Key frame and skeleton extraction for deep learning-based human action recognition

Hai-Hong Phan
hongpth@lqdtu.edu.vn

Trung Tin Nguyen
tinnt@lqdtu.edu.vn

Ngo Huu Phuc
phucnh@lqdtu.edu.vn

Nguyen Huu Nhan
Nhan9ckl@gmail.com

Do Minh Hieu
dmh29112000@gmail.com

Cao Truong Tran
truongct@lqdtu.edu.vn

Bao Ngoc Vi
ngocvb@lqdtu.edu.vn

Data Science Research Group
Faculty of Information Technology
Le Quy Don Technical University, Ha Noi, Vietnam

***Abstract*—In this paper, we propose an efficient framework for action recognition in videos with key frame extraction and deep learning architectures, named KFSENet. First, we propose a key frame extraction technique in a motion sequence of 2D frames based on gradient of optical flow to select the most important frames which characterize different actions. From these frames, we extract key points using pose estimation techniques and employ them further in an efficient Deep Neural Network (DNN) in order to learn the action model. In this way, the proposed method be able to remove redundant frames and reduce the length of the motion. We only consider the remaining essential informative frames in the process of action recognition, thus the proposed pipeline is sufficiently fast and robust. We evaluate the proposed method intensively on publicly available benchmark named UCF Sport and our self-built HNH dataset in our experiments. We demonstrate that our proposed algorithm achieves state-of-the-art on these datasets.**

***Index Terms*—human action recognition, key frame extraction, skeleton, keypoints**

I. Introduction

Identifying human actions in video is one of the most difficult topic in computer vision. Its applications are from sports training to healthcare and physical rehabilitation interactive entertainment, and video understanding [1]. To do that, we need to process a huge amount of information from the input video. There are two approaches to solving this problem. The first method is used focuses on improving precision by creating larger architectures [2]. A number of other studies, using the latter as the direction we use, focus on the problem of more concise, standard action recognition. This reduces the computational cost of action recognition but also increases accuracy and enables real-time execution.

Action recognition can be divided into three steps of input data processing, action representation and action classification. Improving any of the steps above can effectively increase recognition accuracy and speed.

978-1-6654-0435-8/21/$31.00 ©2021 IEEE

With input data processing, many studies have shown that videos contain highly temporally redundant data, making it easier to skip parts without losing much information and some action classes in standard datasets do not require motion or temporal information to be identified. Only a few key frames are used as inputs to the identification [3] [4] [5] [6].

One of the most efficient action representation methods in recent years is the use of skeleton data. Skeleton-based action recognition is essentially a timed human pose motion analysis [7], in which posture is recognized from a sequence of joints over time. The joints can be obtained from different input data; it can be in the form of a simple RGB video or a spatio-temporal joint trajectory acquired by means of a sensor system or an estimated 3D skeleton from image data or a hybrid system. This is a very effective method suggested by many authors to simplify the network structure [8] [9] [10].

In this paper, we propose a new framework, named key frame and skeleton networks (KFSENet), that allows to extract key frames from a motion sequence based on optical flow, a traditional but highly efficient technique; and adopt pose estimation techniques and deep models in order to category actions. More specifically, we built an algorithm that uses a gradient to filter out the most significant frames with the biggest difference in the sequence as the key frames. Although there have been many studies using deep learning models to obtain key frames, most of the current methods often have the disadvantage of large model size and slow execution speed [11] [12] [13]. Besides, we also integrate deep network models for skeleton-base action representation and action classification. Thus, the proposed framework be able to capture the characterization of each action and receive benefit from pose estimation in order to recognize actions using deep networks.

To thoroughly evaluate the proposed approach, we experiment on a popular benchmark dataset containing 10 action classes, namely UCF Sport [14] and a self-built data set containing six classes of action. We analyze the effectiveness of our proposed key frame extraction algorithm on overall action

video recording in terms of accuracy and time complexity. We compare our approach with other modern approaches and conclude that our proposed pipeline is not only fast enough computationally, but also achieves high efficiency in terms of accuracy. The test results on the UCF Sport dataset give over 93% accuracy and on our HNH dataset achieves approximately 99%. We report all the particulars of experiments for the configuration of the deep neural network architectures in Section 4. Our main aim is to simplify the model while still providing high efficiency in action recognition. Figure 1 illustrate the pipeline of our method.

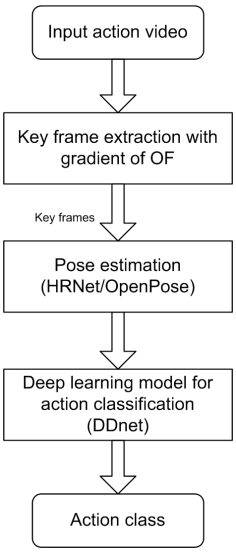

Fig. 1. Illustration of the pipeline of our proposed method

The remaining paper is organized as follows: Section 2 reviews related works in human action recognition base on skeleton and deep learning models; Section 3 describes the proposed method, while Section 4 shows the experimental results and discussions; and finally conclusions are given in Section 5.

II. Related work

In this subsection, we briefly review the existing literature closely related to the proposed model to deal with the two main issues on human skeleton-based action recognition. The first problems is feature representation of skeleton joints and the modeling of temporal dynamics to recognize human actions. Then, the efficient approaches detect key frames in videos.

A. Feature representation of skeleton joints and the modeling of temporal dynamics to recognize human actions

In computer vision, many studies have taken an interest in developing deep learning methods for the analysis of human activity. Some have focused on improving Feature representation of skeleton joints. Others have focused on improving modeling of temporal dynamics.

A multi-level representation of Fisher Vectors and other skeleton-based geometric features is used for hand gesture recognition [15]. Sawant described a systematic method to recognize human activities in real time using Openpose and Long short-term memory networks [16]. A new approach showed by Sun *et al.* [17] that deep convolutional neural networks have achieved the state-of-the-art performance in pose tracking on the PoseTrack dataset. They used novel architecture, namely High Resolution Net (HRNet).

To solve the modeling of temporal dynamics, a lot of authors have used various techniques such as hierarchical RNN [18], Temporal Sliding LSTM [19]. To make skeleton-based action recognition model smaller, faster, Yang *et al.* [20] showed that using a lightweight network Double-feature Double-motion(DDNet) can achieve a super fast speed.

B. Key frame extraction

Early researchers have proposed various key frame extraction methods by several strategies.

Some of them used deep learning approach for this task. Xiang *et al.* [5] used a deep two-stream ConvNet for key frame detection in videos that learns to directly predict the location of key frames. In 2020, he *et al.* [4] continued propose an automatic self-supervised method to select key frames. In a video comprising a two-stream ConvNet and a novel automatic annotation architecture able to reliably annotate key frames in a video for self-supervised learning of the ConvNet. Another method for filtering key frames by aggregating the frames instead of looking at them one by one was introduced by Gowda *et al.* [21]. Hashim Yasin *et al.* [3] used keyframe-based approach for 3D Action Recognition Using a Deep Neural Network and performed extensive experiments on the benchmark MoCap datasets CMU [22] and HDM05 [23].

Besides the modern methods mentioned above, the traditional methods continue to be studied. Choutas *et al.* [24] introduced novel representation named PoTion (human joint heatmaps) that gracefully encodes the movement of some semantic keypoints. Laura *et al.* [25] said that "Most of the top performing action recognition methods use optical flow as a "black box" input". According to them, optical flow is useful for action recognition because it is invariant to appearance, optimized to minimize end-point-error and accuracy at boundaries and at small displacements is most correlated with action recognition performance.

Based on the analysis of the above studies, in this study we propose a new algorithm based on gradient of optical flow to find out the most important key frames in the video. These key frames then are taken as input for feature representation models of skeleton joints and action classification models. In the feature representation of skeleton joints step, we apply HRnet [17] network and OpenPose [26] to take advantage of 2D image skeleton detection. To achieve real-time motion classification we adpot DDNet [20]. Research results show high efficiency for classifying actions in video.

III. The proposed method

In this section, we describe our pipeline of learning model for action classification, KFSENet, including key frame extraction, human pose estimation model, and the deep model

for training action in detail. At test time, only the key frames of a test video are passed through pose estimation and deep network classification model.

To detail our pipeline, three main steps are following:

(1) Key frame Extraction: From video frame sequences, histogram of an optical flow function is used to extract the most important frames which characterize each action and distinguish other actions. This helps to not only show that the significant frames capture the important parts of the video but also that the testing is faster as compared to passing all frames though the deep models latter.

(2) Skeleton Features Extraction: We propose to use a pose estimation model to extract 2D skeletal joints (as known as human keypoints) from key frames extracted from Step 1. We chose the High-Resolution Network (HRNet) [17] 2D pose estimation model as the core model as this model achieved the best performance on the COCO 2019 Keypoint Detection Task dataset. We also apply OpenPose [26] to extract skeleton and compare with using HRNet. The feature vectors which represent human postures are built by the coordinates of the skeleton joints. The most significant postures for each activity are selected.

(3) Feature Learning from Skeleton Data and action classification: We apply DDNet [20] with low computational complexity and parameters. DDNet gives higher action recognition performance and shows its generalization on our experiential datasets.

A. *Key frame extraction algorithm*

Key frame extraction is an important method to summaries a long image sequence. The key frames summarize the contents of a video. To select the best frames, we utilize optical flow displacement fields between successive frames to identify local minima/maxima of motion in a video.

We calculate the key frames in two steps: The first step is computing the pixel intensity changes between the two frames using Dense Optical Flow introduced by Gunnar Farneback [27] and aggregate motion over two directions (horizontal and vertical) at each pixel as a motion metric $M(t)$ for frame t:

$\mathbf{M(t)} = \sum_i \sum_j |OF_x(i;j;t)| + OF_y(i;j;t)|$

where $OF_x(i;j;t)$ is the horizontal component of optical flow at pixel $(i;j)$ in frame t, and similarly for vertical component. It can be seen that optical flow tracks all points over time, thus the sum is an estimation of the amount of motion along consecutive frames.

In the second step, we analyze the metric as a function of time to extract key frames at the minima/maxima of motion. The gradient of this function characterizes the change of motion between consecutive frames and consequently the local minima/maxima would remark significant activities between sequences of poses. An example of this gradient change from a UCF Sport video [28] is shown in Figure 2.

The number of local extreme depends on the content of the video. Hence, complex activities or events would have more local extreme, whereas simpler ones may have less. Thus, to capture the most significant changes of action, we choose k frames corresponds to k the highest local extreme. Figure 3 illustrates selected key frames for a walking video from UCF Sport dataset.

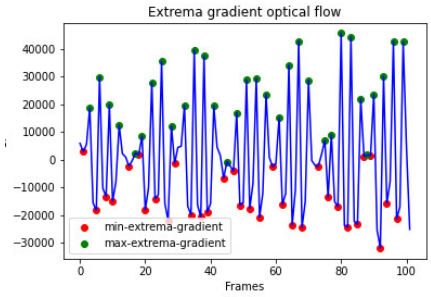

Fig. 2. Illustration of gradient function for a walking video from UCF Sport dataset. The red dots are local minima gradient while the green are local maxima gradient.

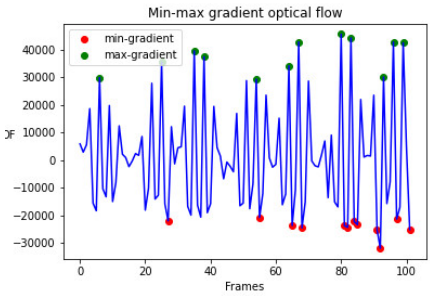

Fig. 3. Illustration of selected key frames for a walking video from UCF Sport dataset. The red/green dots are local minima/maxima selected.

In this way, the proposed algorithm sums up motion from successive frames of a video, thus it captures the motion information to better represent the action. One more advantage of our method is compact. The synthetic of optical flow is simple yet efficient, and a whole video is compressed into a few single frames. Thus, testing is faster and more accurate as compared to passing all frames though the deep neuron networks in latter steps.

B. *Skeleton extraction for feature representation of action*

In basic terms, a human pose estimation technique discover the layout of a human joints and body parts in an image. Fortunately, there are public resources nowadays that clarify the concept of human posture estimation in a basic and brief way.

Human pose estimation schemes identify a person in the image and estimate the coordination of his joints (or keypoints). Therefore, there are two possible approaches to estimate pose: Bottom-up and top-down pose estimations. The bottom-up approach to begin with finds the keypoints and after that maps them to different people in the image, whereas the top-down approach to begin with uses a mechanism to detect

people in an image, put a bounding box area around each person instance and after that estimate keypoint configurations within the bounding boxes. Top-down strategies depend on detection persons separately and ought to appraise keypoints for each individual, in this manner they are high computational cost since they are not genuinely end-to-end systems. By differentiate, bottom-up methods strategies begin localizing identity-free keypoints for all the people in an input image through anticipating heatmaps of diverse anatomical keypoints, taken after by gathering them into person individual, this viably makes them much faster.

From the above analysis, we propose to apply pose estimation models to extract 2D skeletal joints (i.e. keypoints) from image frame sequences. We adopt the High-Resolution Network (HRNet) [17] 2D pose estimation model and OpenPose [26] as the core model as these models achieved the best performance on the COCO 2019 Keypoint Detection Task dataset. In this way, the features of action are compressed into a vector of keypoints in the selected key frames which better represent for action rather than all of original frames.

Multi Person Pose Estimation is a more difficult because there are multiple people in an image. Over years, there are many works focused on solving this problem. HRNet is a state of the art neural network for human pose estimation. HRNet (High Resolution Net) [17] uses the top-down method, the network is built for estimating keypoints based on person bounding boxes which are detected by another network (FasterRCNN) during inference/testing. During training, HRNet employments the annotated bounding boxes of the given dataset. The innovation in the network is the high resolution representation of the input data is maintained. Then it is combined in parallel with high to low resolution subnetworks, while keeping efficient computational cost and lower parameters. However, the running time tends to increase with the number of people in the image and make the realtime execution a challenge.

Meanwhile, OpenPose belongs to the top-down scheme. It is a realtime multiple-person detection library, and it's the first time that any library has appeared the capability to detect human body joints, face, and foot keypoints (up to 135 keypoints) on single images. Thanks to Gines Hidalgo *et al.* for making this project successful. In OpenPose, the author gives an bottom-up approach where the body parts are recognized by the model and a last parsing is utilized to extract the posture estimation results. This approach can decouple the running time complexity from the number of individuals within the image.

In our experiments, we also reveal that HRNet obtains the higher performance than OpenPose on several datasets. However, OpenPose achieves higher speed than HRNet, especially when the number of people in video increases.

C. DDNet for action recognition

In spite of the fact that skeleton-based activity recognition has obtained great success in recent years, most of the existing strategies may suffer from a large model size and slow execution speed. To deal with this issue, we analyze skeleton sequence properties and utilize a Double-feature Double-motion Network (DDNet) [20] for skeleton-based action recognition. By using a lightweight network structure, DDNet can reach a super fast speed. By employing robust features, our pipeline achieves the state-of-the-art performance on UCF Sport and HNH datasets.

IV. Experiment results

In this section, we first introduce experiment setting in Section 4.1. We then detail the results in our experiments on datasets: UCF Sports [14] and self-built HNH dataset in Section 4.2.

A. Experiment settings

The following implementation details are set for our experiments: By several test cases where the recognition rates are calculated training/validation datasets with different parameters, we found that the optimal parameters, as follows:

• Key frame Extraction : we compute a dense optical flow using the Gunnar Farneback's algorithm [27]. We calculate the pixel intensity changes of two successive frames, resulting in an image with highlighted pixels, after changing over to HSV format for clear visibility. The algorithm accumulates magnitude and direction of optical flow from a set of the flow vectors. Then, we compute motion over two directions as a motion metric $M(t)$ for frame t. In order to select key frames, the proposed method calculates gradient of optical flow function $M(t)$ to find out local minima/maxima of motion in a video. Finally, 24 significant frames are selected which take from 10% to 20% total of frames in video. This leads to significantly reduce computational cost for next steps.

• Skeleton Features Extraction : The selected key frames are then extracted 2D skeletal joints using two pose estimation models HRNetW48 [17] and OpenPose [26] individually. HRNet with 48 channels and 384 x 288 resolution input images is used, and outputs 17 keypoints. OpenPose resizes input image to 368 x 368 resolution and extracts 18 keypoints estimation.

• DDNet [20] for action recognition: We adopt DDNet with learning rate = 1e-4, number of epochs = 600.

Experiments are implemented on Laptop core i7 8GB RAM, GPU: GTX 1660ti, CPU: AMD Ryzen 4800HS. GPU reached 100% utilization during inference.

B. Experiment results

1) Experiments on UCF Sport dataset: The UCF Sports is a moving camera dataset. This dataset contains ten classes from 150 sport videos with 720 x 480 image resolution. In comparison other datasets, the video sequences hold the higher resolution and more challenging. The number of clips per each action is not the same. Figure 4 illustrates example frames of UCF sport dataset.

On the dataset, we evaluate performance using the Leave-One-Out (LOO) cross-validation scheme test. This scenario

Fig. 4. Illustration of frames in videos from UCF Sport dataset.

takes out one sample video for testing and trains using all the remaining videos of an action class. This is performed for every sample video in a cyclic manner, and the overall accuracy is obtained by averaging the accuracy of all iterations.

Figure 5 shows the confusion matrix of the proposed algorithm on the UCF Sport dataset. It can be seen that the main error factor comes from between "Rise-Horsing" and "Swing-Side", "Golf-Swing". The similarity spacial-temporal changing of these action is the main reason of confusion. Experimental results show that many action classes achieve the absolute accuracy (100%). In addition, the different categories do not have any confusion. It means that our method demonstrate the efficiency and robustness.

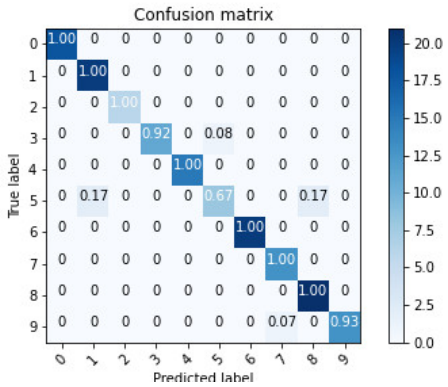

Fig. 5. Illustration of frames in videos from UCF Sport dataset.

The proposed method achieves the state of the art with accuracy of 93.1%. From experiment results in Table I, our method outperforms a lot of methods as [29]–[33]. In comparison to 3D-CNN [34] with the higher computational cost, KFSENet is lower computational complexity. The experiment results also demonstrate the significant contribution of key frame extraction algorithm in our system not only improve approximately 4% of accuracy but also reduce computational expense for pose estimation and action recognition in the latter steps.

2) Experiments on HNH dataset: In this fashion, we introduce a novel human action dataset, HNH moving camera dataset. The dataset contains 180 clips with several image resolutions. Each clip is labeled according to one of six action classes: Shaking hands, Greating, Walking, Sitting, Dancing, Standing. The dataset is divided into a training set of 140 clips and a test set of 40 clips different from the training set. The clips are recorded for both indoor and outdoor and performed by several actors. Shaking hands videos include two persons, remaining action videos contain one person. In comparison to other datasets, the video sequences hold the higher resolution and quite challenging. The number of clips per action is not the same. Figure 6 shows example frames of our HNH dataset.

Fig. 6. Illustration of video frames in our HNH dataset.

TABLE I
PERFORMANCE EVALUATION (%) OF DIFFERENT METHODS ON THE UCF SPORTS

Method	Accuracy(%)
Harris3D+HOG/HOF [29]	78.1
Dense+HOG/HOF [29]	81.6
Kovashka *et al* [30]	87.3
ConvNet (Le *et al*) [31]	86.5
ST-SIFT+HOG3D [32]	80.5
MBH+Dense trajectories [33]	84.2
3D-CNN+LSTM [34]	93.9
KFSENet without key frame	**89.3**
KFSENet with key frame	**93.1**

TABLE II
PERFORMANCE EVALUATION (%) OF DIFFERENT METHODS ON THE HNH DATASET

Method	Accuracy(%)
Openpose + DDNet	92.8
HRNet + DDNet	94.4
KFSENet	**98.9**

From the experiment results in Table II, the KFSENet achieves 98.9% on the dataset. It is worth noting that the accuracy rate is significant improved when integrating our key frames extraction algorithm. The experiment results demonstrate the HRNet gives the higher accuracy than OpenPose whereas Openpose obtains higher frame rate than HRNet.

V. Conclusions

In this paper, we introduce an efficient pipeline based on key frame extraction and deep learning models named KFSENet such that they are more efficient in both terms of higher performance and lower computational time. The proposed method to begin with extract the most important frames from a motion sequence using gradient of optical flow. These key frames then are extracted keypoints using pose estimation methods and employ them further in a Deep Neural Network (DNN) to classify actions. Therefore, we can eliminate unnecessary frames and decrease the computational expense. We demonstrate the efficiency of proposed method on publicly available benchmark and our HNH dataset. In future work, we will evaluate KFSENet on other challenging action datasets as well as other applications, such as gesture recognition and group action recognition.

References

[1] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A survey of vision-based human action evaluation methods," *Sensors*, vol. 19, no. 19, p. 4129, 2019.

[2] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *arXiv preprint arXiv:2002.05907*, 2020.

[3] H. Yasin, M. Hussain, and A. Weber, "Keys for action: An efficient keyframe-based approach for 3d action recognition using a deep neural network," *Sensors*, vol. 20, no. 8, p. 2226, 2020.

[4] X. Yan, S. Z. Gilani, M. Feng, L. Zhang, H. Qin, and A. Mian, "Self-supervised learning to detect key frames in videos," *Sensors*, vol. 20, no. 23, p. 6941, 2020.

[5] X. Yan, S. Z. Gilani, H. Qin, M. Feng, L. Zhang, and A. Mian, "Deep keyframe detection in human action videos," *arXiv preprint arXiv:1804.10021*, 2018.

[6] H.-H. Phan, N.-S. Vu, V.-L. Nguyen, and M. Quoy, "Action recognition based on motion of oriented magnitude patterns and feature selection," *IET Computer Vision*, vol. 12, no. 5, pp. 735–743, 2018.

[7] D. Gong, G. Medioni, and X. Zhao, "Structured time series analysis for human action segmentation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1414–1427, 2013.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[9] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[10] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[11] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat, "3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track," in *Proceedings of the Workshop on 3D Object Retrieval*, 2017, pp. 33–38.

[12] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[13] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2904–2913.

[14] UCF, "Ucf sports action data set," 2020, available online: https://www.crcv.ucf.edu/data.

[15] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[16] C. Sawant, "Human activity recognition with openpose and long short-term memory on real time images," EasyChair, Tech. Rep., 2020.

[17] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[18] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[19] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1012–1020.

[20] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM Multimedia Asia*, 2019, pp. 1–6.

[21] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, "Smart frame selection for action recognition," *arXiv preprint arXiv:2012.10671*, 2020.

[22] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," 2007.

[23] CMU, "Cmu motion capture database," 2019, available online: http://mocap.cs.cmu.edu.

[24] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7024–7033.

[25] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *German Conference on Pattern Recognition*. Springer, 2018, pp. 281–297.

[26] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[27] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[28] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR, 2008 IEEE Conference on*. IEEE, 2008, pp. 1–8.

[29] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, pp. 124–1.

[30] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for har," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on 2010*. IEEE, 2010, pp. 2046–2053.

[31] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *CVPR 2011 IEEE Conference on*. IEEE, 2011, pp. 3361–3368.

[32] M. Al Ghamdi, L. Zhang, and Y. Gotoh, "Spatio-temporal sift and its application to human action classification," in *European Conference on Computer Vision*. Springer, 2012, pp. 301–310.

[33] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[34] S. Arif, J. Wang, T. Ul Hassan, and Z. Fei, "3d-cnn-based fused feature maps with lstm applied to action recognition," *Future Internet*, vol. 11, no. 2, p. 42, 2019.