Pose-Enhanced Relation Feature for Action Recognition in Still Images

Jiewen Wang and Shuang Liang(✉)

School of Software Engineering, Tongji University, Shanghai, China
{wjwlaservne,shuangliang}@tongji.edu.cn

Abstract. Due to the lack of motion information, action recognition in still images is considered a challenging task. Previous works focused on contextual information in the image, including human pose, surrounding objects, etc. But they rarely consider the relation between the local pose and the entire human body, so that poses related to the action are not fully utilized. In this paper, we propose a solution for action recognition in still images, which makes complete and effective use of pose information. The multi-key points calculation method is carefully designed for generating pose regions that explicitly includes possible actions. The extensible Pose-Enhanced Relation Module extracts the implicit relation between pose and human body, and outputs the Pose-Enhanced Relation Feature which owns powerful representation capabilities. Surrounding objects information is also applied to strengthen the solution. Through experiments, it can be found that the proposed solution exceed the state-of-the-art performance on two commonly used datasets, PASCAL VOC 2012 Action and Stanford 40 Actions. Visualization shows that the proposed solution enables the network to pay more attention to the pose regions related to the action.

Keywords: Action recognition · Human pose · Relation networks

1 Introduction

Action recognition is a core task in the field of computer vision, and it has a number of applications in real life like video surveillance and motion sensing games. In the study of the task, still images based action recognition tends to receive less attention, since still images don't have motion or temporal information like videos, which makes it difficult to capture useful features. In order to fully dig out the information in still images, the researchers set their sights on contextual cues other than the human body. At present, poses and surrounding objects are usually used like in [12,19]. Due to the development of pose estimation and object detection technology, it's not difficult to obtain these features.

However, while applying these contextual features, the further relation between them and human body features is seldom considered in existing methods. For surrounding objects, the situation is slightly better, because the interaction between human body and objects is generally obvious; but for pose feature,

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13141, pp. 154–165, 2022.
https://doi.org/10.1007/978-3-030-98358-1_13

since its relation with the human body is not obvious, it is processed in an independent branch in most cases [4]. And addition or concatenate is performed to simply fuse it before classification. Applying these methods sometimes leads to incorrect predictions, especially when there is no object information to assist.

Fig. 1. An example of "similar pose, different action". The actions of the characters on the left/right are different (applauding/jumping), but the local poses (such as the right hand) show higher similarity.

An example that may lead to a false prediction is shown in Fig. 1: Although the people in two images perform different actions, the pose of a certain part may be similar. In this case, the local pose feature becomes a misleading item, and the entire body feature is needed for correct predictions. This reveals that the partial pose and the entire human body are not isolated, and the combination of the two may be able to better eliminate misleading information.

The self-attention mechanism provides a good foundation for dealing with this situation. Self-attention is widely used in deep learning works. It emerged from natural language processing [15] and soon impacted computer vision tasks like object detection [10]. Its main function is to learn the correlation between features, which is exactly what we need. In the action recognition scenario, self-attention mechanism can be applied to model the correlation between the contextual features and the human body feature.

Inspired by self-attention mechanism and its application in action recognition, we propose a solution for action recognition in still images. The contributions of this work can be summarized as follows:

- The multi-key points calculation method is designed to convert pose key points into bounding box format pose regions. The generated pose regions include the actions that may be reflected by the joints, which is helpful to establish the connection between the actual action and the pose region.
- We propose a novel Pose-Enhanced Relation Module to model the relation between the local pose and the entire human body. This relation is implicit

in the Pose-Enhanced Relation Feature output by the module. To the best of our knowledge, the proposed method is the first one to focus on this relation. Moreover, the Pose-Enhanced Relation Module can be easily extended.
– The proposed solution has obtained competitive results in experiments. The results reach new state-of-the-art performance on the Stanford 40 Actions [18] test set and the PASCAL VOC 2012 Action [3] validation set. Ablation study and visualization illustrate the effectiveness of the proposed solution.

2 Related Work

2.1 Action Recognition in Still Images

Consistent with many computer vision tasks, action recognition in still images has undergone a transition from traditional methods to deep learning methods. But what remains the same is the idea of assisting the task with contextual information. Traditional solutions usually use hand-crafted features and combine them with machine learning methods [17]. The neural network doesn't need to manually design features, so researchers pay more attention to the use of features and the mining of deep-level features. Gkioxari et al. [8] mainly captures the key areas in the image, and doesn't care about its specific meaning. Zhao et al. [19] and Fang et al. [4] use pose features explicitly to assist action recognition. New discoveries in computer vision have also been introduced to the task, such as the attention mechanism [16]. Due to the lack of motion information in still images, there is even a way to generate hallucinated motion information for images [6].

2.2 Pose Estimation

Pose estimation is an important pre-work for action recognition. It aims to obtain the key points corresponding to human joints and connect them. Pose estimation has also entered the era of deep learning, and its methods can be roughly divided into top-down methods like [2] and bottom-up methods like [13]. The former relies on the pre-generated human part, and the latter starts from the joints. Generally, the top-down methods perform better in accuracy, while the bottom-up methods control the inference time better in multi-person pose estimation.

Multi-person pose estimation is more challenging because both the inference speed and the accuracy between different people should be considered. Related method [5] is used in the pose regions generation section of this paper.

3 Approach

In this section, we will show the pose regions generation method and pose-enhanced relation module.

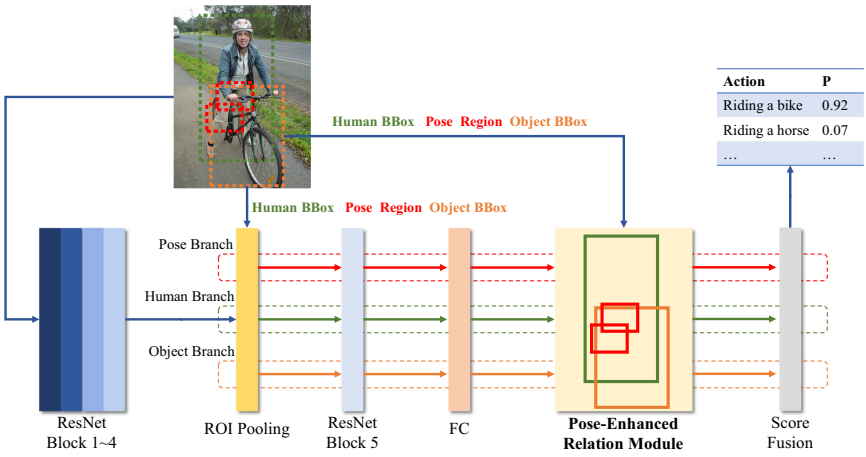

Fig. 2. The overall framework of the proposed method.

3.1 Framework Overview

The overall framework of the proposed method is presented in Fig. 2. The image, human bounding box, objects bounding boxes, and pose regions are taken as inputs, and the final output is the probabilities of each action. Among three kinds of bounding box formatted information, human body bounding box is provided by datasets, objects bounding boxes can be detected by Faster R-CNN [14], and pose regions will be generated by the method that will be introduced later.

The image will be sent to the feature extractor first. Here we choose ResNet [9] because of its efficiency. It can be replaced with other backbones since it's not the key part. ROI Pooling [7] is applied to obtain the instance feature map. Here the input includes the three kinds of bounding boxes introduced earlier. Therefore, the ROI Pooling layer will output 3 branches. They are sent to the ResNet Block5 and an FC layer, which will generate the basic features on human body, pose and objects. Then these basic features are input into the Pose-Enhanced Relation Module and the enhanced relation features will yield. Finally, the relation features are fused with the basic features and participate in the final prediction.

3.2 Pose Regions Generation

The concept of pose regions has been mentioned in Sect. 3.1. The pose estimation network can recognize the human pose, but its output format is the key points, which is difficult to use directly, so further transformation is required. In brief, it takes two steps to complete this process. First, the pose estimation network is applied to generate pose key points. Then a certain number of pose regions are calculated through the combination of key points. The final format of pose regions is the coordinate form consistent with the bounding box.

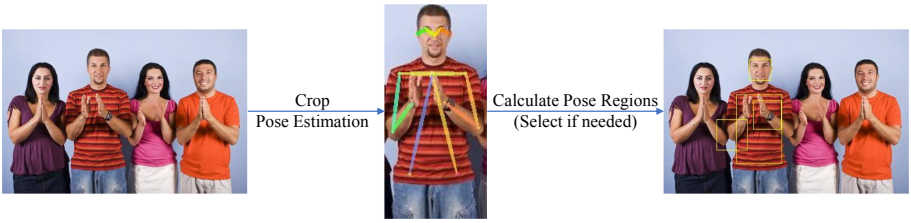

Fig. 3. The process of pose region generation.

The process of generating pose regions is shown in Fig. 3. Specifically, the first step should be cropping the required person according to its bounding box. Then AlphaPose [5] which can handle multi-person scene is applied to estimate poses, and finally 10 pose regions are calculated through key points.

There are two details to pay attention to: First, although most irrelevant characters have been removed by cropping, it is still possible to recognize multiple poses. If this happens, only the pose with the highest confidence will be chosen. In addition, AlphaPose cannot estimate the pose for a few images. There are many reasons, such as: the image is too small, the image is too blurry, or the human body is severely occluded. One solution is to use blank regions instead, but this may adversely affect the recognition results. In the proposed method, if this happens, the human bounding box will be divided into 16 parts, and then one or more (up to 4) parts will be randomly allocated to each pose region. In this way, there are always useful regions being sent to the network.

In order to better reflect the actions embodied by the pose in pose regions, when calculating the coordinates of pose regions, we try to avoid the situation where only one key point is used to calculate a region. For example, the "right elbow" region in Fig. 3 contains information about the bending of the elbow, and 3 key points are used to calculate it. We call it multi-key points calculation method. The key points for calculating each region are illustrated in Table 1.

Table 1. 10 pose regions and corresponding key points.

Region	Key points
Head	Nose, Left/Right ear
Body	Left/Right shoulder, Left/Right hip
Left hand	Left wrist, Left elbow
Right hand	Right wrist, Right elbow
Left elbow	Left wrist, Left elbow, Left shoulder
Right elbow	Right wrist, Right elbow, Right shoulder
Left foot	Left knee, Left ankle
Right foot	Right knee, Right ankle
Left knee	Left knee, Left ankle, Left hip
Right knee	Right knee, Right ankle, Right hip

3.3 Pose-Enhanced Relation Module

As mentioned in Sect. 1, the structure which deals with contextual features should be easy to extend and has strong learning capabilities. Therefore, inspired by [12], the Context-Enhanced Relation Structure shown in Fig. 4 is applied. It can be seen that the structure is symmetrical. The main advantage of such structure is that it not only considers appearance information, but also captures position information through location encoding which uses trigonometric functions to handle the relative position of the bounding boxes.

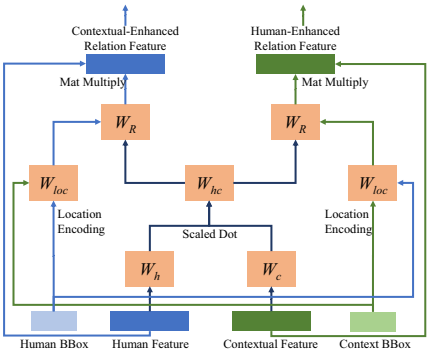

Fig. 4. Details of context-enhanced relation structure.

The final relation feature shown in Fig. 4 can be computed as follows:

$$f_R = fc(\sum(W_R \cdot f_{h\,or\,c})) \tag{1}$$

Here, W_R is the relation weight, which contains attention information from both location and appearance (reflected in W_{loc} and W_{hc} respectively). fc is a fully connected layer. h, c in the subscript represent human and context.

Context-Enhanced Relation Structure is powerful, but it can only handle two types of contextual features. In the proposed solution, the pose regions generated in Sect. 3.2 are added to enhance this module. This is different from [12].

There are two main considerations for adding pose regions:

First, which features need to be combined with the pose feature? Only combining human body feature, or combining both human body feature and objects feature? Our choice is the former. The reason is that the relation between objects and poses is not obvious, and it may have been implicit in the human-object feature. So it's redundant to repeat learning and add useless parameters.

Second, when there are only two types of relation features, the weights of them can be set to 1 for symmetry. However, the addition of the pose feature creates two more relation features. How should the ratio of these relation features be set? In the proposed solution, the final ratio for each relation feature is:

$$f_{ho} : f_{hp} : f_{oh} : f_{ph} = 1 : 1 : 2 : 2 \tag{2}$$

Here h, o, p represent human/object/pose. Larger weights are set on the latter two, mainly to force the network to pay more attention to contextual cues. The structure of the pose-enhanced relation module is shown in Fig. 5. f_{ho} and f_{hp} are merged, so the module outputs 3 types of relation features.

In addition, the idea of Multi-head Attention [15] is also applied in the proposed solution. As presented in Fig. 5, multiple pose-enhanced relation modules (the actual number is 16) are employed to accept different part of the features. The advantage is that different modules can focus on the learning of its own input part, so as to better capture the features of different part in the input.

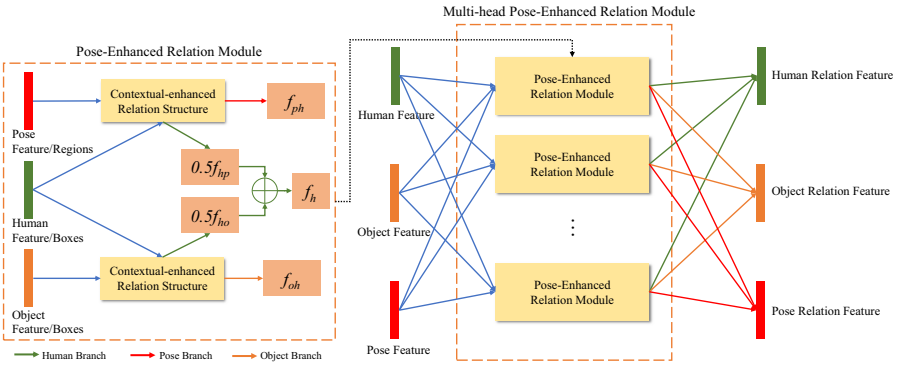

Fig. 5. Details of pose-enhanced relation module and its multi-head form.

3.4 Score Fusion and Loss Function

As shown in Fig. 5, 3 types of relation feature are generated finally. They will be added to the corresponding basic feature and then input into a fully connected layer to obtain the final classification score. For each action a, the score is:

$$Score(a) = Score^{a}_{human} + max^{a}_{object\,1...m} + max^{a}_{pose\,1...10} \tag{3}$$

Here m equals to the number of the surrounding objects, and the max operation selects the highest score of multiple objects/poses for action a. After getting the classification score, $softmax$ is used to convert it to probability and compute the loss. For convenience, cross entropy loss is applied, whose expression is:

$$loss = -log(\frac{exp(Score((a^{gt})))}{\sum_a exp(Score(a))}) \tag{4}$$

4 Experiments

In this section, we introduce some experimental details and evaluate the proposed solution on two public datasets.

4.1 Experiment Settings

Consistent with what was mentioned above, we use ResNet50 as the backbone network and, like most computer vision tasks, initialize the parameters with the pre-trained weights on ImageNet. In terms of the optimizer, we use the SGD optimizer, set the momentum to 0.9 and the weight decay to 5×10^{-4}. In terms of the learning rate, we use cosine annealing method to control the attenuation of the learning rate so that it can be reduced from 3×10^{-5} to 1×10^{-6} in 20 epochs. The collection of three kinds of bounding boxes has been introduced earlier, that is, human bounding boxes are provided by the datasets, objects bounding boxes are detected by Faster R-CNN, and pose regions are generated using the method proposed in Sect. 3.2. All experiments are done on an NVIDIA TESLA K80 GPU, and the deep learning implementation framework is MXNET [1].

4.2 Comparison with Other Methods

PASCAL VOC 2012 Action and Stanford 40 Action are two classic public datasets for action recognition in still images, and both datasets provide human bounding box annotations. Therefore, the proposed solution is evaluated on them.

PASCAL VOC 2012 Action. Due to the limitation of the test set, we choose to evaluate the proposed solution on PASCAL VOC 2012 validation set which is fully public. The results are shown in Table 2.

Table 2. Average precision (%) on the PASCAL VOC 2012 validation set.

Method	Jumping	Phoning	Playing instrument	Reading	Riding bike	Riding horse	Running	Taking photo	Using computer	Walking	mAP
R*CNN [8]	88.9	79.9	95.1	82.2	96.1	97.8	87.9	85.3	94.0	71.5	87.9
Attention [16]	87.8	78.4	93.7	81.1	95.0	97.1	**96.0**	85.5	93.1	73.4	87.1
Part action [19]	89.6	86.9	94.4	88.5	94.9	97.9	91.3	87.5	92.4	76.4	90.0
Object relation [12]	89.2	89.8	96.5	87.6	**98.2**	**99.1**	92.3	**91.6**	95.2	**79.2**	91.9
Ours	**92.3**	**91.4**	**96.9**	**89.1**	97.2	99.0	92.1	90.8	**95.5**	78.9	**92.3**

According to the results in Table 2, the proposed solution has achieved new state-of-the-art performance on the PASCAL VOC 2012 validation set. In terms of specific actions, jumping, reading and phoning gain the largest increase (at least 1.5 AP%), which directly drives the growth of the overall mAP. For the two actions of playing instrument and using computer, the proposed solution still maintains the lead, but the advantage is smaller. For the two actions of walking and riding horse, although the proposed solution is not optimal, it is extremely close to the state-of-the-art. The remaining few actions lag behind the state-of-the-art by more than 0.8 AP%. To sum up, the proposed solution obtains excellent classification results on the PASCAL VOC 2012 validation set.

Stanford 40 Action. Stanford 40 Action is a larger dataset than PASCAL VOC 2012 Action. Its training set has 4000 images while the test set has 5532. The results of the proposed solution on the Stanford 40 Action test set are shown in Table 3. Since the dataset has 40 classes, it's impossible to list them all here.

Table 3. Mean average precision (%) on the Stanford 40 Action test set.

Method	mAP
Human mask loss [11]	91.1
Part action [19]	91.2
Object relation [12]	93.1
Ours	**93.2**

In the experiment on the Stanford 40 Action test set, the proposed solution has surpassed the current state-of-the-art, and it has a more considerable increase compared with the previous method using pose features. It should be noted that we reproduced the experimental results on the source code of [12], and the best result is 92.8% mAP. The proposed solution is built on this source code, and when similar settings are employed, it can reach 93.2% mAP. Therefore, the proposed solution may be able to surpass the current state-of-the-art more if implementation differences are considered.

In addition, the above comparisons are all performed under ResNet50, and stronger backbone is not used here. According to the result of [12], applying a stronger backbone may further improve the result.

Ablation Study, Visualization and Analysis. In general, the proposed solution has added two parts compared with previous methods: one is to apply the pose regions generated by the key points, and the other is to fuse the pose regions with the human body feature through relation module. In order to verify the effectiveness of these two parts, ablation experiments were conducted on the PASCAL VOC 2012 validation set, and the results are reported in Table 4. It should be noted that the baseline of the experiment is exactly the scheme in [12]. The solution of directly adding pose features to the classifier and fusing pose features with the relation module will be verified.

Table 4. Ablation study on the PASCAL VOC 2012 validation set.

Method	Jumping	Phoning	Playing instrument	Reading	Riding bike	Riding horse	Running	Taking photo	Using computer	Walking	mAP
-	89.2	89.8	96.5	87.6	**98.2**	**99.1**	**92.3**	**91.6**	95.2	**79.2**	91.9
+pose	91.9	89.5	96.2	87.8	97.3	98.9	92.2	90.7	**95.5**	76.3	91.6
+pose, +relation	**92.3**	**91.4**	**96.9**	**89.1**	97.2	99.0	92.1	90.8	**95.5**	78.9	**92.3**

According to the results, it can be found that if the pose feature is just simply added to the classifier, the overall mAP would be lower than the group without the addition. From the perspective of a single action, the group of simply adding pose feature has a drop in the classification results of almost all actions, but jumping and using computer are exceptions. Take a look at the example in Fig. 1, we guess that pose information may be helpful for classifying actions that are not related to objects, but it is limited to certain actions, because the result of the walking action has declined. Obviously, the way of fusing pose information with human body feature through the relation module is more effective.

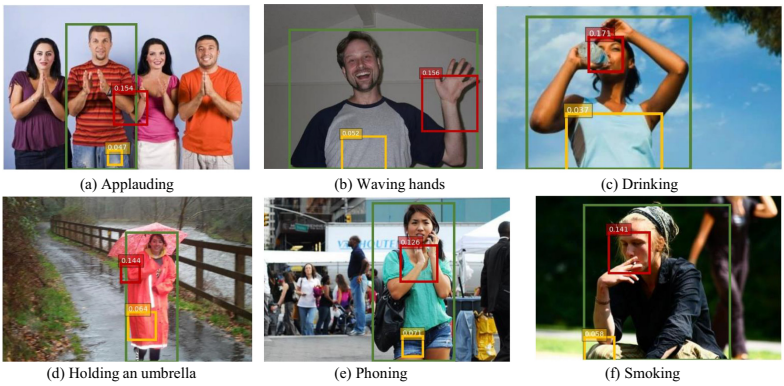

Fig. 6. Some visualization results on the Stanford 40 Action dataset. The green box is the human bounding box. The red box is the pose region with the highest pose relation weight. The yellow box is the pose region with the lowest pose relation weight. The relation weight value is marked on the top of the pose region. (Color figure online)

We analyze the role played by the pose through visualizing the pose relation weight. In general, the proposed solution makes a certain pose region get the focus of the network, and usually this region is the key to action classification from the perspective of human intuition. If there is no interaction with the object, the network will pay attention to the pose region where the action is performed, as shown in Fig. 6 (a)(b); if the action is associated with objects, the network will pay attention to the pose region related with the object, as shown in Fig. 6 (c)–(f). Correspondingly, the pose region which is not related to the action gets a lower pose relation weight, consistent with the expectation.

In order to further explore what type of action the proposed solution is suitable for, the top 3 and bottom 3 of the results on the Stanford 40 Action test set are listed in Table 5.

From the results, it can be seen that the classification results are generally better for images with clear objects and pose information. The result is relatively poor for actions like Waving Hands which has no surrounding objects. Such action may be confused with others. For the other two actions with poor results, we guess that it is because they don't follow a fixed action pattern. For example,

Table 5. Top 3/bottom 3 classification results on the Stanford 40 Action test set.

Top3	AP	Bottom3	AP
Playing violin	99.9	Pouring liquid	77.9
Holding an umbrella	99.7	Waving hands	79.6
Riding a horse	99.9	Taking photos	82.6

when pouring liquid, the hand movement may be upward or downward, which may cause misclassification. How to make the network adapt to such variability is still a problem that needs further exploration.

5 Conclusion

In this paper, we propose a solution that makes full use of various contextual features especially pose to assist action recognition in still images. By generating pose regions with action cues and fusing them with the human body feature in the Pose-Enhanced Relation Module, the proposed solution solves the problem of insufficient pose information usage in previous methods, and has achieved leading results on public datasets. Based on the analysis of the action classification results, adapting the solution to different action patterns may be a problem worth studying in the future.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62076183, Grant 61936014, and Grant 61976159; in part by the Natural Science Foundation of Shanghai under Grant 20ZR1473500 and Grant 19ZR1461200; in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; in part by the National Key Research and Development Project under Grant 2019YFB2102300 and Grant 2019YFB2102301; in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100; and in part by the Fundamental Research Funds for the Central Universities. The authors would also like to thank the anonymous reviewers for their careful work and valuable suggestions.

References

1. Chen, T., et al.: MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274 (2015)
2. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1831–1840 (2017)
3. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
4. Fang, H.-S., Cao, J., Tai, Y.-W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 52–68. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_4

5. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343 (2017)
6. Gao, R., Xiong, B., Grauman, K.: Im2FLow: motion hallucination from static images for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5937–5947 (2018)
7. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
8. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with R*CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1080–1088 (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3588–3597 (2018)
11. Liu, L., Tan, R.T., You, S.: Loss guided activation for action recognition in still images. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11365, pp. 152–167. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_10
12. Ma, W., Liang, S.: Human-object relation network for action recognition in still images. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2020)
13. Papandreou, G., Zhu, T., Chen, L.-C., Gidaris, S., Tompson, J., Murphy, K.: PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 282–299. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_17
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 91–99 (2015)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
16. Yan, S., Smith, J.S., Lu, W., Zhang, B.: Multibranch attention networks for action recognition in still images. IEEE Trans. Cogn. Dev. Syst. **10**(4), 1116–1125 (2017)
17. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 17–24. IEEE (2010)
18. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: 2011 International Conference on Computer Vision, pp. 1331–1338. IEEE (2011)
19. Zhao, Z., Ma, H., You, S.: Single image action recognition using semantic body part actions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3391–3399 (2017)