INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume3, Issue2)

Available online at www.ijariit.com

Identification of Human Actions in Video Database

Smita S. Patil
BNMIT, Bengaluru
smitapa3@gmail.com

Santosh Saraf
KLSGIT, Belagavi
santoshsaraf@git.edu

Abstract: ***The Human activity detection is a more highly sought research as they are used for the Surveillance, Healthcare system, content-based search, and interactive game applications. The Human activity recognition is carried out by three main stages namely object segmentation, feature extraction, their representation and human action detection using different algorithms. This paper is an attempt to identify the human activities such as running, jogging, clapping and hand waving in a video database using the method of a bag of features which is the extension of a bag of words. The Size Invariant Feature Transform (SIFT) based bag of SIFTs is used for the local feature extraction from the images that are chosen for the human activity detection. After finding the SIFT features for both the database and the query video, the query video features are compared with each of the database video features, the distance is found using Euclidean Distance method and the K-Nearest Neighbour classifier is used to find the classified category from the list of the video. MATLAB based implementation is done with various activities identified and their efficiencies are compared.***

Keywords: Features Extraction, Size Invariant Feature Transform (SIFT), Bag of Features (BOF), k- Nearest Neighbour (k-NN).

I. INTRODUCTION

Many videos are created due to the availability of different video recording technologies, these videos are from many areas such as surveillance, health care system, filmmaking etc. Due to the increased collections of videos, the effective search becomes difficult. Presently, a search of a video in video databases is done with a text-based search where in the video is returned based on the text given by a user. This type is not an accurate method of search for two reasons. Firstly the text information given by the video uploader depends on the assessment of the uploader. Secondly, the information present in the video cannot be represented in a text form.

An alternative method called content-based video retrieval (CBVR) which is an extension of content-based image retrieval (CBIR). Example – a video query is given by the user, to search, a CBVR returns with the more accurate results. In CBVR the concentration is on human actions which are different from retrieval. Human action recognition has the interest in the various research areas. Generally, in the video, the identification of the human action is processed in three steps namely segmentation of human object, extraction of feature and their representation and detection of activity. First, the properties of an object like shapes, body motion, colour, poses are extracted from an object which is separated from the video and are represented and then to get human actions an algorithm is performed on the extracted feature of an object.

The first step, to get the required target object, segmentation of objects is done on video, actually on each frame of it. There are two types of object segmentation depending on the status of the camera either it is stationary or moving. The two types of segmentation are stationary camera segmentation and moving camera segmentation.

In the stationary camera segmentation, the position of the camera and the object viewpoint are fixed at some angle. There are many methods to perform static camera segmentation such as background subtraction, GMM, statistical modeling, by tracking models etc. The popular method is background subtraction [1-3] as this method is simple and efficient.

In the moving camera segmentation, both the camera and the target object is in motion along with the variation in the background. This type of segmentation is difficult than static type as in the moving segmentation type in addition to moving target object, the camera is also in motion with a change in background. The methods used in this type are the temporal difference and optical flow technique. The common method is temporal difference method [4], [5] wherein the difference between successive frames is performed.

In the second step, properties like shapes, colours, poses and body motion of a segmented object are extracted and represented. The features are mainly classified into four methods namely frequency transform, space-time information, local descriptors and body modeling.

The third step is human action detection algorithm and is used to identify the actions of human based on the feature extraction of the target object. They are classified as discriminative models, generative models, dynamic time warping (DTW) and others.

II. RELATED WORK

Authors [6] developed a technique, used for search and detection of human actions called random forest based voting, is considered to be robust and efficient which overcomes the challenges present in search and locating human actions in a dense situation and proposed a scheme coarse to fine sub volume search which is faster over existing video branch and bound method which overlooks the problem of running the branch and bound search many times for accuracy and computation cost. For the interactive search, the proposed method can be applied and the search was progressed only after some few numbers of rounds of relevance feedback.

Author [7] proposed an algorithm which is scaled invariant and is efficient against noise. The algorithm is used for retrieval of a video from a large video database. The local features of a large video database are extracted before a query search and are compared with features of the query, candidates are found. Relevance feedback is used for the ranking of candidates. They say that the system is efficient in time and space and used the noise test videos for relevant feedback. They concluded that method has poor performance as it is unable to remove spatially distinct background noise from the result which may not rank the candidates properly.

Authors [8] proposed Hidden Markov Model (HMM) to recognize the human actions which use a bottom-up approach which is based on features because the concentration is more on learning strengths and time scale invariability. The model is made to learn for different categories of human actions and then the improved training sequences are found. These sequences are used to match with the given image features. In the proposed method the image containing human actions are transferred to feature vectors by extracting the features of each image. The feature vectors are assigned to a symbol called code word saved in the codebook. In the learning stage, HMM model variables are trained for a particular feature and the best match is obtained by comparing. The recognition performance decreases as the training sequences are decreased. The method deals with 2D images and can be extended to 3D objects.

Author [9] described to determine features of the local image which changes in both time and space by an interest point detector. These points give important and unique events. Different interest points represent different events and helpful in classification. The problem of classification is addressed using scale adapted descriptor which is used for the representation of the video. In particular, the author has shown the detection of walking people in various situations by spatiotemporal interest point. These points also help in splits and joining of images.

Authors [10] have shown an algorithm to recognize the viability of doing behaviour and proposed an extended spatiotemporal interest point detector which is a different method than 2D interest point detector which is not a proper method for spatiotemporal feature points such as eye opening or knee bending. They have developed and analysed different descriptors. The representation proposed by the author is enough for the identification and efficient with respect to change in the data and shown better results in various domains using the proposed algorithm.

The author [11] described SIFT key points are distinct and useful incorrect matching of key points from a big database of different key points and also described methods for object recognition using the key points. Presents a method which is used to extract the invariant features of an image and these features remain same even though the object views position is changed such as image scale and rotation is called SIFT. The recognition is done by comparing features of an object to the features of objects of the database using Euclidean distance method and author discussed, a fast nearest neighbour algorithm which is used to perform fast in a bigger dataset. The SIFT used in extracting features covering the entire image hence helps in finding the small objects in an image. For a best match of features, minimum numbers of features are to be matched. The method discussed by the author is sensitive means a small change in descriptor will change the change in the features.

Authors [12] proposed a method to represent video sequence contents called content-based browsing and is efficient and best to the existing techniques of fast forward and rewind. This method saves time, bandwidth and expenses.

III.BLOCK DIAGRAM APPROACH

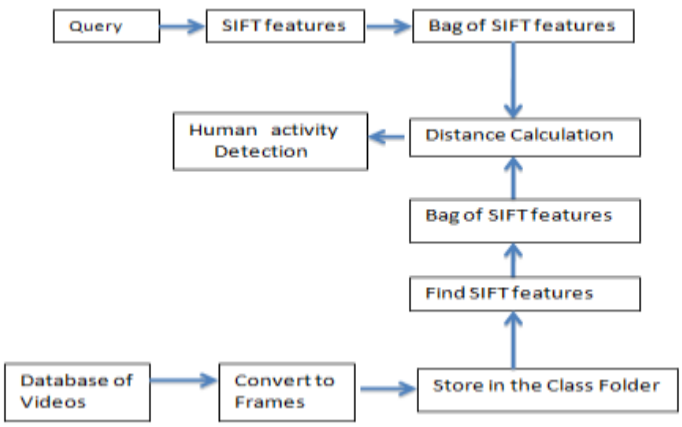

Fig 3.1 Block diagram of overview of the system

The block diagram Fig. 3.1 shows the flow of the project. Database of videos is converted into frames and is stored in the folder according to the classification. The features of the all the frames are found by SIFT algorithm. Once the SIFT features are found then Bag of SIFT features is made. Whenever a query is given the above 3 steps are performed similarly. Once Bag of SIFT features of the query is found then distance is calculated between Bag of SIFT features of the query and the database features using Euclidean distance method. Identification and classification are done by KNN nearest neighbor classifier. The algorithm is as follows:

STEP 1: Video Frame conversion

Frames can be obtained from a video and converted into images using MATLAB function as shown in fig .3.2

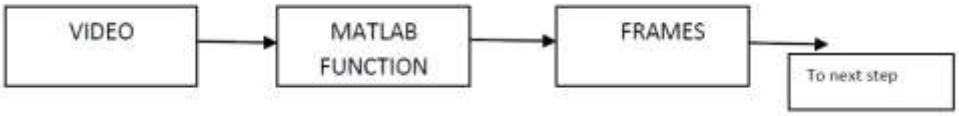

Fig 3.2- Block diagram to show video frame conversion

STEP 2: Saving Frames in the Folder

The frames that are taken from the training and the testing videos are made to store in the folders that are with the frames of the particular class as shown in Fig 3.3

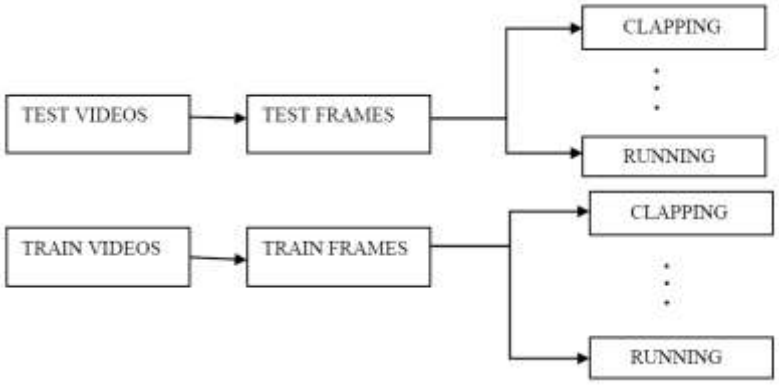

Fig 3 .3- Block diagram to save frames in the folder

STEP 3: Getting image file paths of both Train Image and test image

The image paths of the images that are stored in the folders for both the train and the test images are taken and given to the main file as shown in fig3. 4.

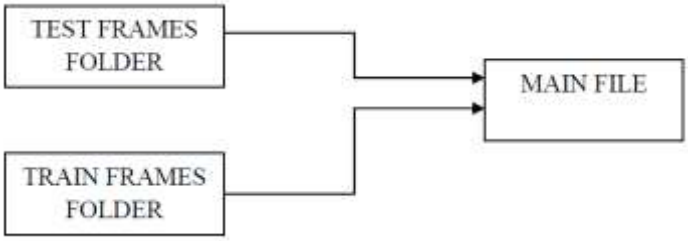

Fig 3.4- Block diagram for providing path to the main file

STEP 4: Computing Bag of SIFTs to both the train and the test image

The features of the image are extracted with the help of an algorithm called SIFT (Scale Invariant Feature Transform).Once features have extracted the Bag of SIFT for both train and test image are found.

We can build BoF with SIFT features as follows.

1. Getting the set of bags of features.

For each feature point, get the SIFT descriptor by extracting the features of all the images which are present in the bigger set of the database. Then for a number of bags, the feature descriptors set are grouped and then bags are trained with feature descriptors which are grouped. To do this job k means algorithm can be used and visual vocabulary is obtained.

2. For given image/video frame getting the BoF descriptor

Here we need to extract SIFT feature points of the given frame and for each point of the feature, the SIFT descriptor is obtained. Vocabulary created in the first step is matched with feature descriptors and the histogram is found.

STEP 5: Finding Euclidean Distance

The importance of Euclidean distance is that it can be embedded with many image classification technique namely PCA, KNN, SVM etc. Here the distance between the bags of SIFT feature of query and database images is found, which will be further used to classify the images using k-NN classifier.

STEP 6: k-NN algorithm for classification and detection of actions
K-NN is the simplest algorithm of all other algorithm used for classification of objects. Here the object is classified based on the major voting of its neighbors where the object belongs to a class among k nearest neighbors. Here once the Euclidean distance is calculated then assigning to a particular class is done by this algorithm and hence the action is detected and classified as jogging or clapping, hand wave or running.

IV. RESULT AND DISCUSSION

We performed an experiment on a database containing videos of different actions. Each category of videos contains more than 500 frames and we could identify the four activities: running, jogging, clapping and hand waving. The parameters like FRR (False Rejection Ratio), FAR (False Acceptance Ratio) and TSR (Total Success Rate) are calculated for different test and train ratios and the tabulation are shown below. We found the system performance is efficient for the threshold (test to train ratio) of 60:40.

FAR, FRR and TSR are calculated as follows.

$$FAR=\frac{Numeberofacceptedimposterclaim}{Totalnumberofimposteraccesses}*100\ldots\ldots\ldots (4.1)$$

$$FRR=\frac{Numeberof\ rejected\ genuineclaim}{Totalnumberofgenuineaccesses}*100\ldots\ldots\ldots (4.2)$$

One more performance parameter, Total Success Rate (TSR) is defined using FAR and FRR which is known as verification rate of the system which is given as follows.

$$TSR=1-[\frac{FAR+FRR}{Totalnumberofaccesses}]*100\ldots\ldots\ldots\ldots (4.3)$$

The train videos are converted into frames and saved in a folder. Train frame of running is selected from the folder, converted into grayscale image for the further processing which is as shown in fig 4.1. Thresholding is done on the gray scale image which converts it into binary image as shown in fig 4.2

Fig 4.1-Original frame of running from a video in running class

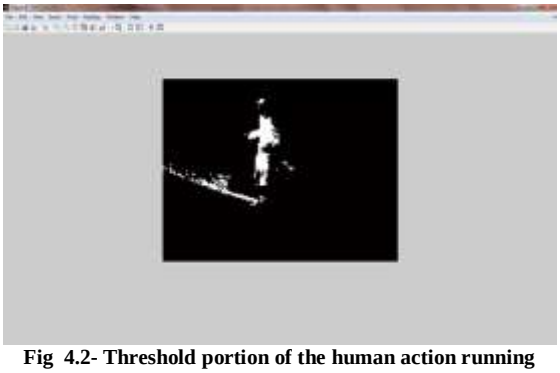

Fig 4.2- Threshold portion of the human action running

For running action identification system is performed efficiently when the threshold is 60:40 with TSR 95% is shown in Fig 4.3 and Table 4.1.

Table 4.1-Human Action Identification (Running)

Threshold	FRR	FAR	TSR
20:80	1	0	53%
30:70	0.13	0.47	64%
40:60	0.07	0.66	84%
50:50	0.04	0.67	86%
60:40	0	0.89	95%

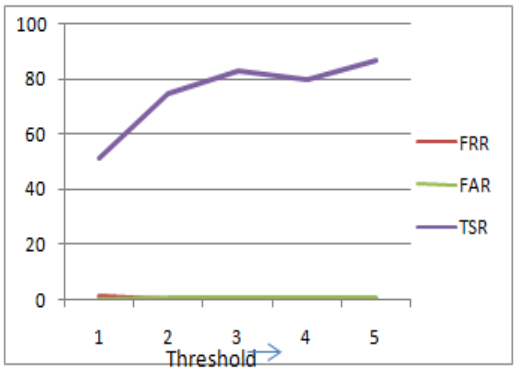

Fig 4.3- Graph showing variation in FRA, FAR and TSR for different Threshold (Running)

Grayscale image for running class is shown in Fig 4.4 and binary image for the same in Fig 4.5.

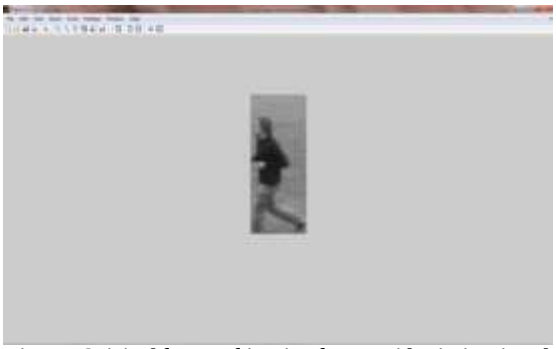

Fig 4.4-Original frame of jogging from a video in jogging class

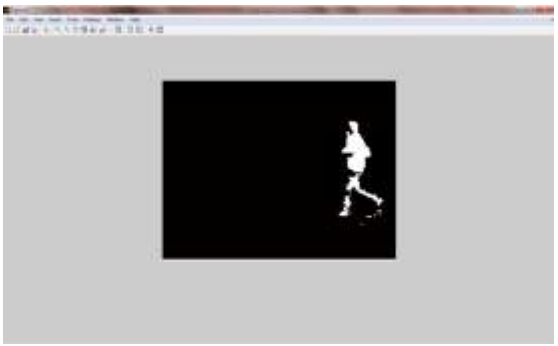

Fig 4.5- Threshold portion of the human action jogging

For jogging action identification system is performed efficiently when the threshold is 60:40 with TSR 87% is as shown in Fig- 4.6 and Table 4.2

Table4. 2-Human Action Identification (Jogging)

Threshold	FRR	FAR	TSR
20:80	1	0	51%
30:70	0.166	0.577	75%
40:60	0.009	0.66	83%
50:50	0.004	0.60	80%
60:40	0	0.83	87%

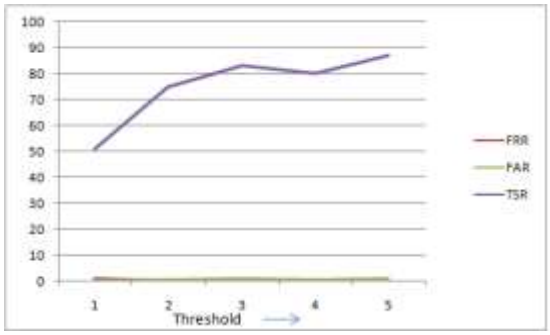

Fig 4.6- Graph showing variation in FRA, FAR and TSR for different Threshold(Jogging)

Gray scale image for Hand clapping is shown in Fig 4.7 and binary image for the same in Fig 4.8.

Fig 4.7-Original frame of hand clapping from a video in clapping class

Figure 4.8- Threshold portion of the human action hand clapping

For hand clapping action, identification system is performed efficiently when the threshold is 60:40 with TSR 91% is as shown in Fig- 4.9 and Table 4.3

Table4.3-Human Action Identification (Hand clapping)

Threshold	FRR	FAR	TSR
20:80	1	0	50%
30:70	0.16	0.444	75%
40:60	0.1	0.60	81%
50:50	0.04	0.64	80%
60:40	0	0.8	91%

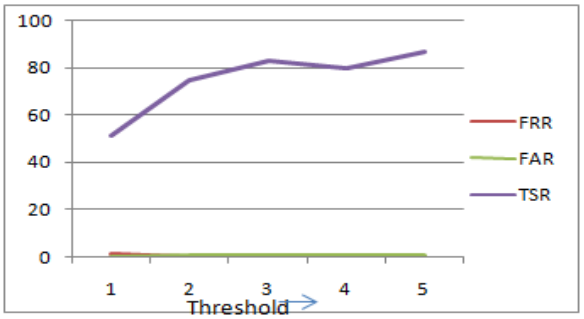

Figure 4.9- Graph showing variation in FRA, FAR and TSR for different Threshold (Clapping)

Grayscale image for Hand waving is shown in Fig 4.10 and binary image for the same in Fig 4.11.

Fig 4.10-Original frame of hand waving from a video in hand wseaving class

Figure 4.11- Threshold portion of the human action hand waving

For hand-waving action, identification system is performed efficiently when the threshold is 60:40 with TSR 85% is as shown in Fig- 4.12 and Table 4.4

Table 4.4-Human Action Identification (Hand waving)

Threshold	FRR	FAR	TSR
20:80	1	0	50%
30:70	0.15	0.43	74%
40:60	0.07	0.56	80%
50:50	0.07	0.57	79%
60:40	0	0.7	85%

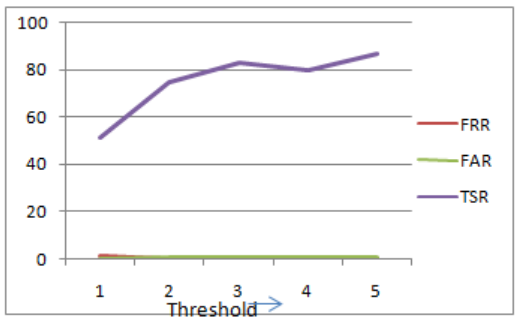

Fig 4.12- Graph showing variation in FRA, FAR and TSR for different Threshold(Hand waving)

CONCLUSION

The Bags of SIFTs and the KNN classifier based implementation were carried out for the identification of human action such as clapping, hand waving, running and jogging using MATLAB simulation. Given the query video, the system searches for the query in the video database and classifies into the class in which it belongs to. The results were positive and the proposed method is efficient. The different performance parameters like FAR, FRR and TSR for all the four classes of videos are calculated and performance found better with the threshold of 60:40.

REFERENCES

[1] Wren, Christopher Richard, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. "Pfinder: Real-time tracking of the human body." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19, no. 7 (1997): 780-785.

[2] Cucchiara, Rita, Costantino Grana, Massimo Piccardi, and Andrea Prati. "Detecting moving objects, ghosts, and shadows in video streams." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25, no. 10 (2003): 1337-1342.

[3] Seki, Makito, Hideto Fujiwara, and Kazuhiko Sumi. "A robust background subtraction method for changing the background." In *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.* pp. 207-213. IEEE, 2000.

[4] Murray, Don, and AnupBasu. "Motion tracking with an active camera." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 16, no. 5 (1994): 449-459.

[5] Kim, Kye Kyung, Soo Hyun Cho, Hae Jin Kim, and Jae Yeon Lee. "Detecting and tracking moving object using an active camera." In *Advanced Communication Technology, 2005, ICACT 2005. The 7th International Conference on*, vol. 2, pp. 817-820. IEEE, 2005.

[6] Yu, Gang, Junsong Yuan, and Zicheng Liu. "Unsupervised random forest indexing for fast action search." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 865-872. IEEE, 2011.

[7] Jones, Simon, and Ling Shao. "Rapid localisation and retrieval of human actions with relevance feedback." In *Computer Analysis of Images and Patterns*, pp. 20-27. Springer Berlin Heidelberg, 2013.

[8] Yamato, Junji, Jun Ohya, and Kenichiro Ishii. "Recognizing human action in time-sequential images using hidden markov model." In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92. 1992 IEEE Computer Society Conference on*, pp. 379-385. IEEE, 1992.

[9] Laptev, Ivan. "On space-time interest points." International Journal of Computer Vision 64, no. 2-3 (2005): 107-123.

[10] Dollár, Piotr, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. "Behaviour recognition via sparse spatio-temporal features." In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65-72. IEEE, 2005.
[11] Lowe, David G. "Distinctive image features from scale-invariant key points. *"International journal of computer vision* 60, no. 2 (2004): 91-110.
[12] Arman, Farshid, Remi Depommier, Arding Hsu, and M-Y. Chiu. "Content-based browsing of video sequences." In Proceedings of the second ACM international conference on Multimedia, pp. 97-103. ACM, 1994.
[13] Shao, Ling, Simon Jones, and Xuelong Li. "Efficient search and localization of human actions in video databases." Circuits and Systems for Video Technology, IEEE Transactions on 24, no. 3 (2014): 504-512.

