed Architecture

4. Experiments

4.1. Datasets

Collected via a Microsoft Kinect sensor, a wearable inertial sensor, which includes an accelerometer and a gyroscope, and video camera, the UTD-MHAD dataset contains 27 different actions performed by 8 subjects (4 females and 4 males) [1]. Each subject was asked to repeat each action 4 times. After removing three corrupted sequences, the dataset contains 861 sequences. There are 4 different types of data, the Depth dataset, Inertial dataset, Skeleton dataset, and RGB dataset. We also validate our model on another dataset NTU RGG + D [16] which contain 600 classes and 17 subjects.

4.2. Data Preprocessing

We use the 3D skeleton data as x, y, z rectangular coordinates and generate descriptors for the depth data. For depth images we stack different frames of same action together and then make small images dataset. For the model easiness we also rescale the images and convert into grey scale by removing the coloured background (see Fig. 3). For the skeleton data we directly input the x, y, z rectangular coordinates and after passing these from temporal attention module the CNN part takes the output as sequence of images.

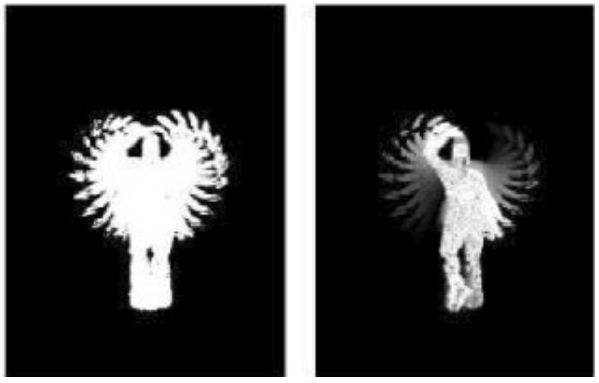

Figure 3. An example of prepossessed depth image

4.3. Experimental Analysis

We focus mainly on optimizing our model for the UTD-MHAD dataset due to its complexity with multiple modalities and the fact that many state-of-the-art methods have struggled to get above 90% score using more than one modality[10]. On skeleton data alone, our model with attention branch only scores a comparatively high accuracy result even though we do not convert the skeleton data into any descriptor like the others do[10]. This makes our model ready for raw skeleton data from any dataset, and cuts

down the time drastically involved in generating descriptors. Our model was implemented in Python using Keras with a Tensorflow backend. We trained the model on a PC with NVIDIA GTX 1080 GPU with Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz with 16 GB of RAM and NVIDIA GTX 1080 with 8 GB of RAM.

In our experiments, we empirically set learning rate, $lr = 0.01$ for Adam optimizer, with a 0.2 factor reduction on plateau. We also set $\beta_1 = 0.9$, $\beta_2 = 0.999$, $s = 1e\text{-}08$, decay=0.0 with categorical loss entropy. More details about our results on the datasets we evaluated our model on and comparative experimental studies with state-of-the-art literature are presented next. Each skeleton sample is scaled to 50×50 and a sample size of 431 was used as training for the UTD-MHAD dataset, leaving 430 for validation[10]. The selection strategy used was cross-subject with odd subjects selected for training and even subjects for validation. For the STCM of the attention temporal branch, we chose ResNet-18[4] after successful trials proved that it was the fastest to train and also provided the best results. For larger networks like DenseNet-161[6] we upscaled each skeleton data to 224×224 to prevent negative dimensions. However, we found that large networks did not train so well as our STCM. Also, the use of spherical coordinates instead of rectangular coordinates was explored and the results from the latter proved superior to the former. The next step in our analysis was to incorporate the depth maps branch to our end to end model. Originally, the data for depth maps from the UTD-MHAD are represented as $240 \times 340 \times T$ where T is the number of frames in the action under investigation. In this case, we trained different CNN's on the depth data only to get the best model. The models tested were ResNet-18, ResNet-50[4], modified AlexNet[11] and the last model used was a mix of a TARM followed by ResNet-18. After training on the multiple models, the mixture of TARM and ResNet-18 produced better results, although we fed our attention module only spatial depth map descriptors. The final step is to apply a score fusion technique such as average, multiply or maximum at the final stage to improve our predictions.

To investigate the effect of using different state-of-the-art models on the UTD-MHAD dataset predictive accuracy, we experiment with each them branch wise on skeleton and DMI descriptors[10]. The results are summarized in the next section. In the future, we hope to make this model our own by amplifying it's effectiveness with an improved architecture after thorough experimentation, and publish the results to the community.

5. Results

The results summarized in Table 5 shows the improvement made on training the raw skeleton data compared with the result from [10]. As can be seen, our proposed architecture has a higher validation accuracy. The training and

Method	Accuracy
Score Fusion Model[10]	50.00%
Our Model	76.25%

Table 1. COMPARISON OF THE PROPOSED METHOD WITH EXISTING METHOD ON UTD-MHAD DATASET

accuracy over 230 epochs are also shown in figures 6 and 7 shows that the addition of a 4th TARM for depth data to the model has better results (81.46%) compared to that of just using 3 TARMs with skeleton data. Although the DMI descriptor does not have a temporal aspect, we can see the improvements made. In table 5, we can observe the validation accuracy for various state-of-the-art models trained solely on the depth maps data and our model that trains on the depth maps using TARM+ResNet-18. The confusion

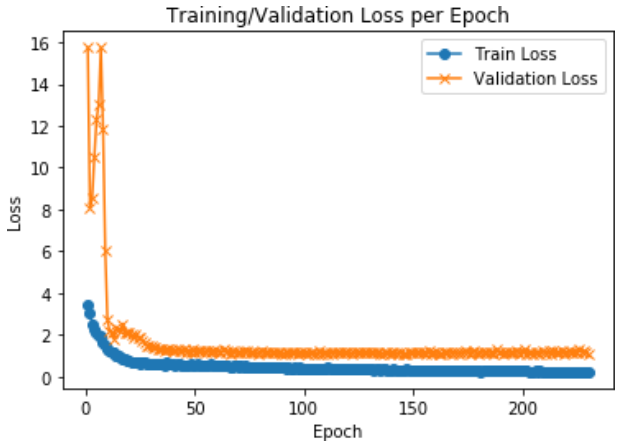

Figure 4. TRAINING AND VALIDATION LOSS FOR OUR MODEL WITHOUT DEPTH DESCRIPTOR CHANNEL

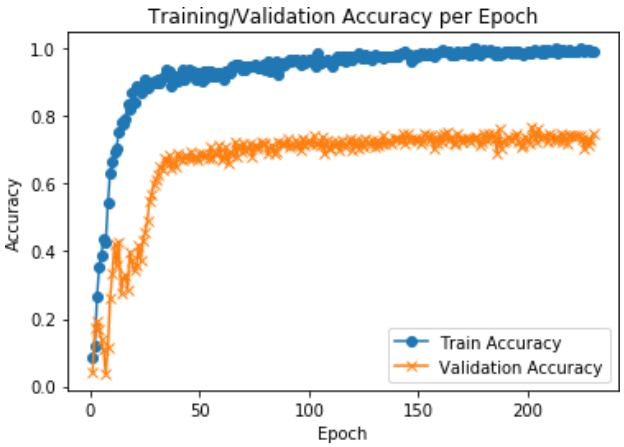

Figure 5. TRAINING AND VALIDATION ACCURACY FOR OUR MODEL WITHOUT DEPTH DESCRIPTOR CHANNEL

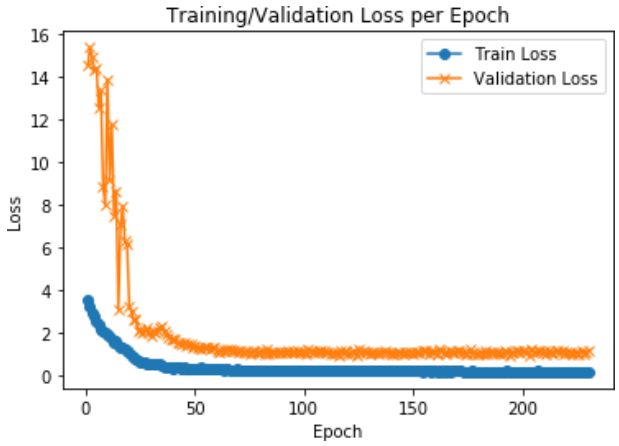

Figure 6. TRAINING AND VALIDATION LOSS FOR OUR MODEL WITH DEPTH DESCRIPTOR CHANNEL

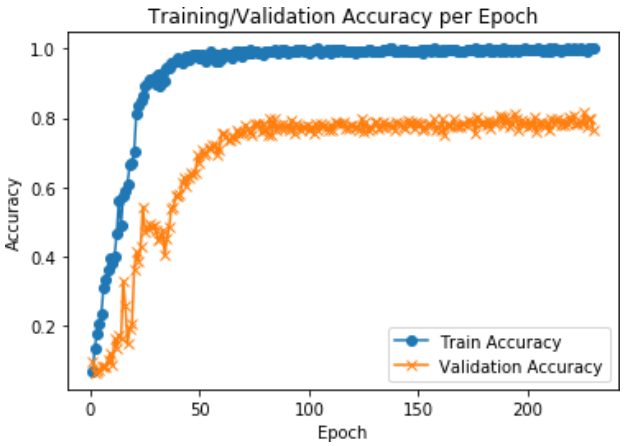

Figure 7. TRAINING AND VALIDATION ACCURACY FOR OUR MODEL WITH DEPTH DESCRIPTOR CHANNEL

Method	Accuracy
ResNet-18	33.41%
ResNet-50	36.56%
Modified AlexNet	34.31%
Ous (Modified TARM + ResNet-18)	49.20%

Table 2. COMPARISON OF THE PROPOSED METHOD USING TARM ON DEPTH DESCRIPTOR WITH EXISTING METHOD ON UTD-MHAD DEPTH DESCRIPTOR IMAGES

matrix is shown in figure 8 is for the results of training our model on UTD-MHAD. This dataset is much more challenging and can be observed in the results. From the confusion matrix we can see that the proposed method can not distinguish some actions well, for example, jog and walk. A probable reason is that the proposed the depth maps descriptor cannot differentiate both actions effectively.

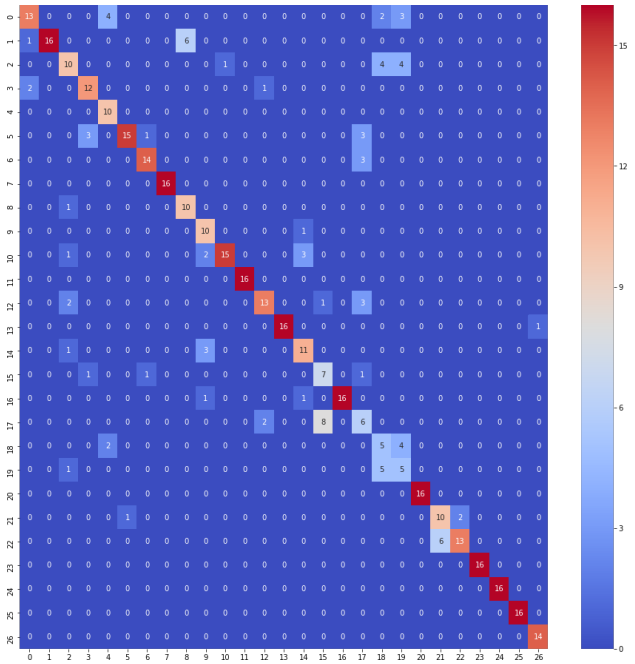

Figure 8. CONFUSION MATRIX OF PREDICTIONS OF OUR MODEL USING SKELETON DATA

6. Conclusion

We propose an end-to-end framework that utilizes skeleton data combined with DMI descriptors for human activity recognition. One angle we explored, compared to other state-of-the-art methods, is that we use skeleton data without any prepossessing and a CNN attached to extract the spatio-temporal features. Using a multi-modal structure, the depth maps branch is added to augment the accuracy of our results as shown in the Results section. Without enough time to fully explore the improvements made from addition of score fusion, we can clearly see that addition of some form of fusion is more likely to improve our results as seen in[10][12][5]. Effective training of the deep network with in our environment is the major setback, besides time limitations. We were, thus, unable to fully train our model on the NTU RGB+D data, even after taking ample time to acquire the huge sized datasets[16]. We, however, foresee promise in being able to train the skeleton data and depth maps together as a unified model as [21] was able to do so with just skeleton data. The depth maps is usually more enormous than the skeleton data, and processing them as descriptors might be an added bottleneck, which we hope to remove by augmenting the TARM to deal with depth data as well. Another possible future work, we will look at is the employ of multi-headed self attention as TARM instead of the added LSTM. We tried implementing the model for the proposed architecture, but still lacks tweaking of hyperparameters to give significant results. Finally, instead of using the carte-

sian coordinates in the raw form, we can project them onto planes instead, to give a more defined spatio-temporal structure for both TARM and SCTM. As already seen in [12], this strategy improves the results, although in the mentioned paper they added the use of generating complex descriptors. The same idealogy will be applied to depth data. In the end, our solution improves the results of validation and also reduce training time since the parameters are fewer and mostly non-costly.

References

[1] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.

[2] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM, 2016.

[3] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[5] Y. Hou, Z. Li, P. Wang, and W. Li. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):807–811, March 2018.

[6] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, July 2017.

[7] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[8] 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*, 35(1):221–231, 2012.

[9] W. Jiang and Z. Yin. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1307–1310. Acm, 2015.

[10] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng. Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1–14, 2018.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] C. Li, Y. Hou, P. Wang, and W. Li. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24(5):624–628, May 2017.

[13] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[14] F. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115,2016.

[15] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[16] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[17] 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.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[20] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[21] C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, and J. Liu. Memory attention networks for skeleton-based action recognition. *arXiv preprint arXiv:1804.08254*, 2018.

[22] M. Yang, K. Kpalma, and J. Ronsin. A survey of shape feature extraction techniques, 2008.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[24] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

[25] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.