ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

A Comparative Study and Review for Human Action Recognition

Brenda M[1], Arun P.S[2]

M.Tech, Electronics and Communication Engineering, Mar Baselios College of Engineering and Technology, Trivandrum, India [1]

Assistant Professor, Electronics and Communication Engineering, Mar Baselios College of Engineering and Technology, Trivandrum, India [2]

Abstract: This paper presents a study and review of state-of-the-art research papers on human action recognition. One of the challenging issue is the process of recognizing and understanding of human actions from videos owing to large variations in human appearance, pose changes, scale changes etc. The most important approach for human action recognition is to extract features from videos as representations.It is a main area of computer vision approach.The main applications include surveillance systems, patient monitoring systems, and a number of systems that involve interactions between persons and electronic devices such as human-computer interfaces. Almost all applications require an automatic recognition of high level activities. A brief overview of various state-of-the-art research papers on human action recognition is discussed here.Action Recognition based on Sparse Representation is one of the latest method for a higher recognition accuracy and improving the performance.For this any Datasets can be used Weizmann Human Action Dataset, UCF Sports dataset, Ballet Dataset.

Keywords: Background subtraction, segmentation, K-SVD Random Projection, Gaussian Smoothing.

I. INTRODUCTION

Humans can easily understand actions in a complex scene by using visual system. One of the main purposes is to make machines to analyse and recognize human actions using motion information as well as different types of information.There are three important processing stages present in an action recognition system; they are human object segmentation, feature extraction and representation, activity detection and classification algorithms. Three main action recognition systems are present; single person action recognition, multiple person action recognition, and abnormal action recognition and crowd behaviour. There are four stages for action detection. The human object is segmented out from the video series first [8]. The different features of the human object such as shape, silhouette, colours, poses, and body motions are then properly extracted into a set of features. Thereafter, an action detection or classification algorithm is applied on the features that are extracted inorder to recognize the various human activities.The starting and ending times of all occurring activities from an input video must be detected for recognition of human actions [8]. Several important applications can be constructed for the recognition of complex activities. Automated surveillance systems in public places such as airports, railway stations, bus stations, stadiums etc.require detection of abnormal and suspicious activities as against normal activities. For example, an airport surveillance system must be able to automatically recognize suspicious activities like ‘a person leaving a luggage’ or ‘a person placing his luggage in a dust bin’, ‘terrorists trying to spy’, ‘illegal activities at unknown time’ etc [8].

Recognition of human actions makes it possible the real-time monitoring of patients, children, and elderly persons. There are various types of human actions present. Depending on their complexity, they generally categorize human actions into four different levels: gestures, actions, interactions, and group activities [8].These complexities make the research topic application oriented and challenging. There are different action recognition systems present and some are reviewed here. Action recognition using sparse representation is a recent method. These representations can be built by decomposing signals over elementary waveforms which are chosen from a family called dictionary. Signals carry large amount of data which contain both relevant and irrelevant information where relevant information is difficult to be obtained. In sparse representations few coefficients contain the relevant information. These representations can improve pattern recognition, compression and noise reduction. Also they are robust against missing data and distortions which finally provides a compact representation useful for action recognition.

II. REVIEW OVER PREVIOUS PAPERS

There are three action recognition systems being discussed here.

A. Human Actions based on Motion Information

The motions of the people in the scene are tracked continuously. The subjects from the images are extracted using background subtraction.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

The positions of the body and the hands are extracted from silhouettes. The detection of head and hand positions is achieved by silhouette boundary shape analysis [3]. Pre-processing steps are carried out inorder to help extracting the required body parts more efficiently. The motion trajectories of body parts are then recorded for analysis.

The original sequence of data defined by the window size is compressed and features are extracted by two computational methods: Principal Component Analysis and Independent Component Analysis. The extracted features are then passed through support vector machines for classification learning.

While modelling the background, each image is subtracted from a model of the background scene and thresholding the resultant difference image to determine the foreground pixels. The pixel intensity of a completely immovable background in the indoor can be effectively modelled with a normal distribution. This model can relatively adapt to slow changes in the scene by recursively updating the model.

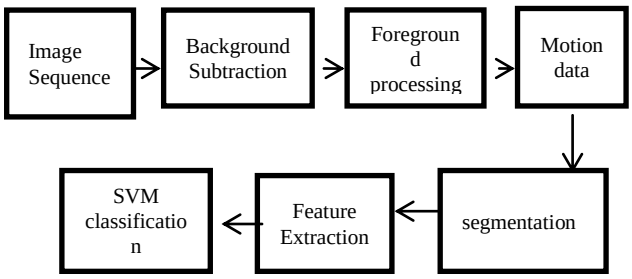

Fig. 1 Block Diagram

B. Full Body Tracking based Human Action Recognition

The action recognition system includes three basic stages: action capture, action representation and action classification [2].

Stage 1: Multiple video streams are simultaneously captured from static calibrated cameras. And foreground/background segmentation is performed on each after background subtraction method. Then volumetric representation sequences of the object to be tracked are created.

Stage 2: An effective method called 3D joints sequence is used. Full body tracking is adopted to track subject's connecting parts. However the subject's large degrees of freedom and the nonlinearity of the dynamic system bring many difficult problems. To overcome this, a tracking approach is developed which fuses the body part segmentation and adaptive particle filter. Then extract the movement feature and configuration feature.

Stage 3: Each action is represented using two HMM's. One is a conventional HMM for improvement feature and other is an exemplar based (non-parametric) HMM for configuration feature.

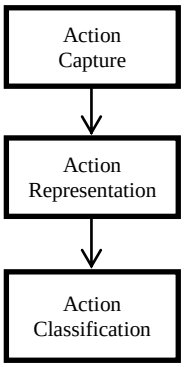

Fig. 2 Flow Diagram

C. Hidden Markov Model

A Hidden Markov Model (HMM) (Yamato et. all's work-1992) is specified by three terms $\emptyset=(\pi,A,B)$ [16]. Action is represented in terms of hidden states.A human is assumed to be in one state at each time frame, and each state generates an observation.In the next frame, the system transitions to another state considering the transition probability between states. Once transition and observation probabilities are trained for the models, activities are commonly recognized by solving the `evaluation problem'. The evaluation problem is a problem of calculating the probability of a given sequence (i.e. new input) generated by a particular state-model. If the calculated probability is high enough, the state model-based approaches are able to decide that the activity corresponding to the model occurred in the given input.The first term is the initial probability of hidden states. The second term is the transition matrix which specifies a transition probability from one hidden state to another hidden state. The third term is the observation matrix which specifies the probability of the observed symbol given a hidden state. There are three fundamental problems present in HMMs they are Evaluation, Decoding and Training and are commonly solved by the Forward/ backward algorithm, Viterbi algorithm and Baum-Welch re-estimation algorithms, respectively. The key parameter of a HMM is the number of states that will be used for each model. Recognition is based on identifying the maximum likelihood path of states that generated the sequence of observations under testing (e.g. a gesture) from all available classes.

D. Methodology

The detection and representation of interest points in videos are the main steps of Action Classification. By Interest point detection the video from a volume of pixels is being reduced to sparse at the same time they are descriptive set of features[4]. Usually, interest points are enough to capture the relevant information to recognize arbitrary human activities. The detected interest points with motion can be kept to get spatio-temporal interest point. The action classification includes four stages; interest point detection, feature description, codebook construction and bag-of-words feature representation and classification stage.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

E. Basic System Model

E.1: Interest Point Detection:

A feature detector finds out the points in the video where features are going to be extracted. These points are called as Spatio-Temporal Interest Points (STIPs). A STIP is a point in space and time (x, y, t) which has high saliency. High saliency indicates that there are high amounts of changes in the neighbourhood of the point. This shows a large contrast change in the spatial domain, yielding a Spatial Interest Point (SIP). However for the temporal domain saliency occurs when a point changes over time, and during this change occurs at a SIP, the point is taken to be a STIP [4]. High spatial saliency occurs all around the contour of the person when there is difference in appearance between the person and the background. A FAST [12] detector is applied or SIFT detector inorder to detect interest point on the first frame (image) to identify candidate features; an optical flow computation between the two frames at a scale very appropriate to the candidate feature (for SIFT detector) is done. This is done to eliminate those candidates that are not in motion. Detector subsequently scans through every frame of the video (overlapping pairs) to identify key points in each frame.

E.2: Feature Description:

A descriptor is a feature that is extracted to describe both shape and motion around an interest point so it plays an important role in action recognition [4]. A certain number of algorithms for interest point descriptor in the spatial domain and temporal domain have been compared and discussed. By [12] Chen has revealed that incorporating explicit motion detection gains a much better result in human action recognitions he also suggests that appearance and motion can be analysed independently. An algorithm called Motion SIFT (MOSIFT) is addressed which discover meaningful key points by SIFT.

E.3: Codebook construction and bag-of-word feature representation:

"A bag of word model represents images as orderless collections of local features". In Bag of Words, "document is represented as a normalized histogram of word counts". The features extracted from a set of training images are clustered and codebook is built to represent the dictionary [4]. One of the simplest methods is K-means clustering being performed over all the vectors. The procedure for generating a Bag of Word image representation can be summarized as follows: (1) Build Vocabulary (codebook): Extract features from all images in a training set. "Vector quantize, or cluster, these features into a "visual vocabulary," where each cluster represents a "visual word" or "visual codebook" or "codes"." (2) Assign codes: Extract features from an image. Euclidean distance or a related strategy is used to assign the features to the closest codes in the vocabulary. (3)Generate Term Vector: The counts of each code that appears in the image are recorded to create a normalized histogram which represents a "code vector." This code vector is named as the Bag of Features representation of the image.

E4: Classification:

In classification stage many Discriminative models are used known as non-linear Support Vector Machine (SVM) classifier. SVM maximizes the distance between the points of two classes.

F. Action Recognition using Sparse Representations

This is a recent method which consists of four stages namely: computation of the spatio-temporal motion descriptors, dimensionality reduction, learning overcomplete dictionaries using lower dimensional descriptors, and classification using dictionaries

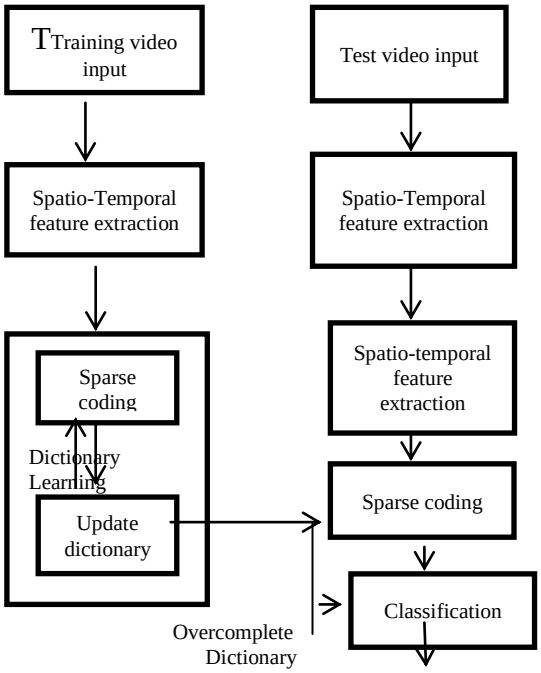

Fig. 3 Flow Diagram [5] Detected activity

F.1 Computation of Spatio-Temporal Features

The first stage is to develop a rich spatio-temporal representation for each action sequence. To get a good number of features we use cuboid descriptors. To determine the spatio-temporal features, some response function has to be computed at every point. Patches are extracted at regular intervals from videos [7]

In order to lower the computational load though thekeypoints are detected in spatial domain, the temporal information is retained. Every keypoint is connected with a spatio-temporal cube. Each cube captures the local space-time changes of the signal which ultimately represents a motion pattern.

These spatio-temporal cubes are extracted from all temporal segments of the video.And to obtain a descriptor 2D Gaussian blurring are performed on each cube. This consequently increases the robustness against noise and other uncertainties due to imperfect segmentation. Finally these cubes are smoothed enough to obtain small temporal variations.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

F.1.1 Feature Extraction

Steps:

- Divide the video sequences into a number of subsequences.
- Define the length of video sequence
- Define frames per subsequence
- Get the first frame
- Detect spatial keypoints at the first frame
- 2D keypoint detector is used.
- Take only luminance value
- Find the derivative masks and image derivatives
- Set threshold 10% of the maximum value.
- Find local maxima on 3x3neighbourhood
- Find local maxima greater than threshold.
- Build interest points.
- Get rid of the points close to the edge of the image.
- Obtain good keypoints
- Store the keypoint locations.
- Extract patches cantered at those points at every frame of a subsequence.
- Rearrange the patch vectors

F.1.2 Computation of descriptors

Steps:

- Get the patch size and patch volume.
- Guassian smoothing done.
- Remove mean
- Compute the three moments(variance, skewness, kurtosis)
- Combine them into one vector

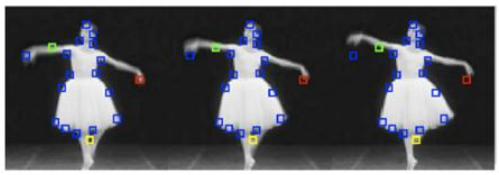

Fig. 4 Segment consisting of 3 frames (2D keypoint detector)[7]

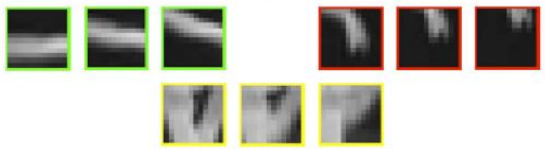

Fig. 5Patches are extracted around each key point at each frame Three space time cubes (green, red, yellow) associated with Key points [7]

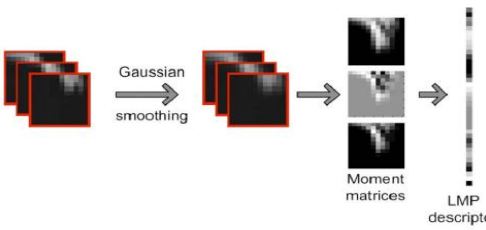

Fig.6 Conversion of cube to a descriptor. Guassian smoothing is done first, three moments are found thereafter, all these moments are concatenated to form the descriptor [7].

F.2 Dimensionality Reduction

All the descriptors are highly dimensional; if they are used with original dimension will need more than 2000 dictionary atoms to be learned for a compact and sparse representation.

This will limit the computational speed. So we need to reduce the dimensionality here a powerful tool named as random projection is used [7].

Consider a set of descriptors obtained from a video sequence. This set can be described as a matrix. These set of descriptors are projected onto an n-dimensional subspace by pre-multiplying the descriptor matrix D by a random matrix R.

The Random Projection reduces the matrix to:

$$Y=RD \qquad (1)$$

Y- Reduced Data Matrix

F.3 Dictionary Learning

Dictionary learning is done for everyaction class that is available in the training data. If there are jdifferent activity classes in training data, then create j numberof action specific dictionaries, Dj (sub-dictionaries) [5].

After creating all the action specific sub-dictionaries combine them all toform an over complete structured dictionary, D. An over complete dictionary D, that leads to sparserepresentations can be pre-designed. The success of these dictionaries depends on how they are suitable for the test data. An over completedictionary, D designed for a particular training data is moresuccessful than a commonly used pre-designed dictionaries [5].

K-SVD is an iterative method that alternates betweensparse coding of the training data which is based on the current dictionary and a process of updating the dictionary [5].

K-SVD algorithm sweeps through all the columnsand will use the most recently updated values from theprevious step.

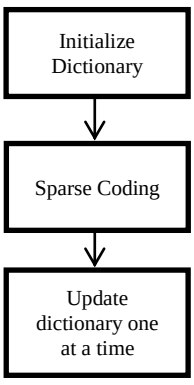

Fig. 7 K-SVD Algorithm [5]

DOI 10.17148/IJARCCE.2015.4652

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

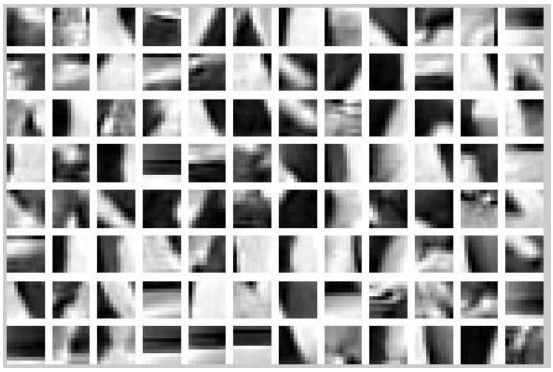

Fig. 8 An overcomplete Dictionary [5]

Finally action classification is done in the test video.

G. Real time applications

Several computer vision systems have been designed to recognize activities in real-time scenarios particularly because most of the proposed algorithms are far from being real-time. This computational difficulty must be met in order for an action recognition method to be applicable for real-world applications like surveillance systems, human-computer interfaces, intelligent robots, and autonomous vehicles.

Recently, different real-time human action recognition systems have been developed, and some of them are reviewed here. The main concept behind most of the approaches is to raise the performance of the algorithms by simplifying them. The Bayesian posterior probability calculation of Lv et al. [2004] searches for an optimal set of features from a large number of features for action recognition. A dynamic programming algorithm has been developed to find a set of features that maximizes the detection accuracy on the training data. PETS 2004 dataset has been utilized for the testing [6]. Many other applications are present for Surveillance Systems:

- In homeland security
- Crime prevention
- Military target detection

III.CONCLUSION

Action recognition of humans is an important area of research in computer vision with applications in various diverse fields. The surveillance application is very common these days in today's world as the tracking and monitoring people has become an important part of everyday activities. An overview of the current approaches to human action recognition is presented here. In this paper, the methodologies are studied that have been previously explored for the recognition of human action. We live in a world where experimental systems aredeveloped and practically deployed at airports and other public places. In coming years more and more, such systems will be put into development. Further, today's environment for human action recognition is completely different from the scenario during the last years. The cameras were mostly fixed cameras and without new advancements. But the cameras of today can be mounted on different types of moving platforms such as a moving car or a truck to an unmanned aerial vehicle (UAV) [6]. The camera system can be attached to a global positioning system to pin-point its location. The action recognition from a moving platform meets many challenges. Noise, tracking, and segmentation issues are some other difficulties as problem for the recognition of action. The recognition of the action being performed becomes very difficult if the tracking algorithm does not extract the desired object. The future direction of research is obviously expected to have much development in action recognition taking into consideration the challenging tasks present ahead. Of all the applications that are related with action recognition the pressing applications are the surveillance and monitoring of public facilities like train stations, bus terminals, underground subways or airports, monitoring patients in a hospital environment or other health care facilities, monitoring activities in country territories and boundaries. Finally after review over all the different past methods for action recognition, a recent method using sparse representation is studied in detail. Through experimental analysis this particular method is found to be producing better performance in terms of recognition accuracy.

ACKNOWLEDGMENT

The Authors would like to thank the Dr. Guillermo Sapiro for his video lectures that really helped to understand the difficult concept 'Sparse Coding'. Authors would also like to thank Dr. Tanaya Guha for providing the publicly available source code for experimental analysis on the performance of the recent method.

REFERENCES

[1] K. Lee and Y. Xu,"Modelling Human Actions from Learning", in Proc. IEEE International Conference on Intelligent Robots and Systems, pages 2787-2792, 2004.

[2] G Junxia, D Xiaoqing, W. Shengin and W. Youshou ,"Full Body Tracking-Based Action Recognition", in Proc. IEEE Workshop on Computer Vision, Berlin, pages 817-829, 2008.

[3] Z. Zhang, Y. hu, S. Chan, and L.T. Chia, "A new representation for human action recognition," in Proc.IEEE Workshop on Computer Vision, pages 817-829, 2008.

[4] Amira Ali Bebars and Elsayed E. Hemayed , "Comparative study for feature detectors in human activity recognition," Cairo University, Egypt.

[5] Dipti Killedar and Sreela Sasi, "Human Activity Detection using Sparse Representation," in Proc. IEEE, 2014.

[6] J. K. Aggarwal and M. S. Ryoo, Human Activity Analysis: A review, ACM, 2010

[7] Tanaya Guha and Rabab Kreidieh Ward, "Learning Sparse Representations for Human Action Recognition," in Proc. IEEE Transactions On Pattern Analysis And Machine Intelligence,Vol. 34, No. 8, August 2012

[8] .Akila.K and Chitrakala.S, A Comparative Analysis of Various Representations of Human Action Recognition in a Video, IJIRCCE, Vol. 2, Issue 1, January 2014

[9] Haocheng Shen, Jianguo Zhang and Hui Zhang, Human Action Recognition by Random Features and Hand-Crafted Features: A Comparative Study, 2013.

[10] Michael B. Holte,Mohan M. Trivedi and Thomas B. Moeslund"Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments," in Proc. IEEE Journal Of Selected Topics In Signal Processing,Vol. 6, No. 5, September 2012.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

BIOGRAPHIES

.

Brenda M received her B.E. Degree in Electronics and Communication Engineering from Anna University, Chennai and now she is currently doing her M.Tech.Degree in Telecommunications Engineering under Kerala University at Mar Baselios College of Engineering and Technology, Trivandrum. Her area of interests includes Image Processing, Video Processing, Pattern and it'sapplications.

Arun P. S. received his B.Tech degree and M.Tech degree in electronics and communication engineering at College of Engineering, Trivandrum from Kerala University, India in 2007 and 2009, respectively.He joined Mar Baselios College of Engineering and Technology in 2011as an Assistant Professor in the Department of Electronics and Communication Engineering and his area of research is video summarization, video coding etc. He is a lifelong member of ISTE and has published several papers in various national and international journals and conferences.

DOI 10.17148/IJARCCE.2015.4652