A Deep Human Action Representation For Retrieval Application

Mohsen Ramezani[1]*, Fardin Akhlaghian Tab [2] and Farzin Yaghmaee [3]

Abstract— **Human action retrieval as a challenging research area has wide-spreading applications in surveillance, search engines, and human-computer interactions. Current methods seek to represent actions and create a model with global and local features. These methods do not consider the semantics of actions to create the model, so they do not have proper final retrieval results. Each action is not considered a sequence of sub-actions, and their model is created using scattered local or global features. Furthermore, current action retrieval methods ignore incorporating Convolutional Neural Networks (CNN) in the representation procedure due to a lack of training data for training them. At the same time, CNNs can help them improve the final representation. In the present paper, we propose a CNN-based human action representation method for retrieval applications. In this method, the video is initially segmented into sub-actions to represent each action based on their sequence using keyframes extracted from the segments. Then, the sequence of keyframes is given to a pre-trained CNN to extract deep spatial features of the action. Next, a 1D average pooling is designed to combine the sequence of spatial features and represent the temporal changes by a lower-dimensional vector. Finally, the Dynamic Time Wrapping technique is used to find the best match between the representation vectors of two videos. Experiments on real video datasets for both retrieval and recognition applications indicate how created models for the actions can outperform other representation methods.**

Index Terms— Action, Deep features, Key-frame, Sub-action, CNN.

I. INTRODUCTION

Human action retrieval as a new research field is applicable in different real-life domains such as search engines, surveillance cameras, and patient movement monitoring [1-6]. In these applications, movements of body parts captured in videos must be tracked to represent and analyze motions [7]. In fact, for retrieving actions, they must be represented by a vector that compares them and indicates similar videos to each other as retrieved ones [8, 9]. Different features seek to model the moving parts of the human body to represent human actions in the video. Some methods seek to represent human actions using global features of the human action, and others use local features to represent human actions [7, 10-12].

Global features are used in different studies to represent the structure of motions. In fact, these features can be used for analyzing the appearance or shape of the human body when a human is playing the action [13-15]. There are different introduced global features such as low-level mesh features [2, 16, 17], extracted human-centered regions [18, 19], Epipolar geometry-based features extracted from different views [20], and binary silhouette [21, 22]. These features are used by different methods like Hidden Markov Model [17], sequence of prototypes [19], feature matching [20, 23], histogram of body poses [24], and spectral coding [25] to represent the captured human action. Yamato et al. [17] introduced an HMM-based method to learn human actions: feature vectors are low-level mesh features of binary frames after background subtraction. Efros et al. [18] extracted interest regions with the human in the center, and optical flow features in regions represent the motions of the main human action. In another study [20], to cover the motions of both human and camera when two moving cameras are capturing the human action, Epipolar geometry is initially applied on both views to improve capturing match score between similar actions via a fundamental temporal matrix.

Moreover, Lin et al. [19], Shao and Chen [24], Shao et al. [22], and Zhu et al. [21] used silhouette extraction to describe the shape of the human body using binary frames. Lin et al. [19] clustered the silhouettes to create a sequence of action prototypes. Finally, the similarity of two different videos is measured based on the number of overlapping prototypes. On the other hand, Shao et al. [22] extracted the silhouettes to represent body poses. Their statistical distribution and temporal relationship in the correlogram of body poses represent the actions and match them. Furthermore, Shao and Chen [24] also extracted the silhouettes as body poses. Their binary vectors are given to the Bag of Words method for creating a histogram of body poses as a representation of human action.

Global feature-based retrieval methods perform better on clean datasets than on real videos. These methods can successfully represent the structure of bodies and that their motions constitute a human action. However, background subtraction, tracking, and occlusion problems significantly decrease the accuracy of global-feature-based retrieval methods on real videos. Moreover, the steps of global-feature-based methods need a lot of time to be performed. Thus, faster methods are needed to represent captured human action in real

1- Department of Computer Science, University of Kurdistan, Sanandaj, Iran
2- Department of Computer Engineering University of Kurdistan Sanandaj, Iran
3- Department of Electrical and Computer Engineering Semnan University Semnan, Iran
Corresponding author: m.ramezani@uok.ac.ir

videos with more accuracy and less execution time.

Local features that are more robust to noise and variations are used for extracting independent patches from videos to describe inside motions [7, 26-28]. The patches are usually around specific points with important movements called Spatio-temporal Interest Points (STIPS). Thus, local feature-based methods have two main steps: STIP detection and patch description [7]. There are different used STIP extraction methods in retrieval studies such as Dollar detector [1, 3, 12, 29-31], SIFT detector [32-34], 3D Harris [13, 14, 35], and 3D-SIFT [36]. In order to describe the motion of STIPs, several descriptors are introduced in different studies that are gradient descriptor [29, 37], HOG3D [38-40], 3D-SIFT [36], 3D-Visual-Word [41], HOG/HOF [35, 42, 43], Fractal based Motion Pattern descriptor [1], and Laplacian Pyramid Coding [15]. As the most common combination in retrieval works, the Dollar method extracts local feature points; the gradient descriptor is used for describing the patch (e.g., cuboid) around each detected point. A bag of Visual Words represents the action from described patches [1, 6, 31, 39, 44, 45].

The dollar detector seeks to calculate a response value for each point of the video using a 2D Gaussian kernel filter, the first filter applied on the spatial axis and the 1D Gabor filter as the second one used for the temporal axis [29]. Moreover, Harris and SIFT feature point detectors are used in different retrieval studies [32, 33] to detect the local feature points in the video. These detectors are modified by Laptev [35] and Scovanner et al. [36], respectively (i.e., 3D Harris and 3D-SIFT) to detect better points by considering the temporal axis in the calculation.

HOG3D as a description method is applied on detected feature points to create histograms of oriented Spatio-Temporal gradient on three *x*, *y*, and *time* dimensions [46]. Laptev et al. [47] proposed HoG/HoF descriptor which is used in different studies [42, 43, 48, 49]. This descriptor considers a grid of cuboids that Histogram of oriented Gradient and Histogram of Optical Flow are calculated for each cuboid, and after normalizing them, they are concatenated into one vector as the final representation of the human action. Shao et al. [15] used a set of band-pass-filtered components for each video sequence as Spatio Temporal Laplacian Pyramid Coding descriptor to represent structural and motion information of each action. Spatio Temporal Pyramid Match (ST Pyramid) [50] and Vocabulary-Guided Pyramid Match (VG Pyramid) [51] are other methods that seek to represent actions by modifying the BoW method using multi-resolution histograms. Ramezani and Yaghmaee [1] proposed Fractal based Motion Pattern descriptor, which is applied to detected points by a Dollar detector to find the complexity of motions' patterns. This method finds a vector for each detected STIP, and a Bag of Words is used to represent the action. In addition, they proposed another method (namely 4-Directions) for representing each vector using the resultant vector of the motions constituting the action [26].

Note that other studies use detected local feature points directly, without describing them, such as the Spatio-temporal distribution of points [52]. It should be noted that local feature-based methods do not have acceptable performance on the clean datasets because they cannot consider the pose of human bodies, leading to overlaps between actions with similar motions. Moreover, some methods seek to use local and global features that, besides more needed time, have not significantly increased accuracy [11].

On the other hand, such local and global features are used in different methods to create action models for action recognition applications. For example, in a recent study, Afza et al. [53] seek to represent actions using a parallel HOG, geometric, and silhouette framework. Recently most action recognition methods benefit deep structures for representing and learning action models. In different studies, some random key frames are extracted, and deep spatial features of these frames are given to an Auto-Encoder or LSTM structure for creating the final model [54-55]. In another study, spatial features of random keyframes are gained by the VGG19 network. They are combined with gradient features to be used by the Naïve Bayes classifier for recognition tasks [56]. Dai et al. [57] also proposed an LSTM structure to extract the final Spatio-temporal model of actions based on optical flow and CNN-based features for achieving better recognition results. Furthermore, Tu et al. [58] used only some regions of the random keyframes based on the human body's appearance. Motion saliency is extracted from these regions to be used by a CNN for classifying the actions.

This paper introduces a novel deep video representation method based on the sequence of sub-actions to be used in retrieval applications. Similar to other methods, some keyframes are considered here to be used instead of all video frames in the representation procedure to save the method's execution time. But, unlike the state-of-the-art retrieval and recognition methods, used keyframes are not selected at random, and they would be selected in a manner to consider all sub-action (i.e., motions executed during the video). In fact, actions are a sequence of some sub-actions, and some sub-actions may be executed in a fast way, and if keyframes are selected randomly, no keyframe may be chosen from such sub-action. Here, input videos are divided into different sub-actions (i.e., episodes) using the R-value calculated using the Dollar detector as used by Ramezani et al.[44]. In other words, each detected sub-action contains a part of the main motions that form the action. Then, keyframes are selected from the achieved sub-actions to ensure that all sub-actions, regardless of their length, are considered in the modeling procedure for achieving a complete semantic model.

After selecting keyframes, frame-level deep spatial features are extracted by a deep structure, i.e., Convolutional Neural Networks (i.e., CNN), as proper spatial features, including both local and global viewpoints to the keyframes [54,59-60]. Here, a proper pre-trained CNN model, VGG-16, is used to extract deep features from the video to solve the challenge of lacking training data in retrieval applications. The output of the VGG-16 for each key-frames is a vector that can be considered as the spatial model of the frame. Then, a high-dimensional vector is created by concatenating the outputs of the VGG-16 network for keywords. The changes along the created vector indicate the action model during the time axis. The created vector is given to a one-dimensional Average-Pooling tool for modeling the temporal changes of the action. Thus, one vector is finally created that contains the spatial and temporal model of human action. Comparing action vectors is the basis of finding similar videos to the query video in an action retrieval system. As

videos of one action category may be executed differently and it is not clear what sub-action is captured at the beginning of the video, Dynamic Time Wrapping (DTW) is applied to the final representations of two videos to find their best correlation and calculate their similarity. Clearly, most similar videos to query one is found and retrieved.

The proposed method is compared to other state-of-the-art action retrieval methods. It should be noted that improving the retrieval methods can be considered hard work because comparing the created models directly without benefiting a learning algorithm to be trained based on the different models of different action categories. In our experiments, the proposed model for the action retrieval application is also given to a learning algorithm for classifying the videos into action categories of used datasets. Experiments show that the proposed action model can be a proper model for the action recognition task. By modifying the model creation step for the action recognition task, much better results would be achieved than the current results for the recognition task. Thus, we seek to use two types of experiments to indicate the superiority of our model rather than others. The contributions of this paper can be summarized as follows:

1- Segmenting video into sub-actions to be considered for extracting keyframes from all motions included in the action regardless of sub-action length. This step helps the method create a semantical action model by considering the sequence of sub-action keyframes.

2- Utilizing spatial features of actions that are extracted by a deep network to create proper final models for an action retrieval system. The proper models are achieved by considering local and global changes in the network using convolutions performed in the network.

3- Utilizing a pooling tool for creating the final Spatio-temporal model of the action based on the outputs of the deep network for keyframes.

4- Incorporating a dynamic model matching approach using DWT for finding the best accommodation between two action models.

5- Comparing the proposed action retrieval method with the state-of-the-art methods indicates the superiority of created model for action using real videos in different datasets. Moreover, the created action model is given to a learning algorithm to learn models and classify actions. The classification results are compared to recently introduced human action recognition studies to indicate the model's superiority.

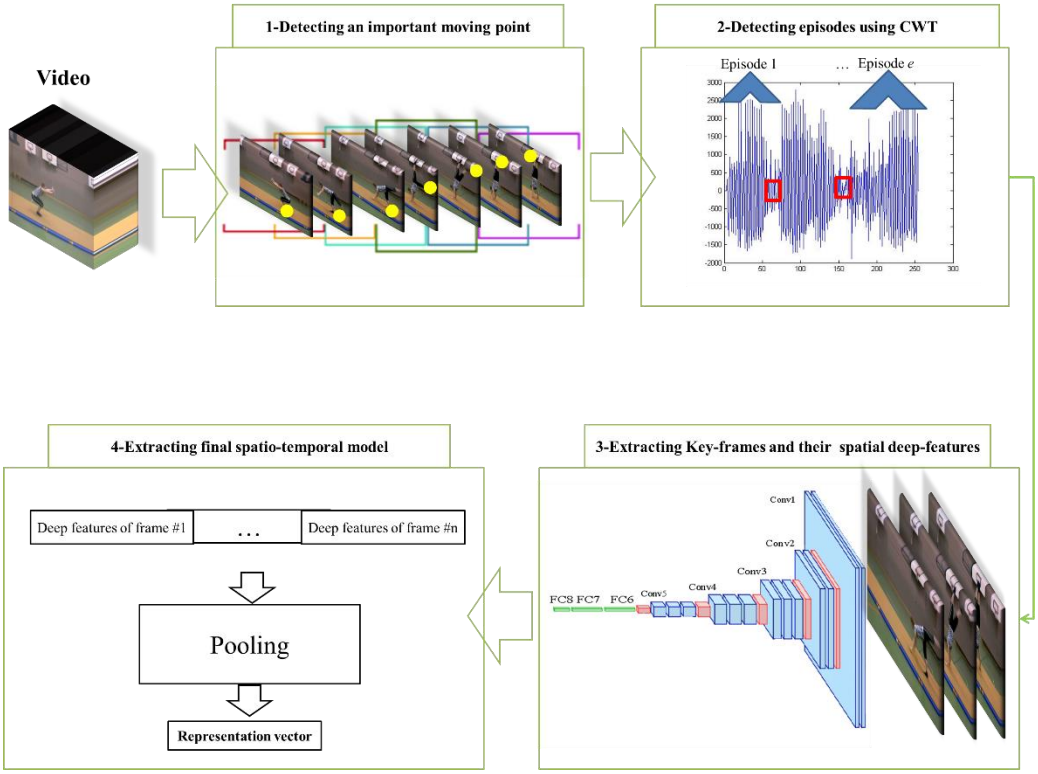

Fig. 1. The proposed feature extraction framework

Conv 1a Conv1b	Max Pooling	Conv 2a Conv2b	Max Pooling	Conv 3a Conv3b	Conv 3c	Max Pooling	Conv4a Conv4b	Conv 4c	Max Pooling	Conv5a Conv5b	Conv 5c	Max Pooling	FC	FC	FC8
3*3				3*3	3*3		3*3	1*1		3*3	1*1		Inner Product	Inner Product	Inner Product
1,1				1,1	1,1		1,1	1,1		1,1	1,1				
64				256	256		512	512		512	512		4096	4096	1000

Fig. 2. The architecture of the pre-trained VGG-16 CNN model

The rest of this paper is organized as follows. In section two, the proposed method is presented in detail. The proposed method is evaluated in section three by describing the experimental results of the used datasets. Finally, the method is concluded in section four.

II. PROPOSED METHOD

In this paper, deep CNN features are extracted as spatial models from the frames of each action to be used for modeling the action independently. These features are then modeled temporally using a pooling tool, and the final model is used in the retrieval application. Note that this method is applied to specially extracted keyframes from time episodes of the video. Fig.1 shows the framework of the proposed deep feature extraction. Here, deep feature extraction would be applied only on special frames (i.e., key-frames) to extract deep spatial features. Unlike most other methods, which process randomly selected keyframes, this method selects keyframes that represent features of all motions in the human action and are distributed among motions properly. To this end, different motions of the human action must be captured in sub-actions (i.e., time episodes). Thus, keyframes are selected from all episodes to prevent lacking important spatial features that may be ignored during the random keyframe selection.

Here, the introduced method by Ramezani and Yaghmaee [44] detect time episodes of the action that contain different motions constituting the main action. As the first step of our framework, the response function is used for calculating R-value for each point in the video. Then, the most important point with a high R-value is selected and tracked during the action playing to indicate where the motion model of this point changes. The response function which is used here is as follows:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad (1)$$

where g indicates the kernel of Gaussian, moreover, h_{ev} and h_{od} are filters of Gabor which are as follows:

$$\mathrm{h}_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\pi^2} \quad (2)$$

$$\mathrm{h}_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\pi^2} \quad (3)$$

As shown in Fig.1 (Step 2), the energy diagram of the point is calculated via mapping the motion diagram of the point using Continuous Wavelet Transform (CWT). Then different time episodes are detected using Dynamic Time Wrapping for keyframe selection. Hereafter, the keyframes are selected in a distributed form between different time episodes. This selection considers frames of all episodes (motions of the action) to extract deep spatial features from all parts of important motions. Thus, it can be considered a good representation with enough details of the human action with fewer needed frames. After detecting the keyframes of each action, deep spatial features must be extracted using a deep pre-trained model. The used deep CNN model parameters have been trained on various datasets. Here, the output of the fully connected FC8 layer of pre-trained VGG-16 CNN is extracted as a deep spatial feature of each keyframe. Fig. 2 indicates the architecture of the used pre-trained VGG-16 CNN. In this Figure, the first row shows the layer name, the second row shows the Kernel size, the third row shows the Stride and padding size, and the fourth indicates the channels. For example, this model contains 3×3 kernels when 1 stride convolutional layer exists to have fewer parameters in layers.

Thus for each keyframe, a 1×1000 vector is created by the VGG-16 CNN as a spatial feature at the corresponding time. Clearly, the extracted features related to spatial dimension and the temporal information of actions are lacking. To this end, a temporal model of the input action must be created by considering the changes in its spatial feature vectors over time. The created vectors of the input action are then concatenated based on the keyframes' order in the video. Let *n* indicate the number of keyframes. The concatenated vector would have a dimension of $1\times(n\times1000)$. The concatenated vector is given to a 1-dimension average-pooling to create the final Spatio-temporal model of the input action. The final representation vector of each video will have a dimension of $1\times(\frac{n\times1000}{d})$, where *d* is the compression rate of pooling.

After creating the model of each action independently, to retrieve similar videos to a query one, the final vector of all videos must be compared with the final vector of the query. It is unclear that the video starts with which sub-action and videos of one action category may have different orders of sub-action at the beginning of the video (see Fig. 3). Thus, as it is important to find the best match between different representation vectors, matching vectors is performed using Dynamic Time Wrapping. The similarity of the vectors is also calculated by this method. Then, most similar videos to query one are selected as retrieved.

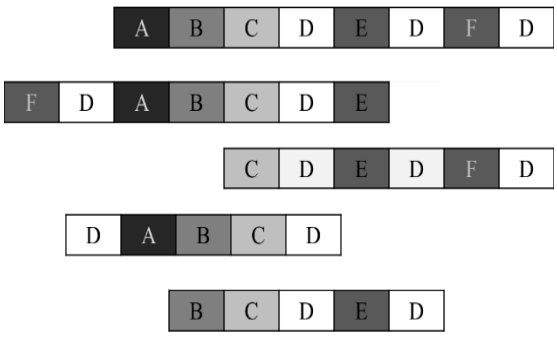

Fig. 2. The effect of different timelapse on five videos' final representatives (motions in the video)

III. EXPERIMENTAL RESULTS

This paper uses a vector of spatial and temporal deep features to describe a video's captured action. The created vector includes the information of all motions to represent each action in the best form. In this section, we first describe the used datasets and the implementation details. Then, the created vector for each video (i.e., action) is evaluated in retrieval and recognition tasks. It should be noted that the video modeling method is introduced to be used in retrieval tasks, but it can achieve acceptable action recognition results

by incorporating a simple classifier. Thus, we would evaluate the modeling manner in both retrieval and recognition tasks.

A. Datasets

The proposed method is applied to three real datasets, i.e., UCF Sport, UCF YouTube (UCFYT), and HMDB datasets. UCF Sports dataset consists of 150 real videos from 9 action categories captured from sports scenes. This dataset has different actors, backgrounds, viewpoints, and scenes. UCF YouTube is another used dataset that varies in actor, background, viewpoint, and scene, similar to the UCF Sports dataset. This dataset contains 1600 videos in 11 action categories captured from realistic actions on YouTube videos. The action categories are basketball shooting, biking/cycling, diving, golf, swinging, horse riding, soccer juggling, swinging, tennis, diving, trampoline jumping, volleyball-spiking, and walking with a dog. As a large-scale dataset with 6849 real clips in 51 action categories, HMDB is also used for evaluating the proposed method. In this paper, 2241 videos that relate to human actions from this dataset are processed that are gathered into 19 action categories as handstand, golf, jumping, flicflac, pull up, kick-ball, clap hands, climb stairs, dive, fall on the floor, push up, run, sit down, sit up, somersault, stand up, turn, walk and wave.

B. Implementation Detail

The experiments are run on a computer system with Intel Core i7 and 16GB DDR3 memory working under Microsoft Windows 10 operating system. We implemented the method using the python language and TensorFlow, widely used for machine learning applications. Here, the VGG-16 pre-trained model is considered for extracting the spatial features of frames. The weights= "imagenet" is called to fetch VGG-16 with the weights relating to the ImageNet dataset. Here, similar to other studies [46,32-33], σ and τ are considered to be 2.4 and 1.7, respectively, to detect the most important STIP in each video during the episode finding step. In fact, these parameters are the size of the Gaussian and the Gabor filters, respectively. After finding episodes, 10 key-frames are extracted from each video in a distributed form on episodes that a 1×1000 vector represents the deep spatial feature of each selected keyframe. Then, these vectors are concatenated to create a 1×10000 vector. Due to having a value of 5 for parameter d, a 1×2000 vector is created by the pooling as the representation of the human action captured in the video. It should be noted that other studies such as Reference [12] and Reference [30, 44] use 1×11271 and 1×2673 vectors to represent actions.

For the retrieval task, all videos are considered as queries independently and removed from the dataset in a one-by-one fashion to retrieve the top 20 similar videos to the query one. The number of retrieved videos from the corresponding category to the query video indicates accuracy. Each dataset is divided into Training and Test parts for the action recognition task with an 80:20 ratio; the first part is used for training a classifier, and the remaining videos are considered in the test step. Action recognition experiments are iterated five times in which each video would fall into test videos for one time.

C. Comparisons of the Retrieval Task

Each video in the dataset would be considered a query in an independent procedure performed to test methods. Each considered query video is eliminated from the dataset, and the top 20 similar videos among the remaining ones to this query are retrieved. Note that the rate of the number of retrieved videos to the number of all videos is 20 to 100. The proposed method is compared to Fractal based [1], BoW [12], ST Pyramid [50], VG Pyramid [51], and 4-direction [30] methods as the most accurate introduced action retrieval methods.

Table I compares the state-of-the-art retrieval methods based on the average accuracy of these methods on used datasets. The proposed method has about 2.8 percent better accuracy on average than the second-best method, the Fractal based pattern representation method.

TABLE I.
Comparing the Total Accuracy of the Proposed and Other Retrieval Methods

Method	Average accuracy on UCFYT and UCF-Sport	Average accuracy on all used datasets
BOW	37.45	29.73
VG Pyramid	40.6	-
ST Pyramid	44.1	-
4-Directions	44.75	34.6
Fractal based Pattern	49	39.33
Proposed Method	**52.55**	**41.7**

Fig.4 shows the accuracy of retrieving videos for different categories of the UCF-Sport dataset on the main diameter of the matrix. The total accuracy of the proposed method on this dataset is about 0.59. Better results are achieved for action categories with exclusive motions. For example, running as an action with common motions with other actions has the least accuracy. Moreover, Fig. 5 indicates that the proposed method performs significantly better than the other state-of-the-art methods. Thus, extracted deep features can be considered suitable for representing sports videos as a clearer dataset than the others.

Similar to Fig. 4, Fig. 6 represents the accuracy of the proposed method for different action categories of the UCF-YouTube dataset on the main diameter of the matrix. The proposed method leads to considerable overlaps between the action categories with similar motions, indicating the proper achieved model. For example, the accuracy of the proposed method for biking and riding-horse categories as categories with similar motions is 0.275 and 0.308, respectively. In contrast, the overlap of these two categories during retrieving their videos is about 0.16. The considerable overlap between action categories with similar motions is due to the successful models created based on the sequence of motions. Such performance helps discriminate different actions that have unique motions. For example, tennis with unique motions rather than other action categories has the best result with 0.8 accuracy. It should be noted that golf can

be considered the most similar action category to tennis, which has 0.11 overlap, and the overlap of tennis videos with other categories is near zero.

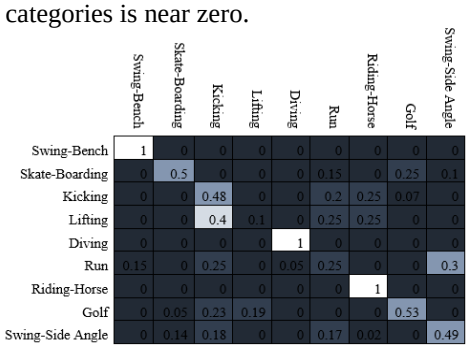

Fig. 3. Confusion matrix of overlaps between categories of UCF-sport dataset using the proposed method.

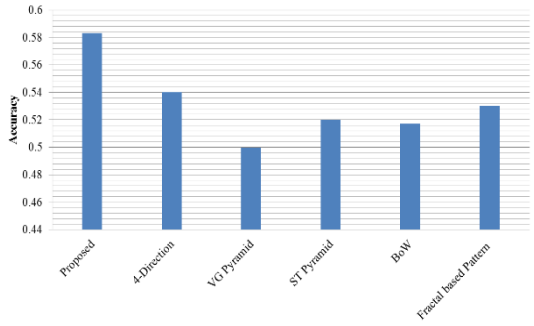

Fig. 4. Comparing the proposed method with other state-of-the-art representation methods on UCF-Sport.

Fig. 7 also indicates that the proposed method performs better than the other representation methods. The proposed method's average accuracy on UCFYT dataset is 0.461 while, this value for Fractal based [1], BoW [12], ST Pyramid [50], VG Pyramid [51], and 4-direction [30] methods are 0.45, 0.23, 0.36, 0.31, and 0.35 respectively. The second-best method has 0.45 accuracy, which is about 1.1 percent worse than the proposed method, which can be considered a considerable improvement for retrieval application that doesn't use a learning algorithm. The model can be learned better by a learning algorithm compared to other recent action recognition methods.

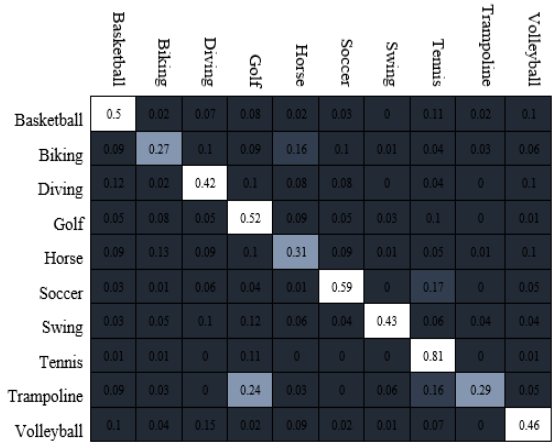

Fig. 5. Confusion matrix of overlaps between categories of UCFYT dataset using the proposed method

For the HMDB as a large-scale dataset, Fig. 8 shows the performance of the proposed method on different action categories. Similar to the previously considered datasets, those categories with unique motions like clap-hand and sit-up have better retrieval accuracy, i.e., 0.36 and 0.38, respectively. On the other hand, those categories, like Flic-Flac and Hand-Stand, which have similar motions to other action categories, have less retrieval accuracy, that is 0.18 and 0.195, respectively. The created model by the proposed method leads to overlap between action categories that have similar motions, such as golf and handstand, because of their similar hand motions, which their overlap is about 24 percent.

Fig. 9 compares the proposed method with others to indicate the superiority of our method rather than most of the other methods on real datasets such as HMDB. The proposed method has similar accuracy to the Fractal based Pattern representation method while it significantly performs better than others.

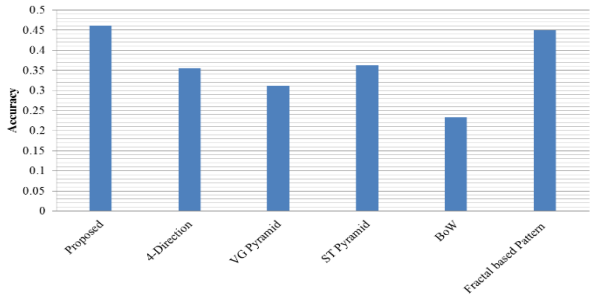

Fig. 6. Comparing the proposed method with other state-of-the-art representation methods on the UCFYT dataset

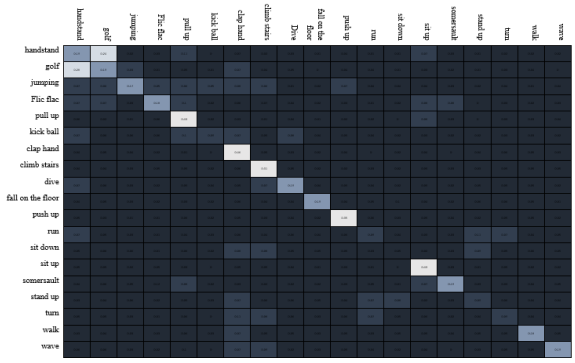

Fig. 7. Confusion matrix of overlaps between categories of HMDB dataset using the proposed method

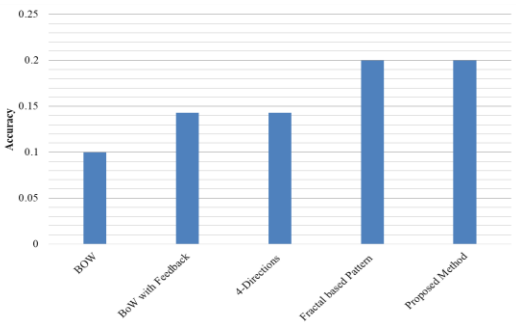

Fig. 8. Comparing the proposed method with other state-of-the-art representation methods on the HMDB dataset

D. Comparisons on Recognition Task

Besides the improvements in the retrieval task, the proper performance of the method in modeling the motions that appeared in true overlaps gained between the action categories with similar motions can help to learn algorithms to discriminate action categories in action recognition tasks. To this end, we use a quadratic SVM classifier (as used in Reference [54]) to learn the created models by the proposed method and classify actions. Here, the output of the pooling step is considered as the input vector of the SVM classifier. The 5-fold cross-validation is considered to run the method several times for achieving more reliable results than one time run. The ratio of the training volume to testing data is 80:20. The results of the proposed method are compared to the recent action recognition methods, as indicated in Table 2. The proposed model+SVM versus the recently published deep methods by Muhammad et al. [55], Afza et al. [53], Khan et al. [56], Dai et al. [57] and Tu et al. [58] have the best performance on average. The second-best method for all datasets is the method introduced by Muhammad et al. [55], with 92.53 percent accuracy on average, while this value for the proposed method is about 93.4.

Our structure for action representation in recognition tasks can be modified using Auto-encoder instead of the current pooling tool to help achieve much better results than our current results. But, the current results show our proper representation method, which is presented for the action retrieval application in representing action. The main reason for such performance relates to discriminating action categories that have dissimilar motions by the model (as shown in retrieval experiments) and discriminating different action categories with similar motions by the classifier, which can easily learn the patterns embedded in the final representation vector.

TABLE II.
Comparing the total accuracy of the proposed model with the SVM classifier with other action recognition methods

Method	UCFYT	UCF-Sport	HMDB
Muhammad et al. [55]	98.30	99.1	80.2
Afza et al. [53]	94.5	99.3	76.9
Khan et al. [56]	**99.4**	98	-
Dai et al. [57]	-	97.53	76.3
Tu et al. [58]	-	-	71.17
Proposed Method	**99.4**	**99.5**	**81.35**

The proposed method performs better than the state-of-the-art methods on different datasets, and its performance on the HMDB as a large-scale dataset is acceptable. It should be noted that the proposed method has lower-dimensional vectors rather than other methods to decrease the vector comparing time. Moreover, this method captures global structure using deep spatial feature extraction and considers local changes by tracking structures over time. In comparison, the Fractal based Pattern method uses the local changes of some STIPs in each video to represent the human action that needs more execution time. Hence, the proposed method is a little faster than the Fractal based Pattern, whose average time for representation is 42 and 49 seconds, respectively.

IV. CONCLUSION

The present paper proposes a novel deep feature-based representation framework to model the human actions used for retrieval applications. Videos in different datasets are independently given to this framework to model the captured human action using deep spatial features. To efficiently model the changes over the time axis, some keyframes are extracted from all sub-actions of the human action, which contain different action motions. Here, a pre-trained VGG-16 CNN is used for creating deep spatial features of actions. Then, these deep spatial features are concatenated and modeled by Average pooling as the final Spatio-Temporal representation of the captured human action in the video. Finally, Dynamic Time Wrapping is used for matching vectors and calculating the similarity of representation vectors. Experiments on benchmark datasets, i.e., UCFYT, HMDB51, and UCF-Sport datasets, indicated the efficiency and accuracy of our method rather than the other representation methods used for retrieval application.

Moreover, the model created by our method is given to a classifier to be compared to other models introduced for recognition tasks. The model created by the proposed method can be learned properly by quadratic SVM for achieving comparable results to the recently introduced deep action recognition methods. This model successfully finds sub-actions of different actions and represents actions by the extracted keyframes from them. To this end, as the future work of this method, sub-action would be modeled independently, and their sequence would be learned as a proper model of the action to be used in action recognition tasks. We would use deep structures for representing sub-actions and then define each action as the sequence of modeled sub-actions for achieving good discrimination of different human actions.

V. REFERENCES

[1] Ramezani M, Yaghmaee F. Motion pattern-based representation for improving human action retrieval. Multimedia Tools and Applications. 2018 Oct 1;77(19), pp:26009-32.

[2] Veinidis C, Pratikakis I, Theoharis T. Unsupervised human action retrieval using salient points in 3D mesh sequences. Multimedia Tools and Applications. 2019 Feb 1;78(3), pp:2789-814.

[3] Qin J, Liu L, Yu M, Wang Y, Shao L. Fast action retrieval from videos via feature disaggregation. Computer Vision and Image Understanding. 2017 Mar 1;156, pp:104-16.

[4] Ding S, Li G, Li Y, Li X, Zhai Q, Champion AC, Zhu J, Xuan D, Zheng YF. Survsurf: human retrieval on large surveillance video data. Multimedia Tools and Applications. 2017 Mar 1;76(5), pp:6521-49.

[5] Zhang L, Wang Z, Yao T, Mei T, Feng DD. Exploiting spatial-temporal context for trajectory-based action video retrieval. Multimedia Tools and Applications. 2018 Jan 1;77(2), pp:2057-81.

[6] Zong M, Wang R, Chen X, Chen Z, Gong Y. Motion saliency-based multi-stream multiplier ResNets for action recognition. Image and Vision Computing. 2021 Mar 1;107:104108.

[7] Ramezani M, Yaghmaee F. A review on human action analysis in videos for retrieval applications. Artificial Intelligence Review. 2016 Dec 1;46(4), pp:485-514.

[8] Zhao S, Chen L, Yao H, Zhang Y, Sun X. Strategy for dynamic 3D depth data matching towards robust action retrieval. Neurocomputing. 2015 Mar 5;151, pp:533-43.

[9] Naeem HB, Murtaza F, Yousaf MH, Velastin SA. T-VLAD: Temporal vector of locally aggregated descriptor for multiview human action recognition. Pattern Recognition Letters. 2021 Aug 1;148:22-8.

[10] Jiang X, Zhong F, Peng Q, Qin X. Action recognition based on global optimal similarity measuring. multimedia Tools and Applications. 2016 Sep 1;75(18), pp:11019-36.

[11] Liu X, Li Y. Research on human action recognition based on global and local mixed features. In2014 International Conference on Mechatronics, Control and Electronic Engineering (MCE-14) 2014 Mar. Atlantis Press.

[12] Jones S, Shao L, Du K. Active learning for human action retrieval using query pool selection. Neurocomputing. 2014 Jan 26;124, pp:89-96.

[13] Junejo IN, Dexter E, Laptev I, Perez P. View-independent action recognition from temporal self-similarities. IEEE transactions on pattern analysis and machine intelligence. 2010 Mar 18;33(1), pp:172-85.

[14] Junejo IN, Dexter E, Laptev I, PÚrez P. Cross-view action recognition from temporal self-similarities. European Conference on Computer Vision 2008 Oct 12 (pp. 293-306). Springer, Berlin, Heidelberg.

[15] Shao L, Zhen X, Tao D, Li X. Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics. 2013 Jul 31;44(6), pp:817-27.

[16] Veinidis C, Pratikakis I, Theoharis T. Querying 3D mesh sequences for human action retrieval. In2014 2nd International Conference on 3D Vision 2014 Dec 8 (Vol. 2, pp. 33-40). IEEE.

[17] Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using hidden Markov model. InCVPR 1992 Jun 15 (Vol. 92, pp. 379-385).

[18] Efros AA, Berg AC, Mori G, Malik J. Recognizing action at a distance. Innull 2003 Oct 13 (p. 726). IEEE.

[19] Lin Z, Jiang Z, Davis LS. Recognizing actions by shape-motion prototype trees. In2009, IEEE 12th international conference on computer vision 2009 Sep 27 (pp. 444-451). IEEE.

[20] Yilmaz A, Shah M. Matching actions in the presence of camera motion. Computer vision and image understanding. 2006 Nov 1;104(2-3), pp:221-31.

[21] Zhu F, Shao L, Lin M. Multi-view action recognition using local similarity random forests and sensor fusion. Pattern recognition letters. 2013 Jan 1;34(1), pp:20-4.

[22] Shao L, Wu D, Chen X. Action recognition using correlogram of body poses and spectral regression. In2011 18th IEEE International Conference on Image Processing 2011 Sep 11 (pp. 209-212). IEEE.

[23] Choi J, Jeon WJ, Lee SC. Spatio-temporal pyramid matching for sports videos. InProceedings of the 1st ACM international conference on Multimedia information retrieval 2008 Oct 30 (pp. 291-297).

[24] Shao L, Chen X. Histogram of Body Poses and Spectral Regression Discriminant Analysis for Human Action Categorization. InBMVC 2010 (pp. 1-11).

[25] Shao L, Liu L, Yu M. Kernelized multiview projection for robust action recognition. International Journal of Computer Vision. 2016 Jun 1;118(2), pp:115-29.

[26] Ramezani M, Yaghmaee F. Retrieving human action by fusing the motion information of interest points. International Journal on Artificial Intelligence Tools. 2018 May 21;27(03):1850008.

[27] Sharif M, Khan MA, Zahid F, Shah JH, Akram T. Human action recognition: a framework of statistical weighted segmentation and rank correlation-based selection. Pattern Analysis and Applications. 2020 Feb;23(1), pp:281-94.

[28] Sahoo SP, Ari S. On an algorithm for human action recognition. Expert Systems with Applications. 2019 Jan 1;115, pp:524-34.

[29] Dollár P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse Spatio-temporal features. In2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance 2005 Oct 15 (pp. 65-72). IEEE.

[30] Ramezani M, Yaghmaee F. A novel video recommendation system based on efficient retrieval of human actions. Physica A: Statistical Mechanics and its Applications. 2016 Sep 1;457, pp:607-23.

[31] Chen S, Sun Z, Zhang Y, Li Q. Relevance feedback for human motion retrieval using a boosting approach. Multimedia Tools and Applications. 2016 Jan 1;75(2), pp:787-817.

[32] Shao L, Jones S, Li X. Efficient search and localization of human actions in video databases. IEEE Transactions on Circuits and Systems for Video Technology. 2013 Aug 6;24(3), pp:504-12.

[33] Jones S, Shao L. Action retrieval with relevance feedback on YouTube videos. InProceedings of the Third International Conference on Internet Multimedia Computing and Service 2011 Aug 5 (pp. 42-45).

[34] Jiang YG, Li Z, Chang SF. Modeling scene and object contexts for human action retrieval with a few examples. IEEE Transactions on Circuits and Systems for Video Technology. 2011 Mar 17;21(5), pp:674-81.

[35] Laptev I. On space-time interest points. International journal of computer vision. 2005 Sep 1;64(2-3), pp:107-23.

[36] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. InProceedings of the 15th ACM international conference on Multimedia 2007 Sep 29 (pp. 357-360).

[37] Jones S, Shao L. Unsupervised spectral dual assignment clustering of human actions in context. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014 (pp. 604-611).

[38] Klaser A, Marszałek M, Schmid C. A Spatio-temporal descriptor based on 3d-gradients, 2008.

[39] Jones S, Shao L. Content-based retrieval of human actions from realistic video databases. Information Sciences. 2013 Jul 1;236:56-65.

[40] Zhen X, Shao L, Tao D, Li X. Embedding motion and structure features for action recognition. IEEE Transactions on Circuits and Systems for Video Technology. 2013 Jan 16;23(7), pp:1182-90.

[41] Ji R, Yao H, Sun X. Actor-independent action search using spatiotemporal vocabulary with appearance hashing. Pattern Recognition. 2011 Mar 1;44(3), pp:624-38.

[42] Yu G, Yuan J, Liu Z. Unsupervised trees for human action search. InHuman Action Analysis with Randomized Trees 2015 (pp. 29-56). Springer, Singapore.

[43] Páez F, Vanegas JA, González FA. Online multimodal matrix factorization for human action video indexing. In2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI) 2014 Jun 18 (pp. 1-6). IEEE.

[44] Ramezani M, Yaghmaee F. Eliminating the Repetitive Motions as a Preprocessing step for Fast Human Action Retrieval. In2019 9th International Conference on Computer and Knowledge Engineering (ICCKE) 2019 Oct 24 (pp. 26-31). IEEE.

[45] Barnachon M, Bouakaz S, Boufama B, Guillou E. A real-time system for motion retrieval and interpretation. Pattern Recognition Letters. 2013 Nov 1;34(15), pp:1789-98.

[46] Tang J, Shao L, Zhen X. Human action retrieval via efficient feature matching. In2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance 2013 Aug 27 (pp. 306-311). IEEE.

[47] Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In2008 IEEE Conference on Computer Vision and Pattern Recognition 2008 Jun 23 (pp. 1-8). IEEE.

[48] Paez F, Vanegas JA, Gonzalez FA. An evaluation of NMF algorithm on human action video retrieval. InSymposium of Signals, Images and Artificial Vision-2013: STSIVA-2013 2013 Sep 11 (pp. 1-4). IEEE.

[49] Bulbul MF, Jiang Y, Ma J. Human action recognition based on DMMs, HOGs and Contourlet transform. In2015 IEEE International Conference on Multimedia Big Data 2015 Apr 20 (pp. 389-394). IEEE.

[50] Choi J, Jeon WJ, Lee SC. Spatio-temporal pyramid matching for sports videos. InProceedings of the 1st ACM international conference on Multimedia information retrieval 2008 Oct 30 (pp. 291-297).

[51] Grauman K, Darrell T. Approximate correspondences in high dimensions. Advances in Neural Information Processing Systems 2007 (pp. 505-512).

[52] Bregonzio M, Gong S, Xiang T. Recognising action as clouds of space-time interest points. In2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20 (pp. 1948-1955). IEEE.

[53] Afza F, Khan MA, Sharif M, Kadry S, Manogaran G, Saba T, Ashraf I, Damaševičius R. A framework of human action recognition using

length control features fusion and weighted entropy-variances based feature selection. Image and Vision Computing. 2021 Feb 1;106:104090.

[54] Ullah A, Muhammad K, Haq IU, Baik SW. Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Future Generation Computer Systems. 2019 Jul 1;96, pp:386-97.

[55] Muhammad K, Ullah A, Imran AS, Sajjad M, Kiran MS, Sannino G, de Albuquerque VH. Human action recognition using attention-based LSTM network with dilated CNN features. Future Generation Computer Systems. 2021 Jun 24.

[56] Khan MA, Javed K, Khan SA, Saba T, Habib U, Khan JA, Abbasi AA. Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimedia tools and applications. 2020 Mar 14:1-27.

[57] Dai C, Liu X, Lai J. Human action recognition using two-stream attention-based LSTM networks. Applied soft computing. 2020 Jan 1;86:105820.

[58] Tu Z, Xie W, Qin Q, Poppe R, Veltkamp RC, Li B, Yuan J. Multi-stream CNN: Learning representations based on human-related regions for action recognition. Pattern Recognition. 2018 Jul 1;79:32-43.

[59] Berlin SJ, John M. Particle swarm optimization with deep learning for human action recognition. Multimedia Tools and Applications. 2020 Feb 16, pp:1-23.

[60] Wang J, Shao Z, Huang X, Lu T, Zhang R, Lv X. Spatial-temporal pooling for action recognition in videos. Neurocomputing. 2021 Sep 3;451:265-78.