



A Discriminative STGCN for Skeleton Oriented Action Recognition

Lin Feng¹, Qing Yuan², Yang Liu¹(✉), Qianxin Huang², Shenglan Liu¹,
and Yingping Li²

¹ School of Innovation and Entrepreneurship, Dalian University of Technology,
Dalian 116024, Liaoning, China

ly@dlut.edu.cn

² Faculty of Electronic Information and Electrical Engineering,
Dalian University of Technology, Dalian 116024, Liaoning, China

Abstract. Action recognition plays a fundamental role in many applications and researches, including man-machine interaction, medical rehabilitation and physical training. However, existing methods realize action recognition mainly relies on the background. This paper attempts to recognize the actions only through the motions. Hence, skeleton information is utilized to realize action recognition. To fully utilize the skeleton information, this paper proposes a discriminative spatio-temporal graph convolutional network (DSTGCN) for background independent action recognition. DSTGCN not only pays attention to the spatio-temporal properties of the motions, but focuses on the inner-class distributions of the actions. Experiments result on two motion oriented datasets validate the effectiveness of the proposed method.

Keywords: Action recognition · Graph convolutional network · Spatiotemporal analysis · Discriminative analysis

1 Introduction

Human action recognition, which is playing a significant role in many applications such as video surveillance, and man-machine interaction [2], has raised the great attention in recent years.

There are many approaches attempting to analysis that under the dynamic circumstance and complicated background. In lots of cases, background information deserved serious consideration. For example, when a person's hand moves to his mouth, it's difficult to distinguish what he's doing. The question will become easy if there is a cup in the person's hand, cause of additional information is provided by the background.

This study was funded by National Natural Science Foundation of Peoples Republic of China (61672130, 61972064), The Fundamental Research Funds for the Central Universities (DUT19RC(3)012, DUT20RC(5)010) and LiaoNing Revitalization Talents Program (XLYC1806006).

However, it would be not effective and even negative that putting the background information together in certain cases. For instance, in figure skating, a person shows a wide range of exaggerated movements for performing. The changing background will disturb the action analysis, therefore, skeleton data without background information is more appropriate in pure action recognition.

Earlier conventional methods [5, 15] treat skeleton data as vector sequences, which could not fully express the interdependency among the joints. Unlike recurrent neural networks (RNN) and convolutional neural networks (CNN), graph convolutional networks (GCN) treats skeleton data as graphs that could fully exploit the relationships between correlated joints.

GCN shows excellent performance in skeleton-based action recognition. However, most previous works [16, 19] pay little attention on feature maps output by the network and there's room for improvement on datasets with unbalanced categories. Therefore, we proposed a new approach to solve them. We use the focal loss [10] instead of the cross entropy (CE) loss to adapt unbalanced categories. The focal loss can give different weights to different categories according to difficulty of recognition. Above that, we added the center loss [18] working for feature maps to make better distinction and make the network more robust.

In this paper, 1) we modify the loss function from the CE loss to the focal loss to make network more adaptable to the datasets with unbalanced categories. 2) We add the center loss on deep features to make better distinction. 3) On two datasets for skeleton-based action recognition, our methods exceeds the state-of-the-art on both.

2 Related Work

2.1 RGB-D Based Action Recognition

RGB-D based human action recognition has attracted plenty of interest in recent years. Due to RGB-D sensors such as Kinect, RGB data and depth data, which encoding rich 3D structural information, could easy to be obtained. Previous works [8, 17, 20] leads the discovery of the information from visual features and depth features. Instead of considering two modalities as separate channels, SFAM [17] proposed the use of scene flow, which extracted the real 3D motion and also preserved structural information. [20] proposed a binary local representation for video fusion. BHIM [8] represents both two features in the form of matrix that including spatiotemporal structural relationships. Those RGB-D based methods focus on finding an appropriate way to fuse two features.

2.2 Skeleton Based Action Recognition

With the development of deep learning, lots of methods based on conventional networks have been proposed, which learn the features automatically. Some RNN-based methods [11, 15] and CNN-based methods [7, 12] have achieved high performance on action recognition. Unlike the above methods, GCN-based methods [9, 16, 19] treat skeleton data as graph which could exploit the relationships

between correlated joints better. ST-GCN [19] is the first to apply GCN on skeleton-based action recognition. 2s-AGCN [16] is an approach to adaptively learn the topology of the graph. AS-GCN [9] made attempts to capture richer dependencies among nodes. Those GCNs automatically learning with information of node location and structure.

3 The Proposed Approach

3.1 Graph Construction

Depending on devices and algorithms, skeleton data are usually represented as the sequence of 2D/3D coordinates. A joint is established connection with others by the graph along both the spatial and temporal dimensions. We construct the graph with joints according to the method in ST-GCN [19]. As shown in the left sketch of Fig. 1, a spatiotemporal graph is composed of the node set N and the edge set E . The node set N contains all the joint coordinates in a sequence. And the edge set E , composed of the spatial edge set E_S and the temporal edge set E_T , represents that how the nodes connected with others. For the spatial dimension, nodes connected with others as their natural connections in a frame. For the second subset E_T , nodes make connections among frames. The temporal edges connect the same nodes between adjacent frames.

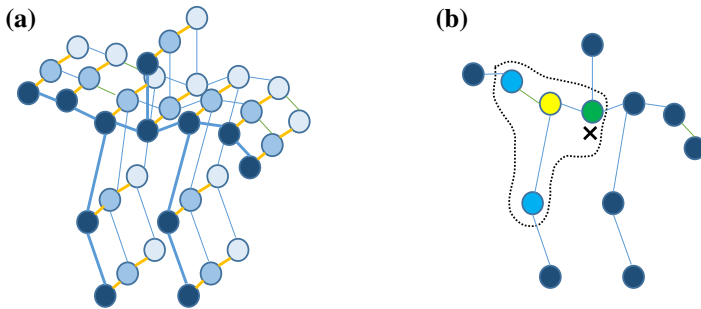


Fig. 1. (a) The spatiotemporal graph used. (b) The spatial configuration partitioning strategy.

3.2 Skeleton Oriented GCN

Deep graph convolutional network could be constructed based on the graph above. ST-GCN [19] consists of the ST-GCN blocks, which contains a spatial graph convolution and a temporal graph convolution.

The spatial graph convolution on node v_i could be formulated as [19]

$$f_{out}(v_i) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{in}(v_j) \cdot w(l_i(v_j)) \quad (1)$$

where f_{out} denotes the output feature and f_{in} denotes the input feature. B_i denotes the set of nodes which connected with node v_i . w is the weight function, which is a little different from original convolutional operation, but both provide the weights for input. The difference is that the number of nodes in the neighbor set B_i is unfixed. To solve that, we use the spatial configuration partitioning strategy, proposed in ST-GCN [19]. As shown in the right sketch of Fig. 1, the block cross represents the gravity center of the skeleton. According to the distance to the block cross, the strategy divide the set B_i into three subsets. The normalizing term Z_{ij} denotes the cardinality of the subset which contains the node v_j . In fact, the feature map of the network could be represented as a $C \times T \times N$ tensor, where C denotes the number of channels and T denotes the length of frame sequences. N denotes the number of nodes in a frame. For the spatial configuration partitioning strategy, the Eq. 1 is transformed into

$$f_{out}(v_i) = \sum_j A_j^{-\frac{1}{2}} A_j A_j^{-\frac{1}{2}} f_{in} W_j \otimes M_j \quad (2)$$

where A_j , a $N \times N$ tensor, denotes the divided adjacency matrix. Note that $\sum_j A_j = A + I$, where A denotes the adjacency matrix, and I is an identity matrix. $A_j^{ii} = \sum_k (A_j^{ik}) + \alpha$ is a diagonal matrix designed for normalized. α is set as 0.001 to avoid A_j^{ii} being zero. W_j is the weight matrix, representing the w function. M_j is an attention matrix, which denotes the importance of nodes. \otimes denotes the element-wise product.

In the temporal dimension, we can easily apply graph convolution like traditional convolution. We chose a certain number of frames before or later than the frame to make the number of the neighbors fixed. Therefore, the temporal kernel size could be determined and the convolution operation could be applied in the temporal dimension.

3.3 Loss Function

Focal Loss. The focal loss, an improved version based on the CE loss function, aims to overcome the difficulties due to the imbalance among categories. The formula for calculating the CE loss for binary classification is Eq. 3.

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise,} \end{cases} \quad (3)$$

and we define p_t as

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases} \quad (4)$$

and $CE(p, y)$ can be written as

$$CE(p_t) = -\log(p_t) \quad (5)$$

Based on the CE loss, [10] proposed the focal loss:

$$L_F(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \quad (6)$$

where α_t is a weighting factor to address the imbalance among categories. The factor $(1 - p_t)^\gamma$ could dynamically scale the loss. We set $\gamma > 0$, and the factor could automatically reduce the weight of easy examples and increase the weight of hard examples. Therefore, we considered that the focal loss is more suitable to the small-scale datasets, and our experiments proved that.

Center Loss. For making the deeply learned features more discriminative as shown in Fig. 2 and making network more robust, we add the center loss [18] in our work. The center loss, which could be formulated as Eq. 7, makes features discriminative by minimizing the intra class variance.

$$L_C = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|_2^2 \quad (7)$$

Where c_{y_i} denotes the deep features center of the y_i th class, and c_{y_i} is dynamically updated based on mini-batch as the deep features changed. The center loss is proposed for face recognition task, due to separable features are not enough, discriminative features are needed. We considered that it will work for action recognition as well, and we proved that in our experiments.

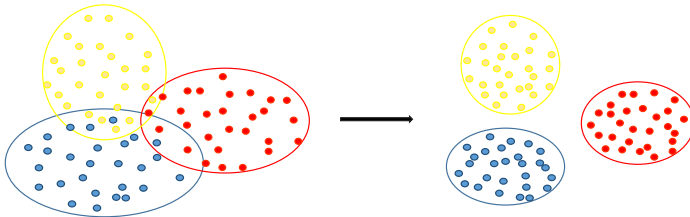


Fig. 2. The center loss function makes deep features more discriminative.

4 Experiments

In this section, we evaluate the performance of our approach and compare with some state-of-the-art methods on two human action recognition datasets: FSD-10 [14] and RGB-D human video-emotion dataset [13]. We evaluate the performance of approaches by top-1 classification accuracies on the validation set.

We use SGD as optimizer in all models, the batch size is set as 64. The learning rate is set as 0.1 and reduced by 10 in epoch 150 and 225. We use $L_S = L_F + \lambda L_C$ as loss function in our methods, and use the CE loss for comparison.

4.1 Evaluation on FSD-10

FSD-10. FSD-10 [14], a skating dataset consists of 1484 skating videos covering 10 different actions manually labeled. These video clips are segmented from performance videos of high level figure skaters. Each clip is ranging from 3 s to 30 s, and captured by the camera focusing on the skater. Comparing with other current datasets for action recognition, FSD-10 focuses on the action itself rather background. The information of background even bring negative effect. We divided FSD-10 into a training set (989 videos) and a validation set (425 videos). We train models on the training set and calculate the accuracy on the validation.

Comparisons and Analysis. For proving that the loss function $L_S = L_F + \lambda L_C$ is more suitable to FSD-10 than the CE loss, we run 2 groups of comparative experiments on FSD-10. The one is based on ST-GCN [19]: we first train ST-GCN with the CE loss, after getting the results, train it again with the loss function L_S . The other group is training on DenseNet [4] with the same operation. Besides, we compared the accuracy with the I3D [1], the STM [6] and the KTSN [14]. Table 1 give the result of our experiments. Both on ST-GCN and DenseNet, we see that the loss function L_S give a better performance than the CE loss on FSD-10.

Table 1. The result of our experiments on FSD-10.

Methods	Top-1 accuracy (%)
I3D (Resnet-50)	62.55
I3D (Resnet-101)	78.82
STM	66.25
KTSN	82.58
ST-GCN + CE loss	84.00
ST-GCN + L_S	87.52
DenseNet + CE loss	84.71
DenseNet + L_S	85.17

4.2 Evaluation on RGB-D Human Video-Emotion Dataset

RGB-D Human Video-Emotion Dataset. RGB-D human video-emotion data-set [13] consists of over 4 thousands RGB video clips and 4 thousands Depth video clips, covering 7 emotion categories. Each clip is around 6 s, containing the whole body of the actor. The background is green, without any information for recognition. The training set has 741 skeleton data, and the validation set has 644. We train models on the training set and calculate the accuracy on the validation.

Comparisons and Analysis. We performed our experiments on the video-emotion dataset for proving that the loss function L_S is suitable to the small-scale datasets for action recognition. Like the comparative experiments on FSD-10, we also run 2 groups of experiments based on ST-GCN [19] and DenseNet [4]. Besides, we compared the accuracy with the MvLE [13] and the MvLLS [3], the methods based on multi-view for recognition on this dataset. Table 2 give the result of methods. We see that the loss function is work on the video-emotion dataset as well, and our methods perform better than the state-of-the-art on this dataset.

Table 2. The result of our experiments on the video-emotion dataset.

Methods	Top-1 accuracy (%)
MvLE	41.00
MvLLS	37.97
ST-GCN + CE loss	54.96
ST-GCN + L_S	55.27
DenseNet + CE loss	52.32
DenseNet + L_S	53.72

5 Conclusion

In this paper, we adapted the center loss and the focal loss to the human action recognition. We use the focal loss aims to overcome the difficulties due to the imbalance among categories. And we consider it’s more suitable to the small-scale datasets with unbalanced categories. We add the center loss to learn more discriminative features and to make better distinction on deep features. We performed our experiments on the FSD-10 [14] and the RGB-D human video-emotion dataset [13], and our methods achieved the state-of-the-art performance.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
2. Duric, Z., et al.: Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proc. IEEE **90**(7), 1272–1289 (2002)
3. Guo, S., et al.: Multi-view laplacian least squares for human emotion recognition. Neurocomputing **370**, 78–87 (2019)
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

5. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 221–231 (2012)
6. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: STM: spatiotemporal and motion encoding for action recognition. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2000–2009 (2019)
7. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3288–3297 (2017)
8. Kong, Y., Fu, Y.: Bilinear heterogeneous information machine for RGB-D action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1054–1062 (2015)
9. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3595–3603 (2019)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017)
11. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
12. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recogn.* **68**, 346–362 (2017)
13. Liu, S., Guo, S., Wang, W., Qiao, H., Wang, Y., Luo, W.: Multi-view laplacian eigenmaps based on bag-of-neighbors for RGB-D human emotion recognition. *Inf. Sci.* **509**, 243–256 (2020)
14. Liu, S., et al.: FSD-10: a dataset for competitive sports content analysis. *arXiv preprint arXiv:2002.03312* (2020)
15. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+ D: a large scale dataset for 3D human activity analysis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019 (2016)
16. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035 (2019)
17. Wang, P., Li, W., Gao, Z., Zhang, Y., Tang, C., Ogunbona, P.: Scene flow to action map: a new representation for RGB-D based action recognition with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604 (2017)
18. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
19. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455* (2018)
20. Yu, M., Liu, L., Shao, L.: Structure-preserving binary representations for RGB-D action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(8), 1651–1664 (2015)