Enhancing Human Action Recognition through Temporal Saliency

Vida Adeli
Department of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
vida.adeli@mail.um.ac.ir

Ehsan Fazl-Ersi
Department of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
fazlersi@um.ac.ir

Ahad Harati
Department of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
a.harati@um.ac.ir

***Abstract*—Images and videos have become ubiquitous in every aspects of life due to the growing digital recording devices. It has encouraged the development of algorithms that can analyze video content and perform human action recognition. This paper investigates the challenging problem of action recognition by outlining a new approach to represent a video sequence. A novel framework is developed to produce informative features for action labeling in a weakly-supervised learning (WSL) approach both during training and testing. Using appearance and motion information, the goal is to identify frame regions that are likely to contain actions. A three-stream convolutional neural network is adopted and improved by proposing a method based on extracting actionness regions. This results in less computation as it is processing only some parts of an RGB frame and also interpret less non-activity related regions, which can mislead the recognition system. We exploit UCF sports dataset as our evaluation benchmark, which is a dataset of realistic sports videos. We will show that our proposed approach could outperform other existing state-of-the art methods.**

Keywords—Action recognition, Motion, Region proposal, Convolutional Neural Networks, Actionness.

I. INTRODUCTION

Human action recognition is among the most fundamental topics in computer vision research community [1, 2]. The capability to automatically understand human actions in real world videos provides variety of applications ranging from intelligent surveillance and interactive systems to animation synthesis and content based retrieval. Despite decades of research, human action recognition remains a challenging problem in realistic videos.

More specifically, action recognition can be defined as the problem of determining what type of action is being performed by analyzing the content of an unknown video sequence and assigning it to a category among a set of predefined class labels. The complexity of motion patterns embedded in a video sequence makes it more complicated in comparison with recognition problems in still images. Spatial domain and visual cues capture the static appearance information about the scene. Whilst these attributes help to distinguish between certain action classes, they might result in confusion and error. For example, the motion pattern for "walking" and "running" classes is the only cue that can be used to distinguish the class labels, and without exploiting temporal extent of the action, they may be easily confused with one another. Therefore, we can say the visual cues of appearance and motion are two complementary elements and their fusion at the early stage of the algorithm can help to estimate high potential actionness regions.

Efficient modeling of actions is critical for recognizing human actions. In order to obtain reliable features, which are truly representative of actions, one may require to discard the background context and non-activity regions. However, in some cases background information can help to certainly prune a number of irrelevant classes from decision. For example, the observation of white snowy background can be effective in recognizing the class "Ice skating". Most existing localization approaches inclusively remove the effects of background cues and try to propose bounding boxes of regions containing the principal action [3-5]. To this end, this paper seeks to extract high potential regions of video frames, which are more likely to entail a generic action instance by the means of motion patterns. At the same time, it proposes a representation that can utilize sufficient and informative amount of background context along with foreground information.

Additionally, it is important how the motion information is used in the recognition process. Inspired by the promising performance of Convolutional Neural Networks (CNN), the majority of existing frameworks make use of a two-stream CNN approach for spatial and temporal domain [1, 3, 6, 7]. They process motion cues separately from appearance information to produce a representation for video. Conversely, the key insight in this paper lies in how motion cues can be exploited to obtain main components of an action for fitting a better spatial representation into description methods such as CNN. Our contributions are two-fold: First, a novel method is proposed to estimate actionness regions of video frames incorporating motion information. This method moves beyond proposing a single bounding box for an action. Instead, it tries to obtain all action components. This results in a better recognition outcome, especially for actions that are composed of multiple human interaction or interaction of human and objects. Second, a representation is proposed that makes use of background context information in a general manner with less prominence compared to foreground. We will show that our approach could outperform many existing methods.

The rest of this paper is organized as follows. In section II a review of the related work on action recognition is provided. Section III describes the proposed action recognition model. Section IV introduces the dataset and discusses the experiment results; and finally in section V the paper is concluded by summarizing the proposed method and obtained results, and outlining potential directions for future work.

II. RELATED WORK

There is a considerable amount of literature that aims at human action recognition. In this section, we try to provide a general overview of existing methods and position our work with respect to the current state of the literature.

Action recognition approaches can be typically divided into two categories: handcrafted feature representations, which mostly make use of local features, and deep neural network models.

A. Handcrafted feature models

There are two main points of view in using handcrafted features for action recognition. One may use a *holistic representation*, which is a global representation of human body or movements, while some others make use of *local features*. Space-Time Volume proposed in [8] falls into the holistic category, where a 3D space-time volume is created by stacking silhouette of a person along temporal dimension. Some holistic methods represent each activity with a two-dimensional template image. Motion Energy Image (MEI) and Motion History Image (MHI) [9] are examples of these template images that are created from a sequence of foreground patterns and indicate history of motion occurrences. However, it is obvious that holistic approaches learn a general schematic estimation of an action and cannot capture finer details of it. This shortcoming pushed the investigation towards local models.

Many studies have addressed human action recognition problem by extracting space-time local features [10-15]. The primary work of [15] on Spatio-Temporal Interest Points (STIPs) detects regions of an image that have significant changes in both appearance and motion. Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) are exploited in [11] to describe spatio-temporal features. It then applies a bag of features model (BoF) to encode video descriptors. Following by this approach as the representation stage, a SVM classifier is used to make action predictions. In later works, cube local video features are proposed as the 3D version of existing methods using temporal information as their third dimension. Examples of these methods are HOG3D [14], Harris3D [16], Cuboids [17].

Despite nearly acceptable achievement of handcrafted features, there exist many shortcomings, like being too rigid in capturing possible variations of actions. This pushed the research towards deep models.

B. Deep Models

Recently, Convolutional Neural Networks have achieved significant success in various fields such as image classification. Inspired by this remarkable progress, researchers have attempted to utilize these networks into video analysis applications such as action recognition. There have been considerable efforts to incorporate temporal information of videos into CNNs [6, 18-20]. We review these approaches in two categories: *Two-stream* and *spatio-temporal networks*.

A prominent class of deep neural networks is the two-stream approach, introduced in [6]. The structure of this network is composed of two parallel channels, one for spatial components to explicitly capture information about appearance, scene and individuals, and the other for temporal components, indicating motion between consecutive frames. Eventually, the final prediction is obtained by the fusion of these two streams' output. The majority of works in action recognition adopt a two-stream technique, in order to engage temporal information into the process [1, 3, 4, 7, 21].

Another approach in using CNNs for action recognition is to train a single network, by providing structures that use temporal information. These networks are called spatio-temporal networks. 3D convolutional networks are examples of these methods, introduced by Ji et al. [18]. Obviously, these networks accept predefined numbers of frames. They can only process short video clips and can encode local, short-term motion patterns. Moreover, it is unclear if they can capture adequate amount of motion information from videos. To exploit temporal cues for classifying video sequences, some studies use another example of spatio-temporal networks such as recurrent neural networks (RNNs) [22] or their extension, called long short term memory (LSTM) [23].

In comparison with the two-stream method, both of previously mentioned approaches involve complicated processing and share a large number of parameters that should be tuned. This entails the need for a very large training data that may be costly to produce. Therefore, it is essential to build up frameworks that can learn temporal information of videos without the cost of enormous training data. As these networks do not discriminate between foreground and background information, we try to feed these networks with informative regions of video frames using motion cues.

III. OVERVIEW OF PROPOSED METHOD

This paper moves beyond just recognizing video frames and tries to propose a method that can estimate regions of interest in each frame using motion and appearance information. A method is proposed that disposes the need for manual user provided annotations by estimating per frame ROI patches, containing eminent components of an action. It is obvious that these informative proposals as the input of a CNN, provides superior representation of an action and improves classification accuracy.

Our approach consists of four main steps: The first step utilizes both appearance and motion to estimate actionness potentials in which a probability value is assigned to each pixel, indicating their confidence of being foreground. In the second step, high potential ROI patches are selected from each frame and the second representation is constructed in order to incorporate adequate amount of background information. Then CNN features are extracted using the three streams of information, two for appearance models and the third for optical

flow. Subsequently, in the last step, class labels are assigned to each video.

A. Actionness map estimation

To estimate a measurement for each pixel, indicating the probability of belonging to the foreground, we follow a temporal saliency method proposed by Papazoglou et al. [24] for video object segmentation. First, the algorithm computes optical flow between two consecutive frames and based on motion information of each pixel and its neighbors, motion boundaries are estimated. Then it determines for each pixel, whether or not they are within that boundary. Motion boundaries are those pixels of frames that have abrupt changes in their optical flow field.

Two conditions are contemplated to capture boundaries. First, the gradient magnitude of optical flow in a pixel can indicate the probability of being motion boundary. If the optical flow is larger than a threshold, then the pixel is likely to be a motion boundary. Due to the inaccuracy of computing optical flow, a second factor is considered that removes the effect of moving camera motions. It uses the gradient of frame, inspecting the difference between moving direction of a pixel and its neighbors. If a pixel moves differently from all its neighbors, then the pixel is likely to be a motion boundary. Considering b_p^m and b_p^θ, respectively, as the first and second factor for each pixel *p*, the final constraint (b_p) is the combination of these two terms:

$$b_p = \begin{cases} b_p^m & if\ b_p^m > T \\ b_p^m \,.\, b_p^\theta & if\ b_p^m \leq T \end{cases} \quad (3)$$

Where *T* can be estimated as the threshold in which the value of b_p^m is considered acceptable. Finally, applying a threshold with value of 0.5 on b_p, provides a binary motion history image. However, there is still a need for appearance information to compute foreground probability of each pixel, on account of the fact that some parts of the person's body may be immovable and motionless in some frames of the video. Consequently, an appearance model is considered by fitting a GMM over RGB values and exploring whether a pixel is inside the produced motion boundaries using point-in-polygon (PIP) problem. This generates an inside-outside map for each frames of the video. Eventually, combing these two models, using the summation of their logarithmic probabilities, presents the likelihood of each pixel belonging to the foreground.

B. Selecting actionness patches

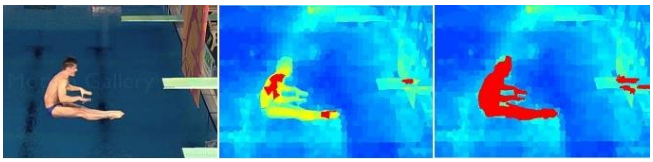
The main idea behind the proposed method is to use motion information embedded in video with the purpose of providing informative inputs to the network both during train and test stages. The algorithm mainly explores those regions where actions are happening. The goal is to decide which windows in frames are appropriate for better describing the video. Using the foreground probability map described in pervious section, the algorithm tries to select high potential areas of frames resembling the problem of sampling over spatial dimension.

For this purpose, a methodology similar to hysteresis thresholding is proposed. Two different thresholds are considered and are applied to the foreground probabilities. This leads to creation of three classes of pixels: weak, strong and in-between pixels. Pixels below the lower threshold are weak pixels, which can be safely discarded due to the low probability of being a part of an action. Pixels above high threshold are strong pixels that should be retained and added to the set of pixels which have high potential of containing an action. Finally, for in-between pixels, which are pixels between low and high threshold, they are classified as strong if they are connected to a strong pixel directly, or by less stringent indirectly, by the means of other previously connected in-between pixels. They can be considered as continuations of strong areas and can be added to the set of selected pixels. Correspondingly, these three types of pixels form three classes of weak, strong and in-between regions in frame. In-between regions are those parts of human body or object that may not have an abrupt motion but have relatively good potential of being foreground due to their almost high probability and also closeness to a strong region. Using two thresholds benefits from two principal advantages: first, it insures selecting only high potential pixels. Those pixels that do not have a significant value and are not nearby a strong region are assumed as noisy pixels and are discarded. Second, not only strong regions are selected but also those in-between regions that are probable to be a part of an action are chosen. This provides a better generalization of algorithm to find high potential regions. Fig. 1 shows the improvement clearly where at first small parts of human body are selected and then the detection is improved using the second threshold. In Fig. 2 the detected high potential pixels for three different actions are illustrated.

After specifying high probable pixels, the algorithm detects connected regions and calculates the average of foreground probabilities within each region. Then the maximum value of the mean probabilities is selected and a fixed size window is considered over that pixels to produce a candidate ROI for that frame. For the last step, all pixels inside the chosen ROI are discarded from further processing at this stage. This procedure is repeated until a specific number of ROIs is acquired. The point to notice is that those ROI patches that have more than seventy percentage overlap with previously selected ROIs are eliminated. Fig. 3 presents extracted patches for three different actions. It can be seen that the proposed algorithm could not only detect the person performing action but also it could identify all components of actions including objects and interaction of other individuals.

C. Incorporating background information

The method proposed in the previous section, is capable of identifying important components of an action. As stated in the introduction, in many cases, background can provide rather useful information about the activity. There are two ordinary trends in exploiting background information. Some methods try to learn the original frames. These methods do not discriminate between ROI and background information. In contrast, in the second category of methods the ROI is typically extracted by leveraging background subtraction, tracking or pose estimation



(a) (b) (c)

Fig. 1. Example of improvement using two thresholds. (b) and (c) shows the heat map of foreground probabilities. The red colors are those pixels that are selected. (a) Original frame of diving class. (b) Above high threshold regions. (c) Rectified selected regions using second threshold.

(a) (b) (c) (d) (e) (f)

Fig. 2. Visual examples of detected regions for three action classes, (a) Lifting (c) Walking (e) Kicking

(a) (b) (c)

Fig. 3. Examples of Detected ROI patches. (a) Both person and the object involved are detected. (b) The person performing the action is detected. (c) All individuals interacting in action are detected.

approaches, where the background is completely ignored. However completely discarding background is not recommended, while it can be helpful of recognizing actions.

Therefore, this paper suggests another representation that can take advantage of background information in decision-making to complement the previously described patch-based representation. In this representation, those pixels with a probability Pr_i higher than a specific value are extracted and then the smallest rectangular window W_f, bounding them is generated. In the next step, a new representation is built where pixels inside the window remain intact in the new representation and the remaining pixels become blurred using Gaussian filters with standard deviation parameters σ proportional to the inverse of Pr_i for each pixel P_i and the distance (Dist) to the selected window W_f:

$$\sigma \sim \mathrm{Dist}(P_i, W_f) \sim \frac{1}{Pr_i} \tag{4}$$

This results in a representation in which background information has less impact and power to respond to the CNN filters since the strong structures in the background have been suppressed. Fig. 4 illustrates the proposed weighted representation.

D. Descriptor computation and classification step

This paper utilizes a convolutional neural network to perform the task of automatic feature learning. Following the two-stream convolutional networks used in [3], we make use of a same architecture but using it as a three-stream network, one temporal stream for the optical flow fields and two spatial streams for the two proposed appearance representations. Eventually, a nonlinear SVM classifier is used as the last step of proposed framework for the final classification.

IV. EXPERIMENTS AND RESULTS

We conduct experiments on the real-world UCF sports dataset to demonstrate the effectiveness of our action recognition model in real scenarios and compare our approach against a number of state-of-the-art methods. Details about the dataset and the experiments performed are given below.

A. UCF Sports Dataset

UCF sports dataset contains 150 videos categorized into 10 action classes. Videos are collected from real sports broadcasts. The recognition task in this dataset is challenging because of a wide range of scenes, viewpoints and camera movements. We use the standard training and test split that is used by many researches and suggested in [25].

B. Implematation details

We implement our deep learning tasks based on the Caffe open source toolbox. The optimal value of the two thresholds used in section III for selecting actionness patches, have been specified through experiments on a validation set. For the lower threshold the value of 0.5 and for the higher threshold the value of 0.7 is selected.

To capture CNN features, a three-stream CNN model is employed. We make use of AlexNet architecture pre-trained on UCF Sports dataset, for the two special streams. The inputs of one of this spatial networks are our ROI patches and the other are the proposed Blurred representation. To capture motion patterns, our temporal stream network bears a close

Fig. 4. Examples of two frames and their blurred weighted representation for actions "Swing bench" and "Walking".

resemblance to the network architecture and setups used in [3]. We adopt the state-of-the-art motion-CNN proposed by [3], which is pre-trained on UCF sport and UCF101 dataset. To construct the input of motion-CNN, optical flow is computed between each consecutive frame, using the method of [26]. Then the x-component, y-component and the magnitude of optical flow are stack to form a 3D image, where the third dimension is the magnitude value. This 3D image is submitted as the input of motion-CNN.

The features are then extracted from fc7 layer [3] of three CNNs, representing action specific regions and motion cues. Afterwards, a temporal statistical pooling is applied in order to obtain fixed length feature vectors per video. Precisely explaining, video features are constructed by aggregating all frame descriptors of each motion and Blurred representation, and all patch descriptors separately using average, max, median and variance aggregations. This results in three descriptors with n=4*4096 dimensions.

Eventually, the three learnt spatial and temporal features are fused to produce a final classifier for action recognition. We use a nonlinear multi-class SVM classifier using LibSVM, with an RBF kernel function to predict action category labels of each video. Features of three representations are fused in the kernel space, using their similarity measures.

C. Exprimental results

We evaluate our method on UCF Sports Action dataset. We choose the mean per-class recognition accuracy as our evaluation criteria. In order to comprehensively evaluate the performance of our method, we try to compare it with state-of-the-arts baselines. TABLE 1 summarizes the result of action recognition accuracies in different works across this dataset. From the results, it is clear that our proposed method improves the accuracy of action recognition in UCF dataset and achieves excellent performance in comparison with other competing state-of-the-art results. The confusion matrix for UCF sports dataset of our method is shown in Fig. 5.

V. EXPRIMENTAL DISCUSSION

The experimental results highlighted the impact of motion cues in order to estimate regions of action interests in videos. We attribute the improvement made by the proposed method to three main reasons. Firstly, it uses the motion embedded in the video to estimate high potential areas of actions, under the strong assumption that appearance cues are not the only information that can be used to distinguish an action. Indisputably, motion information can also provide constructive evidences in determining where an action may have occurred. Using motion in early stage of the algorithm, contributes in a better estimation of actionness regions. Secondly, the proposed method moves beyond finding a single bounding box for each action and tries to find almost all probable components of an action. This exert less constraint on recognizing different types of actions, especially those actions that consist of multiple human interactions or the involvement of some object in the action. Finally, we do not suppress the effect of background information in decision making. We propose a framework that can exploit the gist of background cues holistically, while still giving foreground information significantly more weight.

The only two videos that our method could not label correctly belong to the "running" and "skateboarding" classes. Delicately inspecting these two videos, we came to an interesting discovery. They are the two videos that neither have adequate appearance cues nor significant motion. They are both misclassified as "walking". As previously mentioned, the "running" videos do not have obvious differences in the appearance with "walking". Moreover, in this particular input the video of person who is running is very slowed down. Therefore, it was very hard for the recognition system to classify it correctly. Similarly, the viewpoint that the "skateboarding" video is recorded causes the skateboard to not be observable by the camera. Furthermore, it does not contain any specific motion. There is just a person coming towards the camera steadily and unidirectionally, similar to a person who is walking. While in training videos for the "skateboarding" class, all of the samples contain some kind of spiral movements or at least the skateboard is obviously visible. However, the proposed system managed to identify the location of action perfectly in both videos and consequently the problem might due to the shortcoming of recognition phase. We feel strongly that the shortcoming may be as a result of imperfect training, which can be handled by providing adequate samples of different viewpoints and conditions of movements in the training stage.

VI. CONCLUSION AND FUTURE WORK

This paper has highlighted the importance of motion cues in order to attain representative regions of a video sequence in a weakly-supervised learning manner. To achieve an informative

TABLE 1. COMPARISON OF AVERAGE RECOGNITION ACCURACY WITH STATE-OF-THE-ART METHODS ON UCF SPORTS DATASET.

	Recognition accuracy (%)										
	HOG/HOF [11]	***HOG3D [14]***	***Kovashka et al [27]***	***Wang et al. [28]***	***IDT+FV [29]***	***WOC+AdaBoost [30]***	***H-FCN [4]***	***RANK-POOL-CNN [20]***	***Tubelets [5]***	***R-CNN [21]***	***Our Method***
UCF-Sports	82.6	85.6	87.27	86	88	90.67	82.7	87	80.24	91.49	**95.74**

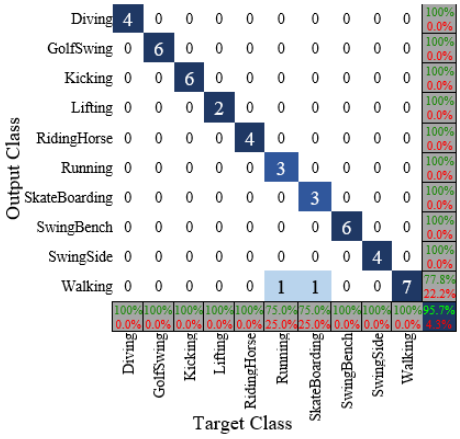


Fig. 5. The confusion matrix of recognition on the UCF Sports dataset

representation of a video, we suggested an approach that utilizes motion information. We proposed two different representations. In the first one, high potential patches of video containing components of an action are extracted. In the second representation, we tried to include background information. This provides a gist of background structure as the input of the recognition system. These two representations are passed through a three-stream CNN, where the third stream is the optical flow fields. From the experimental results, we illustrate that our approach outperforms the state-of-the-art methods for action recognition on the UCF Sports dataset. We also presented an analysis of our experimental results.

As a potential direction for future work, we intend to investigate a novel mechanism to efficiently aggregate sets of frame level features into discriminative and fixed-size video descriptors.

REFERENCES

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732.

[2] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 3169-3176.

[3] G. Gkioxari and J. Malik, "Finding action tubes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 759-768.

[4] L. Wang, Y. Qiao, X. Tang, and L. Van Gool, "Actionness estimation using hybrid fully convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2708-2717.

[5] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek, "Action localization with tubelets from motion," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 740-747.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933-1941.

[8] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, pp. 1395-1402.

[9] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on pattern analysis and machine intelligence, vol. 23, pp. 257-267, 2001.

[10] 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, 2009, pp. 124.1-124.11.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1-8.

[12] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on, 2009, pp. 1996-2003.

[13] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 2929-2936.

[14] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in BMVC 2008-19th British Machine Vision Conference, 2008, pp. 275: 1-10.

[15] I. Laptev, "On space-time interest points," International journal of computer vision, vol. 64, pp. 107-123, 2005.

[16] I. Sipiran and B. Bustos, "Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes," The Visual Computer, vol. 27, pp. 963-976, 2011.

[17] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, 2005, pp. 65-72.

[18] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 35, pp. 221-231, 2013.

[19] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3218-3226.

[20] B. Fernando and S. Gould, "Learning end-to-end video classification with rank-pooling," in International Conference on Machine Learning, 2016, pp. 1187-1196.

[21] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in European Conference on Computer Vision, 2016, pp. 744-759.

[22] 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.

[23] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in International Conference on Machine Learning, 2015, pp. 843-852.

[24] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1777-1784.

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

[26] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," Computer Vision-ECCV 2004, pp. 25-36, 2004.

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

[28] Y. Wang, Y. Li, and X. Ji, "Human Action Recognition Based on Global Gist Feature and Local Patch Coding," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, pp. 235-246, 2015.

[29] H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 3551-3558.

[30] X. Yan and Y. Luo, "Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier," Neurocomputing, vol. 87, pp. 51-61, 2012.