Action Recognition of Basketball Players Based on PWC-Net and 3D CNN

Jinyu Na, Kaijie Shen,and Chen Xu

Abstract—Basketball is one of the most popular exercise in the world. The statistics analysis of basketball matches can provide coaches and audiences with valuable information. However, players' behavior in matches contains a lot of complicated information, it takes a lot of manpower and material resources to process these data and obtain valuable information. Therefore, it is becoming more and more important to continuously improve the statistics of basketball matches. Statistical indicators of various data are an important basis for measuring the performance of players. These indicators can be obtained by analyzing the actions of players in basketball match videos.
This paper introduces a fusion model based on PWC-Net and 3D convolution neural network(3D CNN), which can quickly and accurately classify the players' sports actions (3-pointer, free throw, layup, dunk, steal and other 2-pointer) in basketball match videos. This paper counted the accuracy of each action and analyzed the factors affecting the accuracy.

Index Terms—Machine learning, Action Recognition, Optic Flow, PWC-Net, 3D CNN.

1 Introduction

This paper is mainly based on the mechanical learning of basketball players' action recognition, through the basketball game video to analyze the players' action on the field. This paper focuses on the action recognition of players, because with the development of society, more and more people pay attention to basketball, which makes the action analysis of players in basketball matches an urgent matter. Traditional statistics are realized manually, which has the disadvantages of low efficiency and poor success rate. This paper chooses 3D CNN in mechanical learning and PWC-Net in optical flow estimation algorithm to study, and finally realizes efficient and fast player action recognition. Compared with other types of action recognition, there are interferences on information extraction caused by individual motion occlusion, complex background information and motion blur in basketball video scenes. Therefore, this paper chooses the optical flow estimation algorithm to extract the player's action features, and relieves the influence of interference information in feature extraction by extracting inter-frame optical flow information. PWC-Net is a kind of optical flow estimation algorithm, which has the characteristics of small model parameters and fast operation speed. On the basis of the estimated playground, this paper uses 3D CNN to realize the spatio-temporal modeling of sequence data and extract the spatio-temporal information in the video. 3D CNN can extract the spatio-temporal characteristics of the video more specifically than 2D CNN, and then uses softmax to realize the prediction probability of basketball semantic event categories.

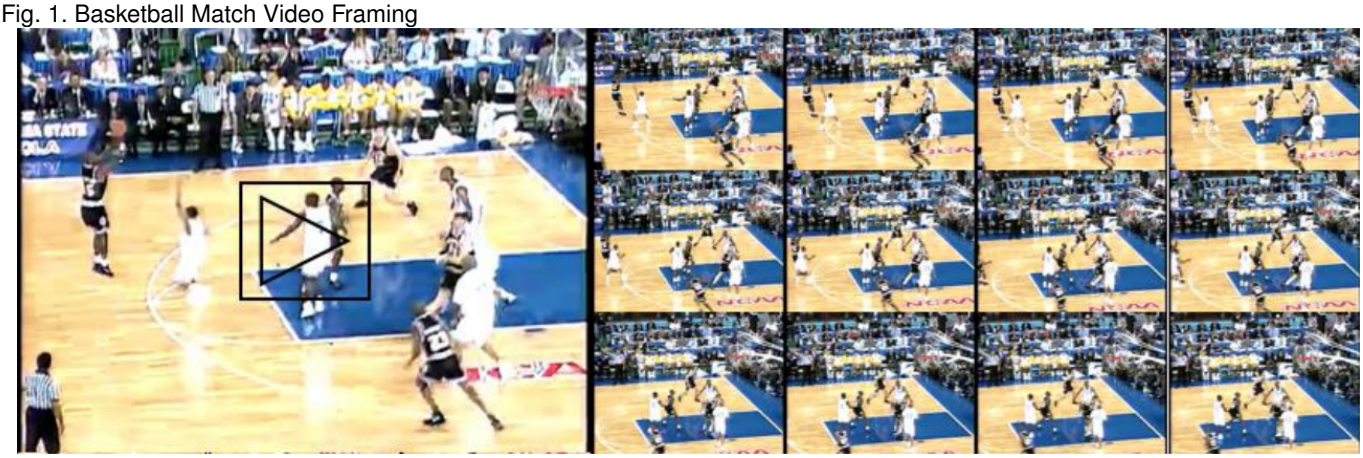
2 Data Source

The data set of this paper selects 53 NCAA basketball matches on YouTube. These videos were randomly divided into 20 training videos and 33 test videos. The input of PWC-Net is two consecutive pictures. Therefore, we need to frame the downloaded video file to obtain the continuous picture sequence we need. In this paper, each video is divided into segments with a time interval of 4 seconds by annotating the boundary, and they are subsampled to 6fps. Then, 6 types of events were identified and marked manually, namely 3-pointer, free throw, layup, steals, dunks and other 2-pointer.

3 Feature Extraction

Inter-frame optical flow field estimation based on PWC-Net[1] is one of the current mainstream methods. PWC-Net is a compact and effective convolution neural network optical flow model. Its model

Fig. 1. Basketball Match Video Framing



is small in size, easy to train, and fast in running speed. Its innovation lies in the characteristic pyramid, warping and the design of cost volume. In a learnable feature pyramid, PWC-Net uses the current optical flow estimation to wrap the convolution neural network features of the second image. Then, a cost volume is constructed by using the features of the first image and the warping features, and the convolution neural network estimates the optical flow by processing the cost volume.

3.1 PWC-Net Model

Inter-frame optical flow field estimation based on PWC-Net[2] is one of the current mainstream methods. PWC-Net is a compact and effective convolution neural network optical flow model. Its model is small in size, easy to train, and fast in running speed. Its innovation lies in the characteristic pyramid, warping and the design of cost volume. In a learnable feature pyramid, PWC-Net uses the current optical flow estimation to wrap the convolution neural network features of the second image. Then, a cost volume is constructed by using the features of the first image and the warping features, and the convolution neural network estimates the optical flow by processing the cost volume.

Given two input images I_1 and I_2, we generate L-level pyramids of feature representations. At the l_{th} level, we warp features of the second image toward the first image using the $\times 2$ upsampled flow from the $l + 1_{th}$ level:

$$c_w^l(x) = c_2^l(x + up_2(w^{l+1})(x))$$

where x is the pixel index and the upsampled flow $up_2(w^{l+1})$ is set to be zero at the top level.

Next, we use the features to construct a cost volume that stores the matching costs for associating a pixel with its corresponding pixels at the next frame [3]. We define the matching cost as the correlation between the features of the first image and the winding features of the second image [4]:

$$cv^l(x_1, x_2) = \frac{1}{N}(c_1^l(x_1)^T c_w^l(x_2))$$

where T is the transpose operator and N is the length of the column vector $c_1^l(x_1)$.

Optical flow estimator is a multi-layer CNN. Its input are the cost volume, features of the first image, and upsampled optical flow and its output is the flow w^l at the l_{th} level.

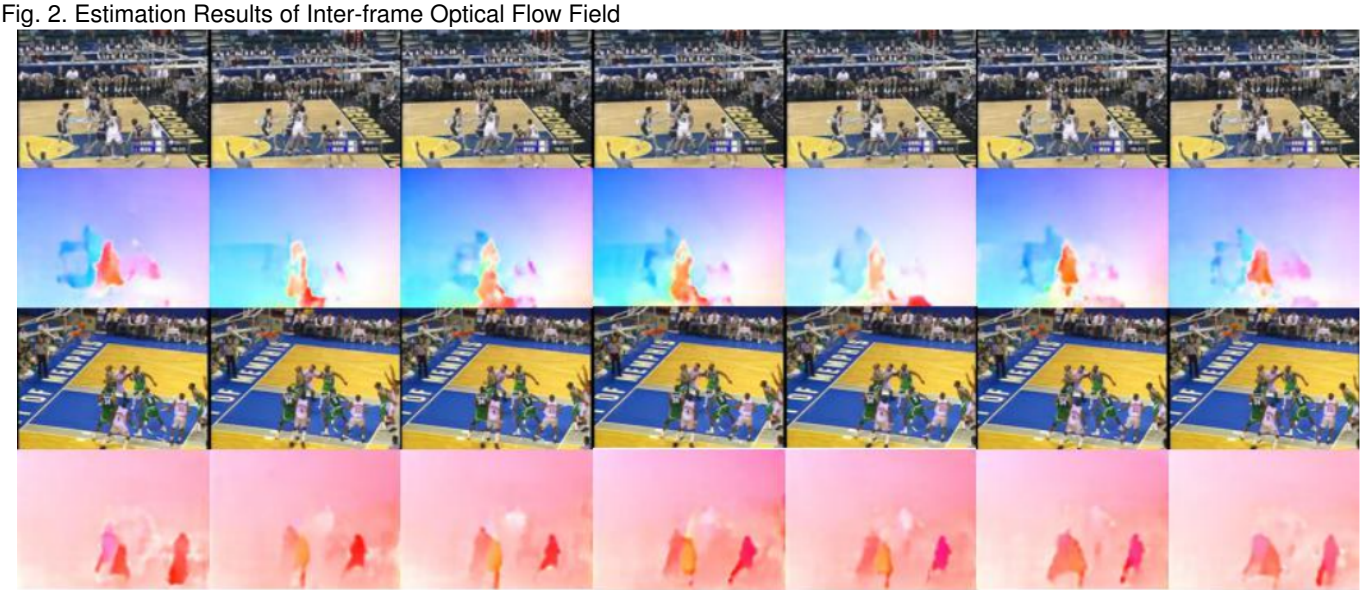
The context network takes the estimated flow and features of the second last layer from the optical flow estimator and outputs a refined flow.

Let Θ be the set of all the learnable parameters in our final network, which includes the feature pyramid extractor and the optical flow estimators at different pyramid levels (the warping and cost volume layers have no learnable parameters). Let w_Θ^l denote the flow field at the lth pyramid level predicted by the network, and w_{GT}^l the corresponding supervision signal. We use the same multiscale training loss proposed in FlowNet [5]:

$$\tau(\Theta) = \sum_{l=l_0}^{L} \alpha_l \sum_x \left| w_\Theta^l(x) - w_{GT}^l(x) \right|_2 + \gamma \left| \Theta \right|_2$$

where $|\cdot|_2$ computes the $L2$ norm of a vector and the second term regularizes parameters of the model.

Fig. 2. Estimation Results of Inter-frame Optical Flow Field



3.2 Visualization of Optical Flow Estimation Results

The optical flow estimation result is visualized as the input of 3D CNN hereinafter. Optical flow can also be defined as the distribution of apparent velocities of movement of brightness pattern in an image.In the picture after optical flow visualization, different colors indicate different motion directions, such as light blue for the right, blue for the upper right, green for the lower right, red for the left, purple for the upper left, and brown for the lower left. While depth indicates the speed of movement. The darker the color, the faster the speed. The lighter the color, the slower the speed. As shown in Fig. 2:

4 Action Recognition

4.1 3D CNN

The research object of this paper is the player's actions in the basketball match video, which includes the time and space features, while 3D CNN is more suitable for learning the time and space features than traditional convolution neural network. Therefore, this paper chooses 3D CNN model as the processing model of photo streaming pictures for the classification of players' actions. The overall network structure in this paper includes 8 convolution layers, 5 pooling layers, 3 full connection layers, and 2 BN layers. The activation function of the model is ReLU, the parameter optimizer is Adam optimizer, and the initial learning rate is set to 0.001.

4.2 Model Test

In the video test phase, video framing is carried out on the video of 33 matches in the test set, the obtained continuous picture sequence is taken as the input of PWC-Net, the inter-frame optical flow field estimation of 33 basketball matches is extracted, and then the inter-frame optical flow field estimation result is color coded and converted into [0-255] three-channel RGB images. The input of 3D convolution neural network takes 16 consecutive frames of data as a sequence unit, and the resolution of the data is 112×112. In the initialization stage of test data, 16 frames of pictures after visualizing the optical flow field in the sequence are sampled at equal intervals and the image size is adjusted to 112×112, so that the data dimension of the test set is $1 \times 16 \times 112 \times 112$. The updated dimension test set data is taken as the input of the model, and the output result is the probability distribution of 6 types of events. The test process is shown in the Fig. 3.

Fig. 3. Testing Process

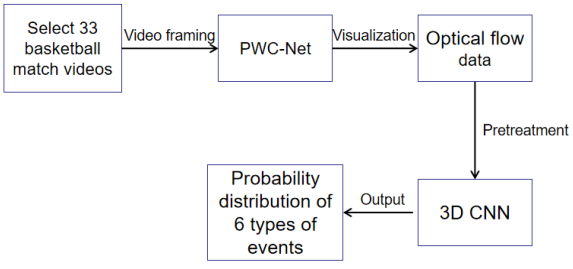


TABLE 1
Basketball Action Recognition Test Results

Action Event		Actions Classification						Accuracy/%
		3-pointer	Free throw	layup	Other 2-pinter	dunk	steal	
Action Recognition Results	3-pointer	**255**	0	2	105	1	1	70.05
	Free throw	4	**54**	3	4	2	0	80.60
	layup	44	0	**101**	130	2	4	35.94
	other 2-pointer	154	0	33	**176**	3	2	47.83
	dunk	2	0	7	5	**4**	0	22.22
	steal	3	0	3	5	0	**406**	97.36

5 Results and Analysis

The results of the test set after passing through the feature classification model are shown in Table 1. Table 1 is the confusion matrix of the six sports actions output by the feature classification model. From left to right, the confusion matrix respectively represents three-pointer, free throw, layup, other two-pointer, dunk and steals. From top to bottom, it respectively represents three-pointer, free throw, layup, other two-pointer, dunk and steals and accuracy.

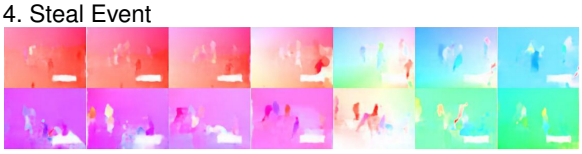
In this paper, the test results are measured by the accuracy rate. The calculation formula of the accuracy rate is as follows:

$$Accuracy = \frac{Number\ of\ correct\ samples}{Total\ number\ of\ samples}$$

The accuracy rate of recognizing the steal is the highest. This is because the occurrence of steals will cause the players of both sides to reverse their movement direction. Defensive players will attack immediately after successful steals, while offensive players will return to their own courts to defend immediately after steals. This reversal of movement direction is reflected in the color coding transformation for players in the picture after optical flow visualization. Secondly, when there is a steal, the camera will also reverse its direction of motion. In basketball matches, the lens will follow the movement of the ball and move in the opposite direction if the ball right is transferred. This is reflected in the change of the background color of the picture after optical flow visualization, which is often from one extreme to the other, such as red (left) to blue (right), as shown in Fig. 4 below. Finally, this kind of color transformation appears in almost all the captured optical flow pictures, with strong regularity.

Free throws are the second most accurate event. During the free throw, the fouled player stands on the free throw line to make a free throw, while the other players of both teams wait for rebounds near

Fig. 4. Steal Event



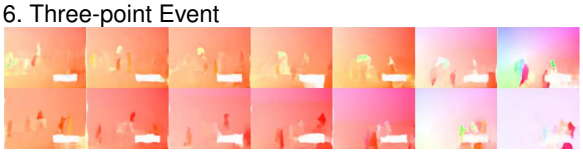
the free throw line. When a player makes a free throw, other players are still. The lens focuses on the area where the free throw is located. As both the players and the basketball are in an approximately stationary state, the lens is also stationary. Therefore, the color of the player and the background in the photo after optical flow visualization is white, as shown in fig. 5. The free throw incident also has a strong regularity. In most free throw incidents, the background of the picture is mainly white.

Fig. 5. Free throw event

The event with the third highest accuracy rate is the 3-pointer. The player's three-point shot is divided into two steps, i.e. before the three-point shot and after the three-point shot. Before shooting a 3-point shot, the camera places the player in the center of the camera and moves along with the player's movement. After the three-point shot is made, the camera moves to the vicinity of the basket and zooms to determine if the goal is scored. Therefore, the background color of the first half of the picture sequence after optical flow visualization is the same, and the background color of the second half of the picture sequence is different. The player's movement rule is different before and after the 3-point shot. Before the shot, the defensive player is in a one-on-one defensive state. After the shot, the defender made a quick move to the basket because he wanted to get the rebound. Before and after the

shot, the player's movement will change drastically, which is reflected in the color darkening in the photo of optical flow visualization, as shown in fig. 6.

Fig. 6. Three-point Event



The accuracy of layups and other two-point shots is not high, because their movement patterns are similar to each other. The layup occurred mainly in the area under the basket, while the other two points were also likely to occur in the area under the basket. As can be seen from table 1, nearly half of the layup events are identified as other two points, which is due to the same movement pattern of the two. It can be seen from fig. 7 that there is no obvious distinction between the two events and there is no obvious feature for classification.

Fig. 7. Layup event and other two-pointer event

The accuracy rate of dunking is the lowest, because the samples of dunking are few and it is difficult to learn robust features. The small number of dunking training videos results in poor robustness of dunking events. Secondly, the movement pattern of dunking is similar to that of layup. As can be seen from Table 1, a considerable number of dunking events are identified as layups. Dunk events are mainly concentrated in the area under the basket, which is similar to the area where the dunk event occurs, as shown in Figure 5 below. The picture in the first line is easy to identify because the shot was enlarged and focused on the basket at the moment the player dunked. The picture of the second line is not easy to identify because the lens is not zoomed at this time, which is similar to the layup picture of the first line in fig. 8, and thus is identified as layup.

Fig. 8. Dunk event

6 Conclusion

Based on PWC-Net and 3D CNN, this paper realizes the action recognition of basketball players. The network structure, principle, testing process and results are described in detail. The prediction results of 6 kinds of events are analyzed in detail, and the high accuracy of steals, 3-pointers and free throws and the low accuracy of layups, other 2-pointers and dunks are explained by combining the players' movement patterns, optical flow charts and sample numbers. The research work in this paper mainly focuses on the data preprocessing and neural network model training process. The interference factors in the image are eliminated through preprocessing, and then the feature classification model is obtained through training the neural network. Finally, the feature classification model is verified by selecting 33 NCAA basketball matches on YouTube as test sets. Compared with other optical flow estimation algorithms, the inter-frame optical flow estimation algorithm based on PWC-Net has more advantages in model size and running speed, but there is still some room for improvement in the effect of optical flow visualization. Many photos after optical flow visualization are very blurred, as shown in Figure 1 below. In several pictures in fig. 9, it is difficult to distinguish the position of the players and their movement patterns with naked eyes. the players are integrated with the background, which may be caused by discontinuous front and back frames or violent shaking of the lens.

Fig. 9. Blurred Visualized Pictures

Motion recognition based on 3D CNN is more accurate than other motion recognition technologies, but it is not accurate enough in some similar events, such as layups and other two-point shots. The similar aspects of these events make it difficult for the neural network to identify the difference between the two, thus making classification difficult. Secondly, the number of samples is also a key factor affecting the recognition efficiency. Too few samples make the robustness of neural network training poor. The solutions to these two situations are:

- choose a more stable lens or a higher resolution picture.
- Optimize the network structure of PWC-Net and 3D CNN.

- Expand the database and increase the number of samples.

ACKNOWLEDGMENTS

Thanks to Professor Erdogmus for the support and instruction.

REFERENCES

[1] Sun D, Yang X, Liu M Y, et al. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. *Densely connected convolution networks. In IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017.

[3] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. *Fast cost-volume filtering for visual correspondence and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), 2013.

[4] J. Xu, R. Ranftl, and V. Koltun. *Accurate optical flow via direct cost volume processing. In IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017.

[5] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, et al. *FlowNet: Learning optical flow with convolution networks. In IEEE International Conference on Computer Vision* (ICCV), 2015.