Human Action Recognition by Conceptual Features

Ghazal Shamsipour, Jamshid Shanbehzadeh, and Abdolhossein Sarrafzadeh

***Abstract*—Human action recognition is the process of labeling a video according to human behavior. This process requires a large set of labeled video and analyzing all the frames of a video. The consequence is high computation and memory requirement. This paper solves these problems by focusing on a limited set rather than all the human action and considering the human-object interaction. This paper employs three randomly selected video frames instead of employing all the frames and, Convolutional Neural Network extracts conceptual features and recognize the video objects. Finally, support vector machine determines the relation between these objects and labels the video. The proposed method have been tested on two popular dataset; UCF Sports Action and Olympic Sports. The results show improvements over state-of-the-art algorithms.**

***Index Terms*—Computer Vision, Human Activity Recognition, Convolutional neural networks, Support vector machine**

I. Introduction

Recently, analyzing and understanding human action or activity become an interesting topic because of two factors. First, the advancement of technology and the increase of low cost and powerful imaging equipment result in exponential growth of video creation. Second, the development of a large number of programs including human-robot interaction, human-computer interaction, intelligent video surveillance, face analysis, object tracking, video-processing and video annotating, robotics, smart aware-house, rehabilitation center, video games and a variety of systems that involve interactions between human and computer [1]-[2]-[3]-[4].

Human behaviors are analyzed according to gesture, action and activity Gesture or elementary action includes automatic and simple movement such as hand raising or foot forwarding. Action is a series of gesture that temporarily put together and they describe the entire body. Finally, the activity is a series of actions which include interactions and group activities [1]-[2]-[5]. Also, interactive activities include human-object or human-human interactions (See Fig. 1). This paper focuses on human-object interaction to understand the human activity.

Manuscript received Nov 20, 2016.

Ghazal Shamsipour is a Ph.D. student with the Department of Computer Engineering, Islamic Azad University of Najafabad, Esfahan, I.R. Iran (e-mail: gh.shamsipour@gmail.com).

Jamshid Shanbehzadeh is an Associate Professor with the Department of Electrical & Computer Engineering, Faculty of Engineering, Kharazmi University (Tarbiat Moallem University of Tehran), Tehran, I.R. Iran (phone: +98 26 34550002; fax: +98 26 34569555; e-mail: jamshid@khu.ac.ir).

Abdolhossein Sarrafzadeh is a Professor and Director, High Tech Research Network, Unitec Institute of Technology, Auckland, New Zealand (Phone +64 9 815 4321 ext. 6040; email: hsarrafzadeh@unitec.ac.nz).

This work is the outcome of Shamsipour's M.Sc thesis at Kharazmi University

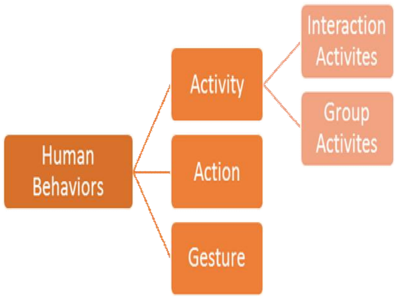


Fig. 1. Human behaviors based on complexity level.

Although, human action and activity recognition has started since 1973 [1], unsolved issues have been remained such as view point, clutter, diversity of actions, actor movement variations, high cost computing and memory requirement. One of the main problems in this area relates to the variation of human actions and activity.

As previously mentioned, we have three categories for human behavior. Also, each person has his own style [2]. Therefore, designing a high performance recognition system for all categories of human behaviors is complicated. A solution is to implement system to recognize a limited number of actions or activities.

This paper focuses on improving the recognition performance by limiting the variation of actions by understanding human-object interaction. It employs only three randomly selected video frame instead of all the frames and, a pre-trained Convolutional Neural Network (CNN) via ImageNet pictures extracts the high level and conceptual features and recognize the video objects and, Support Vector Machine (SVM) understands the action by determining the relation between the objects and labels the video.

The rest of this paper is organized as follows. Section 2 reviews the human action or activity recognition systems. Section 3, explains the theoretical background; convolutional neural networks and support vector machine. Section 4 explains the proposed system. Section 5 present the simulation results and comparison with other along the same systems. The final section concludes the paper and talks about future activities.

II. Literature Review

Figure 2 shows the general form of a human action or activity recognition systems where it consists of four sections: input, detection, tracking and recognition [1]-[2]-[6] Although action and activity are semantically difference, but in most cases it can be considered that there is no different and they are equal [7]. Next, we discuss each section briefly.

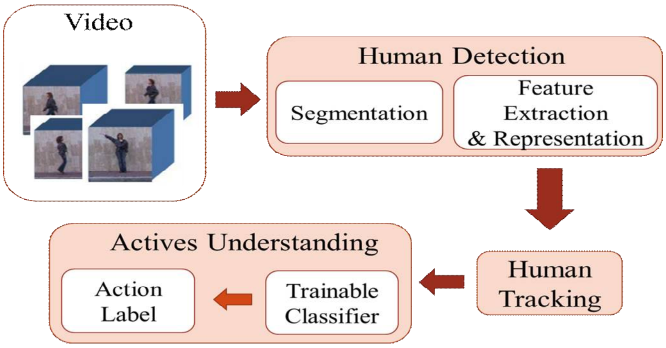


Fig. 2. Human action recognition system.

A. Input

A recognition system receive input data such as video or still image from visual devices such as cameras. This paper focuses on systems that the location of camera is unimportant and the input is video.

B. Detection

The important section in this system is detecting human. This section separates the main object in video frames. The detection is performed by data representation and the features extracted should be robust against small video changes such as human appearance, rotation and displacement, lighting, partial human latency, occlusion, point of view or run operation. The system have to fix its human of interest [1]-[4]-[6]-[8]. Recently, researchers use more hierarchal structure for action recognition.

C. Tracking

After finding the human's location in each frame, the system locates a few points of human body. We use these points to track the human body and extracting the movement pattern. This part distinguishes human's movements from other objects' movements [1]-[6] Morris et all [9] divided the object tracking algorithms into 6 groups based on region, contour, feature, model, hybrid and optical flow.

TABLE I
PRESENTS THE RESULTS OF METHODS

First author	Algorithm
Ji, 2013 [11]	Develop a 3D convolutional neural network. It is an extension of 2D CNN. Extracts features from both spatial and temporal dimensions.
Simonyan, 2014 [12]	Represent a two-stream architecture of CNN. *Spatial stream* captures appearance information from still frames and *Temporal stream* extracts optical flow to capture motion between frames.
Wang, 2015 [13]	Presented model called trajectory pooled deep convolutional descriptor (TTD). He used hand-crafted features to extract trajectories and a two stream CNN for extracting feature maps.

D. Understanding

The role of understanding is analyzing the motion pattern and find the best description of actions and activities. Understanding human behavior is classification the human action and activity. The classification matches an unknown data with a group of sample reference and labeling human action and activity. The learning methods involves supervised, unsupervised and semi-supervised model [6]-[10]. Table 1 present the several popular algorithms on human action recognition which used deep neural networks for action recognition.

III. BACKGROUND THEORY

In This section introduces the sub-systems of the proposed system including CNN and SVM. This introduction makes understanding this paper more easily. The role of CNN and SVM are extracting suitable features and classification of action and activity respectively.

A. Convolutional Neural Networks

Alex Krizhevsky and colleagues used CNN to classify 2.1 million high-resolution images into 1000 classes. CNN improves the classification performance compared to the other methods significantly and it overcomes overfitting which is observed in most learning issues. Figure 3 shows the structure of CNN employed by Krizhevsky and colleagues. This network can run in parallel on GPUs [14].

CNN composes of 8 layers, 5 convolutional layers and 3 fully connected layers. The small local parts of the input are captured by the convolutional layer with a set of local filters, and the pooling layer preserves the invariant features. Top fully connected layers combine inputs of all features to perform the classification [14]-[15]. This hierarchical organization generates good results in image processing and it had been applied to large-scale visual recognition tasks because low-level edge information leads to complex representations such as mid-level cues like edge intersections or high-level representation like object parts. Also, this method is convenient for extracting dependent_local features and scale invariant feature [16]-[14].

B. Support Vector Machine

Vapnyk introduced primary SVM in 1963 [17] SVM is a supervised learning algorithm for classification and regression and, it is widely used in object detection & recognition, content-based image retrieval, text recognition, biometrics, speech recognition [18]. The advantages of SVM are the training simplicity, high interoperability and, it works well with low samples and high dimensional input space. Experimental results have shown that this method often outperforms competing clustering methods [18]-[19]-[20].

SVM classifies data by finding the best hyperplane that separates all data points into two classes. The training data is a set of data points $x_i \in R^d$ along with their categories $y_i = \pm 1$,

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

where each value corresponds to a data class. The equation of a hyper-plane is:

$$g(x) = W^T X + b \tag{1}$$

Where g(x) is a linear function, $W \in R^d$ and b is a real number. A positive or negative value of g(x) means that the data will have y_i=+1 or y_i=-1 respectively.

In cases where it is difficult to find a simple hyperplane to separating data, we simply increase the data dimension by a kernel. The commonly-used kernel functions are Linear, Polynomial, Gaussian and Sigmoid [18]-[19]-[20].

IV. The Proposed Method

Figure 3 shows the block diagram of the proposed method. This algorithm consists of three steps including: random selection of video frames, object and human recognition and understanding human action.

1) Random selection of video frame

At first we have to find all the video objects. We can use one video frame instead all the frame to find the video objects. However, we may lose video objects in situations where we may face with occlusion or fast video changes. Figure 4 presents such an example in a diving video. Overcoming these problems, we divide a video into three parts and, randomly select a frame of each part as the input of CNN to detect the video objects. The advantages of this random selection are the improvement of computation complexity and reducing the memory requirement.

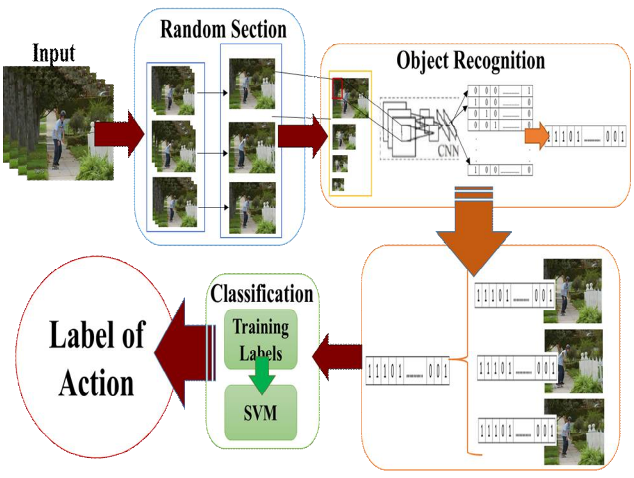
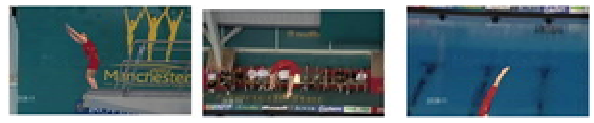


Fig. 3. The block diagram of the proposed system including random video frame selection, object recognition by CNN and action recognition according to objects of the frame by SVM.



A: The first frame consist of human and platform	B: The thirty-fourth frames shows only human	C: The Fifty-first frame consist of part of human and water

Fig. 4. Three frames of diving movie; fast changes of movement results in different objects in each frame.

2) Object and human recognition

This section employs a pre-trained CNN with ImageNet [21] to recognize all video objects. This CNN processes 400 images with 16x16 dimensions per-second and it detects one object in each process per image. The results shows that CNN classifies objects more accurately than along the line algorithms [16]-[22]-[23]-[25].

Normally an image consist of different objects with different sizes. Here, we use two methods to find image object; partitioning image into small sizes and image resizing. Partitioning an image to small parts gives us the chance to find an object regardless of its location in the image. Image resizing provides the opportunity to find small and large object and.

An image is partitioned into 16x16 sub-images where each part overlaps with its neighbor with an 8x16 block as it is shown in Figure 5. CNN finds a specific object in a 16 * 16 sub-image. The final result of CNN for each sub-image is a vector. The vector size is equal to the total number of objects. Each element of the vector corresponds to an object and the existence of this element will be shown by one the vector otherwise by zero (See Fig. 5).

The size of objects are different (some of them are big such as horse and some of them are small such as shoes). For a better representation, we convert the original image into 4 different sizes in a pyramid form. In this case, big and small objects will be detected at small and large size version of the image respectively. Figure 6 presents an example where an image with original size of 128x128 is changed into 64x64, 32x32, and 16x16. The tree is a small object and this can be detected in 128x128 size image and the horse is a large object and it can be detected in 16x16 size one.

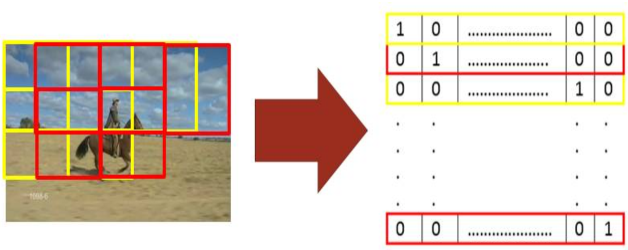


Fig. 5. Input and output of CNN.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig. 6. Changing the size of the original image for CNN.

In this example, the output of CNN will be 337 vectors for one frame by adding the vectors of different image sizes from 12x128 to 16x16. The output vectors of different size of an image are combined into one vector (See Fig. 7).

As we select three frames of three parts of a video the above process is performed for all of the selected frames and we have a final vector for each of the three frame and finally, the three vectors are converted into one vector by union function. Figure 8 shows this process

3) Understanding the relationship between objects of an image

Up to now, the objects and human has been recognized. For example if the input video is lifting, the output of the second stage is a vector that includes the objects of video such as human, shoes, barbells or other heavy weights and etc. In this step, we used support vector machine with RBF kernel.

First the classification is trained with 70 samples. After that each vector of the frame is mapped onto the correct class and SVM determines the label of interactions between human and objects.

V. Experimental Results

This paper employs a combination of CNN and SVM to understand human action. CNN has been employed to extractthe image objects and SVM to understand the action according to the relation of image objects. Therefore, in experiments, we have to show that a combination of these algorithms will improve the understanding accuracy.

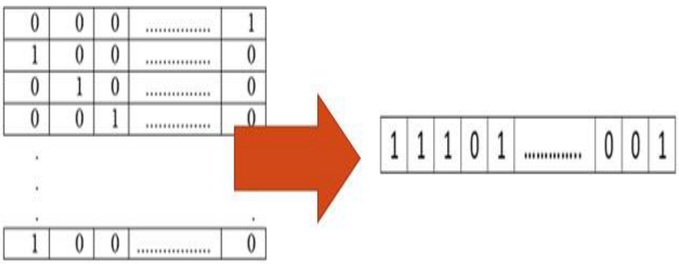


Fig. 7. All the output vectors of different image sizes of a frame generated by CNN are combined into one vector for each frame.

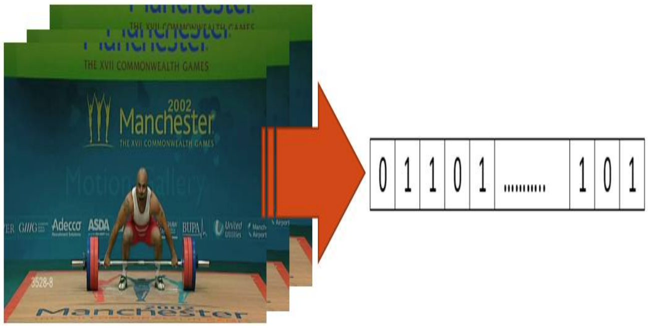


Fig. 8. The final output for a video is a vector which is generated by combination of the vectors of three frames of that video.

This paper performs experiments to evaluate the accuracy of the proposed algorithm. The evaluation shows the final output with the reference video and compares the final results of the proposed method with the algorithms along the same line based on ground-truth videos. All the experiments were implemented with Matlab 2015 under OS Windows 8 and were run on an ASUS computer with CPU 2.0 GHz with 8 GB RAM. Next, the data set of images for evaluation, the evaluation criteria and the final outcomes are explained in detail.

A. Datasets

We have evaluated the proposed algorithm on the known benchmark dataset the Olympic and UCF sports action. These datasets consist of athletes practicing different sports and human use the objects (except in some videos like running, jumping and walking). So, this datasets are good for discovering human-objects interaction.

Fig. 9. The sample of UCF sports action datasets.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II.

PRESENTS THE RESULTS OF METHODS ON UCF SPORTS ACTION

First author	Algorithm	Accuracy
Wang, 2009 [27]	He tested various detector on UCF sports and best result is obtained by the dens sampling.	85.6
Le, 2011 [28]	Design a way to learn invariant spatio-temprol directly from video data. He consider dense sampling for his algorithm.	86.5
Kovashka, 2010 [29]	Learn the shapes of space-time feature neighborhoods.	87.2
Ravanbakhsh, 2015 [30]	compute CNN features for each frame and map them into a short binary code space	88.1

1) UCF Sports Action

The UCF Sport Action datasets [7]-[25]-[26] consists of a set of sports activities which is collected from TV channels such as BBC and NSPN in 2008. This type of video datasets generally contains background clutter, changes in object appearance, scale, illumination conditions, and viewpoint. It has been used for applications such as: action recognition, action localization, and saliency detection [2]-[6]-[7]. This datasets includes 150 videos from 10 different sports, the running and walking are considered as one sport: Diving (14 videos), Golf Swing (18 videos), kicking a ball (20 videos), lifting (6 videos), horse riding (12 videos), running (35 videos which 22 videos is for walking) skateboard (12 videos), swing-bench (20 videos) and swing-side (13 videos). Figure 9 shows the sample of all 10 actions.

2) Olympic Sport datasets

Olympic Sports dataset [31] includes videos of various sports practicing and it was created in 2010. This data set is obtained from YouTube video.

Fig. 10. The example of Olympic sports datasets.

TABLE III

PRESENTS THE RESULTS OF METHODS ON OLYMPIC SPOTRS

First author	Algorithm	Accuracy
Jain, 2013[33]	Cluster dense trajectories. Decompose visual motion into dominant and residual motions both for extracting trajectories and computing descriptors.	83.2
Li, 2013 [34]	Recognize different camera motion types Separate foreground and background motion.	84.5
Wang,2013 [35]	Multiple descriptors. Densely sample feature points in each frame. Track features in the video based on optical flow.	84.9
Gaidon, 2014 [36]	Compute the approximate motion of the background plane. Generate stabilized videos. Extracting dense trajectories.	85.5
Wang, 2015 [37]	Use camera motion estimation for dense trajectory features. Using Multiple descriptors and dense optical flow.	85.8

Also it has high intra-class similarity between videos for certain classes. It includes 50 videos from 16 different sports: high-jump, long-jump, triple-jump, pole-vault, discus, hammer, javelin, shot put, basketball, bowling, tennis-server, platform diving, springboard, snatch, clean-jerk and vault [7]-[32]. In this paper, because we focused on human-objects interactions, 12 sports are used. The high, long and triple-jump are considered as a jump class. The platform diving and springboard are considered as a Dive class. The snatch and clean-jerk are considered as a snatch class. Figure 10 presents the example of all 12 sports.

B. Evaluation criteria and the results

The evaluation criteria is accuracy and it can be calculated by finding four factors as explained next. The confusion matrix presents all the factors.

1. True Positive (TP): Number of positive cases correctly identified.
2. False Positive (FP): Number of negative cases incorrectly classified as positive.
3. False Negative (FN): Number of positive cases incorrectly classified as negative.
4. True Negative (TN): number of negative cases correctly identified.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

	Diving	Golf-swinging	Kicking	Swinging on High Bar	Riding	Running	Swinging on Bench	Lifting	Skateboarding	Total	Recall
Diving	100	0	0	0	0	0	0	0	0	100	100%
Golf- swinging	0	89	2	0	0	4	0	0	5	100	89%
Kicking	0	5	93	0	0	2	0	0	0	100	93%
Swinging on High Bar	5	0	0	89	0	1	5	0	0	100	89%
Riding	0	0	0	1	97	2	0	0	0	100	97%
Running	0	2	9	0	0	88	0	0	1	100	88%
Swinging on Bench	0	0	0	0	0	0	100	0	0	100	100%
Lifting	0	0	0	0	0	0	0	100	0	100	100%
Skateboarding	0	0	4	0	14	0	0	0	82	100	82%
Total	108	83	116	84	111	102	108	100	88	900	
Precision	95.24%	92.71%	86.11%	98.89%	87.39%	90.72%	95.23%	100%	93.18%		

(a)

	Basketball Layup	Discus Throw	Hammer Throw	Shot Put	Bowling	Jump	Tennis Serve	Javelin Throw	Vault	Pole Vault	Dive	snatch	Total	Recall
Basketball Layup	100	0	0	0	0	0	0	0	0	0	0	0	100	100%
Discus Throw	0	73	6	9	0	2	0	4	4	2	0	0	100	73%
Hammer Throw	0	11	70	10	0	3	0	3	3	0	0	0	100	70%
Shot Put	0	6	8	70	0	1	0	5	6	2	0	0	100	70%
Bowling	2	0	0	3	97	0	0	0	0	0	0	0	100	97%
Jump	0	2	1	3	0	87	0	0	2	0	5	0	100	87%
Tennis Serve	0	0	0	0	0	0	100	0	0	0	0	0	100	100%
Javelin Throw	0	6	6	5	0	4	0	77	2	0	0	0	100	77%
Vault	0	3	1	1	0	1	0	3	89	2	0	0	100	89%
Pole Vault	0	0	0	0	0	6	0	0	4	88	2	0	100	88%
Dive	0	0	0	0	0	0	0	0	0	0	100	0	100	100%
snatch	0	0	0	0	0	0	0	0	0	0	0	100	100	100%
Total	102	101	92	101	97	104	100	92	110	94	107	100	1200	
Precision	98.04%	72.28%	76.09%	69.31%	100%	83.65%	100%	83.70%	80.91%	93.62%	93.46%	100%		

(b)

Fig. 11. (a) is a confusion matrix for UCF Sport Action. (b) is a confusion matrix of Olympic Sport. Closely related activities such as Hammer Throw and Shot Put (in UCF Sport) or running and kicking (in Olympic Sport) tend to be more easily confused.

The accuracy is:

(TP+TN)/ (TP+TN+FP+FN)

Table. 4 presents the results of proposed methods and some previous work on these two datasets. It can be seen that the proposed algorithm outperformed on these datasets. As mentioned, our system worked like object recognition system and detected all objects in a video. We show that learning features directly from data is an important research direction and it is also powerful in recognition tasks. Three frames, instead of all frames, have been used and caused the cost of computing and memory requirement are decreased. Temporal features are omitted and human's motion won't be tracked in following frames. So, the performance and accuracy are increased.

TABLE IV

COMPARING OF RECOGNITION ACCURACY ON UCF AND OLYMPIC SPORTS

UCF Sport Action		UCF Sport Action	
Methods	Accuracy	Methods	Accuracy
Souly, 2015 [38]	85.1	Niebles, 2010 [31]	72.1
Wang, 2009 [27]	85.6	Jain, 2013 [33]	83.2
Le, 2011 [28]	86.5	Li, 2013 [34]	84.5
Kovashka, 2010 [29]	87.2	Wang, 2013 [35]	84.9
Ravanbakhsh, 2015 [30]	88.1	Gaidon, 2014 [36]	85.0
Wang, 2011 [39]	89.1	Wang, 2015 [37]	85.8
Weinzaepfel, 2015 [40]	90.5	Our model	87.5
Our model	93.1		

Figure 11 shows the average performance of our method over 100 runs for UCF Sport Action and Olympic Sport. In some cases, because of the apparent similarity of the environment or objects in the videos, the videos of different categories in the layer of classification, are confused with each other.

VI. Conclusion

This paper presented an algorithm for human action recognition that suitable in situations where we have human-object tntractions in videos. This method collects UCF Sport Action and Olympic Sport videos have several kind of sports that human use the objects, and then we employ a supervised learning algorithm where the underlying network trained by image-net data to train the system for objects recognition. The employing of learning method, significantly improves the performance in vision tasks. The accuracy improvement comes from the using three frames instead of all video frames, the learning algorithm and the approch for objects recognition. Our approach does not require segmentation, tracking of humans, pruning of motion or static features, or stabilization of videos. The results show that, UCF Sport Action(87.5%) and Olympic Sport(93.1%), the proposed method provides better performance in comparison to other state-of-the-arts results.

References

[1] E.J. Amirbandi and G. Shamsipour, “Exploring methods and systems for vision based human activity recognition,” in 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC). pp. 160-164, 2016. IEEE

[2] R. Poppe, “A survey on vision-based human action recognition. Image and vision computing,” 2010. 28(6): pp. 976-990.

[3] J. Aggarwal, and L. Xia, “Human activity recognition from 3d data: A review.” Pattern Recognition Letters, 2014. 48: pp. 70-80.

[4] S. Ke, R., et al., "A review on video-based human activity recognition. Computers," 2013. 2(2): pp. 88-131.
[5] T.B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis." Computer vision and image understanding, 2006. 104(2): pp. 90-126.
[6] S. Vishwakarma, and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance." The Visual Computer, 2013. 29(10): pp. 983-1009.
[7] J.M. Chaquet, E.J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition." Computer Vision and Image Understanding, 2013. 117(6): pp. 633-659.
[8] K. Grauman, and B. Leibe, "Visual object recognition." Morgan & Claypool Publishers, Synthesis lectures on artificial intelligence and machine learning 5, no. 2, pp.1-181, Apr. 2011.
[9] B.T. Morris, and M.M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance." Circuits and Systems for Video Technology, IEEE Transactions on, 2008. 18(8): pp. 1114-1127.
[10] M. Fiaz, and B. Ijaz. "Vision based human activity tracking using artificial neural networks." in Intelligent and Advanced Systems (ICIAS), 2010 International Conference on. 2010. IEEE.
[11] 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, 2013. 35(1): pp. 221-231.
[12] K. Simonyan, and A. Zisserman. "Two-stream convolutional networks for action recognition in videos." in Advances in Neural Information Processing Systems, pp. 568-576. 2014.
[13] L. Wang, Y. Qiao, and X. Tang. "Action recognition with trajectory-pooled deep-convolutional descriptors." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4305-4314. 2015.
[14] A. Krizhevsky, I. Sutskever, and G.E. Hinton. "Imagenet classification with deep convolutional neural networks." in Advances in neural information processing systems, pp. 1097-1105. 2012.
[15] J. Schmidhuber, "Deep learning in neural networks: An overview. Neural Networks," 61: pp. 85-117. 2015.
[16] R. Girshick, J. Donahue, T. Darrell, and Malik, J. "Rich feature hierarchies for accurate object detection and semantic segmentation." in Computer Vision and Pattern Recognition (CVPR), pp. 580-587. 2014 IEEE.
[17] V. Vapnik, "The nature of statistical learning theory (Book style)." 2013: Springer Science & Business Media."
[18] D. Meyer, and F.T. Wien, "Support vector machines." The Interface to libsvm in package e1071, 2015 Aug 5.
[19] M.P. Brown, P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M., and Haussler, D "Support vector machine classification of microarray gene expression data." University of California, Santa Cruz, Technical Report UCSC-CRL-99-09, 1999.
[20] S. Abe, "Support vector machines for pattern classification (Book style)." Vol. 53. 2005: Springer.
[21] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets." arXiv preprint arXiv:1405.3531, 2014.
[22] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.
[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9. 2015
[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229, 2013.
[25] K. Soomro, and A.R. Zamir, "Action recognition in realistic sports videos," in Computer Vision in Sports. 2014, Springer. pp. 181-208.
[26] M. Rodriguez, "Spatio-temporal maximum average correlation height templates in action recognition and video summarization." 2010.
[27] 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, pp. 124-1. BMVA Press, 2009.
[28] Q.V. Le, Zou. Will Y, Yeung Serena Y, and Ng. Andrew Y. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis." in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3361-3368. IEEE, 2011.
[29] 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. IEEE.
[30] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis. "Action Recognition with Image Based CNN Features." arXiv preprint arXiv:1512.03980, 2015.
[31] J.C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in Computer Vision–ECCV 2010. 2010, Springer. pp. 392-405.
[32] Ibrahim, M., S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition." arXiv preprint arXiv:1511.06040, 2015.
[33] M. Jain, H. Jégou, and P. Bouthemy. "Better exploiting motion for better action recognition." in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference, pp. 2555-2562. 2013.
[34] W. Li, et al. "Dynamic pooling for complex event recognition." in Computer Vision (ICCV), 2013 IEEE International Conference, pp. 2728-2735. 2013.
[35] L. Wang, Y. Qiao, and X. Tang. "Mining motion atoms and phrases for complex action recognition." in Computer Vision (ICCV), 2013 IEEE International Conference, pp. 2680-2687. 2013.
[36] A. Gaidon, Z. Harchaoui, and C. Schmid, "Activity representation with motion hierarchies." International journal of computer vision, 2014. 107(3): pp. 219-238.
[37] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition." arXiv preprint arXiv:1504.05524, pp.1-20, 2015.
[38] N. Souly, and M. Shah, "Visual Saliency Detection Using Group Lasso Regularization in Videos of Natural Scenes." International Journal of Computer Vision, pp. 1-18, 2015.
[39] 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, pp. 3169-3176. IEEE, 2011.
[40] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. "Learning to track for spatio-temporal action localization." in Proceedings of the IEEE International Conference on Computer Vision, pp. 3164-3172. 2015.