International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019,ISSN (Online): 2395-566X

Automated Human Action Recognition System Using Machine Learning

Sagar Goyal
Dept. of Computer Science and Engineering
VIT Chennai
sagar.goyal2015@vit.ac.in

Susmitha Shereef
Dept. of Computer Science and Engineering
VIT Chennai
susmitha.shereef2015@vit.ac.in

Abstract - **Human action recognition is an important research topic, nowadays because of security and safety constraints. An action recognition is helpful in gesture recognition, posture detection, video surveillance and many more important activities that are desirable in today's world. Action recognition can help in reducing the criminal acts or in detecting any kind of that activity at correct time so that appropriate action against that activity can be taken. There are various conventional action recognition methods that recognize actions in a video directly which can often give incorrect results because analyzing videos directly requires great speed in detection of actions. However, detecting images is much simpler and easier job to do. So, we convert an input videos into static image frames for easy analysis and interpretation of actions. There are many problems in current action recognition methods which face deficiency in labelled videos or some methods use lot of labelled videos in training dataset that further face the problem of over-fitting the training data and inaccuracy in detecting actions in online testing dataset. Hence, in this project, we use a semisupervised learning approach to make a balance between labelled and unlabeled videos and utilizing both simultaneously that overcomes the problem of overfitting the data and also provide the best results in least amount of time. The experimental results show that our action recognition method works for online testing dataset and gives accurate results with very good performance and is better than any other action recognition method.**

Keywords- **HOG feature extraction, Key frame extraction, KTH dataset, Video Surveillance, Labelled videos, Indexing, Histogram, Gesture Recognition, Semi supervised Learning, Video Analysis, KNN Algorithm**

I. INTRODUCTION

We have considered the Kth dataset for our project for getting input videos for actions like jogging, walking, hand-clapping, etc. We have referred 'Robust Feature-Based Automated Multi-View Human Action Recognition System' research paper for our project. Also, this dataset consists of videos that are shot very professionally and under great supervision. The videos are shot in proper lighting with proper orientation from different angles and at various distances by multiple cameras simultaneously in order to facilitate the ease in detection and recognition of action performed by a person in a video.

Any occlusion or poor video quality and object discrepancy can give incorrect results. Therefore, the video should be shot under great supervision and the cameras should be handled with care to get the desired output. The action recognition system has a huge scope nowadays with the immense advancement in technology and computer industry which provide internet facilities. Hence, video action recognition has become an important aspect of research because of its essential real world applications like video surveillance, video indexing, annotating the videos, etc. However, the videos must be shot very carefully to do the desired job. So, occlusion, low camera quality, camera shaking, improper background, etc must be avoided while taking the video as an input for testing the action recognition method.

Also, there is a significant need for security and safety from criminal activities at public places, in this vast world. So, action recognition system can easily detect any criminal action after analyzing the video at public places, which can prove to be very helpful in making a crime-free area. For example, if a person is taking out a gun then this criminal action can be detected easily and then alerts can be provided immediately to the general public and appropriate actions can be taken at the right

© 2019 IJSRET

time and correct place. However, it is an extremely challenging job to detect every human action in videos. There are various present conventional action recognition methods that recognize actions in a video. In these traditional methods, we first extract the histogram range from videos to detect the edges. Then also ranges are shown in a bar graph vector with BOW illustration i.e. bag of words. Following this, we do the classification for vectors like SVM (Support vector machine) which work for videos that are shot very carefully and are completely simple and straight. But, they fail miserably, in case, any noise or disturbance step into video accidentally or non-accidentally.

So, they won't work well in case of improper video quality and shooting, inadequate background, poor camera angle, etc. Also, these traditional methods won't be generalized well in order to predict the accurate result of actions for new testing video dataset. Therefore, we built a system which focuses on important part of actions like human stance, human objects, look and many more so as to give accurate results even when there is a disturbance or any kind of noise. This focus helps to clearly obtain the object from its background and obtain the object's features easily by making a check on each and every human move in a video. This will make the interpretation and analysis of human action, very easy.

The conventional action recognition methods require a lot of training with huge dataset of input videos with a problem of overfitting of the data which is not advisable for real world implementation. Also, enough labelled videos are not available in real-world. Hence, we use semisupervised learning which will not only improve the accuracy of the results, but will also speed up the process of recognition of actions. The semantic information from human brain is used and interpreted to classify it and give labelled videos. Various studies also show that considering both labelled and unlabeled videos help in improving the performance for human action recognition. This is what is called as semi-supervised learning in which we do classification and prediction for some dataset and directly label the videos for some dataset of input videos for action recognition system.

II. LITERATURE SURVEY

1. MoSIFT Recognizing Human Actions in Surveillance Videos By Ming-yu Chen and Alex Hauptmann The aim of this paper is to create robust human video action recognition system which could be applicable for real world surveillance videos. For obtaining compact and descriptive representation while doing video analysis and recognition of the motion in the video we will be utilizing spatio-temporal features around points of interest. What is been observed is that many of the approaches which is existing today have a tendency of extending spatial descriptions by combining it with temporal component according to the appearance descriptor that is capable of capturing information contained in the motion implicitly.

The concept of bigram model had also been introduced in order to construct a correlation between local features. By this way we would be capturing more global structure of actions. Because of this method advancement of the state of art result take place on the KTH Dataset which yields an accuracy of about 95.8%. We had implemented our method to about hundred hours of surveillance video data during the TRECVID Event Detection task.This event has brought a very promising results about the recognition of human actions where the input is real world surveillance videos.

2. Observing Human-Object Interactions- Using Spatial and Functional Compatibility for Recognition By Abhinav Gupta, Member, IEEE, AniruddhaKembhavi, Member, IEEE, and Larry S. Davis, Fellow, IEEE The appropriate interpretation of images and videos which has human interacting with different variety of objects is a daunting job. This requires the involvement of of correct understanding of the scenes or events which are used for analysing of human movements, recognition of manipulating objects and observation of the effect of the human movements on different type of objects. Conduction of all of the above mentioned perceptual tasks can be done independently but the rate of recognition only improves when consideration of interactions among them are taken into account.

But according to the methods mentioned in this research paper, the methods will be going beyond many of the traditional methods. These methods will be applying spatial and functional constraints on each and every perceptual elements which are being involved in the coherent semantic interpretation. These kind of constraints will helps us in the recognising of objects and recognising of actions though the shapes which are obtained is not enough discriminative.

3. Visual Event Recognition in Videos by Learning from Web Data By Lixin Duan Dong Xu Ivor W. Tsang School of Computer Engineering Nanyang Technological University , JieboLuo Kodak Research Labs Eastman Kodak Company, Rochester, NY, USA In this paper, the approach proposed is about visual event recognition framework for a particular set of domain of consumers videos. This is being done by a technique called leveraging majority of the web videos which was loosely labelled and these videos can be obtained from anywhere like YouTube. In the first step, this paper has presented a aligned space-time pyramid matching technique which is used

© 2019 IJSRET

to measure the distances between any two video clips of the required domain. In this method each and every video clip is being divided space-time volumes over multiple levels. The last approach which is used in the paper is that of cross-domain learning method and this approach can also be called as Adaptive Multiple Kernel Learning (A-MKL). Many type of extensive experiments that had took place in the past demonstrated about the effectiveness of the framework which is proposed in this paper. This is due to the fact that the technique in this paper had the requirement of minimum number of labelled consumer videos which are obtained by doing leveraging of web data. 3.4 Action Recognition Using Nonnegative Action Component Representation and Sparse Basis Selection.

By Haoran Wang, Chunfeng Yuan, Weiming Hu, Haibin Ling, Wankou Yang, and Changyin Sun. The approach which has been used in this paper is the proposal of a system that has been build using high-level action units and is used for the representation of actions of humans in videos. By using this units as a base the paper has proposed a novel sparse model that is being used for human video action recognition. According to this paper, there has about three interconnected approaches which is used by the approach in the paper. The first approach is that we will be developing a new kind of context-aware spatial-temporal descriptor for the which can be called as locally weighted word context.

During learning of dictionary for representation of action, the sparse model introduced through this research paper captures the fact that actions which are being performed from the same class has similar units in common. The idea that has been used in this research paper had been done evaluation on different kind of datasets which are publicly available. The result of experiments and that of analysis had clearly demonstrated the effectiveness of algorithm proposed in this research paper.

5. Recognizing human actions in still images: a study of bag-of-features and part based representations. The algorithms related to human video action recognition is commonly addressed in the domain of interpretation of video. On the other side many of commonly performed actions of humans like reading of a book, playing of a guitar or writing of notes provides a very natural description for a majority of still images.

Besides the above mentioned actions, there are certain actions in the videos like taking of photographs which are by nature static and it has the requirement of static cues for the recognition methods. The main concept whose investigation is done through the paper is that of the role of the background scene context. It is also been demonstrated that the action recognition can be further improved by the following methods:-

- By making combination of the statistical and that of part-based representations.
- By integration of person-centric description along with that of background scene context.

III. PROPOSED SYSTEM

Figure 1 Series of actions required The approach which we have used in the project is that we, first, extracted the image (static) feature of images and key frames obtained from the videos.

We, then, computed the computational efficiency and for computing it we extracted key frames by using an algorithm called shot boundary detection algorithm. The procedure of the key frames extraction is that first, we calculate the colour histogram of consecutive five frames. After that, the histogram which we calculated earlier is substracted from the previous frame. The final step is that we assigned that the frame is a shot boundary if it satisfies the condition that the subtracted value is larger or smaller than our empirically set threshold value. If it is larger, we have consider it as a shot. The middle frame of the shot frames is termed as key frame.

At the same time, video or the motion feature is extracted. This video feature is obtained from video domain. After extraction of video feature, it is combined with that of image feature which we had obtained earlier. After this compilation process our image feature has now become a subset of combined feature. According to our project, we have used Kernel Principal Component Analysis (KPCA) technique. In KPCA method, we have mapped two features, namely combined features, which we had obtained earlier and image features.

© 2019 IJSRET

This method actually explores the basic principal knowledge of the Hilbert spaces which we had mapped. So, we can say that our proposed method of training is more better and has more efficiency. We also made adaption of knowledge on shared space of common features and then, it is used for optimization of classifier A. For making proper usage of unlabeled videos, we introduced the concept of semi-supervised classifier AB.

We have also trained this semi-supervised classifier AB based on the information contained in the heterogeneous features available in the domain of video. After this, we completed an appropriate integration of the above mentioned two classifiers to form a joint optimization framework. The recognition results which we obtained finally, by testing of the input videos, are further improved. For further improvement, we fused the results obtained from classifier A and semi supervised classifier AB.

1. Hog Feature Extraction Algorithm

- Transform color image into gray scale for applying various image processing concepts that can be applied to 2D images only.
- Calculate the Luminance gradient at every pixel point.

The luminance gradient is a vector value that changes with the change in the luminance due to changes in magnitude and orientation.

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2}$$

$$\theta(x,y)=\tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right)$$

where,

$$-\frac{\Pi}{2}<\theta<\frac{\Pi}{2},$$

M=magnitude, Θ= Orientation and L=Luminance value of the pixel

- Then, the histogram of those gradient orientations is calculated for every pixel. This will be different for different features, so, is used for feature extraction.

2. Knn Algorithm

- k is defined first to ensure how many neighbors are close to the new sample video.
- We all k neighbors of that data point that are closest to it.
- The common classification out of these k closest neighbors is find out.
- This classification is assigned to the new sample video.

IV. IMPLEMENTATION AND RESULTS

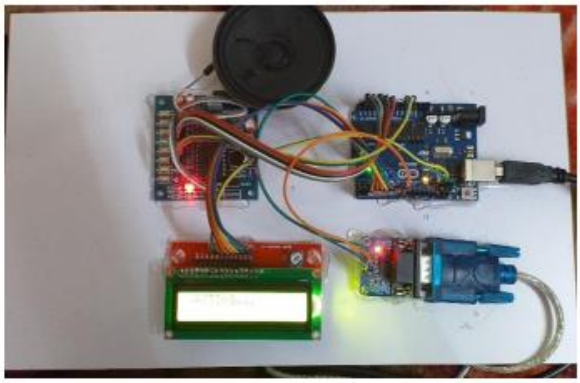

Fig.2 Hardware Connections with LCD, Arduino, APR and MAX232 Connector.

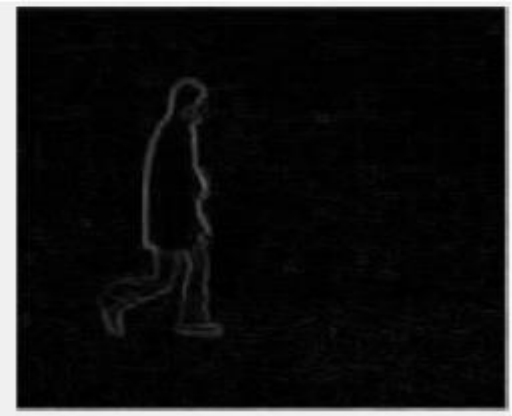

Fig. 3 Jogging Action Detection in a Video(Magnitude Information).

The results have shown that this semi-supervised learning framework based algorithm is not only efficient and reliable, but also has achieved great accuracy and performance, when the labeled training samples given are very less. The experimental results show that our action recognition method works great for online testing dataset and gives accurate results with a very good performance and is better than any other action recognition method. CONCLUSION The video action recognition method has shown better results in terms of recognition and retrieval of data for analysis of video from images.

Information like posture, object and appearance which are semantically similar and relatable are used for better action recognition performance and accuracy. The actions considered are boxing, walking, running, jogging, clapping and hand waving. The final desired output is shown in LCD display (the hardware part) successfully and the audio voice of the output is also provided using APR Module for easy understanding of

© 2019 IJSRET

a common lay man. The bilinear interpolation method defines feature vectors for each action and creates an appropriate knowledge database for classification. The semi-supervised learning framework which is used in recognizing human actions, uses both labelled and unlabeled data in video domain. The KNN classification produces good results with correct prediction of the action.

REFERENCES

[1] J. Wu, Y. Zhang, and W. Lin, "Good practices for learning to recognize actions using FV and VLAD," IEEE Trans. Cybern., to be published.

[2] H. Wang et al., "Action recognition using nonnegative action component representation and sparse basis selection," IEEE Trans. Image Process., vol. 23, no. 2, pp. 570–581, Feb. 2017.

[3] A. Khan, D. Windridge, and J. Kittler, "Multilevel Chinese takeaway process and label-based processes for rule induction in the context of automated sports video annotation," IEEE Trans. Cybern., vol. 44, no. 10, pp. 1910–1923, Oct. 2017.

[4] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion mapsbased local binary patterns," in IEEE Winter Conference on Applications of Computer Vision (WACV), 2017.

[5] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in ACM International Conference on Multimedia (MM), 2015.

[6] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning", Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp. 28-35, 2015.

[7] A. Sambath and A. Nirmala, "A survey on multimodal techniques in visual content-based video retrieval", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.

[8] B. Afsari and R. Vidal. Distances on Spaces of High Dimensional Linear Stochastic Processes: A Survey. Springer International Publishing, 2014.

[9] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In Proceedings of the International Symposium on Computer Vision, pages 265–, 2015.

[10] O. Zoidi, A. Tefas, and I. Pitas, "Visual object tracking based on local steering kernels and color histograms," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 5, pp. 870–882, 2015.

© 2019 IJSRET