Received: 31 January 2021 | Revised: 24 April 2021 | Accepted: 3 May 2021

IET Image Processing

DOI: 10.1049/ipr2.12265

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Latent label mining for group activity recognition in basketball videos

Lifang Wu[1] | Zeyu Li[1] | Ye Xiang[1] | Meng Jian[1] | Jialie Shen[2]

[1] Beijing University of Technology, Beijing, China

[2] School of Electronics, Electrical Engineering and Computer Science, Queen's University of Belfast, Belfast, UK

Correspondence
Ye Xiang, Beijing University of Technology, 100 pingleyuan, Chaoyang District, Beijing, China.
Email: xiangye@bjut.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61702022, 61802011, 61976010; Beijing Municipal Education Committee Science Foundation, Grant/Award Number: KM201910005024; Beijing University of Technology Ri Xin Cultivation Project

Abstract

Motion information has been widely exploited for group activity recognition in sports video. However, in order to model and extract the various motion information between the adjacent frames, existing algorithms only use the coarse video-level labels as supervision cues. This may lead to the ambiguity of extracted features and the omission of changing rules of motion patterns that are also important sports video recognition. In this paper, a latent label mining strategy for group activity recognition in basketball videos is proposed. The authors' novel strategy allows them to obtain the latent labels set for marking different frames in an unsupervised way, and build the frame-level and video-level representations with two separate levels of supervision signal. Firstly, the latent labels of motion patterns are digged using the unsupervised hierarchical clustering technique. The generated latent labels are then taken as the frame-level supervision signal to train a deep CNN for the frame-level features extraction. Lastly, the frame-level features are fed into an LSTM network to build the spatio-temporal representation for group activity recognition. Experimental results on the public NCAA dataset demonstrate that the proposed algorithm achieves state-of-the-art performance.

1 | INTRODUCTION

Content-based sports video analysis has been attracting significant attentions from the field of computer vision, owing to its widespread applications in real world [1–4]. Among all the related directions in content-based analysis for sports videos, effective group activity recognition has wide range applications in facilitating athletic training improvement, fast video browsing and accurate video retrieval [5, 6]. For the popular broadcast basketball videos, the task of group activity recognition becomes more significant. This task, however, may easily suffer from some severe difficulties, including the frequent interactions between players, cluttered background and high similarity among different categories, thus is quite challenging.

To solve the challenges in group activity recognition, the motion patterns across video frames are mostly extracted, which can both help avoid the interference of background noise and discover the intrinsic distinctions between group activity classes. The motion patterns, as a particular and effective modality of data, are actually composed of global motion patterns and local motion patterns. Global motion means the camera movement, while local motion basically refers to the players' movement. In a video clip, there are often some different kinds of motion patterns with fixed changing rules, which can be exploited as the intrinsic and discriminative features for recognizing the specific group activity class. Take the class of the three-point as an example, there are both different types of global and local motion patterns, the global camera motion includes panning or tilting firstly and zooming in on the basket lastly, meanwhile, the local motion includes one player in a certain region moving vigorously and others moving slightly. Both these different motion patterns and their typical changing rules can be very helpful to define the group activity.

Existing works utilize the motion patterns for group activity recognition mainly in two ways: (i) take an optical flow calculation method to estimate the motion fields, from which the features representing the motion information are extracted by a neural network model; (ii) employ a 3D CNN to implicitly extract the motion information across different frames for improving the recognition accuracy. For example, some

algorithms [7–9] take the two-stream neural networks to extract the two kinds of features separately. Thereinto, one stream works on extracting the appearance features from the RGB frames, the other stream works on extracting the motion features from the motion fields. The classification results from the two streams are added lastly with equal weights. There are also some other algorithms [10–13], which utilize the 3D CNN to extract the spatial and temporal features simultaneously. By using the filters that can operate across different frames, the motion information is extracted implicitly.

For both (i) and (ii), only one single kind of supervision signal, that is, the video-level labels, is used for the construction of feature representation. This may lead to a deficiency that the embeddings are unable to represent the rich and various motion information associated with the sequential multi-frames within one same video clip. Besides, for (ii), the 3D CNN framework normally requires high computational complexity and is weak at modeling long term information variation, which may lead to the omission of significant motion features and the reduction of recognition accuracy.

In view of these problems, this paper aims at mining the latent labels of motion patterns from the frames and further combining two levels of supervision signal to obtain the effective spatio-temporal representation. Specifically, we explore an unsupervised way to mine the latent labels for frames, including the motion field estimation, statistic features extraction and hierarchical clustering steps. Each cluster is tagged with a latent label of motion pattern. The generated latent labels are then used as the frame-level supervision signal to train a deep CNN for the frame-level features extraction. Lastly, the frame-level features are fed into an LSTM (Long Short-Term Memory) network, which takes the video-level labels as supervision signals to obtain the spatio-temporal representation. The spatio-temporal representation can thus depict various motion patterns and their particular changing rules more explicitly and effectively improve the group activity recognition performance. The main contributions of this work are summarized as follows:

- We propose a group activity recognition algorithm framework for basketball datasets. Our method combines a latent label mining strategy without additional network structure, which will be more conducive to the industrial deployment of the algorithm in basketball games.
- A two-stage training method combining frame-level labels is used for network learning, which allows CNN to learn more detailed spatial features.
- Extensive experiments on the basketball benchmark dataset namely NCAA have been conducted. The results demonstrate that our method outperforms current state-of-the-art methods for group activity recognition, which indicates the effectiveness of the proposed method.

2 | RELATED WORK

Video understanding and analysis is one of the important task of computer vision. We introduce the related work from two aspects of video action recognition and group activity recognition, respectively.

2.1 | Video action recognition

One major kind of the existing algorithms based on the deep neural networks exploits the 3D CNNs or their variants. The 3D filters or pseudo-3D filters slide in both spatial and temporal dimensions to let the spatial and temporal representations be learned simultaneously. For example, Tran et al. [10] proposed a neural network called C3D, which extended the 2D convolution to 3D convolution to capture the appearance features and the temporal dynamics between consecutive frames at the same time. The C3D, however, can not fully utilize the existing 2D CNN models that have been pre-trained effectively on a large-scale dataset. To this end, Carreira and Zisserman [12] proposed the I3D, which inflated the deep pre-trained 2D CNNs for image classification into spatio-temporal feature extractors. Furthermore, to reduce the computational complexity and prevent the over-fitting problem, several algorithms [11, 13–15] proposed to decouple the 3D convolution filter to a 2D spatial convolution filter followed by a 1D temporal convolution filter to decrease the number of parameters. All these 3D CNNs can enhance the recognition performance. However, they also suffer from high computational complexity.

Another major kind of existing algorithms based on deep neural networks is mainly based on two-stream architecture. One of the typical examples is the model proposed by Simonyan and Zisserman [7], which applies one stream to extract the appearance features and another stream to calculate the motion information from the estimated optical flow [16]. The final predictions for videos were averaged over the two streams, which were trained separately. This framework has attracted significant attentions recently and extended by many works. For example, a network called TSN [17] was proposed to use the two-stream framework to capture the features from the short snippets, which were extracted from the long videos by a sparse sampling scheme. However, there is a problem that in the typical two-stream architecture, the interactions between the frames and the modalities are actually very limited. Some algorithms [18, 19] also noted this and studied the fusion strategies in the middle of the two streams, but the problem still remains open.

Apart from the 3D CNNs and two-stream architecture, the ConvLSTM structure is also frequently used. The ConvLSTM structure basically refers to taking the 2D CNNs and LSTM network to build the frames and videos representation, respectively [20]. Du et al. [21] further introduced the attention mechanism into the ConvLSTM structure, so that the network only paid attention to the areas that were strongly related to the behavior category. Wu et al. [22] took the optical flow as input and combined the sequential CNNs and LSTM for basketball event prediction. Yang et al. [23] proposed a two-stage scheme on the basis of the ConvLSTM structure. Briefly speaking, the ConvLSTM structure-based algorithms usually have less computational complexity, but the performance is uncompetitive.

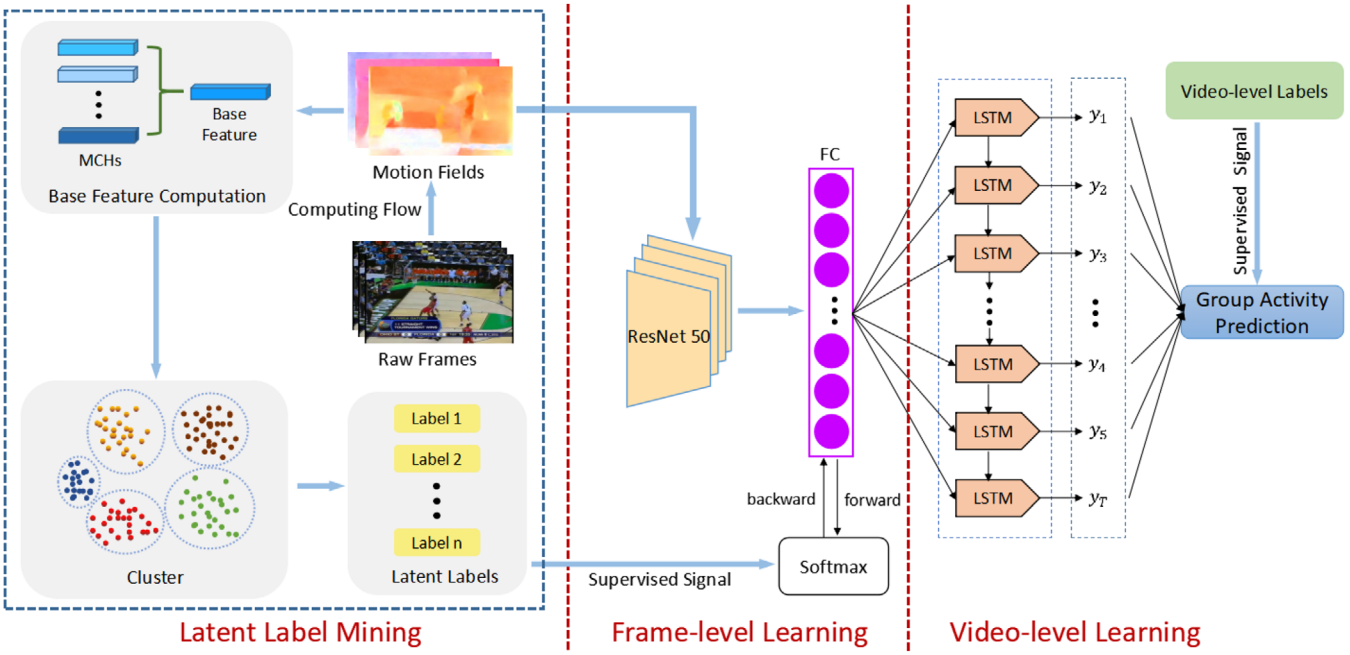


FIGURE 1 The framework of the proposed scheme.

2.2 | Group activity recognition

Group activity recognition methods have attracted more and more attention in recent years due to their important applications in sports event analysis and video surveillance [24, 30, 37]. Traditional group activity recognition methods [25, 26] were mostly based on hand-crafted features and probabilistic graphical models to predict group activities. With the rapid development of deep learning, some methods combined with RNN further improved the recognition performance. Ibrahim et al. [27] used a hierarchical LSTM model to aggregate individual actors information for the understanding of the entire activity. Shu et al. [28] proposed a two-level hierarchy of LSTMs that minimized predicted energy and maximized confidence. Bagautdinov et al. [29] maintained the temporal consistency of each actor in the video by the RNN structure. Wang et al. [31] used LSTM to unify the interactive feature modeling process of single-person dynamics, intra-group and inter-group interactions. Li and Chuah [32] used the method of generating captions for each video frame to infer group activities. Ibrahim and Mori [33] constructed the relationship representation of each person and used it for group activity prediction. PC-TDM [34] proposed a Participation-Contributed Temporal Dynamic Model, which aggregated temporal dynamics of key actors with different participation degrees over time from each person to recognize group activity. In addition to RNN-based methods, there are some other deep learning methods worthy of attention. Azar et al. [35] generated intermediate spatial representations (activity maps) based on individual and group activities to iteratively refine group activity predictions. Wu et al. [36] built a flexible and efficient Actor Relation Graph (ARG) to simultaneously capture the appearance and position relation between actors.

This paper focuses on the ConvLSTM structure. By exploring an unsupervised way to introduce the additional latent frame-level labels into the ConvLSTM structure, we aim to improve the recognition performance with less computational complexity.

3 | THE PROPOSED SCHEME

3.1 | Overview

The proposed scheme for group activity recognition, as shown in Figure 1, mainly contains three key modules: latent label mining, frame-level features extraction and video recognition with temporal representation, respectively. Firstly, the motion fields are estimated for each pair of adjacent frames, from which the latent labels of various motion patterns are mined by an unsupervised algorithm. Secondly, the generated latent labels are used as the frame-level supervision signal to train a CNN model, by which the frame-level features that can help identify various motion patterns are extracted. Finally, an LSTM model, taking the frame-level features as input, is utilized to get the temporal representation for group activity recognition in video clips.

The overall loss for training is

$$L = L_{\text{frame-level}} + L_{\text{video-level}}, \tag{1}$$

in which $L_{\text{frame-level}}$ and $L_{\text{video-level}}$ represent the cross entropy losses for the construction of frame-level and video-level representations, respectively. During the whole scheme, we employ the $L_{\text{frame-level}}$ and $L_{\text{video-level}}$ in the second and third steps separately. Below we will introduce the three steps in detail.

3.2 | Latent label mining

To discover latent labels of various motion patterns in a video clip, the motion field between two consecutive frames is generally estimated at first. With the motion field as basic input, the deep motion information will be easily extracted and depicted. Among existing works, there are many classic optical flow methods for motion field estimation [38–40]. This paper employs the PWC-Net [41], which is an end-to-end convolutional neural network and can make a good balance between the accuracy and computational cost. By using PWC-Net, the motion field with two channels is obtained, representing the pixel displacement in the horizontal (x-component) and vertical (y-component) directions, respectively.

Since a motion field is obtained, an unsupervised algorithm called Motion Characteristic Histogram (MCH) is introduced to map the motion field into statistic features. The MCH mainly focused on summarizing the appearance characteristics of each point in the motion field. Specifically, a point in the motion field is characterized by its motion direction and amplitude. The point can be regarded as stationary when its motion amplitude is less than or equal to 0.2 pixels. To improve the robustness of MCH and alleviate the inference from noise points, the motion field is further divided into $n \times n$ local non-overlapping regions of equal size, the MCH for each local region is computed independently. Lastly, MCHs of all local regions are concatenated in sequence to form the ultimate statistical representation of the motion field.

Based on the statistical representation of the motion field for the frame, a bottom-up aggregation strategy using a hierarchical clustering algorithm is proposed for the latent labels mining. As shown in Algorithm 1, the frames in the training set are clustered into k classes hierarchically by an unsupervised way, each class has a representative vector, that is, the cluster center. The resulting total k classes are namely the desired latent labels set. For each frame in training set, the label of motion pattern could be obtained by the nearest distance between its MCH and all representative vectors, that is, $\{r_1, r_2, \dots, r_k\}$. Sequential frames in one video clip are thus tagged with different labels, delivering rich motion information that can help improve the recognition performance. Several samples of the motion fields tagged with different labels are visualized in Figure 2. We can see that both the motion intensity and motion direction are similar in certain regions for the three motion fields in each row, which have the same latent label.

ALGORITHM 1 Hierarchical Clustering for Frame-level Latent Labels Generation

Input: Statistical feature vectors $\{f_1, f_2, \dots, f_m\}$ for different frames

Output: Representative vectors $\{r_1, r_2, \dots, r_k\}$ for latent labels

// Distance matrix construction
for $i = 1$ to m **do**
  for $j = 1$ to m **do**
    $D_{ij} = \cos(f_i, f_j)$;
    $D_{ji} = D_{ij}$;
  end for
end for
Find the smallest value $D_{i^*j^*}$ in D;
// Hierarchical clustering
$k \leftarrow m$;
Copy $\{f_1, f_2, \dots, f_m\}$ into $\{r_1, r_2, \dots, r_k\}$;
while $D_{i^*j^*} < \theta$ **do**
  Merge r_{i^*} and r_{j^*}: $r_{i^*} = Avg(r_{i^*} + r_{j^*})$;
  for $j = j^* + 1$ to k **do**
    Renumber the cluster r_j as r_{j-1};
  end for
  Delete the j^*-th row and j^*-th column of D;
  $k \leftarrow k - 1$;
  for $j = 1$ to k **do**
    $D_{i^*j} = \cos(r_{i^*}, r_j)$;
    $D_{ji^*} = D_{i^*j}$;
  end for
  Relocate the smallest value $D_{i^*j^*}$ in D;
end while
return remaining cluster centers $\{r_1, r_2, \dots, r_k\}$, *i.e.*, representative vectors for latent labels

FIGURE 2 Visualization of the samples of motion fields tagged with different latent labels. Each row corresponds to the one same latent label. Color hue indicates the motion direction and color value denotes the motion intensity.

3.3 | Frame-level features extraction

The latent labels generated from the previous step actually form the additional high-level semantic information besides initial video labels. These additional labels could act as a frame-level supervision signal for frame-level features extraction. Comparing with a scheme using video labels for both frame-level and video-level features extraction in existing works [22, 23], the scheme using the frame and video labels for frame-level and

video-level features extraction separately is more reasonable by precisely depicting the various frames.

After latent labels as frame-level supervision signals are determined, the frame-level features could be effectively extracted through a deep CNN. This paper employs the common VGG-11 network [42] with batch normalization layers [43], yet makes a minor modification. Specifically, the output dimensions of the first two fully-connected layers in the original VGG-11 are both adjusted to 1024, since the number of classes to be recognized is rather few. After the second fully-connected layer, the frame-level features, that is, output vectors of length 1024, are just extracted.

The cross entropy loss function is widely used in group activity recognition task, so the loss function is defined as follows:

$$L_{\text{frame-level}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} f_{ij}\log(x_{ij}), \tag{2}$$

where n denotes there are n samples in a batch, k is the number of categories of latent labels, f_{ij} means the j^{th} latent label of the i^{th} image, and x_{ij} means the probability that the i^{th} image belongs to the j^{th} latent label outputed by CNN.

3.4 | Video recognition with temporal representation

Frame-level feature extraction by using deep CNN intrinsically produce the spatial representation of frames. To construct the temporal representation for video recognition, this paper takes an additional LSTM network [44]. LSTM has a function of judging the important degree of features at each time node due to its unique memory cells and gate operations. Hence the discriminative features can be just preserved and the redundant information can be just forgotten over time. LSTM also introduces the notion of memory state, which can suppress the vanishing gradient and exploding gradient problems effectively. By integrating the VGG-11 network with frame-level labels as supervision signal and the LSTM network with video-level labels as supervision signal, a new ConvLSTM structure that models the spatio-temporal representation of video clips is established.

At the end of the ConvLSTM structure, a fully connected layer with softmax activation is utilized for group activity classification. Thereinto, the number of neurons is set as the total number of group activity categories. It is worth noting that the cross entropy loss function is still used as the loss function for LSTM network training, which is defined as follows:

$$L_{\text{video-level}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\log(x_{ij}), \tag{3}$$

where n denotes there are n samples in a batch, m is the number of categories of video-level labels, f_{ij} means the j^{th} video-level label of the i^{th} image, and x_{ij} means the probability that the i^{th} image belongs to the j^{th} video-level label outputted by LSTM network.

4 | EXPERIMENTS

In this section, we show that the proposed scheme can significantly improve the performance of group activity recognition on basketball videos. All experiments are conducted on the public NCAA dataset. Firstly, different parameters are tested to find the optimal setting. Using the optimal parameter setting, an ablation study is made to verify the effectiveness of the latent label mining strategy. At last, comparison with the state-of-the-arts is performed.

4.1 | Experiment settings

4.1.1 | Dataset

NCAA is a challenging public dataset for group activity recognition in broadcast basketball videos, which was released by Ramanathan et al. [1]. This dataset is collected from Youtube, in which videos across different venues and different periods of time are contained. There are 257 basketball videos in NCAA, each lasts about 1.5 h long. All videos are randomly split into training, validation and testing set with 212, 12 and 33 videos, respectively. To understand and analyze these videos, the 6 types of group activities are defined, including three-point, free-throw, lay up, two-point, slam dunk and steal, as shown in Figure 3. For each activity, the start and endpoints are marked manually through a crowdsourcing platform, generating the video clip. As shown in Figure 4, each group activity class has different number of video clips. Average length of a video clip is about 45 frames, which covers the essential context for group activity recognition.

4.1.2 | Evaluation metrics

We use two evaluation metrics to measure the recognition effectiveness, accuracy and confuse matrix, which are defined as follows:

- Accuracy is the most basic evaluation index in recognition. It is the percentage of the samples with correct prediction in all the samples, and accuracy is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

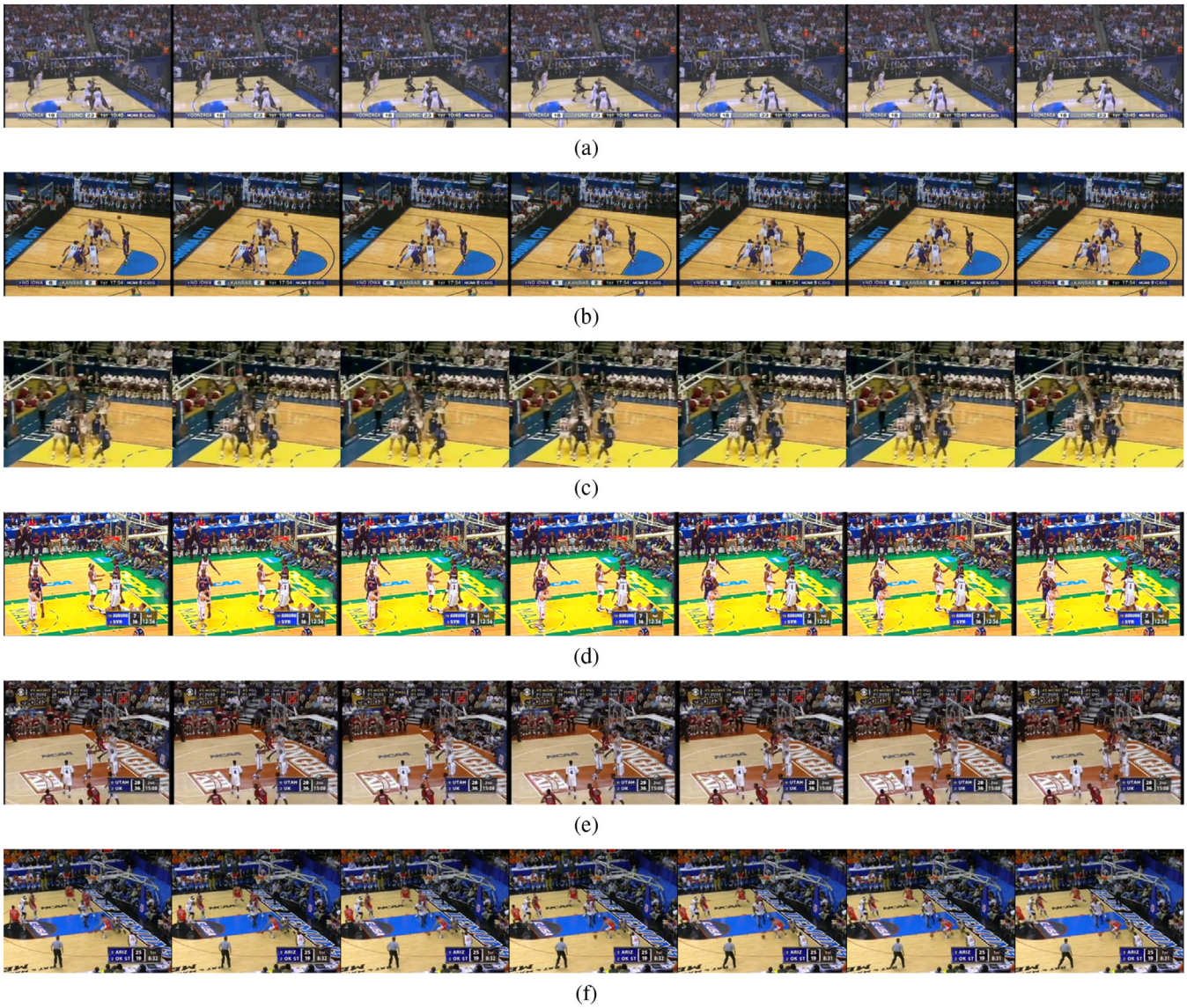


FIGURE 3 Some example points from the NCAA dataset, we randomly selected five consecutive frames from each type of event for display, the first row is three-point, the second row is free-throw, the third row is lay up, the fourth row is two-point, the fifth row is slam dunk, and the last row is steal.

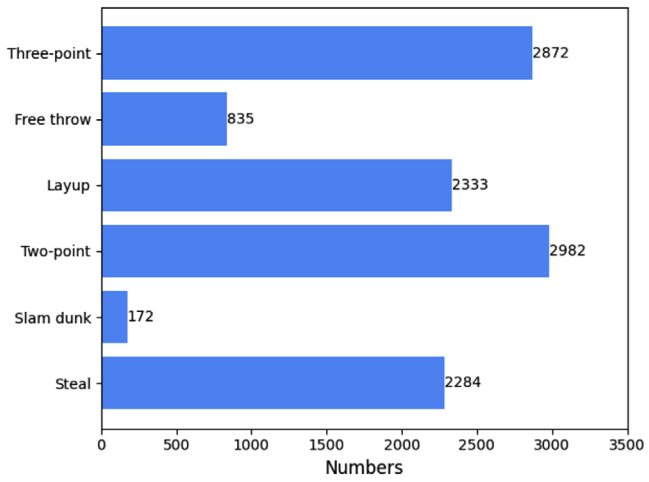


FIGURE 4 The data distribution of the NCAA dataset

where *TP* is the number of samples that were predicted to be positive and were actually positive, *FP* is the number of samples that were predicted to be negative and were actually positive, *FN* is the number of samples that were predicted to be positive and actually were negative, and *TN* is the number of samples that were predicted to be negative and actually were negative.

- The confusion matrix reflects the confusion degree of classification results, and the effect of classification can be seen intuitively through visualization. In general, its vertical axis is the predicted category of the sample, and its horizontal axis is the true category of the sample. Suppose a confusion matrix is $A_{n\times n}$, then A_{ij} represents the number of samples that are actually class i but are identified as class j. The diagonal line A_{ij} represents the number of samples that are actually class i and identified as class j, that is, the number of samples that are correctly predicted.

TABLE 1 Comparison of classification accuracies using different numbers of non-overlapping regions

Regions number	Features dimension	Accuracy (%)
3×3	81	67.10
4×4	144	68.23
5×5	225	66.54

TABLE 2 Comparison of classification accuracies using different merging threshold θ

θ	Categories number	Accuracy (%)
0.4	110	55.61
0.5	44	62.47
0.6	14	68.23

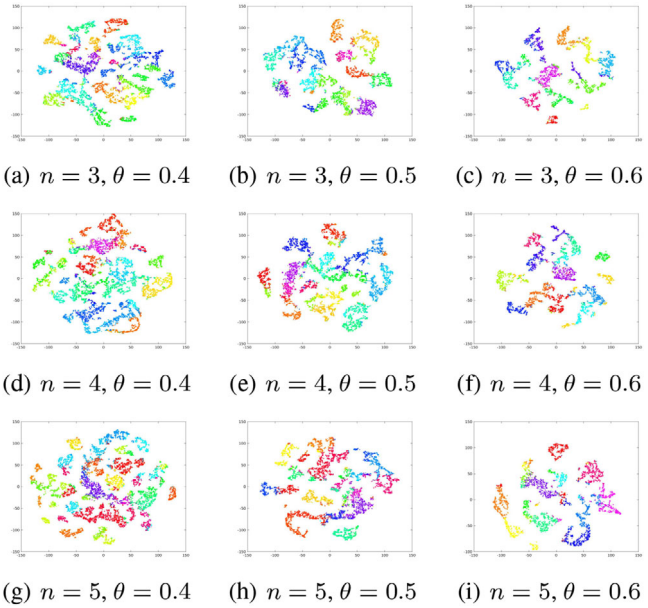
4.1.3 | Implementation details

During the data pre-processing stage, linear transformation is utilized for the motion field, which serves as input for latent label mining and frame-level features extraction. After transformation, element values within the motion field are limited to the range of [0, 255]. Taking the processed motion field as input, the two networks for frame-level and video-level representation construction are sequentially trained. For frame-level features extraction, the VGG-11 network is initially pre-trained on ImageNet dataset [45] and then fine-tuned on the NCAA dataset. During the fine-tuning stage, an oversampling algorithm is used to ensure that the sample numbers for all categories are basically equal within a mini-batch. In order to achieve video-level representation construction, the LSTM network contains 16 units, allowing 16 frames in one video clip as input. For the training of both networks, the optimizer of Stochastic Gradient Descent (SGD) is employed. Specifically, the weight decay is set to 5e-4, the momentum is set to 0.9, and the batch size is set to 64 and 30 for VGG-11 and LSTM, respectively. The whole training process is implemented on the deep learning framework Pytorch with a Nvidia Titan X GPU.

4.2 | How do the parameters influence the performance

4.2.1 | Number of non-overlapping regions

To improve the robustness of features for mining latent labels in Section 3.2, we divide the motion field into $n \times n$ local non-overlapping regions of equal size and compute the statistic features of each local region independently. For an experiment designed to find the optimal setting, different values of n, including 3, 4, 5 are tested, respectively. With the increment of sub-regions from 3×3 to 5×5, the dimension of generating statistic features increases from 81 to 225. It is a common idea that a higher feature dimension can represent the motion field more accurately, which ought to generate better clustering results and achieve higher classification accuracy. However, as shown in Table 1, the accuracy drops when the number of sub-regions increases from 16 to 25. This may be in part due to the existence of noise components in the motion field, such as scoreboard regions and the motion generated by the audiences. In these circumstances, the finer division of the motion field will easily amplify the inference from the noisy component. Hence the best result is obtained by $n = 4$, which is utilized through the following experiments.



(a) $n = 3, \theta = 0.4$ (b) $n = 3, \theta = 0.5$ (c) $n = 3, \theta = 0.6$

(d) $n = 4, \theta = 0.4$ (e) $n = 4, \theta = 0.5$ (f) $n = 4, \theta = 0.6$

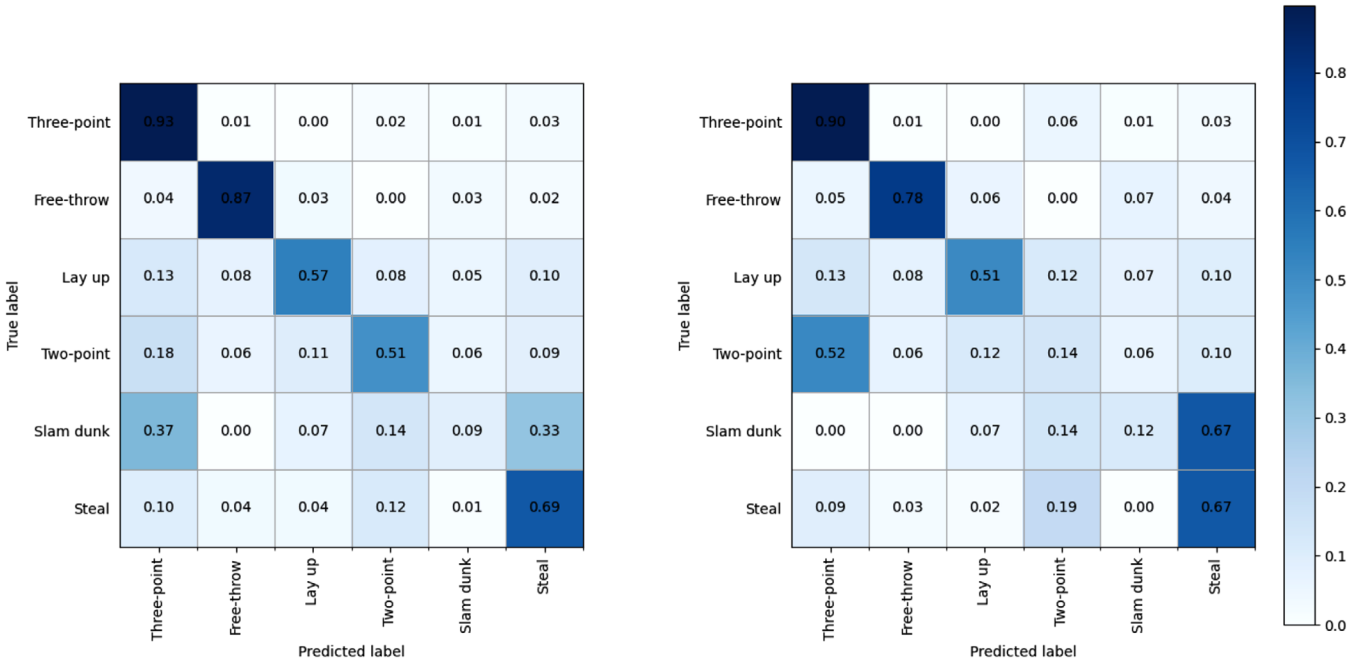
(g) $n = 5, \theta = 0.4$ (h) $n = 5, \theta = 0.5$ (i) $n = 5, \theta = 0.6$

FIGURE 5 Visualization of clustering result of feature embeddings using different n and θ

4.2.2 | Merging threshold θ

In Algorithm 1, a merging threshold denoted by θ is used as the maximal distance between two examples that are allowed to be merged into one cluster. To evaluate the influence of θ, different values including 0.4, 0.5 and 0.6 are tested, respectively. The bigger the θ is, the weaker the merging constraint becomes, thus the fewer categories will be remained. As shown in Table 2, the number of categories reduces from 110 to 14 as the θ changes from 0.4 to 0.6, while the classification accuracy increases progressively from 55.61% to 68.23%. This is probably because among numerous categories, the lack of data for certain categories can easily lead to the over-fitting problem. In consequence, the θ is set to 0.6 for an optimal setting.

We further visualize the clustering result of feature embeddings that belong to different classes of motion patterns. By using the t-SNE algorithm, the result with different n and different θ is shown in Figure 5. It can be seen that the distribution of features is always more regular when $\theta = 0.6$. Based on this, the result in Table 2 also becomes easier to understand. Therefore, by using $\theta = 0.6$, the proposed latent label mining algorithm is believed that it will obtain an effective frame-level feature representation and help improve the video recognition performance.



(a) with latent label mining (b) without latent label mining

FIGURE 6 Comparison of confusion matrices between the frameworks with and without latent label mining

4.3 | Ablation study

We conduct an ablation study for the proposed latent label mining strategy. For the framework without latent label mining, the motion field is directly fed into a ConvLSTM structure, which only takes the video labels as supervision signals to obtain both the frame-level and video-level representation. To make a fair comparison, the deep CNN and LSTM network in ConvLSTM structure is exactly the same across the frameworks with or without latent label mining.

The confusion matrices are compared and shown in Figure 6. We can observe that the proposed framework with latent label mining clearly outperforms the framework without latent label mining for almost all group activity categories except the slam dunk. This may be because the slam dunk owns too few examples, and during the hierarchical clustering, its motion patterns could be easily mixed into those of other categories and then fade out in the average operation. As a whole, the comparison results fully prove the effectiveness of the proposed latent label mining strategy.

4.4 | Comparison with state-of-the-arts

To evaluate the performance of the proposed scheme, we apply the state-of-the-art algorithms for the purpose of comparison. The core results are summarized in Table 3. It is not hard to see that the two lowest accuracies are achieved by GMP based ConvLSTM and On_GCMP, which, respectively, utilizes the typical ConvLSTM structure and the two-stream architecture. The C3D, R(2+1)D, P3D and I3D are all based on the 3D CNNs or their variants, which obviously achieve better performance than GMP based ConvLSTM and On_GCMP. However, the 3D CNNs based algorithms usually require much higher computational complexity. Among the four 3D CNNs based algorithms, C3D is a regular 3D CNN, which is improved by the following R(2+1)D, P3D and I3D. R(2+1)D and P3D both decouple the 3D convolution filter to a 2D spatial convolution filter followed by a 1D temporal convolution filter to reduce the number of parameters. Fewer parameters enable R(2+1)D and P3D to have a deeper structure of neural networks, and obtain better recognition performance than C3D which only has eight convolution layers. The I3D performs even better than R(2+1)D and P3D, and achieves the second-best recognition accuracy.

TABLE 3 Comparison of accuracies with state-of-the-arts on the public NCAA dataset

Method	Accuracy (%)
GMP based ConvLSTM [22]	60.28
On_GCMP [23]	60.96
C3D [10]	65.02
R(2+1)D [11]	65.21
P3D [13]	66.83
I3D [12]	67.56
Iterative optimization-based model [46]	68.80
Ours	70.92

FIGURE 7 Some prediction results of our method

It is mainly due to the fact the I3D takes full advantage of the 2D CNN models that are pre-trained on a large-scale dataset. Iterative optimization-based model decoupled global motion and local motion to get better efficiency, the performance of local motion is 68.80% better than that of the global motion and mix motion, but is still 2.12% lower than our algorithm. Comparing with these different algorithms, our proposed scheme achieves the highest accuracy 70.92%. We surpass both the typical GMP based ConvLSTM and the advanced 3D CNNs based algorithms, which fully demonstrates the effectiveness of the proposed scheme. Besides, we take an unsupervised way to introduce the additional latent frame-level labels into the ConvLSTM structure, hence the computational complexity is relatively low. Figure 7 shows some visualization results of

successful and failure cases of our method. The first five rows are, respectively, successful cases for three-point, free-throw and two-point, while the last row shows the unsuccessful case for lay-up. From the optical flow field, it can be seen that the motion patterns adopted in our algorithm for three classes, that is, three-point, free-throw and two-point, are significantly different, they have latent labels with certain rules of variation, and are thus discriminative for recognition. Therefore, our proposed algorithm can effectively improve the group activity recognition performance. However, there are still some unsuccessful cases. It can be seen that the motion patterns for lay up and two-point are quite similar, which may lead to the misclassification.

5 | CONCLUSION

In this paper, we introduce a latent label mining strategy for group activity recognition in basketball videos. Through an unsupervised hierarchical clustering technique, the latent frame-level labels for categorizing various motion patterns are generated. Taking the latent labels as supervision signals, the frame-level features extracted by a deep CNN are attached with more explicit semantics. Following the deep CNN, a novel LSTM network is utilized to build the final spatio-temporal representation, which reflects the fixed changing rules of motion patterns over time and can significantly improve the recognition performance. Experimental results on the public NCAA dataset fully demonstrate the effectiveness of our proposed scheme. In the future, we plan to introduce the latent label mining strategy into the two-stream architecture for further performance improvement.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (61976010, 61802011, 61702022), Beijing Municipal Education Committee Science Foundation (KM201910005024), and Beijing University of Technology Ri Xin Cultivation Project.

ORCID

Zeyu Li https://orcid.org/0000-0001-9249-9432

REFERENCES

1. Ramanathan, V., et al.: Detecting events and key actors in multi-person videos. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3043–3053. IEEE, Piscataway (2016)
2. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Proceedings of European Conference on Computer Vision, pp. 721–736. Springer, Berlin (2018)
3. Qi, M., et al.: stagNet: An attentive semantic RNN for group activity recognition. In: Proceedings of European Conference Computer Vision, pp. 101–117. Springer, Berlin (2018)
4. Jian, M., et al.: Deep key frame extraction for sport training. Neurocomputing 328, 147–156 (2019)
5. Xie, L., et al.: Dynamic multi-view hashing for online image retrieval. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 3133–3139. ACM, New York (2017)
6. Wang, L., et al.: Enhancing sketch-based image retrieval by CNN semantic re-ranking. IEEE Trans. Cybern. 50(7), 3330–3342 (2020)
7. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of Advances in Neural Information Processing Systems, pp. 568–576. MIT Press, Cambridge, MA (2014)
8. Feichtenhofer, C., et al.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1933–1941. IEEE, Piscataway (2016)
9. Wang, L., et al.: Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of European Conference on Computer Vision, pp. 20–36. Springer, Berlin (2016)
10. Tran, D., et al.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision, pp. 4489–4497. IEEE, Piscataway (2015)
11. Tran, D., et al.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 6450–6459. IEEE, Piscataway (2018)
12. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 6299–6308. Springer, Berlin (2017)
13. Qiu, Z., et al.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of IEEE International Conference on Computer Vision, pp. 5534–5542. IEEE, Piscataway (2017)
14. Luo, C., Yuille, A.L.: Grouped spatial-temporal aggregation for efficient action recognition. In: Proceedings of IEEE International Conference on Computer Vision, pp. 5512–5521. IEEE, Piscataway (2019)
15. Sudhakaran, S., et al.: Gate-shift networks for video action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1102–1111. IEEE, Piscataway (2020)
16. Zach, C., et al.: A duality based approach for realtime tv-l 1 optical flow. Proceedings of Joint Pattern Recognition Symposium, pp. 214–223. Springer, Berlin Heidelberg (2007)
17. Wang, L., et al.: Temporal segment networks for action recognition in videos. IEEE Trans. Pattern Anal. Mach. Intell. 41(11), 2740–2755 (2018)
18. Christoph, R., Pinz, F.A.: Spatiotemporal residual networks for video action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 3468–3476 (2016)
19. Feichtenhofer, C., et al.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4768–4777. IEEE, Piscataway (2017)
20. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2625–2634. IEEE, Piscataway (2015)
21. Du, W., et al.: Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In: Proceedings of IEEE International Conference on Computer Vision, pp. 3725–3734. IEEE, Piscataway (2017)
22. Wu, L., et al.: Global motion pattern based event recognition in multi-person videos. In: Proceedings of CCF Chinese Conference on Computer Vision, pp. 667–676. Springer, Singapore (2017)
23. Wu, L., et al.: Ontology based global and collective motion patterns for event classification in basketball videos. IEEE Trans. Circuits Syst. Video Technol. 30(7), 2178–2190 (2019)
24. Wang, Q., et al.: Detecting coherent groups in crowd scenes by multiview clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42(1), 46–58 (2018)
25. Amer, M.R., et al.: Hirf: Hierarchical random field for collective activity recognition in videos. In: Proceedings of European Conference on Computer Vision, pp. 572–585. Springer, Berlin (2014)
26. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Proceedings of European Conference on Computer Vision, pp. 215–230. Springer, Berlin (2012)
27. Ibrahim, M.S., et al.: A hierarchical deep temporal model for group activity recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1971–1980. IEEE, Piscataway (2016)

28. Shu, T., et al.: Cern: Confidence-energy recurrent network for group activity recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 5523–5531. IEEE, Piscataway (2017)
29. Bagautdinov, T., et al.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4315–4324. IEEE, Piscataway (2017)
30. Wang, Q., et al.: Nwpu-crowd: A large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. 43(6), 2141–2149 (2020)
31. Wang, M., et al.: Recurrent modeling of interaction context for collective activity recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3048–3056. IEEE, Piscataway (2017)
32. Li, X., Choo Chuah, M.: Sbgar: Semantics based group activity recognition. In: Proceedings of IEEE International Conference on Computer Vision, pp. 2876–2885. IEEE, Piscataway (2017)
33. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Proceedings of European Conference on Computer Vision, pp. 721–736. Springer, Berlin (2018)
34. Yan, R., et al.: Participation-contributed temporal dynamic model for group activity recognition. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1292–1300. ACM, New York (2018)
35. Azar, S.M., et al.: Convolutional relational machine for group activity recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 7892–7901. IEEE, Piscataway (2019)
36. Wu, J., et al.: Learning actor relation graphs for group activity recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 9964–9974. IEEE, Piscataway (2019)
37. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vision 129, 225–245 (2021)
38. Dosovitskiy, A., et al.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766. IEEE, Piscataway (2015)
39. Ilg, E., et al.: Flownet 2.0: Evolution of optical flow estimation with deep networks.: In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2462–2470. IEEE, Piscataway (2017)
40. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4161–4170. IEEE, Piscataway (2017)
41. Sun, D., et al.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 8934–8943. IEEE, Piscataway (2018)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:14091556, 2014
43. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint, arXiv:150203167, 2015
44. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)
45. Krizhevsky, A., et al.: Imagenet classification with deep convolutional neural networks. proc. In: Advances in Neural Information Processing Systems, pp. 1097–1105. MIT Press, Cambridge, MA (2012)
46. Wu, L., et al.: Global motion estimation with iterative optimization-based independent univariate model for action recognition. Pattern Recognit. 116, pp. 107925 (2021)

How to cite this article: Wu, L., et al.: Latent label mining for group activity recognition in basketball videos. IET Image Process. 1–11 (2021). https://doi.org/10.1049/ipr2.12265