IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

Review of Multiple Human Action Recognition Techniques

Mr. Dhananjay A. Deshpande[1] and Prof. Snehil G. Jaiswal[2]
M. Tech Student, Department of Electronics and Telecommunication[1]
Assistant Professor, Department of Electronics and Telecommunication[2]
G. H. Raisoni University, Amravati, Maharashtra, India

Abstract: *During the last decades topics such as video analysis and image understanding have acquired a big importance due to its inclusion in applications such as security, intelligent spaces, assistive living and focused marketing. Human action detection is investigated in utilization of artificial intelligence and computer vision. Numerous effective action recognition strategies have demonstrated and the action information are successfully gained from motion videos and still pictures. In order to get equivalent actions, the proper activity information gained from various kind of media like video or picture might be connected. The majority of existing video activity action identification strategies experience the ill effects of inadequate recordings. In this review article we are going to discussed about some earlier human action recognition techniques. In past numbers of researchers, developers dose a nicest work in same specific domain, in this review article/Paper we are going to review their work. During this we are going to find out and conclude the result and working of different Human action recognition techniques. In this process we are going to serve our focus on some specific action resignation techniques like: Action Recognition System using 3D Convolutional Neural Networks, Using Deep Learning, Deep Convolutional Neural Networks, Image Processing.*

Keywords: Action Recognition System, Convolutional Neural Network (CNN), Human Activity Recognition (HAR), Deep Learning (DL), Machine Learning (ML), 3D CNN

I. INTRODUCTION

Understanding human action in videos has received significant research attention in the field of video analysis. Most applications are in summarization, video content retrieval, and human-computer interfaces. Most existing method require manual annotation of relevant portion of action of interest. Human action recognition can be made more reliable without manual annotation of relevant portion of action of interest. This paper presents not only an update extending previous related surveys, but also focuses on a joint learning framework that identify the temporal and spatial extent of action in videos. Dense trajectories are used as local features to represent the human action. It is more fine grained. Action localization is made by learning the temporal and spatial extents of video. Split and merge algorithm allows the segmentation which is followed by training the video.Human action detection in videos are emerging topics in computer vision since understanding the human action helps in management, summarization and retrieval of videos. In the past two decades significant progress has been made with the invention of local invariant features and bag of features representation. The task is challenging due to variations in action performance, background settings and inter-personal differences. To understand the action in the video, there are two things to be noted- Action recognition and action localization. One is 'what action' is performed in the video and the other is 'where the action' of interest is taken place. The problem of assigning videos into several predefined action classes is known as action recognition and action localization is finding the spatio-temporal content of the video. In action recognition we are training the videos into several classes. Training includes both positive and negative samples. After the training phase we can test the videos. In the existing systems, in the training phase we have to manually annotate the relevant part of the action of interest in the video. Manual annotation is tedious, time consuming process and error prone process. So here we introduce an automatic method for finding the relevant portion of the action of interest in the video without human intervention. Different from previous approaches it does not require reliable human detection and tracking as input. There are also

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

such action detection methods that identify which video contents are occupied by the performer of action. introduces a method to identify the temporal extent of the video. But it ignored the spatial context. Ignoring one domain may produce irrelevant content from that domain. The proposed method introduces a joint learning framework for finding the spatial and temporal extent of the action. So that it is easy to find the relevant portion of action in the video and thereby easily recognizing the video. The person location is inferred as latent variables. Temporal smoothness is also enforced along with learning the spatial model. Trajectories are extracted from the video to represent the action. Using dense trajectories for representing the video is finer grained because it is at pixel-level accuracy than single media-based solutions.

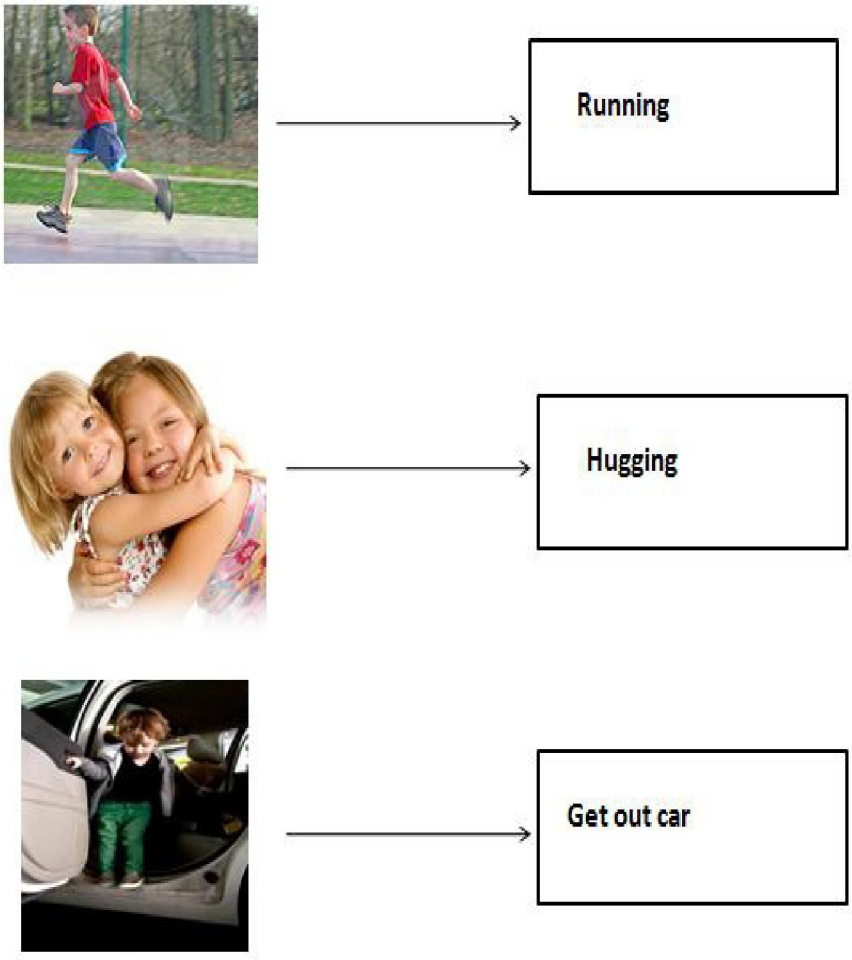

Figure 1. Basic Action detection process and output. The second column shows the name of the action detected from the frames of the videos.

II. LITERATURE REVIEW

For many years human action recognition has been studied well. Most of the action recognition methods require to manually annotate the relevant portion of the action of interest in the video. In recent years it has been studied that the relevant portion of action of interest can be found out automatically and recognize the action. We can review the action recognition methods

2.1 Action Recognition

For representing video, feature trajectories have shown efficiency. But the quality and quantity of these trajectories were not sufficient. As the use of dense sampling came popular for image classification Wang et al. [1] proposed to use

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

dense trajectories for representing videos. Dense points from each frame are sampled and traced them based on displacement information. For improving the performance Wang et al. [2] takes into account the camera motion. The camera motion is estimated by matching feature points between the frames by using SURF descriptors and dense optical flow. Another approach [3] aimed at modelling the motion relationship. The approach operates on top of visual codewords derived from local patch trajectories, and therefore does not require accurate foreground-background separation. Dorr et al. [4] proposed another method for finding the informative regions. They used saliency mapping algorithms. As a new method this paper proposes using a joint learning framework for learning spatial and temporal extents of action of interest.

2.2 Action Detection

Recognition was performed using the Mahala Nobis distance between the moment description of the input and each of the known actions. Recent popular methods which employ machine learning techniques such as SVMs and AdaBoost, provide one possibility for incorporating the information contained in a set of training examples.[4] introduces the Action MACH filter, a template-based method for action recognition which is capable of capturing intraclass variability by synthesizing a single Action. Another method is proposed in [5], multiple-instance learning framework, named SMILE-SVM (Simulated annealing Multiple Instance Learning Support Vector Machines), is presented for learning human action detector based on imprecise action location. Wang et al. [6] used a figure-centric visual word representation. In that localization is treated as latent variable so as to recognize the action. A spatio-temporal model is learned. During the training [7] model parameters is estimated and the relevant portion is identified.[8] proposed an independent motion evidence feature for distinguishing human actions from background motion. Most of the methods require that the relevant portion of the video has to be annotated with bounding boxes. Human intervention was tedious. So, to overcome the bounding box Brendel et al. [9] divides the video into a number of subgroups and then a model was generated that identify the relevant subgroup. This paper introduces a method that learns both spatial and temporal extents for detection improvement. Dense trajectory is used here as local features to represent the human action.

2.3 Same Domain Related Work Analysis

Several recent depth-based approaches have been reported to improve human action recognition accuracy. An action graph based on a sampled 3D representation from a depth map to model the human motion is proposed in. Several 4Ddescriptors have been used to represent the human action. In a histogram of oriented 4D normal (HON4D) used in order to describe the action in 4D space covering spatial coordinates, depth and time. Also represents the depth sequence in 4D grids by dividing the space and time axis into multiple segments. Another 4D descriptor proposed by called Random Occupancy Pattern (ROP) which deals with noise and occlusion combined with sparse coding approaches to increase robustness. Action recognition from different side views has been applied to gain more discriminative features. Generates side view from the front view of the depth map, both views are transformed to DMA (Depth Motion Appearance) descriptor and DMH (Depth Motion History)descriptor. Then, SVM is trained with the two descriptors to classify the action. Recently generate top and side views by rotating 3D points from the front view. The three views are used as inputs to three convolutional neural network models for feature extraction and action classification. In parallel to depth-based approaches, skeleton-based methods also have a huge contribution to the action recognition research area. In, each joint is associated with a Local Binary Pattern descriptor which is translation invariant and provide highly discriminative features. Additionally, a temporal motion representation called Fourier Temporal Pyramid is also proposed in order to model the joints movements. Eigen Joints is a new type of features proposed in to combine action information including static postures, motion and offset features. A framework based on sparse coding and temporal pyramid matching is proposed in for better 3Djoint features representation. A histogram of 3D joint location called HOJ3D in represents the human joint's locations. Then, a posture words are built from HOJ3D vectors and trained using a Hidden Markov Model to classify the actions. In a framework is proposed for online human action recognition using a new Structured Channelling Skeletons feature(SSS) which can deal with intra-class variations including viewpoint, anthropometry, execution rate, and personal style. proposed non-parametric Moving Pose (MP) for low latency human action and activity recognition, the framework considers pose information, speed, and acceleration of the joints in the current frame within a time window. A hierarchical dynamic framework was

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

reported in based on using deep belief networks for feature extraction and encodingdynamic structure into a HMM-based model. Addresses action recognition in videos by modelling the spatial-temporalstructures of human poses. The method improves the pose estimation first, then groups the joints into five body parts. Moreover, data mining techniques have been applied to get spatial-temporal pose structures for action representation and transform the joint coordinates to a 2D image descriptor. A convolutional neural network model is used for action classification from the descriptor. Very recent works: SOS and Joint Trajectory Maps propose a new approach which transforms the skeleton joints trajectories shapes from3D space into three images that represent the front view, the top view and the side view of the joints' trajectory shapes. Three convolutional neural networks extract features from the three images to classify the action. Convolutional neural network is a powerful technique for feature extraction and classification. Recent action recognition approaches started to focus more on using CNN for action classification rather than using SVM. Researchers in deep learning try always to come up with new techniques to improve the CNN architectures and enhance the performance of feature extraction, classification and computation speed. Summarise recent advances in convolutional neural network in term of regularisation, optimisation, Activation functions, loss functions, weight initialization and so on. Recent CNN based action recognition methods are based on using multiple action representations that employ many CNN channels for the processing. In many features concatenation architectures are proposed in order to improve the classification accuracy using multiple sources of knowledge. In spite of the fact that the previous approaches achieved good results, the problem of action recognition is still open and require more robust action representations and feature extraction techniques to improve the accuracy and overcome the weakness of the previously mentioned methods. To this end, the proposed work in this paper investigates the use of both types of data, depth maps and postures to enhance the action recognition throw the power of CNN for feature extraction and classification.

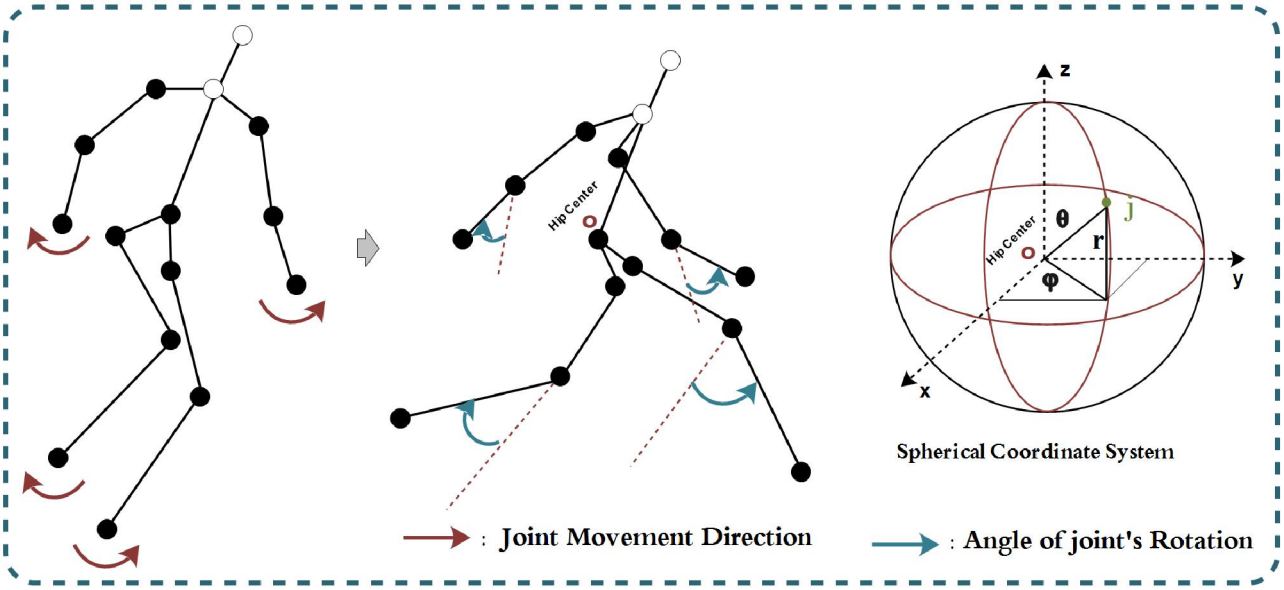

Figure 2. Human body joints motion direction during a running action. The joints motion more subject to a rotation, which makes the spherical coordinate system more suitable to represent the joints movements.

III. PROPOSED SYSTEM FOR MAXIMUM ACTION RECOGNITION TECHNIQUES

3.1 Action Recognition Using Image Processing

Many activity acknowledgment techniques pursued the customary system. Initially, countless movement highlights are extricated from videos. At that point, every single neighbourhood include are quantized in to a histogram vector utilizing back of-words (bow) portrayal. Later, the vector-based-classifiers, e.g., bolster vector machine are utilized to perform acknowledgment in the testing and recording. At a point when the recordings are straightforward, these activity acknowledgment strategies have accomplished promising outcomes. Nonetheless, noises and the uncorrelated data might get added to the bow amid the quantization and extraction of the nearby highlights. In this way, these techniques

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

are typically not powerful and couldn't be used much when the video having significant camera shaking, impediment, jumbled foundation, etc. So as to improve the acknowledgment precision, important parts of activities, e.g., related articles, human appearance, act, etc, ought to be used to form a clearer semantic understanding of human activities. Late endeavours have exhibited the viability of utilizing related items or human postures. These techniques may require a preparation procedure with extensive measure of recordings to get great execution, particularly for true videos. In most cases, human activity inclination can likewise be passed on by still pictures. In proposed an adjustment technique for video action recognition. Not quite the same as the current adjustment methods based on a similar component, our strategy can able to adapt knowledge among spaces that are in various feature spaces. Distinctive highlights can give enhanced performance and thanks to the corresponding attributes. Meanwhile, the adaptability expanded and the adjustment can be conducted between diverse spaces. In request to investigate the nearby complicated structures along with the preparing video information successfully use the unlabelled information in video domain, the adjustment procedure in a semi supervised learning system can be done. Test results show that the calculation isn't just effective but also has better adjustment execution, particularly when just few named preparing tests are given.

3.2 Proposed Work

In the proposed work, the image feature from the pictures and key edges of videos were extracted. Considering computational productivity, the proposed system will separate key edges by a shot boundary detection algorithm. First the video is given as input and then the features are extracted in the form of images and then combined with video feature and preceded to classifier and by using classification techniques the output is generated as shown in figure 3. To start with, the colour histogram of each 5 frames is determined. Second, the histogram is subtracted with that of the earlier frame. Third, when the subtracted value is bigger than the empirically set threshold then the frame will be set as a key frame shot boundary. The frame in the centre of the shot is considered as a key frame only when we get the shot. This method is called shot boundary detection. Meanwhile, the video (movement) is separated from the video domain and joined with the image feature. The picture element is a subset of the combined element. The Kernel Principal Component Analysis (KPCA) technique is used in the proposed system for finding the image features and joined features.

The KPCA strategy says the primary information of the mapped Hilbert spaces. In this way, the preparation procedure is progressively proficient, which makes the Independent Vector Analysis (IVA) increasingly reasonable for true applications. The common features can be obtained by mapping the image feature into a Hilbert space. So as to get the heterogeneous features-ab, the joined features are mapped into another Hilbert space. The information can be adjusted dependent on those shared space with the common features, after it is used to upgrade the classifier-a. So as to make utilization of unlabelled videos, a semi supervised classifier-ab is prepared dependent on the heterogeneous features in video domain. By combining the two classifiers we can get joint optimization framework. The last acknowledgment after effects of testing recordings are improved by combining the consequences of previously mentioned two classifiers. It avoids over fitting and gives good performance even during a few labelled training videos are available.

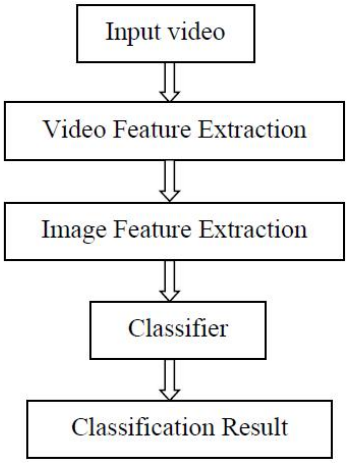

Figure 3. flow chart of the proposed work

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

IV. CONCLUSION

The availability of big data and powerful models diverts the research focus on human actions from understanding the present to reasoning the future. We have presented a complete survey of state-of-the-art techniques for action recognition and prediction from videos. These techniques became particularly interesting in recent decades due to their promising and practical applications in several emerging elds focusing on human movements. We investigate several aspects of the existing attempts from above study present in this article, we can conclude that we have successfully reviewed multiple action recognition techniques successfully.

V. ACKNOWLEDGMENT

We are deeply grateful to all those who contributed to the success of this review research paper. First and foremost, we would like to thank our primary supervisor **Prof. Snehil G. Jaiswal**, for their guidance, support, and encouragement throughout the entire process. Their mentorship and expertise were invaluable in helping us to shape the direction of our review research and to bring our ideas to fruition.

I would also like to thank the organizations and individuals who provided me a support for this review research, including G.H. RAISONI University Amravati, Maharashtra, India. Without their generous contributions, this review research would not have been possible.

Overall, this research project would not have been possible without the support and contributions of so many people. We are deeply grateful to all of those who helped to make this project a reality, and we hope that our findings will make a meaningful contribution to the field

REFERENCES

[1]. H.Wang,A .Klaser,C.Schmid and C-L.Liu, "Action recognition by dense trajectories ," in Proc. IEEE Conf. Comput. Vis.Pattern Recog., Jun.2011, pp 3169-3176.

[2]. H.Wang and C Schmiid , "Action recognition with improved trajectories," in Proc.IEEE Int. Conf.Comput. Vis., Dec 2013, pp 3551-3558.

[3]. Y-G Jiang,Q.Dai,X.Xue,W.Liu and C-W Ngo. "Trajectory-based modelling of human actions with motion reference points," inProc. Eur.Conf .Comput.Vis.,Oct 2012,Vol 7576,pp.425-438.

[4]. M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatiotemporal maximum average correlation height filter for action recognition,"in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun. 2008, pp. 1–8.

[5]. Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Sep.–Oct. 2009, pp.128–135.

[6]. T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in Proc. IEEE Int. Conf. Comput. Vis., Nov. 2011, pp. 2003–210.

[7]. M.Raptis, I.Kokkinos and S.Soatto," Discovering discriminative action parts from mid-level video representations" ,in Proc ,IEEE Conf.Comput.Vis,.Pattern Recog.,Jun 2012,pp.1242-1249

[8]. W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," in Proc. IEEE Int. Conf. Compute. Vis., Nov. 2011,

[9]. M.Jain,J.van Genmert,H.Jegou ,P.Bouthemy and C.Snoek ,"Action localization with tubelets from motion" ,in Proc IEEE Conf, Comput.Vis.Pattern Recog. Jun 2014 pp 740-747

[10]. Caroline Rougier, et.al, "Robust Video Surveillance for Fall Detection Based on Human Shape Deformation", IEEE Transactions on Circuitsand Systems for Video Technology, Vol. 21, No. 5, May 2011. pp. 611-622.

[11]. Ronald Poppe, "A survey on vision-based human action recognition", The Netherlands Image and Vision Computing, vol. 28 (2010), Pp.976–990.

[12]. Jungong Han, et.al, "Enhanced Computer Vision with Microsoft Kinect Sensor: A Review", IEEE Transactions on Cybernetics, Vol. 43, No.5, October 2013. Pp. 1318 – 1334.

[13]. Nicolas Thome, et.al, "A Real-Time, Multiview Fall Detection System: A LHMM-Based Approach", IEEE Transactions on Circuits andSystems for Video Technology, Vol. 18, No. 11, November 2008. Pp. 1522-1532.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, July 2023

[14]. Daniel Weinland, et.al, "A survey of vision-based methods for action representation, segmentation and recognition", Computer Vision andImage Understanding, vol.115 (2011), Pp. 224–241.

[15]. D.M. Gavrila and L.S. Davis "3D model-based tracking of humans in action: a multi-view approach", IEEE. Pp. 73-80.

[16]. B. Ma, L. Huang, J. Shen, and L. Shao, "Discriminative tracking using tensor pooling," IEEE Trans. Cybern., to be published, doi:10.1109/TCYB.2015.2477879.

[17]. L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: A genetic programming approach," IEEETrans. Cybern., vol. 46, no. 1, Jan. 2016, Pp. 158–170.

[18]. A. Khan, D. Windridge, and J. Kittler, "Multilevel Chinese takeaway process and label-based processes for rule induction in the context ofautomated sports video annotation," IEEE Trans. Cybern., vol. 44, no. 10, Oct. 2014, Pp. 1910–1923.

[19]. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in Proc.Brit. Mach. Vis. Conf., London, U.K., 2009, Pp. 124.1–124.11.

[20]. L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," IEEE Trans. Cybern., vol. 44, no.6, Jun. 2014, Pp. 817–827,

[21]. M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," School Comput. Sci., Carnegie MellonUniv., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-09-161, 2009.