International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 08, August-2015

Study on Recent Approaches for Human Action Recognition in Real Time

R. Rajitha Jasmine,
Assistant Professor, IT Dept,
RMK Engineering College,
Chennai, India.

Dr. K. K. Thyagharajan,
Professor, ECE Dept,
RMD Engineering College,
Chennai, India

***Abstract*—The important area in computer vision is human understanding and recognizing actions. The main aim of action recognition is an automatic analysis of various actions from video data. Major video analysis system includes action detection and classification, action tracking, recognizing actions and behavior understanding. Even though, traditional methods have achieved greater success on several human actions. But, still it is a challenging problem to recognize human action. The challenge is to recognize human actions with more accuracy and efficiency in recognition time. The action recognition application includes CCTV, video indexing, patient monitoring systems and HCI systems. In this paper, we focus our attention to various modern approaches to human action recognition in real time.**

Keywords—Sparse tree, Ratio histogram, Code book, Hidden Markov model, Tree based codebook, bag-of-features.

I. INTRODUCTION

Human action recognition is the process of recognizing similar actions from video data. An action is defined as a sequence of human body movements, which consists of several body parts active simultaneously. In view of computer vision technology, the recognition of action is to compare the actual patterns from videos with previously defined patterns and a label is given which is called as action type. The activities of human can be divided into four parts: gestures, interactions, actions and group activities. Major automated systems include feature extraction, action learning and classification, and action recognition and segmentation. An action recognition process includes understanding the human and/or its body parts, tracking, and recognizing specific actions. Consider an example, to recognize a hand shaking activities, two person's arms and hands are first detected and tracked to generate a spatial-temporal description of their movement. This description is compared with existing patterns in the training data to determine the action type. Human actions analysis is equivalent to dealing with sequence of video frames that contain both spatial and temporal information. The recognition time and encoding spatial and temporal information are the challenging features in human action analysis. To solve this challenging problem some modern approaches build to represent human actions.

II. MODERN METHODS FOR HUMAN ACTION RECOGNITION

In the following Sections, we discuss the various challenging methods for action recognition

A. Matching Mixtures Of Curves [2]

The challenging task in the research area is nothing but classifying human actions. Several feature extraction methods are used for describing and recognizing human actions. Efros et al. [10] recognize human actions from low-resolution sports video sequences using the nearest neighbor classifier. The work of Wang et al. [11] is focused on tracking dense sample point from video sequences using optical flow. Leet al. [12] discovers the action label in an unsupervised manner by learning features directly from video data. A high-level representation of video sequences, called Action Bank, is presented by Sadanand and Corso [13]. Each video is represented as a set of action descriptors which are put in correspondence. The final classification is performed by a SVM classifier. Yan and Luo [14] have also proposed a new action descriptor based on spatial temporal interest points (STIP) [15]. In order to avoid over fitting they have also proposed a novel classification technique by combining the Adaboost and sparse representation algorithms. Wu et al. [16] employed, a visual feature using Gaussian mixture models efficiently represents the spatio temporal context distributions between the interest point at several space and time scales. An action is represented by a set of features extracted by the interest points over the video sequence. Finally, a vocabulary based approach has been proposed by Kovashka and Grauman [17]. The main idea was to find the neighboring features around the detected interest points quantize them and form a vocabulary. Raptis et al. [18] proposed a midlevel approach extracting that spatio-temporal features construct clusters of trajectories, which can be considered as candidates of an action, and a graphical model is utilized to control these clusters.

This method presents a learning based framework for action recognition and recognition relies on the description of an action by time series of optical flow motion features. Action recognition using this method consists of two steps: In the learning step, the motion curves representing each action are clustered using Gaussian mixture modeling (GMM). In the recognition step, the optical flow curves of a probe sequence are also clustered using a GMM, then each probe sequence is projected onto the training space and the probe curves are matched to the learned curves using a non metric similarity function. Longest Common Subsequence (LCS) finds the

similarity between the curves. Dynamic Time Warping [26] is used for measuring similarity between two sequences which may vary in time or speed. Alignment between the mean curves is performed using canonical time warping, which allows the spatial- temporal alignment between two human motion sequences. The action recognition accuracy is estimated as 98.3%. A perfect recognition performance is accomplished with a fixed number of Gaussian mixtures.

TABLE 1: Summary list on Matching Mixtures of Curves Model

Year	Author	Concept
2012	Sadanand and Corso	Action Bank, video is represented as a set of action descriptors
2010	Kovashka and Grauman	Vocabulary based approach
2011	X. Wu, D. Xu,L. Duan, J. Luo	A visual feature using Gaussian mixture models efficiently represents the spatio temporal context distributions between the interest point at several space and time scales. Difficult to recognize when more than one person are present in the scene
2012	Raptis et al	Proposed a midlevel approach extracting that spatio-temporal features construct clusters of trajectories

B. *Hierarchical Multi-Channel Hidden Semi Markov Graphical Models* [1]

Real time systems for modeling and recognizing daily activities can have a wide range of applications in surveillance, assistive technologies and intelligent environments. There are two key challenges faced by researchers in this area – (1) Developing robust low-level features for effectively capturing information from images and other sensory data (2) Developing appropriate models for bridging the gap between low-level features and high-level concepts while modeling errors and uncertainty. While several features have been proposed for the first challenge, the second challenge is focused and solved using probabilistic graphical models.

Generative models typically generalize hidden Markov models (HMMs) [19] and are especially useful in situations where we only have a small amount of unlabeled or partially labeled training data. Discriminative models on the other hand generalize conditional random fields (CRFs) and have shown strong performance in action recognition when large amounts of annotated training data is available [20,21]. HMMs and CRFs have been widely used in action recognition and other sequential data analysis applications due to their simplicity and well understood algorithms for learning and inference. The basic HMM/CRF representation suffers from three key limitations: The first-order Markov assumption causes the probability of staying in a state to decrease exponentially with time, which is unrealistic (e.g. a person can keep walking for an arbitrary amount of time). While activities occur at different levels of abstraction, there is no direct way to model this in an HMM/CRF. Use of single variable state representation makes it difficult to model activities involving multiple interacting agents. This method introduces a family of graphical models called hierarchical multi-channel semi-Markov graphical models (HMHSGMs) that simultaneously overcome these limitations. This allows mapping of logic based event definitions to probabilistic graphical models. Several hierarchical, multi-channel architectures for modeling multi agent actions and duration modeling using the semi-Markov model [1] as well as variable transition models [22] are considered. This method also provides an efficient learning and inference algorithm based on local variation approximations. The average action recognition accuracy is estimated as 97.73%

TABLE 2: Summary list on HM-SMGM Model

Year	Author	Concept
2004	H.Bui, D.Phung, S.Venkatesh	Hierarchical hidden Markov model for monitoring daily activities It is difficult to model activities involving multiple interacting agents
2006	M.Chan, A.Hoogs, Perera	Dynamic Bayesian network were used to simultaneously link broken trajectories and recognize complex actions
2011	M.Ryoo, J.Agarwal	Stochastic context free grammars were used for representing & recognizing group activities

C. *Structured Code Book Construction* [4]

Bag of words (BOW) model is widely used to obtain the global representation for action recognition. Primitive features assigned to the closest visual word in vocabulary. The each sample is represented by a feature vector that describes the histogram of words occurrences. The main disadvantage of standard BOW model is that it ignores the structure information such as Spatial & temporal contextual information. Wong et al propose an extension for probabilistic semantic analysis model, which captures both semantic & structural information for human action recognition. Ryoo and Aggarwal propose a spatio-temporal relationship matching strategy for human action recognition. Carso proposes a new high level representation of video to capture semantic information of features. V. Thanikachalam [27] proposes a system using Discrete Wavelet Transform (DWT) descriptors for representation and recognition of human action. We Zhou, Chunheng Wang proposes action recognition through structured code book construction to enrich the spatial and temporal contextual information. Sadanand and Corso [8] propose a new high-level representation of video: action bank, to capture semantic information of features. Zheng et al. [20] propose a co-occurrence matrices descriptor which captures temporal information for human action recognition. Shao et al. [23] propose a correlogram of body poses representation which takes advantage of both the probabilistic distribution and the temporal relationship of human poses for action recognition. Yuan et al. [24] propose a novel data mining method to discover two complementary co-occurrence patterns that are discriminative for visual recognition. They efficiently discover the optimal co- occurrence patterns with minimum empirical errors. The most related work is Zhu et al. [21]. They encode local spatio-temporal features within the sparse

coding framework. The frame work can be divided in to the following steps: first, each local spatial–temporal feature is transformed to a linear combination of a few "atoms" in a trained code-book; then, max pooling is operated on the whole sparse coefficients of local features to obtain the final video representation; since the "atoms" used in Zhuet-al. [21] are built based on local features, they code little spatial and temporal contextual information for video representation while contextual information can prompt the performance further.

This structured codebook construction method [4] used to encode rich spatial and temporal contextual information for human action recognition. The overview of this method is described as follows: first, the interest points are detected from each action video, and sub-volumes around interest points are selected as elementary actions, which denote the movements of local patches, next, the construction of structured codebook is done based on these elementary actions; then, a video is represented by a sparse combination of elementary actions using sparse coding framework; finally, a linear SVM is applied as the classifier to predict the action class. A codebook is learned to quantize input features into visual code words. Classification is then performed on the histograms of code words. Generally a large sized code book is required to obtain high recognition accuracy. An oversized code book leads to high quantization errors and over fitting problems. K – Means clustering is a popular algorithm for code book learning. Feature quantization by a large flat code book such a K-Means is however computationally heavy. Tree based code book [24] have been explored as an alternative to speed up the feature quantization.

The KTH data set, a common benchmark for action recognition research, involves sequences of six action classes taken with camera motions, scale, and appearance and subject variations. The average recognition speed estimated as 95.67%.

TABLE 3: Summary list on SCC Model

Year	Author	Concept
2012	Sadanand and Corso	A new high-level representation of video: action bank, to capture semantic information of features.
2010	Y.Zhu,X.Zhao,Y .Fu,Y.Liu	Encode local spatio-temporal features within the sparse coding framework
2011	L.Shao, D. Wu, X.Chen	A correlogram of body poses representation which takes advantage of both the probabilistic distribution and the temporal relationship of human poses for action recognition

D. Hierarchical Tree Based Approach [24]

A tree based approach can be used for Complex activities, which consist of a variable number of sub-events connected by complex spatial-temporal relations. This method [3] presents hierarchical representations of activity videos in an unsupervised manner. These hierarchies of mid-level motion components are data-driven decompositions specific to each video. A spectral divisive clustering algorithm has been introduced to efficiently extract a hierarchy over a large number of *tracklets* named as local trajectories. This structure used to represent a video as an unordered binary tree. This method model this tree using nested histograms of local motion features. It efficiently computes the structural and visual similarity of two hierarchical decompositions by relying on models of their parent–child relations.

This method proposes a hierarchical *divisive clustering* algorithm which is based on recursive bi-partitioning of an approximate multi-modal spectral embedding of tracklets. The resulting structure, called *cluster-tree* (Duda et al. 2001), used to model a video as an unordered binary tree, called *BOF-tree*. It is represented by nested bag-of-features histograms of local motion features. Using this structured information presents several challenges. First, BOF-trees have a variable number of nodes (motion components), and a structure specific to each video. Second, there is no natural left-to-right ordering of the children of the same parent node. Therefore a positive definite kernel on variable-sized, unordered binary trees has been introduced. It consists in efficiently comparing all their sub-trees by approximating sub-trees with simple edge models, and leveraging the additive structure of BOF trees. Finally this method uses the kernel with support vector machines (SVM) to learn powerful nonlinear activity classifiers.

TABLE 4: Summary list on Tree based Model

Year	Author	Concept
2012	Tang et al	Model the transitions between hidden states and their durations using a semi-Markov Model
2012	Jiang et al	Methods use global tree structures to speed up the computation of a matching score Do not explicitly model feature neighborhoods and co-occurrences.

A single spatial-temporal structure is sufficient to represent a class of action. This method identifies a collection of hierarchical space-time trees from video training data, and finds an action model that create these discovered trees to classify actions in videos.

Hierarchical Space time construction method uses a video which first extract a collection of space-time segments (STSs); each STS is a sub-volume that can be produced by video segmentation, and it may cover the whole human body or a body part in space-time. The hierarchical, spatial and temporal relationships among the STSs have been identified; this transforms a video into a graph. From the graph, a compact set of frequent and discriminative tree structures are learnt with discriminative weights for the tree's nodes and edges. This model uses trees instead of graphs for multiple

reasons: first, any graph can be approximated by a set of spanning trees; second, inference with trees is both efficient and exact; third, trees provide a compact representation as it is easy to account for multiple structures using a single tree by allowing partial matching during inference (which is no more expensive). Partial matching of trees during inference also allows us to effectively deal with variations in action performance and segmentation errors. It is noticed that, as the size of the tree structures increases the trees get more specific, thereby capturing more structural information that can provide greater discriminative power in classification.

a) Advantages:

1. Provides the extracting of rich high-level tree structures that capture space, time and hierarchical relationships among the action words

2. This method utilizes both small structures such as words and pairs and tree structures which achieves better performance in recognizing and localizing human actions benchmark video datasets.

TABLE 5: Recognition Accuracy for various recent methods

Year	Author	Title	Recognition Accuracy %
2014	Michalis Vrigkas, Vasileios Karavasilis, Christophoros Nikou	Matching mixtures of curves for human action recognition	98.3%
2013	Pradeep Natarajan, Ramakant Nevatia	Hierarchical multi-channel hidden semi Markov graphical models for activity recognition	97.73%
2014	Wen Zhou, Chunheng, Wang, Baihua Xiao	Action recognition via structured codebook construction	95.67%
2010	Zhuolin Jiang1, Zhe Lin2, and Larry S. Davis	A Tree-based Approach to Integrated Action, Recognition and Segmentation	100%

III. CONCLUSION

In this paper, the various recent methods for Human Action recognition have been discussed. Research on action recognition is very much important to enable a wide range of real time applications. Based upon the survey, Tree Based Construction Model tends to produce more recognition accuracy when compared to recognition methods like Matching Mixtures of Curves, Hierarchical Semi Markov Model and Structured Codebook Construction.

IV. REFERENCES

[1] Pradeep Natarajan, Ramakant Nevatia, Hierarchical multi-channel hidden semi Markov graphical models for activity recognition, Computer Vision and Image Understanding 117 (2013) 1329–1344
[2] Michalis Vrigkas, Vasileios Karavasilis, Christophoros Nikou ,Matching mixtures of curves for human action recognition 2014
[3] Adrien Gaidon, Zaid Harchaoui,Cordelia Schmid ,Activity representation with motion hierarchies, Int J Comput Vis (2014) 107:219–238
[4] WenZhou, Chunheng Wang, Baihua Xiao, Zhong Zhang Action recognition via structured codebook construction, Signal Processing: Image Communication 29 (2014)546–555
[5] A. A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: Proc. 9th IEEE International Conference on Computer Vision, vol. 2, Nice, France, 2003, pp. 726–733
[6] H. Wang, A. Kläser, C. Schmid, L. Cheng-Lin, Action recognition by dense trajectories, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, United States, 2011, pp. 3169–3176.
[7] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatiotemporal features for action recognition with independent subspace analysis, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2011, pp. 3361 3368.
[8] S. Sadanand, J.J. Corso, Action bank: a high-level representation of activity in video, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012, pp. 1234–1241.15.
[9] I. Laptev, On space–time interest points, Int. J. Comput. Vis. 64 (2–3) (2005)107–123
[10] X. Yan, Y. Luo, Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier, Neurocomputing 87 (2012) 51–61
[11] X. Wu, D. Xu, L. Duan, J. Luo, Action recognition using context and appearance distribution features, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2011, pp. 489–496.
[12] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space–time neighborhood features for human action recognition, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, SanFransisco, CA, 2010, pp. 2046–2053
[13] M. Raptis, I. Kokkinos, S. Soatto, Discovering discriminative action parts from mid-level video representations, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 2012, pp. 1242–1249.
[14] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, in: Proceedings of the IEEE, 1989, pp. 257–286
[15] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Conditional random fields for contextual human motion recognition, in: ICCV, 2005, pp. 1808–1815.
[16] L.P. Morency, A. Quattoni, T. Darrell, Latent-dynamic discriminative models for continuous gesture recognition, in: CVPR, 2007.
[17] P. Ramesh, J.G. Wilpon, Modeling state durations in hidden Markov models for automatic speech recognition, ICASSP'92, 1992, pp. 381–384.
[18] Wen Zhou, Chunheng Wang n, Baihua Xiao, ZhongZhang, Action recognition via structured codebook construction, Signal Processing: Image Communication 29 (2014)546–555;
[19] J.Yuan, M.Yang, Y.Wu,Mining discriminative co-occurrence patterns for visual recognition in:2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),IEEE,2011, pp. 2777– 2784.
[20] F.Zheng, L.Shao,Z.Song, A set of co-occurrence matrices on the intrinsic manifold of humans for action recognition,in: Proceedings of the ACM International Conference on Image and Video Retrieval,ACM,2010,pp.454–461.
[21] Y.Zhu,X.Zhao,Y.Fu,Y.Liu,Sparse coding on local spatial–temporal volumes for human action recognition,in:ACCV,2010,pp.660–671. Comparison with BOW and SCon UCF sports data set by action class. W.Zhouetal./SignalProcessing: Image Communication 29(2014) 546–555 555.

[22] S. Sadanand, J. Corso, Action bank: a high-level representation of activity in video, in: CVPR, 2012.
[23] L.Shao,D. Wu,X.Chen, Action recognition using correlogram of body poses and spectral regression,in: 2011, 8th IEEE International Conference on Image Processing(ICIP), IEEE,2011,pp. 209–212.
[24] Duda, R., Hart, P.,&Stork,D. (2001),Pattern classification, NewYork: Wiley
[25] Thanikachalam V, Thyagharajan K K, "Human Action Recognition Using Accumulated Motion and Gradient of Motion from Video" Proceedings of the Third International Conference on Computing Communication and Networking Technologies ICCCNT 2012, Published in IEEE Explore. DOI: 10.1109/ICCCNT.2012.6395973
[26] Thanikachalam V, Thyagharajan K K, "Human Action Recognition based on motion and appearance", International Journal of Advanced Information Science and Technology (IJAIST), Vol. 7, No. 7, pp. 90-95. ISSN: 2319:2682. (IF 0.135)
[27] V. Thanikachalam and K.K. Thyagharajan, "Human Action Recognition by Employing DWT and Texture" Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, Advances in Intelligent Systems and Computing 325,DOI 10.1007/978-81-322-2135-7_34