International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 8 Issue 9 ǁ 2020 ǁ PP. 54-59

Pose-Attention Networks for 2D Skeleton-Based Action Recognition in Videos

Gaofeng Li[*1] , Songlin Wang[1]
*[*1]College of Phsics and Electonic Informatian, Luoyang Normal University, Luoyang 471934, China;*
Corresponding Author: Gaofeng Li

Abstract
The aim of this paper is developing a lightweight neural network for action recognition in videos. We present a novel pose-attention model for online action recognition in embedded machine and mobile terminal. Hierarchical attention mechanism is respectively incorporated into the recurrent neural networks with pose-attention in sequence and joints-attention in 2D skeleton. The key frames are dynamical weighted and updated from consecutive frames. Joints-attention depends on human joints graph layer for pose estimation. Our model is trained on dataset of UCF and HMDB. Experiments results show that the proposed model effectively realize online action recognition and outperforms state-of-the-art methods.

Keywords: *Action recognition, Pose attention, Skeleton graph.*

Date of Submission: 04-09-2020 Date of acceptance: 19-09-2020

I. INTRODUCTION

Action recognition in videos is a challenging task in machine learning, which is widely applied in video retrieval, intelligent medical treatment, public surveillance [1]. Due to the non-rigid structure of human body and the diversity of appearance, there are still many difficulties for modeling pose in real scene. The dynamics of human motion is complex with the global movement of body and the local articulated motion of joints. Real video is affected by human appearance varying, such as rotation, scaling, deformation, occlusion and viewpoint varying [3]. Conventional methods based on RGBD depth images are becoming research interests [4]. However, two-dimension images from camera still widely exist in CCTV and surveillance. And applications are more applied in embedded and mobile machine rather than workstations and servers.

We propose an action recognition method with pose attention based on human 2D skeleton. The lightweight method reduces complexity and computation of videos. The human skeleton graph gives concise representation of different frames function for pose estimation. Self-attention mechanism is introduced to weight the pose features with frame-attention and joints-attention. We conduct training experiments on UCF and HMDB dataset, and extend to embedded machines such as Jetson Nano and Raspberry Pi.

Pose features are the fusion of typical spatial and temporal information. The descriptors of human pose are developed and based on handcrafted feature, which samples keypoints with corner extraction, such as HOG [5], SIFT [6]. Hand-crafted features especially depend a priori hypothesis of locality for human appearance, and do not always show efficiency of various poses in cluttered scene and different viewpoints. The manual features are separated from action recognition process and more complicated with spatial-temporal information. It is difficult to determine the consistency of features vectors extracting for the same pose of different targets and even the intra-class pose of the same target.

Attention mechanism allows neural networks to selectively pay attention to specific information to improve the ability of model learning [7]. [8] propose self-attention networks to assist neural machine translation instead of CNNs and RNNs. [9] designs hierarchical attention networks for document classification with word level and sentence level. A decoupled spatial-temporal attention network is designed for action recognition [10]. In the frame, observers always pay attention to the places of interest and even carefully watch some details. The relationship between frames is useful to pose estimation. The key-frames in video are the places of interest of action.

II. POSE-ATTENTION MODEL

In the paper, human pose is regarded as the combination of pose sequences rather than convolutional features extracted directly from videos. We prefer to 2D joints data by pose estimator firstly [11]. Action is closely related to pose for human visual understanding. One action can be represented by one posture from a still image, while others maybe be represented by multiple postures in fixed order, for instance, “Walking” versus “LongJump”.

We present the pose-attention model for action recognition in videos as shown in Figure 1. The video data is fed frame by frame into CNN for extracting human joints. The skeleton is input to attention layers for non-local frame attention and local attention. Attention parameters are key pose of frames and joints enhanced by the pose. The continuity between of pose is learned with pose-attention by networks.

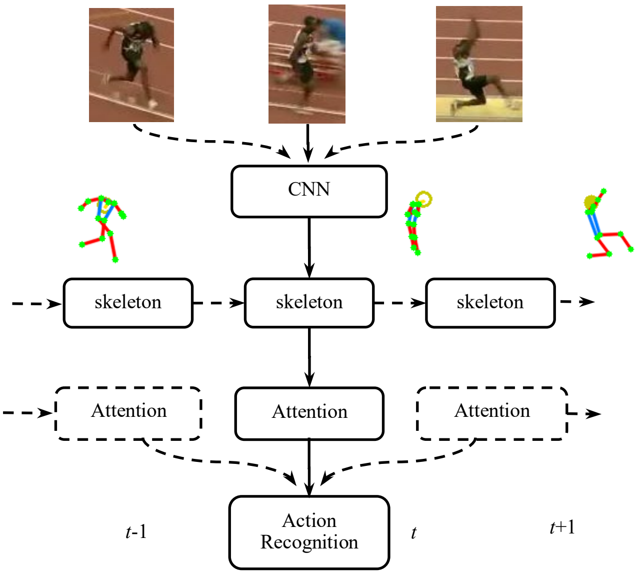

Figure 1: Pose-attention model

2.1 CNN MODULE

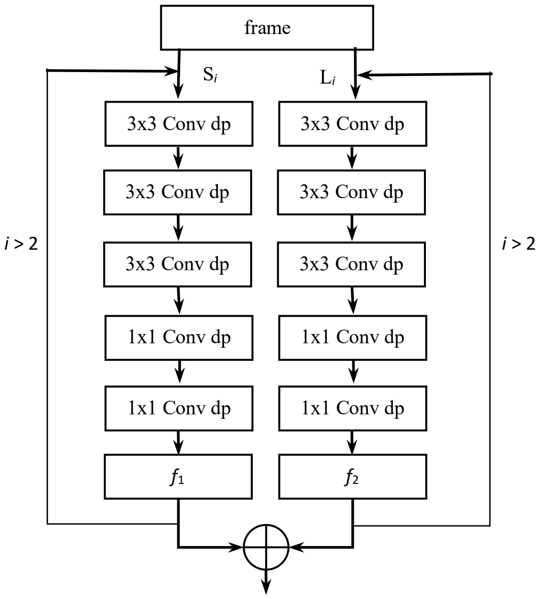

Figure 2: Pose extraction

Extracting the spatial features of human from the frame image, using lightweight convolution network and adopting depth separable convolution method can reduce the number of network parameters and the calculation amount of training [12]. The dimension of input network is 224x224x3, which has 12 layers. The first layer is a standard convolution layer with 24 convolution kernels and a step size of 2. Starting from the second layer, the depth separable convolution is performed, and the output dimension of the last layer is 28x28x384.

After feature extraction by CNN module, the pose module is used for training, and joint response and loss function are calculated in each stage [13]. In this paper, two-way network is used to detect human joint features and limb apparent features, such as direction and gradient. The calculation process is shown in Figure 2. The *S* branch calculates the eigenvalues of human joint points, which is used to learn the corresponding confidence level. The *L* branch calculates the human body component vector, that is, the correlation degree between the relevant nodes. In each stage, the depth of convolution can be separated to replace the standard convolution, and the convolution size of 2 ~ 6 stages is 3x3, which reduces the computation. Conv_dp represents depth separable convolution, including depth convolution of 3x3 and pointwise convolution of 1x1.

2.2 HUMAN GRAPH NETWORK

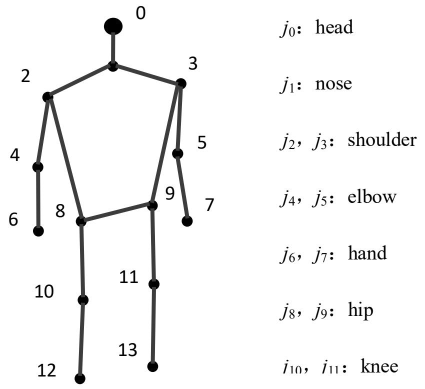

Figure 3: Human graph network

The human graph $G_j(V_j, E_j)$, j=0,…,m. m is the number of joints. j_i is the node of joints graph from pose, in which edge is the limb between two joints, as shown in Figure 3. The adjacency matrix is $A\in\{0,1\}^{N\times N}$. The nodes feature matrix is $X\in\mathbb{R}^{N\times D}$. N is number of joints; D is the number of pose feature. The neural message passing of joints-graph layer is shown as formula (1). W_l is the parameters matrix. σ is the activation function.

$$H_{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H_lW_l\right) \tag{1}$$

We construct the human graph incorporated to networks (HGN), which is constrained by human joints with soft attention and mapped to cluster set. The assignment matrix of layer is $S\in\mathbb{R}^{N\times N}$ in formula (2) to calculate next layer X_{l+1}, A_l in formula (3) and (4).

$$S_l = ReLU(G(A_l, X_l)) \tag{2}$$

$$X_{l+1} = {S_l}^T Z_l \tag{3}$$

$$A_{l+1} = {S_l}^T A_l S_l \tag{4}$$

The output is defined as following (5). We run the model to gain embedding of human graph, then pose-attention network (PAN) on the pose with learned embedding for final output representation.

$$Z_l = PAN(A_l, X_l) \tag{5}$$

2.3 POSE-ATTENTION NETWORK

We incorporate a novel attention method with GRU for improving the performance of action recognition. It is based on the self-attention mechanism with pose-attention and joints-attention. It is appropriate for action recognition and prediction and gives insight into which joints contribute to the action recognition. The architecture of attention networks is shown in Figure 4.

GRU uses reset gate f_t and update gate s_t to track the hidden state of sequences, which control long time information updating. The new state is h_t as in (6).

$$h_t = s_t \odot \tilde{h}_t + (1 - s_t) \odot h_{t-1} \tag{6}$$

It is a linear interpolation between previous state h_{t-1} and new state $\tilde{h}_t$. s_t is updated in (7). $\tilde{h}_t$ is calculated in (8). f_t impact on past information to current state in (9).

$$s_t = \sigma(W_s x_t + U_s h_{t-1}) \tag{7}$$

$$\widetilde{h_t} = \tanh\,(W_h x_t + f_t U_c h_{t-1}) \tag{8}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \tag{9}$$

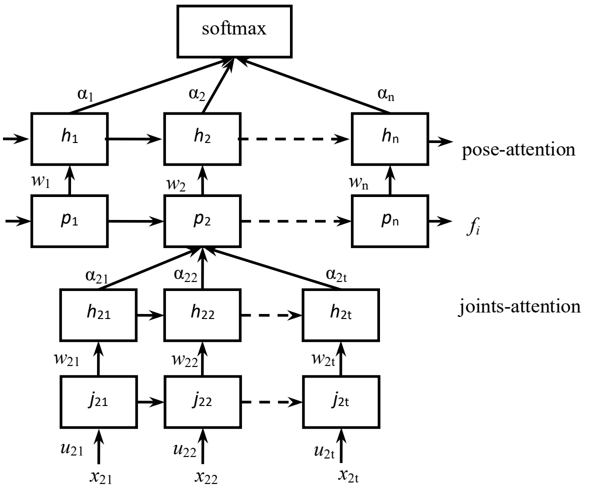

Figure 4: Pose-attention network

We give pose-attention with frame level context vector u_s. The importance of each pose is measured with α_i. as shown in (10). v is the frame vector that summarizes all the information of action in video.

$$\alpha_i = \frac{\exp\,(u_i^T u_s)}{\sum_i \exp\,(u_i^T u_s)} \tag{10}$$

$$u_i = \tanh\,(W_s h_i) \tag{11}$$

$$v = \textstyle\sum_i \alpha_i h_i \tag{12}$$

A part of joints plays an importance on the presentation of pose, other part of joints is omitted by body occlusion or background. Joints-attention is the lower level layer of pose-attention as shown in Figure 3. The joints annotation h_{it} is through dense layer to get j_{it}. The important weight α_{it} is calculated by formula (13) and (14). The joints feature vector s_i is a weighted sum of joints as (15).

$$\alpha_{it} = \frac{\exp\,(j_{it}^T u_w)}{\sum_t \exp\,(j_{it}^T u_w)} \tag{13}$$

$$j_{it} = \tanh\,(W_w h_{it}) \tag{14}$$

$$s_i = \textstyle\sum_t \alpha_{it} h_{it} \tag{15}$$

The loss function is the sum of pose loss and joints loss as shown in (16), where L_{pose} and L_{joints} are the action loss in (17) and the joints loss in (18), and λ_{pose}, λ_{joints} are the coefficients of respective terms.

$$L = \lambda_{pose} L_{pose} + \lambda_{joints} L_{joints} \tag{16}$$

$$L_{pose} = -\textstyle\sum_p \log\,(\sigma(W_c v)) \tag{17}$$

$$L_{joints} = -\textstyle\sum_p \sum_j \log\,(\sigma(W_w s_j)) \tag{18}$$

III. RESULT AND DISCUSSION

We evaluate the model in the paper on UCF and HMDB datasets with P5000 GPU. Although one video of dataset is only annotated as one label, so we divide the video by segments for action recognition. For instance, the action “LongJump” is regard as the combination of two annotations “Running” and “Jumping”. As following, we conduct qualitative and quantitative analysis.

The examples of action recognition in videos are as shown in Figure 5. There are four actions displaying, which labels are "Walking&Dog", "Baseball" and "WallPushup". The pose at each row is shown with skeleton and captioning.

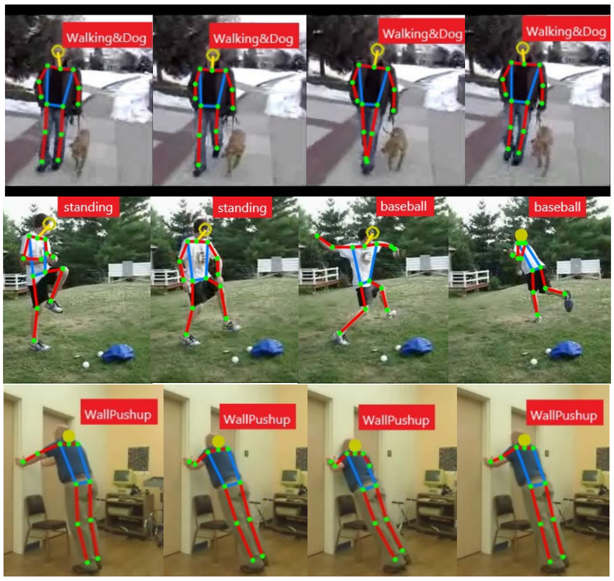

Figure 5: Examples of action recognition on UCF

The accuracy curve and loss function curve of sample training set are shown in Figure 6. The orange curve is accuracy of training. The blue curve is loss value during epoch. The final accuracy of training set is up to 0.93. The model proposed in the paper has good generalization ability and reduces the phenomenon of over-fitting.

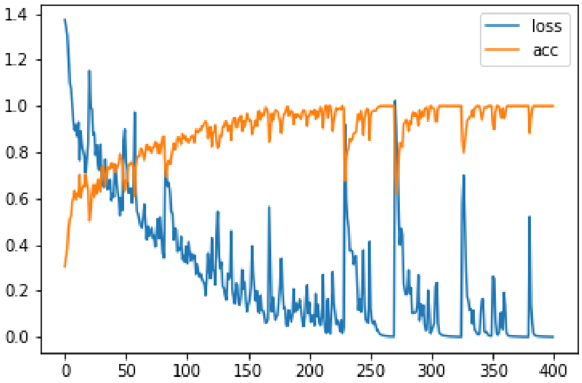

Figure 6: Curves of accuracy and loss

Table 1: Mean accuracy of UCF101 and HMDB.

Method	UCF101	HMDB
C3D	85.2%	51.2%
Two-Stream + LSTM	88.6%	59.1%
Transformation (VGG-16)	92.4%	62.0%
Our model	93.5%	63.3%

We compare our method and state-of-the-art methods on testing set of UCF101 and HMDB. The mean accuracy is shown in Table 1. The method in the paper outperforms state-of-the-art method, such as C3D, Two-Stream + LSTM and Transformation.

IV. CONCLUSION

In the paper, we design a pose-attention networks based on human graph. Hierarchical attention mechanism is incorporated into neural networks with pose-attention and joints-attention in 2D skeleton. Joints attention method emphasizes key joints features learning for action recognition with human graph skeleton. Our method can learn key pose features and predict the action till current frame. We test our model on two popular benchmarks, i.e. UCF101 and HMDB. The experiments show the method outperforms state-of-the-art methods with high accuracy.

Author Contributions: Gaofeng Li: Formal analysis, Methodology, Conceptualization, Writing–original draft, Writing–review & editing. **Songlin Wang**: Data curation, Investigation, Project administration, Resources, Software, Validation, Supervision, Writing–review & editing.

Funding: This work is supported by Henan Province Science and Technology Research Plan Project under Grant No. 172102210408.

Acknowledgments: The authors would like to thank the Associate Editor and the reviewers for their constructive comments that greatly improved the quality of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1]. Aggarwal, J.K.; Ryoo, M.S. (2011) “Human Activity Analysis: a Review”. ACM Computing Surveys, 43, pp1-43.

[2]. Herath, S.; Harandi, M.;Porikli, F. (2017) “Going Deeper into Action Recognition: a Survey”. Image and Vision Computing, 60, pp4-21.

[3]. Zhang, H.-B.; Zhang, Y.-X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.-X.; Chen, D.-S. (2019) “A Comprehensive Survey of Vision-Based Human Action Recognition Methods”. Sensors, 19, 1005.

[4]. Wu, D.; Sharma, N.;Blumenstein, M. (2017) “Recent Advances in Video-based Human Action Recognition Using Deep Learning”. International Joint Conference on Neural Networks, Anchorage, AK, pp2865-2872.

[5]. Henawy, I.; Ahmed, K.; Mahmoud, H. (2018) “Action Recognition Using Fast HOG3D of Integral Videos and Smith-Waterman Partial Matching”. IET Image Processing, 12, pp896-908.

[6]. Yamada, K.; Kimura, A. (2018) “A Performance Evaluation of Keypoints Detection Methods SIFT and AKAZE for 3D Reconstruction”. International Workshop on Advanced Image Technology, Chiang Mai, Thailand, pp1-4.

[7]. Bahdanau, D.; Cho, K.; Bengio, Y. (2015) “Neural Machine Translation by Jointly Learning to Align and Translate”. International Conference on Learning Representations, San Diego, California,.

[8]. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. (2017) “Attention Is All You Need”. Advances in Neural Information Processing Systems, Long Beach, USA, pp5998-6008.

[9]. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. (2016) “Hierarchical Attention Networks for Document Classification”. Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, San Diego, California, pp1480-1489.

[10]. Shi, L.; Zhang, Y.-F.; Cheng, J.; Lu, H.-Q. (2020) “Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action Recognition”. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington.

[11]. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. (2016) “Convolutional Pose Machines”. Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, pp4724-4732.

[12]. M. Sandler, A. Howard, M. Zhu, et al. (2018) “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp4510-4520.

[13]. Zhe Cao, Tomas Simon, Shih-En Wei, et al. (2017) “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. Computer Vision and Pattern Recognition, pp1302-1310.