ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

A Comparative Analysis of Various Representations of Human Action Recognition in a Video

Akila.K [1], Chitrakala.S [2]

Assistant Professor, Dept. of CSE, R.M.K College of Engineering & Technology, Chennai, India [1]

Associate Professor, Dept of CSE, Anna University, Chennai, India [2]

ABSTRACT: The action recognition is an automated analysis of ongoing events and their context from video data. The process of recognizing and understanding of human actions from videos still remains a challenging problem due to the large variations in human appearance, posture and body size within the same class; hence a pattern recognizer needs to be built into the video system. Robust solutions to this problem have applications in domains such as visual surveillance, video retrieval and human–computer interaction. Various features are used to recognize humans and they are categorized into four groups based on the visual representations such as spatio-temporal based visual representation, shape or pose based visual representation, interest point based visual representation and motion or optical flow based visual representation. The performance of these works is analyzed efficiently with the evaluation metrics such as Precision, Recall, F-Measure and Accuracy. It is tested in both controlled and uncontrolled environments. This paper helps to develop novel technologies over the biometric recognition system.

Keywords: Recognition, Spatio-temporal templates, Visual Representation, Optical flow, Interest points.

I. OVERVIEW OF HUMAN ACTION RECOGNITION SYSTEM

Human action recognition in video sequences has become an important research topic in computer vision, whose aim is to make machines to recognize human actions using different types of information, especially the motion information, in the video sequences. This research field has captured the attention of several computer science communities due to its strength in providing personalized support for many different applications and its connection to many different fields of study such as medicine, human-computer interaction, or sociology. Three aspects for human activity recognition are addressed including core technology, human activity recognition systems, and applications from low-level to high-level representation. In the core technology, three critical processing stages are thoroughly discussed mainly: human object segmentation, feature extraction and representation, activity detection and classification algorithms. In the human activity recognition systems, three main types are mentioned, including single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition. Finally the domains of applications include surveillance environments, entertainment environments and healthcare systems.

Detection of human activity is robustly emerging as an investigational tool pushed ahead by the growing needs of the budding industries including indexing of professional and amateur video archives, automatic video scrutiny, and man-machine interface. Many a system is in immediate need of action detection to perform consistently in various and pragmatic video settings. This section offers a vivid and wide account of the common procedures relating to a human action detection method system. Action Recognition is a method of detecting actions that happen in video series, here, in this case, by individuals. Fig.1 elaborates the various stages of the Activity Recognition method. There are altogether four procedures forming part of the action detection structure.

The human object is first segmented out from the video sequence. The characteristics of the human object such as shape, silhouette, colors, poses, and body motions are then properly extracted and represented by a set of features. Subsequently, an activity detection or classification algorithm is applied on the extracted features to recognize the various human activities. Datasets usually consist of several human actions extensively prevalent throughout length and breadth of the practical world. Making effective use of this enables mining out the forefront from the video run. Thereafter the tracking procedure is carried out in an efficient manner. Subsequently the actions are detected on the strength of every video sequence.

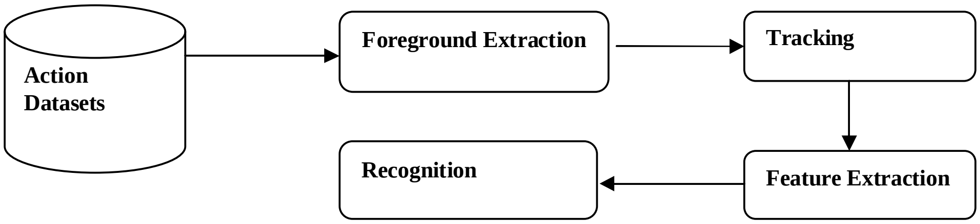

Fig. 1: General Block Diagram for the action recognition of humans

II. SURVEY OVER VARIOUS PAPERS OF HUMAN ACTION RECOGNITION SYSTEM

A survey to effectively furnish a general assessment of human action detection with a helping hand from the several investigational contributions in the literary world is done. These reports, in turn, are classified into 4 groups based on the visual illustrations: (i) Spatio-temporal based (ii) Shape or pose based (iii) Interest point based (iv) Motion based.

A. *Spatio-temporal based visual representation*

Based on hierarchical event descriptions, Juan C. SanMiguel and José M. Martínez [2] have presented an interesting method by using both of the semantic and probabilistic approaches for the real-time single-view event recognition in videos. To deal with uncertainty of the low-level analysis, they have developed a two-layer structure. Hierarchical Bayesian Network (BN) was utilized in short-term layer for recognizing timeless activities having changes in the features of objects and in long-term layer, probabilistically-extended Petri Net (PN) was utilized for finding activities with the temporal characteristics among their counterparts. Uncertainty was handled by the naïve extension of PN.

Using empirical covariance matrices of features Kai Guo [3] have proposed a generic construction for the accurate action recognition in videos. In order to facilitate the localized description of the action, a set of spatio-temporal feature vectors and to represent the action, they have aggregated in a covariance matrix. In their paper, they have included two supervised learning methods with the usage of feature covariance matrices for the recognition of action. First supervised learning method has used a suitable Riemannian metric with the nearest-neighbor classification. The logarithm of training covariance matrices was approximated by the second supervised learning method with a sparse linear combination of the logarithms of training covariance matrices. After that, the sparse co-efficients have been utilized in the process for the identification of action label.

The performance of Action modeling with bag of video words (BoVW) representation has limitations with the factors such as granularity, the size of vocabulary and the space for the clustering of features and words that has high sensitiveness. To deal with these issues, Jingen Liu [4] have developed a technique with the usage of Diffusion Maps embedding for semantic vocabulary learning from video words and for taking advantage of feature vocabularies representative of human actions.. The video words were specified by Pointwise Mutual Information (PMI) vector between that particular word and the training video clips, which are called the mid-level feature. The authors have obtained the semantic meaningful distance measure between the midlevel and highest level features by clustering the mid-level features and also found the relationship between them using diffusion distance measure.

Two feature extraction methods were developed by Georgios Goudelis [8] that examined the Trace transform's capability. At first, they have made an intuitive representation of the capability of Trace for making the differences in the active classes and then calculated the Weighted Trace Transforms (WTTs). First feature was extracted by the Trace transforms from the binarized silhouettes that represented different stages of a single action period using History Trace Templates (HTTs) method. At the final stage, history template only gave the whole sequence containing more spatio-temporal information about the human action. The second feature has involved Trace for the creation of a set of invariant features that specified sequence of actions using History Triple Features (HTFs) method. This method was able to make noise robust features that were invariant for the trace transformations such as translation, rotation, scaling.

A novel and hybrid feature based action recognition method was introduced by the authors Rashid Minhas [9]. Spatio-temporal features and local static features were combined as a hybrid feature which was useful to extract the features using motion-selectivity attribute of 3D dual-tree complex wavelet transform (3D DT-CWT) and affine SIFT local image detector. Visual vocabularies were made using these two features that were given as the input to the ELM. When compared with the classical neural networks, SVM and AdaBoost, the ELM has facilitated classification with high speed.

Different variety of techniques including relevance feedback approaches such as SVMs, ABRS-SVMs and Maximum of similarities have been developed by Simon Jone and Ling Shao [14] for the content based human action retrieval. They have also utilized various means of local feature extraction approaches, soft-assignment clustering, approaches for action specification such as included the Bag-of-words, vocabulary guided and spatio-temporal pyramid matches. The combinations of such works have facilitated the performance result of existing works.

A mined hierarchical spatially and temporally grouping of simple corners was introduced by Andrew Gilbert [17] in order to form complex discriminative compounds of simple 2D-harris corners. The usage of an over-complete feature set was allowed by the Data mining, from which the sparse complex compound features were learnt effectively.

In general, Convolutional Neural Networks (CNNs) are only handling the 2D raw inputs. To handle the 3D raw input also for the action recognition, Shuiwang Ji [19] have proposed a new 3D CNN model. The motion information that was encoded in multiple adjacent frames was captured by extracting the spatial and temporal features. They have also regularized the outputs with high-level features for boosting up the performance level. The results of proposed work was compared with the publishing methods, which showed the promising results in their proposed work with better performance and accuracy for the recognition.

A novel view-invariant action recognition method was proposed by the authors Anwaar-ul-Haq. [20] and they explored the invariance feature of temporal order of action instances. Spatiotemporal features were utilized for ensuring the temporal order during matching. Initially, they had extracted the spatiotemporal features from the video sequences and then fused the features for encapsulating within the class similarity value for the same viewpoints, in which the matching of features across various views were obtained.

B. *Shape or pose based visual representation*

The Recognition of human activities by the usage of small number of labeled data is the recently used application. One proposed framework was Labeled Kernel Sparse Coding (LKSC) and sparse L1 graph, which was introduced by Shuyuan Yang [7] under semi-supervised learning. They have extended the sparse representation classifier (SRC) to empirical kernel projection space. In a parameter-free way, they have made a sparse L_1 graph by computing the labeled as well as unlabeled samples of sparse coding co-efficients as the graph weights.

While taking video shots, same actions may create large intra-class variations in correspondence with visual appearance, kinetic patterns, video shooting and editing styles. Moreover, description of heterogeneous features may have another impact on how to deal with these features of certain problems such as redundancy, complementariness and disagreement. To address these tackles, a Localized Multiple Kernal Learning (L-MKL), which combined the localized classifier ensemble learning and multiple kernel learning classifier's to provide good strength on heterogeneous diverse feature representation, was proposed by the authors Yan Song [10]. A locality gating framework was generated by L-MKL for partitioning the input space of the heterogeneous feature representation into a set of localities of the simpler data structure. Then the recognition of actions were performed by co-ordinating the localized multiple kernel classifiers using gating framework.

Jing Wang and Zhijie Xu [12] have proposed a 3D shape-matching method based on Spatio Temporal Volume (STV) and Region Intersection (RI) for recognizing the actions in videos. The distinctive characteristics and the performance gain of the approach stemmed from a co-efficient factor-boosted 3D region intersection and matching mechanism developed in their proposed work. The approaches for the filtering of STV data have helped to decrease the volumetric pixel's amount that required to be processed in every operational cycle in the system.

Modeling every variation in each domain of videos was a challenging problem, which was solved by Nazli Ikizler-Cinbis and Stan Sclaroff [18] using a generic approach, which gathered the images from Web for learning the actions and annotating it. To construct the human pose classifier, at first they gathered the incremental image retrieval procedures. Their method had not intervened with the human beings, since it was an unsupervised one, which improved the action image retrieval. The temporal ordering of poses was encoded by their proposed “Ordered Pose Pairs” (OPP), which increased the recognition accuracy for the actions.

C. *Interest points based visual representation*

Basically action recognition methods with bags of space-time interest points rely on the discriminative power of individual local space-time descriptors and which avoid the global spatio-temporal distribution of interest point’s related information. Matteo Bregonzio [6] has considered this problem and has proposed a new approach to use the global spatio-temporal distribution information of interest points. In order to process different complementary information in the spatio-temporal distribution of interest point, they have included a Multiple Kernal Learning based feature fusion method.

D. *Motion or Optical flow based visual representation*

In action recognition field, view-invariant action recognition is one of the very popular recognition models. To solve the problems including interest point detection in the image sequence, extraction and description of motion features and computing view-invariant, Kaiqi Huang [1] have proposed a novel discriminative approach, in which view invariants and motion patterns were fused altogether for producing invariance and distinctiveness combination..

The motion field in an action region is spoiled by the background motions, which is the issue while working on the field of action recognition with the moving background. Due to the robustness of Motion History Image (MHI), YingLi Tian [5] has presented a Hierarchical Filtered Motion (HFM) using it as a basic motion representation for the recognition of actions. Initially, they have found the interest points with the positions of high intensities in the MHI as the 2D Harris corners. For the removal of noisy motions, the gradients of the MHI were subjected to a global spatial motion smoothing filter. Smoothed gradients of the MHI were subjected to a local motion field filter at each interest point, which calculated the structure proximity between any pixel in the local region and the interest point to enhance or weakened the motion at a pixel.

A new technique that combined approximate reasoning and sequential machines for the action recognition was introduced by L. Rodriguez-Benitez [11], in which the input data was taken from the segmentation and tracking algorithm. The input data was subjected to fuzzification process. This was allowed the information of low-level and medium-level vision tasks to be managed in order to consolidate all the information that were taken from various vision algorithms into a homogeneous description. The actions were represented by means of the automaton.

By calculating the similarity in motion with the help of compressed domain features, Chuohao Yeo [13] have proposed a unique method in their paper for recognizing and localizing the actions at high speeds, in which the domain features could be extracted with low complexity. The similarity in motion was computed by taking the differences in the motion directions and magnitudes. Their proposed approach was appearance-invariant which needed no prior segmentation and was able to localize the events in both space and time.

A Boosting EigenActions algorithm was proposed by Chang Liu and Pong C. Yuen [15] for the recognition of actions in human beings. Using pixel density function, a spatio-temporal Information Saliency Map (ISM) was computed from a video sequence. The sequence of human actions was segmented from the curve of ISM into a set of periodic motion cycles. Every motion cycle was specified by a Salient Action Unit (SAU), which was exploited to compute the EigenAction using principle component analysis. By using multi-class Adaboost algorithm with Bayesian hypothesis, they made a human action classifier as a weak classifier.

A framework to extract a set of kinematic from the optimal flow was proposed by Saad Ali and Mubarak Shah [16] for the effective recognition of human actions. Divergence, symmetric and anti-symmetric flow fields, principal invariants of flow gradient and rate of strain tensor, and third principal invariant of rate of rotation tensor were taken as the kinematic features. Multiple Instance Learning (MIL) was exploited for the every video was embedded into a kinematic-mode-based feature space and the video coordinates in that space were utilized for the classification using the nearest neighbour algorithm. Their proposed work was very robust for facilitating very good quality improvement in the optical flow.

III. RESULTS AND DISCUSSIONS

Action Recognition in human beings has been evaluated using various datasets for the video scenes in our survey papers. Each dataset has facilitated different activities of human beings, which have been evaluated with the help of every proposed method in the survey papers.. The experimentations of these review papers are carried out in every paper with standard datasets. The datasets which are used in every survey papers are given in the following table I. The experimentations have been performed on both the controlled and uncontrolled environments. It helps us to analyze the results of every method with detailed evaluation.

TABLE I
VARIOUS DATASETS USED IN THE SURVEY PAPERS

Methods Used	Various Datasets Used								
	Wiezmann	**KTH**	**YouTube**	**UCF You tube**	**Hollywood1**	**Hollywood2**	**IXMAS**	**MSR**	**UCF Sports**
A Discriminative Model of Motion and Cross Ratio	✓	✓							
Semantic & Probabilistic approaches									
empirical covariance matrices of features	✓	✓	✓						
Diffusion Maps embedding for semantic vocabulary learning		✓		✓					
Hierarchical Filtered Motion		✓						✓	
global spatio-temporal distribution of interest points	✓	✓							
Labeled Kernel Sparse Coding & sparse L1 graph	✓								
Two feature extraction methods for Exploring trace transform	✓	✓							
3D dual-tree complex wavelet transform & affine SIFT	✓	✓							
Localized Multiple Kernal Learning (L-MKL)			✓			✓			
reasoning and sequential machines									
Spatio Temporal Volume & Region Intersection		✓							
compressed domain features									
RF algorithm				✓		✓			✓
Eigen Actions algorithm	✓	✓							
kinematic features from the optimal flow	✓	✓							
mined hierarchical spatially and temporally grouping of corners		✓			✓	✓			
Ordered Pose Pairs (OPP)			✓	✓					
3D CNN model		✓							
invariance feature of temporal order of action instances							✓		

In general, the metrics such as Precision, Recall and Accuracy of recognition are used as the evaluation metrics in all recognition systems. They are given in detail one by one with the comparison.

According to the values of metrics in the survey papers, the results are analyzed and discussed as follows. Following table II shows the precision values that are obtained with every survey work.

TABLE II
PRECISION VALUES OBTAINED FOR ACTION RECOGNITION IN SURVEY PAPERS

Methods	**3D-CNN** **[19]**	**PbHOG** **[18]**	**Hierarchical Mined** **[17]**	**PWRC+RI+CF** **[12]**	**L-MKL** **[10]**	**Hierarchical Filtered Motion** **[5]**	**Semantic based probabilistic** **[2]**
Precision values (%)	61.49	83	95.7	70	43.14	60	77

According to the table II, we can generate a graph for the precision value as given in the following fig. 2.

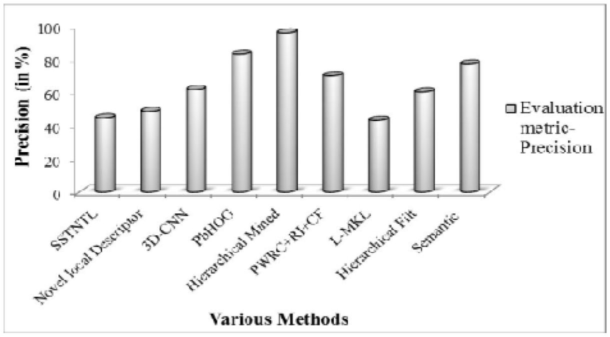

Fig. 2: Precision values for various existing works for human action recognition

It is observe from the table II that the shape or pose based visual representation methods and the spatio-temporal based visual representation methods facilitate good performance results.

TABLE III:
RECALL VALUES OBTAINED FOR ACTION RECOGNITION IN SURVEY PAPERS

Methods	**3D-CNN** **[19]**	**PWRC+RI+CF** **[12]**	**Hierarchical Filtered Motion** **[5]**	**Semantic based probabilistic** **[2]**
Recall values (%)	14.33	60	48	86

Using the data furnished in table III, we may draw a graph for the precision value as per fig. 3.shown below.

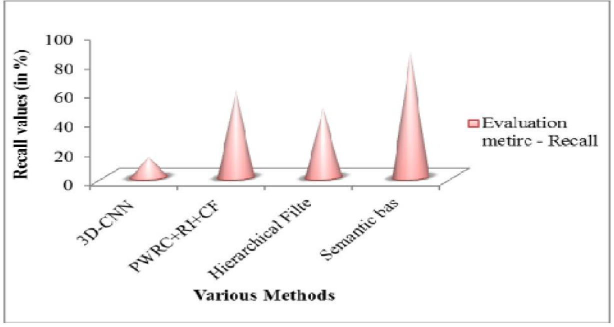

Fig. 3: Recall values for various existing works for human action recognition

The next evaluation metric Recall is evaluated using the various existing methods in various survey papers, which is shown in the above table III and the graphical representation is shown in fig. 3. Here also we observe that spatio-temporal based and shape or pose based visual representation methods give higher recognition results.
Table IV and fig. 4 illustrate the accuracy recognition values for some of the review papers from our survey work.

TABLE IV
ACCURACY VALUES OBTAINED FOR ACTION RECOGNITION IN SURVEY PAPERS

Methods	**wposes+ vposes +OPP [18]**	**3D-CNN [19]**	**Hierarchical Mined [17]**	**PWRC+ RI+CF [12]**	**L-MKL [10]**	**Hierarchical Filtered Motion [5]**	**Spatio-tmp based view invariant [20]**	**Kinematic feature [16]**	**BoCP+Extended-MHI [2]**	**RF Alg [14]**
Accuracy values (%)	52	90.2	54	82	77.91	93.6	89.2	87.7	90.3	96

In conformity with the table IV, we can generate a graph for the precision value furnished in fig. 4 given below.

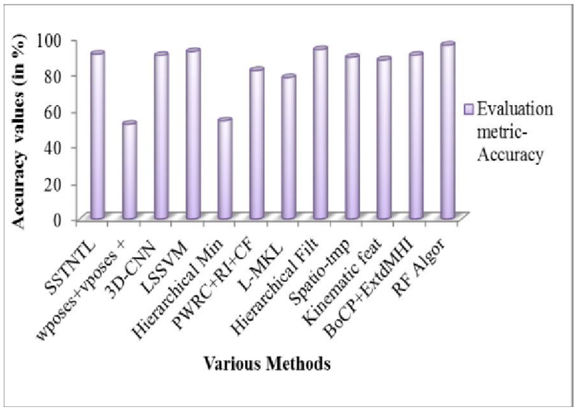

Fig. 4: Accuracy values for various existing works for human action recognition

The RF algorithm only provides high accuracy value among all these methods Then Hierarchical Filtered Motion, LSSVM, SSTNTL, BoCP+Extended-MHI and 3D-CNN provides next high accuracy values. Wposes+vposes+OPP and Hierarchical Mined only facilitate below-par performance result.

IV. OPEN ISSUES IN HUMAN ACTION RECOGNITION SYSTEM

Many issues are still open and deserve further research in Human recognition system.

- For real-world deployment, action recognition systems need to be tested against such real-world conditions include noise, occlusions, shadows, etc. Invariability features – such as viewpoints- front-view or side-view; execution rate - differences in execution rates while performing the same action and anthropometry- those induced by the size, shape, gender, etc.

- Biometric identification is combined with human action detection system for security-sensitive applications. Human gait as a new biometric feature can be employed for personal recognition.

- To pursuit intelligence, increasing the robustness of action/activity detection modules, integration with other modalities such as audio, temperature, motion, and inertial sensors needs to be investigated in a more systematic manner.
- In past decades a systemic-structural analysis of a human activity has been explored, currently to infer what is going to happen, reasoning about the intentions of humans presents a significant intellectual challenge.

    - Generating descriptive sentences from images (or) videos is a further challenge

V. APPLICATIONS

Human action Recognition finds its application in the various fields.

- Surveillance system
    - In homeland security
    - Crime prevention
    - Traffic surveillance
    - Military target detection
    - Entertainment System
- Human-computer interaction
    - Extracting statistics for sport
- Health care systems
    - Patient rehabilitation processes
    - Elderly behaviour monitoring
    - Home abnormal activity

VI. CONCLUSION

Recognition is one of the popular tasks in image processing, while the recognition of human action is a critical one. In this paper, we have surveyed the topic of human action recognition which is categorized into four groups based on the visual representations. The performance of the human action recognition works has been extensively analyzed with the evaluation metrics – Precision, Recall, F-Measure and Accuracy. From this evaluation, we observe that some of the methods have provided efficient performance result. Even though the performance of these methods has yielded good quality results, today’s technology has been developing day by day, which calls for further improvement in the performance.

REFERENCES

1. Kaiqi Huang, Yeying Zhang and Tieniu Tan, "A Discriminative Model of Motion and Cross Ratio for View-Invariant Action Recognition", IEEE Transactions on Image Processing, Vol. 21, No. 4, pp. 2187-2197, April 2012.
2. Juan C. SanMiguel and José M. Martínez, "A semantic-based probabilistic approach for real-time video event recognition", Computer Vision and Image Understanding, Vol. 116, pp. 937–952, 2012.
3. Kai Guo, Prakash Ishwar and Janusz Konrad, "Action Recognition from Video Using Feature Covariance Matrices", IEEE Transactions on Image Processing, Vol. 22, No. 6, pp.2479-2494, June 2013.
4. Jingen Liu, Yang Yang, Imran Saleemi and Mubarak Shah, "Learning semantic features for action recognition via diffusion maps", Computer Vision and Image Understanding, Vol. 116, pp. 361–377, 2012.
5. YingLi Tian, Liangliang Cao,Zicheng Liu and Zhengyou Zhang, "Hierarchical Filtered Motion for Action Recognition in Crowded Videos", IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews, Vol. 42, No. 3, pp. 313-323, May 2012.
6. Matteo Bregonzio, Tao Xiang and Shaogang Gong, "Fusing appearance and distribution information of interest points for action recognition", Pattern Recognition, Vol. 45, pp. 1220–1234, 2012.
7. Shuyuan Yang, Xiuxiu Wang, Lixia Yang, Yue Han, and Licheng Jiao, "Semi-supervised action recognition in video via Labeled Kernel Sparse Coding and sparse L1 graph", Pattern Recognition Letters, Vol. 33, pp. 1951–1956, 2012.

8. Georgios Goudelis, Konstantinos Karpouzis, and Stefanos Kollias, "Exploring trace transform for robust human action recognition", Pattern Recognition, Vol. , pp. 2013.

9. Rashid Minhas, Aryaz Baradarani, Sepideh Seifzadeh and Q.M. Jonathan Wu, "Human action recognition using extreme learning machine based on visual vocabularies", Neurocomputing, Vol. 73, pp. 1906–1917, 2010.

10. Yan Song, Yan-Tao Zheng, Sheng Tang, Xiangdong Zhou, Yongdong Zhang, Shouxun Lin, and Tat-Seng Chua, "Localized Multiple Kernel Learning for Realistic Human Action Recognition in Videos", IEEE Transactions on Circuits And Systems For Video Technology, Vol. 21, No. 9, pp. 1193-1202, September 2011.

11. L. Rodriguez-Benitez, C. Solana-Cipres, J. Moreno-Garcia and L. Jimenez-Linares, "Approximate reasoning and finite state machines to the detection of actions in video sequences", International Journal of Approximate Reasoning, Vol. 52, pp. 526–540, 2011.

12. Jing Wang and Zhijie Xu, "STV-based video feature processing for action recognition", Signal Processing, Vol. 93, pp. 2151–2168, 2013.

13. Chuohao Yeo, Parvez Ahammad, Kannan Ramchandran and S. Shankar Sastry, "High-Speed Action Recognition and Localization in Compressed Domain Videos", IEEE Transactions on Circuits and Systems For Video Technology, Vol. 18, No. 8, pp. 1006-1015, August 2008.

14. Simon Jone and Ling Shao, "Content-based retrieval of human actions from realistic video databases", Information Sciences, Vol. 236, pp. 56–65, 2013.

15. Chang Liu and Pong C. Yuen, "Human action recognition using boosted Eigen Actions", Image and vision computing, vol. 28, pp. 825-835, 2010.

16. Saad Ali and Mubarak Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning", IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 32, No. 2, pp. 288-303, February 2010.

17. Andrew Gilbert, John Illingworth and Richard Bowden, "Action Recognition Using Mined Hierarchical Compound Features", IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 33, No. 5, pp. 883-897, May 2011.

18. Nazli Ikizler-Cinbis and Stan Sclaroff, "Web-Based Classifiers for Human Action Recognition", IEEE Transactions on Multimedia, Vol. 14, No. 4, pp. 1031-1045, August 2012.

19. Shuiwang Ji, Wei Xu, Ming Yang and Kai Yu, "3D Convolution Neural Networks for Human Action Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 1, pp. 221-231, January 2013.

20. Anwaar-ul-Haq, Iqbal Gondal and Manzur Murshed, "On Temporal Order Invariance for View-Invariant Action Recognition", IEEE Transactions On Circuits And Systems For Video Technology, Vol. 23, No. 2, pp. 203-211, February 2013.