Towards Robust Human Activity Recognition from RGB Video Stream with Limited Labeled Data

Krishanu Sarker
Dept. of Computer Science
Georgia State University
Atlanta, GA, USA
ksarker1@student.gsu.edu

Mohamed Masoud
Dept. of Computer Science
Georgia State University
Atlanta, GA, USA
mmasoud1@student.gsu.edu

Saeid Belkasim
Dept. of Computer Science
Georgia State University
Atlanta, GA, USA
sbelkasim@gsu.edu

Shihao Ji
Dept. of Computer Science
Georgia State University
Atlanta, GA, USA
sji@gsu.edu

Abstract—Human activity recognition based on video streams has received numerous attentions in recent years. Due to lack of depth information, RGB video based activity recognition performs poorly compared to RGB-D video based solutions. On the other hand, acquiring depth information, inertia etc. is costly and requires special equipment, whereas RGB video streams are available in ordinary cameras. Hence, our goal is to investigate whether similar or even higher accuracy can be achieved with RGB-only modality. In this regard, we propose a novel framework that couples skeleton data extracted from RGB video and deep Bidirectional Long Short Term Memory (BLSTM) model for activity recognition. A big challenge of training such a deep network is the limited training data, and exploring RGB-only stream significantly exaggerates the difficulty. We therefore propose a set of algorithmic techniques to train this model effectively, e.g., data augmentation, dynamic frame dropout and gradient injection. The experiments demonstrate that our RGB-only solution surpasses the state-of-the-art approaches that all exploit RGB-D video streams by a notable margin. This makes our solution widely deployable with ordinary cameras.

Keywords: human action recognition; computer vision; deep learning; LSTM; limited data; RGB

I. INTRODUCTION

Human action is inherently complex due to the inter-class affinity and intra-class diversity. Recognizing activity is hence a difficult task, which has attracted numerous researchers’ attention [1], [2], [3]. Even though state-of-the-art image classification methods have surpassed human level accuracy [4], performance of methods proposed in the literature for activity recognition/classification is still unsatisfactory, especially methods based on RGB video stream [1].

Despite many efforts, human action recognition from RGB video streams still lacks in accuracy compared to the progresses made in multi-modal data that includes depth enabled RGB-D video. One of the reasons behind this is that multi-modal datasets provide higher quantity of information, and the extracted depth information give precise detection of movement in the scene. However, depth enabled cameras are expensive and require special settings for many possible use-cases of human action recognition. The economical factor and the installation complexity are the main reasons that most of the surveillance systems are using RGB cameras. Therefore, focusing our attention on the more popular RGB videos for detecting and classifying human motion would benefit the users and the real life applications.

Traditionally, the works on RGB video stream are based on handcrafted features [5], [6], [7], [8], [9]. These approaches are highly data dependent. Due to this problem, these methods are very brittle and hard to deploy in real life in spite of higher accuracy they achieve. With the advent of deep learning, methods were proposed where features could be automatically extracted [10], [11], [12], [13]. Successful use of deep learning with image classification inspired researchers to deploy such methods in video classification [2]. These methods use raw RGB frames, often coupled with motion, to learn the temporal features. Even though these methods automate the feature extraction task, they often struggle to gain high performance due to complex background and partial occlusion of subjects in video streams. Hence, more robust, automated action recognition system is yet to be developed.

Approaches based on multiple modalities of data [14], [15], [16], [17], however achieves higher accuracy even with complex actions. In these approaches, skeleton information extracted from depth images are proven very efficient in extracting important features of action. Inspired by this, in the paper we propose to use a technique that aims at separating salient features from the scene by extracting skeleton key-points from RGB-only video streams. This is a distinct departure from all previous approaches that either use raw RGB video stream as input directly or use skeleton key-points extracted from the depth information for activity recognition. Specifically, we utilize Openpose API [18] as a black box to extract the skeleton key-points from each RGB frame. These key-point features are then fed into a Bidirectional Long Short Term Memory (BLSTM) based model to learn the spatio-temporal representations, which are subsequently classified by a softmax classifier.

We use RGB-only modality for our experimental evaluations whereas state-of-the-art methods utilized multiple available modalities (depth, inertia and skeleton data). This essentially reduces training data to one forth for our experiments. Hence, we are dealing with one of the key challenges of deep learning, i.e., training with limited labeled data. To train the deep network effectively, we explore data aug-
Deep learning based approaches for human activity recognition have also been explored extensively [10], [11], [12]. For example, Baccouche et al. [10] propose to use Convolutional Neural Network (CNN) to extract spatial features and then use LSTM to learn the temporal features. Ji et al. [11] present 3D CNN to classify actions which learns inherent temporal structure among the consecutive frames. A two-stream CNN based method is proposed in [12]. In contrast to state-of-the-art handcrafted feature design approaches, deep learning based approaches use an end-to-end learning pipeline and extract feature representations automatically from data. However, these methods often fail to achieve higher accuracy as the high level features extracted from CNN are blurry and incapable of capturing the sharp changes in video streams. This is primarily because convolution and pooling tries to accurately capture the overall structure, while repetitive convolution and pooling operations often ignore the fine-grained details.

In order to solve the aforementioned issues, skeleton information from RGB-D video has been widely studied to improve recognition accuracy [15], [16], [17], [19], [20]. Observations from seminal work by Johansson [21] suggests that a few movement of human joints is sufficient to recognize an action. Recently, Liu et al. [20] propose a CNN based approach leveraging the skeleton data. In [19] the authors propose hierarchical bidirectional Recurrent Neural Network (RNN) to classify the human actions. Methods proposed in [22] and [15] utilize skeleton data on three CNN streams that are pretrained on large ImageNet Dataset [23]. Li et al. [16] use view invariant features from skeleton data to improve over [22] and [15], and they used similar four stream pretrained models. All these methods utilize skeleton data, either extracted from depth data or Kinect. Inspired from these works, we adopt a bidirectional LSTM in our method; instead of extracting skeleton data from depth information as in other methods, we extract skeleton keypoints from RGB frames, which are available in ordinary digital cameras.

In addition, there exist a few CNN and LSTM based approaches for activity recognition from RGB-only data [24], [25]. However, none of them pay special attention to the issue of training deep network effectively on limited labeled data. We emphasize more on algorithmic approaches to address the training issues of deep networks with limited training data to alleviate overfitting and gradient vanishing problems. Enhanced by these techniques, our RGB-only solution is able to surpass the state-of-the-arts that all exploit RGB-D streams.

III. METHODOLOGY

In this section, we present an end-to-end framework for human activity recognition from RGB video containing human silhouette. To make our discussion self-contained, we review some important concepts in the following subsections.

II. RELATED WORKS

Human activity recognition has been extensively studied in the recent years [1], [2], [3]. Most of state-of-the-art methods exact handcrafted features from RGB videos and rely on traditional shallow classifiers for activity classification [5], [6], [7], [8], [9]. For example, Schuld et al. [5] present a method that identifies spatio-temporal interest points and classifies action by using SVMs. Zhang et al. [6] introduce the concept of motion context to capture spatio-temporal structure. Liu and Shah [7] consider the correlation among features. Bregonzio et al. [8] propose to calculate the difference between subsequent frames to estimate the focus of attention. These methods often achieve very high accuracy. However, since handcrafted features are highly data dependent, these methods are not very robust to the change of environments. We instead utilize OpenPose to extract the salient skeleton features from raw RGB frames, which makes the proposed method less data dependent, robust to different environments and therefore widely deployable in real life applications.
A. Overview

Our proposed architecture aims to classify human actions from RGB-only streams to make our approach most amenable to ordinary cameras. We formulate our problem as learning the mapping, \( F : x \rightarrow \ell \), where \( x \) is the raw video and \( \ell \) is the collection of action categories. After training, \( F \) is used to classify the test samples.

Fig. 1 shows the overall pipeline of the proposed method. First, we extract pose key-points of human silhouette from input raw RGB video using OpenPose API [18]. We then preprocess the extracted pose key-points to improve the quality of the feature representations. After preprocessing, we use a variety of data augmentation techniques on the extracted keypoints to increase the training data size (and therefore mitigate the problem of data scarcity). In the end, the augmented training set is used to train our classifier. We used deep BLSTM [26] network coupled with MLP [27] as our classifier. Overfitting is a major drawback for LSTM when dealing with small dataset. In addition to data augmentation, we therefore explored additional regularization techniques, such as dropout and L2 regularization to prevent our model from overfitting. We also propose Dynamic Frame Dropout to reduce the redundant frames from a video and improve the robustness of the BLSTM classifier. To mitigate the vanishing gradient issue of LSTM, we introduce Gradient Injection to improve gradient flow. We will discuss each of these components in greater details in the following subsections.

B. OpenPose

OpenPose [18] is an open source API that can be used to detect the 2D poses of multiple human subjects in an image. The API leverages a novel two stream multi-stage CNN, which facilitates it to work in real time. The methodology proposed in [18] was ranked number one in COCO 2016 keypoints challenge. The input of the architecture is raw RGB image and the output is 15 or 18 pose key-points along with the part joining edges. More details about the architecture and working principle can be found in [18]. In our work, we treat OpenPose as a black box with raw video frames as inputs and 18 pose key-points per person as output.

C. LSTM

Long Short-Term Memory (LSTM) [26] is a descendant of Recurrent Neural Network (RNN) especially designed to adapt long range dependencies when modeling sequential data. RNN, in general, has been proven very successful in modeling sequences that have strong temporal dependency. However, vanishing gradient problem makes Vanilla RNN hard to train [28]. LSTM mitigates this issue by introducing non-linear gates regulating the information flow. In addition, vanilla LSTM can only learn from past contexts, whereas Bidirectional LSTM (BLSTM) [29] can be used to learn both from past and from future context by utilizing forward and backward layers. For human activity recognition task, we found that BLSTM is a more suitable architecture than vanilla LSTM as incorporating long term dependency in both directions in general helps improve learning of sequential data.

D. Preprocessing

The preprocessing step represents the first step of our end-to-end pipeline where the raw video frames are fed into the OpenPose API. The output of OpenPose for each video frame is a matrix of shape \((n_{\text{pose}}, (a, b), c)\). Here, \(n_{\text{pose}}\) is the number of pose key-points, \((a, b)\) is the coordinates of the key-points in Cartesian plane and \(c\) is the confidence score of the respective key-point. To simplify our problem, we put a constraint that each frame can contain at most one person, and hence the value of \(n_{\text{pose}}\) here is 18. When all pose key-points are extracted from a video, we use a filter to set the pose keypoints values that has confidence lower than a threshold value, \(\Theta\), to zero. Later, we mask these zero valued keypoints in order to avoid learning from these points. Afterwards, the pose matrix is flattened and converted into a vector, \(A\), of size \(n_{\text{pose}} \times 2\), excluding the confidence value. We concatenate each pose frame into a 2 dimensional matrix of shape \((n_{\text{frame}}, v)\), where \(n_{\text{frame}}\) is the number of frames in the video and \(v\) is the length of pose vector, \(A\).

E. Proposed Network Architecture

Our proposed deep architecture combines deep BLSTM layers and MLP. We use five consecutive BLSTM layers with dropout to regularize the model training. We utilize Batch Normalization (BN) after each BLSTM layer to keep the data normalized throughout the pipeline. We feed the output of the Deep BLSTM layers to the MLP consisting of two Dense layers. For intermediate hidden BLSTM and Dense layers, we have utilized the Parametric Rectified Linear Unit (PReLU) [30] activation layer. We use the softmax function for the final output layer to produce probabilistic score for each class. Categorical cross-entropy is used to measure the loss of our proposed network. We utilized RMSprop optimizer [31] to minimize the loss function.
IV. Effective Training of BLSTM

It’s challenging to train the BLSTM architecture as the number of parameters can be easily larger than several hundred millions, while the number of training data for the purpose of model parameter estimation is typically very small (e.g., 2-3 orders of magnitude lower). Therefore, special attention is needed to address the effective training of BLSTM, otherwise overfitting, gradient vanishing can quickly plague the learning process. In the following, we discuss a few techniques we explored to train the deep model effectively, with the high level scheme demonstrated in Fig. 2.

A. Dynamic Frame Dropout

We propose to utilize Dynamic Frame Dropout (DFD) to reduce data redundancy. As different actions require different time span, and often there are redundant information in consecutive frames, taking all frames into account actually oculdes crucial information and hampers the learning. Techniques like randomly dropping frames or dropping each n frames etc. are often used in state-of-the-art methods. However, doing so naively may result in loss of important information. Instead of randomly dropping frames, DFD drops frames only if they contain information that are almost redundant to their preceding frames. This not only reduces the computation complexity but also introduces stochasticity and help to regularize the model training in a similar spirit of dropout.

Specifically, we measure the redundancy between two consecutive frames by computing the euclidean distance between their feature vectors, i.e., pose key-points of two consecutive frames. The lower distance corresponds to similarity and the higher distance means these frames actually have meaningful differences. Empirically, we set a cutoff threshold, \( \hat{c} = 15 \). If \( d \) is distance between \( frame_1 \) and \( frame_2 \) and \( d < \hat{c} \), then we drop \( frame_2 \). According to our experiments that follow, this setup of \( \hat{c} \) drops 20 to 25 frames per video that carry information with minimal significance.

B. Gradient Injection

Although LSTM serves as the solution of vanilla RNN for gradient vanishing problem, it itself faces this issue in some degree when training deep model [32]. LSTM many to one architecture is often used as the final layer of network for video classification. This creates a dependency on processing the whole video sequence before we can perform classification. However, a video can often be clearly classified before having to see all the frames till the end. Hence, to improve gradient flow and avoid gradient vanishing problem, we connect the MLP classification layer to the last \( K \) time-steps and this allow the model to classify a video by incorporating multiple step information. When back-propagating error to update model parameters, this also allows the gradient to be propagated earlier in time and mitigate the gradient vanishing problem. We call this technique Gradient Injection (GI). In other words, we utilize many to many architecture of LSTM at the top layer to allow gradients flow from multiple time steps, consequently, reducing the problem of vanishing gradient. Moreover, as outputs from multiple time steps are now available, it creates an ensemble of multiple outputs and reduces dependency on all the video frames.

C. Data Augmentation

Training a deep networks with limited amount of labeled training data is a major challenge in supervised learning paradigm. Our goal of achieving state-of-the-art performance with RGB-only data modality faces the same brick wall: insufficient training data. According to our problem formulation, we only leverage RGB data modality. Data augmentation has been proven very successful in supervised learning for image analysis. Inspired by this, we have explored several data augmentation techniques to solve the data scarcity problem. In our case, instead of the raw input video, we take skeleton key-point features as the input for data augmentation. We use translation, scaling and random noise to augment skeleton data. To keep the augmentation consistent throughout a single sample, we deploy same transformation on each key-point frames of that sample.

In the experiments that follow we evaluate the significance of each of these techniques when training deep networks using limited training data.

V. Experimental Results

The primary goal of this paper is to show that by using RGB-only data modality with limited training data, we can achieve similar or higher accuracy on action recognition task than the state-of-the-arts that use RGB-D video streams. We have tested our proposed method with two widely used datasets, KTH [5] and UTD-MHAD [14]. We focus on UTD-MHAD as this is a complex dataset offering multiple modalities and current state-of-the-art methods utilize data modalities consisting depth information to classify actions. Extensive experiments show that with the effective training techniques discussed in Section IV, i.e., data augmentation,
Figure 3: Accuracy comparison on the UTD-MHAD dataset on our models with different number of LSTM layers.

dynamic frame dropout and gradient injection, the proposed RGB-only solution surpasses the state-of-the-art methods by a notable margin. On RGB-only dataset such as KTH, our method outperforms all the other methods reported in the literature, demonstrating the versatility of the proposed architecture and techniques for human activity recognition.

We implemented our system in Python with Tensorflow backend on a GPU cluster with Intel Xeon CPU E5-2667 v4 @ 3.20GHz with 504 GB of RAM and NVIDIA TITAN Xp with 12 GB of RAM and 3840 cuda cores. In our experiments, we empirically set learning rate, $lr = 5e^{-5}$ for RMSprop optimizer. We report confidence interval based on 50 bootstrap trials. More details about datasets we evaluated our model on and comparative experimental studies with state-of-the-art literatures are presented next.

A. Dataset

KTH [5] is an RGB-only video dataset containing six action classes (walking, running, boxing, hand-waving, and hand-clapping), performed by 25 subjects in various conditions. KTH dataset provides full silhouette figure in all the sequences, which satisfies our requirements of pose-based activity recognition. We have followed the same experimental setup stated in [5].

UTD-MHAD [14] is a multi-modal action dataset containing 27 actions performed by 8 subjects (4 males and 4 females) performing same action 4 times, a total 861 sequences. This dataset provides four temporally synchronized data modalities; RGB videos, depth videos, skeleton positions, and inertial signals from Kinect camera and a wearable inertial sensor. We follow 50-50 train-test split similar to [14]. In the experiments we only use the RGB modality to evaluate our proposed method.

B. Experiments on UTD-MHAD dataset

We first explore the choices of depth of the network. We test our baseline BLSTM model with three settings: 3-Layer, 5-Layer and 7-Layer models. Fig. 3 presents the accuracy of these models on top 1, 3 and 5 categories. Evidently, the 5 and 7 layer models outperform the 3 layer model, and the performances of the 5 and 7 layers models are almost on par. Therefore, the 5-layer model reaches a good accuracy and model complexity balance. We use the 5-layer model architecture in all our following experiments. Notice that the 3-layer model has shown comparative accuracy on top 3 and 5 categories with other two models, indicating that deeper models mainly boost the top-1 accuracy.

To understand the impact of different techniques over the baseline BLSTM model (e.g., DFD, GI and data augmentation), we evaluate the accuracy of the BLSTM model as we enable them in a cumulative fashion. We begin our experiments on UTD-MHAD dataset using our baseline BLSTM model. Then the second model includes DFD; the third one includes both DFD and GI; in the fourth model we use random jittering to augment data, and finally in the fifth and last model we use affine transformation as data augmentation. Fig. 4 shows the comparison among all these models on accuracy and F1 score. As we can see, the baseline BLSTM model itself reaches an accuracy of 87% (about 1% behind the state-of-the-art accuracy 88% [16]); by including DFD on top of baseline, our method achieves an accuracy of 89.06% outperforming state-of-the-art by 1%. On top of that, GI and data augmentation further boost the performance to 91%. An interesting phenomena to observe here is that although GI does not have much effect on top of data augmentation, it helps gaining performance over DFD. In total, utilizing data augmentation we gain 2% accuracy over DFD model, while using random data jittering is less effective and does not improve accuracy.

To investigate the effect of data augmentation on the predictive accuracy, we experiment with incremental data augmentation. The results are summarized in Table I. As can be seen, data augmentation regularizes the model training and helps model avoid overfitting. As a result, when we
Figure 5: Accuracy comparison on the UTD-MHAD dataset. [14], [15], [16], [22] and [33] use depth enabled modalities, while our method use RGB-only modality (confidence interval of our method is also included).

We increase the training dataset by data augmentation, the top-1 error rate is consistently reduced. We also notice that data augmentation does not have much effect on top-3 error rate, indicating that data augmentation mainly boosts correct answers from top-3 positions to top-1 positions.

Table I: Effect of Data Augmentation on the UTD-MHAD dataset.

| Augment Size | Top-1 Error (%) | Top-3 Error (%) |
|--------------|-----------------|-----------------|
| 0            | 10.94           | 4.01            |
| 430          | 9.75            | 3.75            |
| 860          | 9.35            | 3.68            |
| 1290         | 9.09            | 3.69            |
| 1720         | 9.05            | 3.66            |

Finally, we summarize the results of our proposed method with state-of-the-art methods [14], [15], [22], [16], [33] in Fig. 5. Most of these methods use depth or inertia data modalities or both (section II). These data modalities are only available from depth enabled camera and provide more precise information of motions related to actions. On the contrary, we use RGB-only modality to train our model from scratch. It can be seen from Fig. 5 that our method achieves an accuracy of 90.95% which outperforms all the state-of-the-art methods.

C. Experiments on KTH dataset

To further strengthen our hypothesis, we then compare our proposed method with the state-of-the-arts on the RGB only dataset, KTH [5], with the results presented in Fig. 6. We utilized similar training-testing split of the data as suggested in [5] to obtain the reported results. CNN based hybrid model proposed by Lei et al. [34] achieves 91.41% accuracy, which is outperformed by most of the state-of-the-art handcrafted feature based methods. On the other hand, [9], [35], [6], and [8] using handcrafted features achieve competitive accuracy.

Figure 6: Accuracy comparison on KTH dataset with state-of-the-arts. [34] use CNN based method, while [6], [8], [9] and [35] utilize hand crafted features. (confidence interval of our method is also shown).

However, these methods are extremely data dependent; proposed handcrafted feature extractors in these methods cannot robustly work on heterogeneous data. Hence, these methods are not suitable for real world deployment. Our proposed method with data augmentation and dynamic frame dropout achieves 96.07% accuracy, outperforming all the others.

VI. CONCLUSION AND FUTURE WORK

We propose an end-to-end framework that utilizes pose key-points extracted from OpenPose coupled with BLSTM for human activity recognition. A major difference to the state-of-the-art methods is that we use RGB-only modality while all the other methods use RGB-D modality. Effective training of deep networks in our setting is the major technical challenge as we typically have very limited training data and exploiting RGB-only modality exaggerates the difficulty even further. We therefore explore a number of algorithmic techniques like Dynamic Frame Dropout, Gradient Injection and Data Augmentation to train our framework effectively. Extensive experiments demonstrate the effectiveness of our BLSTM model and training methodologies, among which data augmentation is the most effective one. In the end, our RGB-only solution surpasses all the state-of-the-art methods that exploit RGB-D streams. This makes our solution cost effective and widely deployable with ordinary digital cameras.

Our experiments were conducted on the KTH and UTD-MHAD datasets, where there is only one person present per action and whole silhouette is visible. In the future, we would like to extend our method for multi-person datasets where some body parts can be partially occluded, which happen more often in real video surveillance applications.

ACKNOWLEDGMENT

The authors would gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.
REFERENCES

[1] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, “Advances in human action recognition: A survey,” arXiv preprint arXiv:1501.05964, 2015.

[2] S. Herath, M. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” Image and Vision Computing, vol. 60, pp. 4–21, 2017.

[3] C. Chen, R. Jafari, and N. Kehtarnavaz, “A survey of depth and inertial sensor fusion for human action recognition,” Multimedia Tools and Applications, vol. 76, no. 3, pp. 4405–4425, 2017.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[5] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3. IEEE, 2004, pp. 32–36.

[6] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, “Motion context: A new representation for human action recognition,” Computer Vision–ECCV 2008, pp. 817–829, 2008.

[7] J. Liu and M. Shah, “Learning human actions via information maximization,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on., IEEE, 2008, pp. 1–8.

[8] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 1948–1955.

[9] H. Wang, A. Klüser, C. Schmid, and C.-L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” International journal of computer vision, vol. 103, no. 1, pp. 60–79, 2013.

[10] M. Baccouche, F. Mamalet, C. Wolf, Christian Garcia, and A. Baskurt, “Sequential deep learning for human action recognition,” in International Workshop on Human Behavior Understanding. Springer, 2011, pp. 29–39.

[11] 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.

[12] 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.

[13] M. E. Masoud, K. Sarker, S. Belkasim, and I. Chahine, “Automatically generated semantic tags of art images,” in IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017.

[14] 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 Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015, pp. 168–172.

[15] 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.

[16] C. Li, Y. Hou, P. Wang, and W. Li, “Joint distance maps based action recognition with convolutional neural networks,” IEEE Signal Processing Letters, vol. 24, no. 5, pp. 624–628, 2017.

[17] H. Rahmani and M. Bennamoun, “Learning action recognition model from depth and skeleton videos,” in The IEEE International Conference on Computer Vision (ICCV), 2017.

[18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in CVPR, 2017.

[19] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1110–1118.

[20] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” Pattern Recognition, vol. 68, pp. 346–352, 2017.

[21] G. Johansson, “Visual perception of biological motion and a model for its analysis,” Perception & Psychophysics, vol. 14, no. 2, pp. 201–211, Jun 1973.

[22] 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, 2016.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, 2009, pp. 248–255.

[24] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[25] I. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4694–4702.

[26] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] W. S. Sarle, “Neural networks and statistical models,” 1994.

[28] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” IEEE transactions on neural networks, vol. 5, no. 2, pp. 157–166, 1994.

[29] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” Neural Networks, vol. 18, no. 5, pp. 602–610, 2005.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[31] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” COURSEMA: Neural networks for machine learning, vol. 4, no. 2, pp. 26–31, 2012.

[32] W.-N. Hsu, Y. Zhang, A. Lee, and J. Glass, “Exploiting depth and highway connections in convolutional recurrent deep neural networks for speech recognition,” cell, vol. 50, p. 1, 2016.

[33] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, “Action recognition using 3d histograms of texture and a multi-class boosting classifier,” IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4648–4660, 2017.

[34] J. Lei, G. Li, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” IEEE transactions on neural networks, vol. 5, no. 2, pp. 157–166, 1994.

[35] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” Neural Networks, vol. 18, no. 5, pp. 602–610, 2005.