Human Action Recognition Based on Temporal Pose CNN and Multi-dimensional Fusion

Yi Huang[1](✉), Shang-Hong Lai[1], and Shao-Heng Tai[2]

[1] National Tsing Hua University, Hsinchu, Taiwan
jeffreyhuang0823@gmail.com, lai@cs.nthu.edu.tw
[2] Umbo Computer Vision, Taipei, Taiwan
daniel.tai@umbocv.com

Abstract. To take advantage of recent advances in human pose estimation from images, we develop a deep neural network model for action recognition from videos by computing temporal human pose features with a 3D CNN model. The proposed temporal pose features can provide more discriminative human action information than previous video features, such as appearance and short-term motion. In addition, we propose a novel fusion network that combines temporal pose, spatial and motion feature maps for the classification by bridging the gap between the dimension difference between 3D and 2D CNN feature maps. We show that the proposed action recognition system provides superior accuracy compared to the previous methods through experiments on Sub-JHMDB and PennAction datasets.

Keywords: Action recognition · Multi-stream · Fusion · Pose estimation

1 Introduction

In light of the recently launched evolution of deep learning, the research of action recognition, being one of the most widely applicable study in computer vision (e.g., video surveillance, human-computer interaction [11]), has been focused on developing innovative solutions for the subject matter via deep learning [16,21,23,25].

Perhaps the most integral factor in achieving accurate action recognition results resides in the extraction of discriminative temporal clues within videos. Thus, though there exists a cornucopia of information in a video, researchers can never be overly scrupulous about their selection of features when it comes to optimizing the performance of action recognition algorithms.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11012-3_33) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11130, pp. 426–440, 2019.
https://doi.org/10.1007/978-3-030-11012-3_33

Some achieved the capturing of temporal clues via feeding convolutional neural networks (CNNs), both two-stream and classical ones, with optical flow features [15,16,23,25]. Some utilized recurrent neural networks (RNNs) to model the inter-relationship between high-level features extracted from the fully-connected layer of a CNN for action recognition. However, we argue that in general, extracting temporal human pose based feature is the most effective way for human action recognition from videos (Fig. 1).

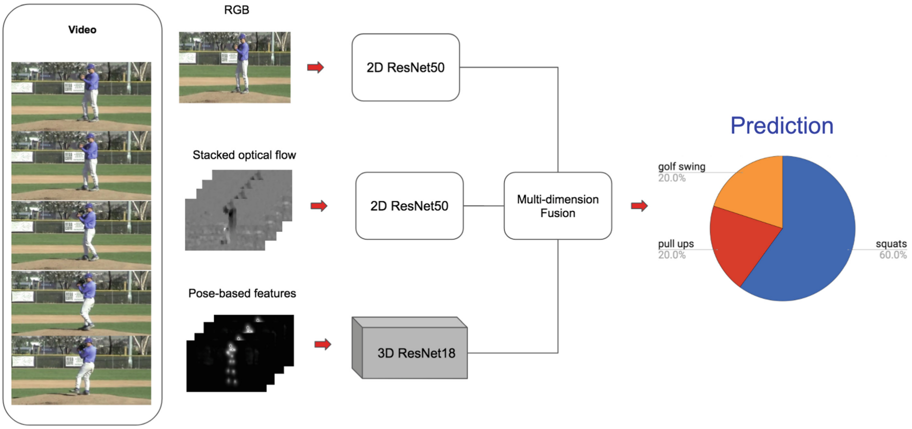


Fig. 1. Overview of the temporal pose-based convolutional neural network with multi-dimensional fusion

In this paper, we propose a 3D CNN network that is capable of exploiting temporal pose features within videos. Our method demonstrates the effectiveness of pose-based features in terms of modeling temporal information for human action recognition. We also develop a multi-dimensional fusion method to fuse the features extracted from 3D pose stream and 2D two-stream architecture, which further enhances the performance of our multi-stream posed-based CNN.

Our main contributions in this paper are summarized as follows:

- We propose a novel 3D temporal pose CNN for utilizing pose-based features to effectively capture the temporal human pose features in videos for action recognition.
- To take the advantages of both 2D and 3D network, we present a simple but highly effective multi-dimension fusion network which bridges the gap between 3D and 2D CNN feature maps and enables our model to leverage 3D temporal pose, spatial, and motion feature maps for human action recognition.
- By conducting extensive experiments, we validate the performance of the proposed framework and show that the proposed multi-stream action recognition system provides superior accuracy compared to the previous methods on Sub-JHMDB [13] and PennAction [26] datasets.

2 Related Work

2.1 Pose-Based Action Recognition

As a kind of high-level visual information, human pose features are exploited in many works with different architectures of pose-based action recognition approaches [2,5,7,12]. [5] introduced a new video descriptor called P-CNN, which was derived from aggregating appearance (RGB) and short-term motion (optical flow) features around different human body parts across the whole video, and then such video descriptor was used to train a linear SVM classifier. Cao et al. [2] proposed to pool 3D deep CNN activations of different segments of a video using joint positions of frames in the video. By aggregating features across segments to form a video level representation, the aggregated result is input to a linear SVM for classification. [7] developed an end-to-end recurrent pose-attention network to leverage pose features with attention mechanism. However, the purpose of these works to utilize pose-based features is to indicate an attention region for other kinds of features, which we believe is not the optimal utilization of pose features. The main difference between the proposed method with previous pose-based methods is that we generate fused joint position maps and directly use as the input of 3D CNN to further model their temporal information.

2.2 Two-Stream-Based Action Recognition

The deep learning approach of action recognition is a very active research area in the past few years [6,8,14,16,17,21–23,25]. Among several standard CNN architectures in action recognition related field, the two-stream CNN approach [21] is simple but highly effective [15]. It leverages the power of two single stream CNN to predict actions in videos: one for modeling the appearance clues in RGB images, and the other stream for capturing short-term motion in optical flow images. Recently, there are several works proposed to enhance the two-stream architecture. [23] proposed a sparse temporal sampling strategy and a series of good practice to further enhance the performance of [21] and make it more efficient. [1] aggregated local convolutional features of the two streams to introduce a new video representation for action classification. [25] included the audio stream and adopted LSTM networks to explore long-term temporal dynamics. In this work, we combine the two-stream architecture with a 3D CNN based pose stream by using a novel fusion method.

2.3 Multi-stream Fusion

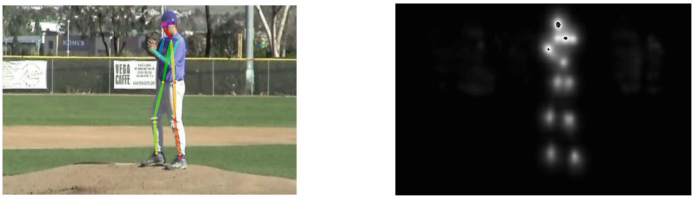
In the original two-stream method [21], since the authors just simply fuse the features with average fusion, which only average the prediction scores of the softmax layer in both streams. [8] improved the original work by fusing the two streams with a single convolutional fusion layer. [25] took action class relationships into account to learn the best fusion weights of different deep neural network streams for different action classes. [16] used multiple 2D convolution layers to model the concatenation of features in different domains. However, all of these previous works only developing the fusion techniques based on the feature maps with the same dimension. In our work, we propose a novel fusion method to fuse the 3D convolutional pose stream with the 2D convolutional appearance stream and the 2D convolutional short-term motion stream by utilizing the proposed feature compression sub-network to bridge the gap of the discrepancy of feature dimensions of the three target CNN streams. The design details will be discussed in Sect. 4. We will compare the proposed multi-stream action recognition method with the state-of-the-art methods on some public datasets, which will be described in Sect. 5.5.

3 Pose-Based Action Recognition

In this section, We propose a CNN based model for action recognition based on using the human pose features. With the help of human pose features and proposed channel-wise convolution techniques. We present a novel way of utilizing pose features for human action recognition. We first briefly introduce the pose estimation method in Sect. 3.1, and then provide the details of the proposed method in Sects. 3.2, 3.3 and 3.4

3.1 Pose Estimation

In this paper, we generate the pose estimation result from a strong bottom-up multi-person pose estimator [3,20,24], which is capable of computing human pose features in different scales and positions with real-time speed. A sample result of human pose estimation is shown in Fig. 2(a). Instead of directly using their pose estimation result, we propose some pre-processing procedure to best utilize the multi-channel human joint position maps.


(a) Pose estimation

(b) Joint position map

Fig. 2. An example of (a) the visualization of original pose estimation result and (b) the joint position map in our action recognition model. (Color figure online)

Joint Position Map: Figure 2(a) shows the pose estimation result of [3,20,24], which utilizes different colors to denote different human joint positions. For the reason that the input dimension to the 3D CNN will be very large with the concatenation of the feature maps in temporal domain, we do not directly apply this result to 3D convolutional neural network, Instead, we take the 15-channel heatmaps of their estimation result and generate a joint position map by using channel-wise convolution, which is shown in Fig. 2(b) as the input of 3D convolutional neural network to extract the spatial and temporal features in human action. The details of the channel-wise convolution will be given subsequently in this section.

3.2 Pose-Based CNNs

In this work, with the belief that high-level pose-based features outperform the mid-level flow-based features [13] for action recognition, we extend the optical flow stacking CNNs [21,23] to human-joint-part stacking CNNs to better capture spatial and temporal clues. The flow diagram of our pose-based 3D CNN model is depicted in Fig. 3.

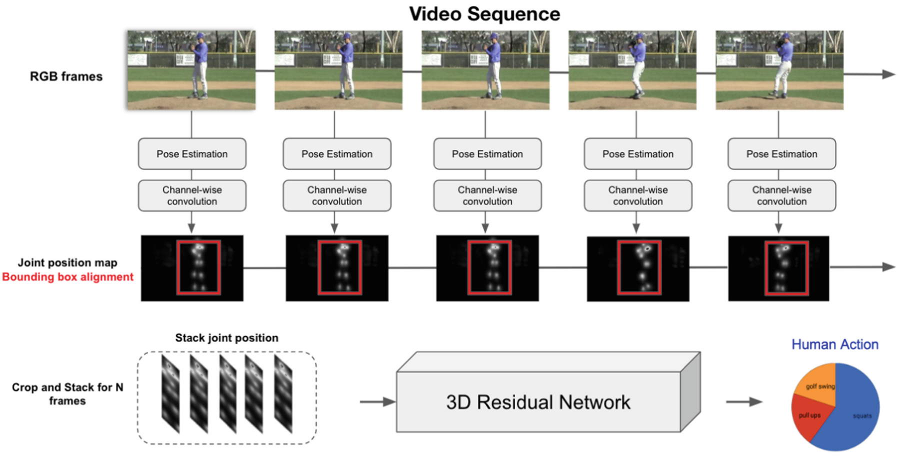


Fig. 3. Flow diagram of the pose-based 3D CNN model

Channel-Wise Convolution: To leverage the information computed for all human parts, we first compute the 15-channel heatmaps to generate the joint position map [13], which can be considered as the linear combination of the probability maps for the corresponding human parts. According to the experimental results in Table 3, we found that utilizing two 3D convolution layers to merge the features from a 15-channel heatmap achieve the highest accuracy in our experiments.

Stacked Joint Position: According to recent success in 3D CNN [9], we utilize the 3D residual network to model the pose-based feature and also compare its performance with a 2D network, which is used in the motion stream of [21] for modeling temporal clues. The details of our experiments will be described in Sect. 5.1. Here, we denote $P_t(u, v)$ to be the probability map of human parts at the point (u, v) at frame t. To model the temporal information of human joint parts across a sequence of frames, we stack the joint position map over L consecutive frames to form a stacked joint position map of totally L input channels. More precisely, let w and h be the width and height of videos. For $u = [1 : w]$, $v = [1 : h]$, $k = [1 : L]$, the input volumes of the pose-based CNNs, I_t^{3d} and I_t^{2d}, at frame t are constructed as follows:

$$\text{For 3D CNN, } \mathrm{I}_t^{3d}(k, 1, u, v) = P_{t+k-1}(u, v) \tag{1}$$

$$\text{For 2D CNNs, } \mathrm{I}_t^{2d}(k, u, v) = P_{t+k-1}(u, v) \tag{2}$$

The main difference between 2D and 3D approaches is that 2D CNNs only model temporal clues at the first convolution layer. On the contrary, 3D CNNs are capable of modeling temporal clues at all convolution layers. In Sect. 5.1, we will have some discussion of the pros and cons between 3D and 2D network.

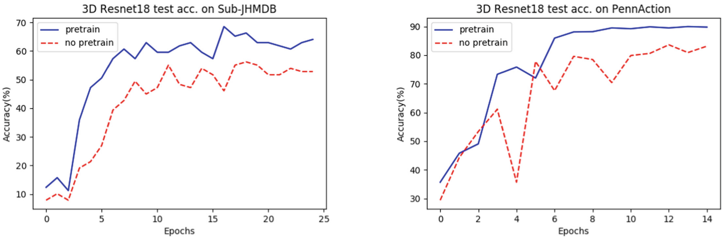
Human Detector: Since the intensity in the human pose heat maps is consistent with the region that contains human, our experimental results in Table 5 show that utilizing ground-truth bounding box to crop the person in action can significantly improve the action recognition accuracy and we also compare it with the state-of-the-art human detection method (Faster R-CNN [19]) in Table 5.

3.3 Transfer Learning for ImageNet Pretrained Weights

We also combine two powerful transfer learning techniques: the cross-modality pre-training [23] and bootstrapping 3D filter from 2D filter [4]. By applying the combination of these techniques on ImageNet pre-trained weights, we are able to transfer the knowledge from image domain to pose domain and bootstrap our pose-based 3D CNN. The combination method is constructed as follows. We first follow [23] to modify the weights of each convolution layer of ImageNet pre-trained model to handle the input of our stacked joint position field. More precisely, we average the weights across the RGB channels and replicate this average by the channel number of 2D pose-based network. Then we apply the idea in [4] to process the weights of 2D pose-based network for utilizing in 3D pose-based network. Since the architecture of 3D residual network is inflating from 2D residual network, for these 3D filters of size $N * N * N$ are formed by expanding an additional dimension from the 2D filters of size $N * N$. Thus, by repeating the weights of the 2D filters N times along the time dimension, and rescaling them by dividing by N, we can generate the weights for pose-based 3D CNN. The transfer learning method not only successfully reduces convergence time but also improves the accuracy of our model. The experimental result is shown in Fig. 4

3.4 Implementation Details

Hyperparameter: In the training stage, we set the initial learning rate of pose stream as 1×10^{-3} and it is divided by 10 when the validation accuracy is



(a) figures/Sub-JHMDB (b) PennAction

Fig. 4. The validation of transfer learning techniques on pose-based 3D resnet18. The blue line denotes the model with transfer learning from Imagenet and the red line denote the model without transfer learning techniques. (Color figure online)

saturated. The weight decay is set to be 1×10^{-4}, momentum is set to 0.9, and we use SGD as the optimizer for training. In the testing stage, we slice each video into a non-overlap 15 frames clips and the prediction score of each video is the average of the prediction of each clip in a video and this proposed framework is trained and tested with mini-batch size 32.

Data Augmentation: To boost the performance of our model, We utilize both spatial and temporal augmentation mechanism to train our model. For spatial augmentation, after generating the stacked joint position maps for a 15-frame video clip, we first resize it into $[15, 1, 128, 128]$ and utilize the random crop technique to crop a $[15, 1, 112, 112]$ joint position map as the input of our pose stream. For temporal augmentation, we follow [4] to pick a starting frame among those that guarantee a desired number of frames in each stack.

4 Multi-dimensional Fusion Network

In the two-stream architecture [21], a video sequence is first preprocessed to obtain the RGB frames and the optical flow maps, and then two independent CNNs are used to compute the spatial and temporal features, respectively. This framework provides the baseline model for action recognition research. Furthermore, there are several different extensions on this architecture for either better fusing the spatial and temporal feature [16] or enhancing the way of modeling temporal clues [23]. While the recognition accuracies reported by these previous works are quite excellent, we argue that precise human joint positions [13] provide very critical features for action recognition. Thus, we propose a pose-based 3D CNN based on Residual Network [10] as an additional stream on a multi-steam framework for action recognition.

4.1 Multi-dimensional Fusion

Due to the excellent performance for the two-stream architecture, we propose to include the pose-based approach into the two-stream framework to construct a

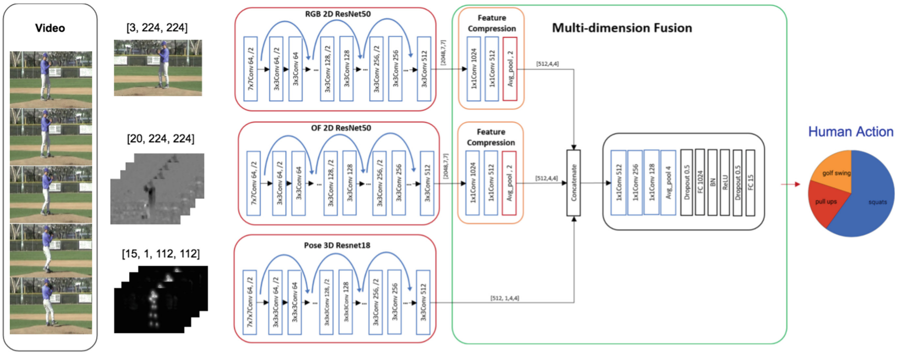


Fig. 5. Flow of the multi-dimensional fusion network: ($N \times N$ conv, k) denotes a convolutional layer with kernel size $N \times N$ and the output filter dimension k.

multi-stream CNN model, which combines spatial, short-term motion and human joint parts motion for action recognition. Motivated by the spatial-temporal fusion networks [8,16], we propose a novel fusion method to take advantage of the pose-based 3D CNNs and the two-stream CNNs such that channel responses of different types at the same pixel position were integrated appropriately. Here we intend to follow [16] to utilize convolution fusion for combining the feature maps for different streams. However, the main difficulty is that the previous fusion methods are only capable of fusing the feature maps of the same dimension. In other words, the previous fusion methods can not fuse feature maps of 2D and 3D networks.

Fusion of 2D and 3D CNN: To overcome the problem of fusing feature maps of different dimensions, we design a multi-dimensional fusion method. Here we demonstrate our techniques with spatial resnet50, motion resnet50 and pose 3D resnet18. Firstly, we follow [8,16] to extract the spatial, motion and pose feature maps before the average pooling layer, whose shapes are $[2048, 7, 7]$, $[2048, 7, 7]$, $[512, 1, 4, 4]$, respectively. Then we design a feature compression network for reducing the feature dimension of spatial and motion feature maps to $[512, 4, 4]$, which can be concatenated with the 3D pose-based feature maps.

Feature Compression Network: There are many existing methods to reduce the dimension of features (e.g. average pooling, max pooling and conv fusion). According to [16], which states that using multiple convolution layers to gradually reduce the dimension is better than directly applying average or max pooling. Hence, we follow the concept to design our network for dimension reduction. The implementation detail is depicted in Fig. 5.

4.2 Implementation Details

Multi-stream Model Input: We use RGB frames and stacked optical flow maps as the input of the spatial and motion streams, which follow the pre-processing procedures used in [21]. For the pose stream, we first apply the human pose computation method in [3, 20, 24] to generate 15-channel heatmaps, and then follow the procedure in Sect. 3.2 to generate the input of the pose-based CNNs.

Hyperparameter: In the training stage, the learning rate of multi-dimensional fusion network is initially set to 1×10^{-3} and divided by 10 when the validation accuracy is saturated. The weight decay is set to be 1×10^{-4}, momentum is set to 0.9, and we use SGD as the optimizer for the training. The testing scheme is similar to that used in the single pose-stream method given in Sect. 3.3. All of these models are trained and tested with mini-batch size 32.

Data Augmentation: For the spatial stream, we first resize an input image to $[256, 256]$ and apply random cropping to crop a $[224, 224]$ sub-image as the training data. For motion and pose streams, we utilize temporal augmentation mechanism proposed in [4] by picking a starting frame among those that guarantee a desired number of frames in each stack. The augmentation method of pose stream remains the same as the techniques described in Sect. 3.3.

5 Experimental Evaluation

To demonstrate the importance of the pose-based features, we evaluate our model on two widely used pose-related action recognition benchmarks; i.e., Sub-JHMDB [13] and PennAction [26] datasets. Furthermore, we use the published evaluation protocol of Sub-JHMDB (split1) and PennAction to report the classification accuracy for both datasets.

Table 1. Comparison of methods based on high-level pose features: SJP, OF, box denote stack joint position, optical flow and cropping by human detection bounding boxes, respectively, and L denotes the number of frames in each stack. Note that in these experiments, we first utilizing average fusion to model the features of each channel of human joint heatmap to test our pose-based method.

Input features	Network	Sub-JHMDB	PennAction
Stacked OF	2D resnet50	45.4	85.4
SJP	2D resnet50	57.3	86.1
SJP+ box	2D resnet50	61.8	89.8
SJP+ box	3D resnet18	67.4	90.0

5.1 Performance of Pose-Based CNN

To investigate the properties of pose-based CNN, we design several methods to utilize high-level pose features in CNN and compare them to the original stacked optical flow methods, which is proposed in [21].

Furthermore, we evaluate the proposed method with different stack lengths (L), which denote the temporal footprints [4] in the videos. In action recognition related field, the temporal footprint is a key factor to design an action recognition algorithm. Therefore, we validate the effectiveness of the temporal high-level features on the two popular benchmarks.

According to Table 1, we find the method that utilizes 3D resnet18 to model the stack joint position maps cropped with the associated ground-truth bounding boxes outperforms the performance of optical flow-based methods and the pose-based methods on 2D network.

Furthermore, we also conduct experiment to determine the best architect of capturing temporal footprints in video. In Table 2, we show that when the number of input frames L, which is the temporal footprint of our model set to 15, gives the better accuracy than the original flow-based methods for both datasets.

The experiment results not only demonstrate that pose-based features perform better for extracting action features than optical flow based features, but also show that 3D resnet18 has superior performance to 2D resent18. We believe the main reason why a 3D network is superior than 2D network is that using only a single convolution layer for modeling human pose features is insufficient to capture temporal action from video.

Table 2. Comparison of the performance of different temporal footprint: SJP, OF, Box denote stack joint position, optical flow and cropping by ground truth bounding boxes, respectively, and L denotes the temporal footprint, which is the number of frames in each stack.

Input features	Network	Sub-JHMDB	PennAction
(a) Temporal footprint L = 10			
Stacked OF	2D resnet50	46.1	87.1
Stacked OF+ Box	2D resnet50	60.8	88.2
SJP+ Box	3D resnet 18	62.5	90.0
(b) Temporal footprint L = 15			
Stacked OF	2D resnet50	41.1	86.3
Stacked OF+ box	2D resnet50	59.5	89.4
SJP+ box	3D resnet18	68.5	91.5

5.2 Channel-Wise Convolution

After we determine the architecture of classification model, we propose an experiment of channel-wise convolution to validate the most effective way to combine different human body part features. With the experimental results given

Table 3. Comparison of different strategies of combining the features of each human part: 3D convN (a, b) denotes a 3D convolutional layer with kernel size $N \times N \times N$, where a is the input filter dimension and b is the output filter dimension.

Human-part pooling method	Sub-JHMDB	PennAction
Average pooling	68.5	90.5
3D conv1 (15,1)	69.2	91.1
3D conv1 (15,7) + 3D conv1 (7,1)	71.4	92.9

in Table 3, We found that using two 3D convolution layers to model the features from all channels of pose estimation result outperforms other methods. Therefore, we demonstrate the effectiveness of this architecture in the proposed pose-based action recognition model.

5.3 Performance of Multi-dimensional Fusion Network

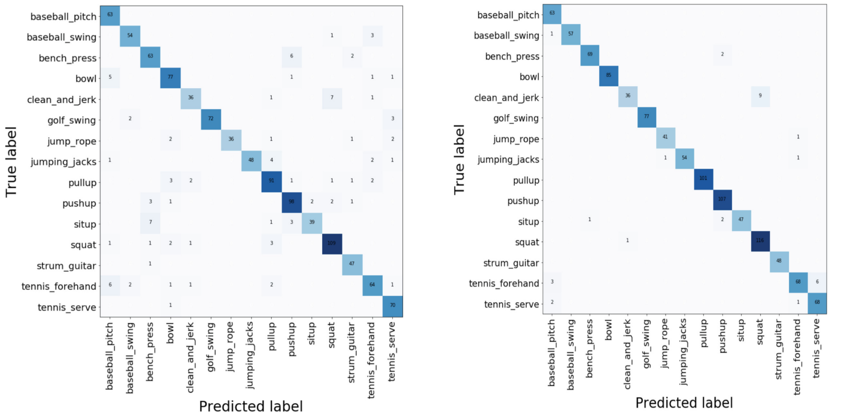
According to the fact that 3D spatial (3 channels) and 3D temporal (2 channels) networks contain more parameters than the proposed 3D pose model (1 channel), which makes them suffer from the over-fitting problem, our proposed method is based on fusing 3D pose stream, 2D spatial stream and 2D temporal stream to achieve superior or comparable performance compared to the state-of-the-art methods. The comparison of different fusion methods is shown in Table 4. Our multi-dimensional fusion framework outperforms the previous average fusion and convolution fusion methods. In addition, We also provide the confusion matrices of the proposed pose-based 3D CNN and multi-stream network in Fig. 6 to demonstrate the effectiveness of the proposed framework.

5.4 Comparison of Performance for Using Ground-Truth and Human Detector Bounding Boxes

To validate the proposed model, we compare the performance of our framework by using the ground-truth human bounding boxes and the bounding boxes obtained from the Faster R-CNN [19] human detector. The results are shown in Table 5. According to the experimental results, when we replace the ground-truth human bounding boxes by those obtained from the state-of-the-art human detector, we have slightly degraded accuracies in our experiment on the PennAction dataset. Therefore, we claim that in the proposed framework, ground truth human bounding box can be replaced by a state-of-the-art human detector with slight accuracy decrease.

5.5 Comparison with State-of-the-art Methods

We also evaluate our pose-based CNN and multi-stream network by comparing performance with state-of-the-art posed-related action recognition methods.



(a) Prediction of pose features (b) Prediction of multi-stream features

Fig. 6. The confusion matrix of the proposed multi-stream fusion is much sparser than the confusion matrix by using the proposed pose-based method. It demonstrates the effectiveness of fusing multiple types of features, and our fusion strategy significantly enhances the performance for human action recognition.

Table 4. Comparison with different fusion schemes: In the fusion part, s, m, p denote spatial, motion and pose streams, respectively. MD-fusion denotes multi-dimensional fusion, conv fusion denotes fusion with convolution layers and average fusion denotes the fusion method proposed in [21], which simply fusion the scores of all streams at the output of the softmax layer.

Stream	Sub-JHMDB	PennAction
Spatial	55.1	80.2
Motion	60.7	87.1
Pose	68.5	91.5
Fusion	*Sub-JHMDB*	*PennAction*
s+m, conv fusion	70.2	92.4
s+p, average fusion	69.7	91.9
s+p, MD-fusion	74.1	95.6
s+m+p, average fusion	71.0	93.7
s+m+p, MD-fusion	78.9	97.6

In Table 6, the result of our multi-dimensional fusion network provides superior performance on the methods based on either hand-crafted features or deep learning approaches. Finally, we successfully justify our argument that with proper modeling and fusion techniques, human pose features can be directly applied to 3D convolution neural networks to model the temporal evolution in videos and significantly enhance the performance of human action recognition.

Table 5. Performance comparison of the proposed framework by using the ground-truth bounding boxes and those obtained from a human detector on PennAction dataset. In this table, gt-Bbox denotes ground-truth bounding box, sf-Bbox denotes the bounding box generated by human detector, and MD-fusion denotes multi-dimensional fusion.

Framework	Bounding box	Performance
Pose-stream	Ground truth	90.5
Pose-stream	Faster R-CNN	90.1
MD-fusion	Ground truth	97.8
MD-fusion	Faster R-CNN	97.6

Table 6. Comparison of state-of-the-art action recognition methods on Sub-JHMDB [13] and PennAction [26] datasets.

State-of-the-art	Stream	Sub-JHMDB	PennAction
Actemes [26]	RGB	-	79.4
pose+NTraj [13]	Pose	75.1	-
SP-AOG [18]	RGB + Pose	61.2	85.5
P-CNN [5]	RGB + Flow + Pose	66.8	-
JDD [2]	RGB + Flow	77.7	87.4
C3D [2]	RGB + Flow	-	86.0
pose [12]	Pose	61.5	79.0
Pose + idt-fv [12]	Pose + Flow	74.6	92.9
RPAN [7]	RGB + Flow + Pose	78.6	97.4
Pose-stream (our)	Pose	**71.4**	**92.9**
Pose+MD-fusion (our)	RGB + Flow +Pose	**78.9**	**97.6**

6 Conclusion

In this paper, we presented a novel multi-stream action recognition method based on fusing 3D pose, 2D spatial and 2D temporal features. We develop a pose-based 3D CNN which integrates multi-channel human joint heatmaps with channel-wise convolution and applied 3D CNN to extract spatial and temporal features at the same time. In addition, we propose a multi-dimensional fusion method that bridges the gap between dimension differences between the 2D spatial, 2D motion and 3D pose feature maps. Our experiments showed the proposed multi-stream CNN model outperforms the state-of-the-art methods on both Sub-JHMDB and PennAction datasets.

References

1. ActionVLAD: learning spatio-temporal aggregation for action classification. In: CVPR (2017)
2. Cao, C., Zhang, Y., Zhang, C., Lu, H.: Action recognition with joints-pooled 3D deep convolutional descriptors. In: IJCAI (2016)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR (2017)
5. Chéron, G., Laptev, I.: P-CNN: pose-based CNN features for action recognition. In: ICCV (2015)
6. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
7. Du, W., Wang, Y., Qiao, Y.: Rpan: an end-to-end recurrent pose-attention network for action recognition in videos. In: ICCV (2017)
8. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016)
9. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3D residual networks for action recognition. In: ICCV (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: a survey. Image Vis. Comput. **60**(Suppl. C), 4–21 (2017)
12. Iqbal, U., Garbade, M., Gall, J.: Pose for action – action for pose. In: FG (2017)
13. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV (2013)
14. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. TPAMI **35**(1), 221–231 (2013)
15. Kay, W., et al.: The kinetics human action video dataset. ArXiv:1705.06950v1 [cs.CV] (2017)
16. Ma, C.Y., Chen, M.H., Kira, Z., AlRegib, G.: TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. ArXiv:1703.10667v1 [cs.CV] (2017)
17. Ng, J.Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: CVPR (2015)
18. Nie, B.X., Xiong, C., Zhu, S.C.: Joint action recognition and pose estimation from video. In: CVPR (2015)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. https://arxiv.org/pdf/1506.01497.pdf
20. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV (2015)
23. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2

24. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
25. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: ACM MM (2016)
26. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: a strongly-supervised representation for detailed action understanding. In: ICCV (2013)