International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 09 Issue 11 ǁ 2021 ǁ PP. 72-81

Recognition of Human Action

S. R. Vidhya[1], T. Jayasree[2]
1. Lecturer, Department of Electronics and Communication Engineering, Kamaraj Polytechnic College, Pazhavilai,
2. Assistant Professor, Department of Electronics and Communication Engineering, Government College of Engineering, Tirunelveli,
Corressponding Author: T.Jayasree

ABSTRACT
This paper presents a hybrid strategy for efficient classification of human activities from a given video sequence. All video sequences are divided with respect to the subjects into a training set, a validation set and a test set. A new set of fused features are extracted from sequences using local binary patterns (LBP) with histogram-oriented gradient (HOG). The classifiers were trained on a training set while the validation set was used to optimize the parameters of each method.In this method, KTH datasets are used for training the KNN classifier for human classification and for testing, we utilized static camera pedestrian videos with 25fps frame rate to achieve high recognition rates. The proposed approach is quite simple and achieves state-of-the-art results without compromising the efficiency accuracy. It shows its suitability for real time applications like surveillance systems, human-computer interaction, and traffic control systems. The extensive experiments on a well-known dataset confirmed the superior performance of our method as compared to similar state-of-the-art methods.

Date of Submission: 02-11-2021 Date of acceptance: 16-11-2021

I. INTRODUCTION

A system which intelligently detects a human from an image or a video is a challenging task of the modern era. From the last decade, computer vision and pattern recognition community concentrated on the human detection largely due to the variety of industrial applications, which include video surveillance, traffic surveillance, human-computer interaction, automotive safety, real-time tracking, human-robot interaction, search and rescue missions, humans' collective behaviour analysis, anti-terrorist applications, pedestrian detection, etc. This research addresses human detection in the recorded videos, which is a challenging task in terms of variations in color, movement, appearance, etc. Furthermore, some other complex problems are also considered such as light variations, poor background, etc.

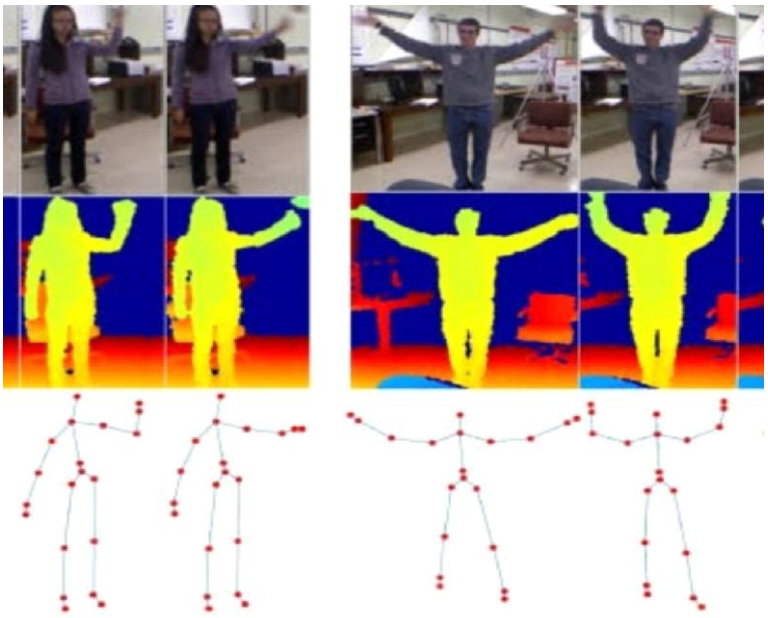


FIG 1HUMAN ACTION RECOGNITION

The recognition of movement can be performed at various levels of abstraction. Different taxonomies have been proposed and here we adopt the hierarchy used by Moeslund et al.: action primitive, action and activity. An action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the movement that is being performed.

Christian Schuldt et al [1] used local space-time features capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. In this paper we demonstrate how such features can be used for recognizing complex motion patterns. Ronald Poppe et al [2] applied vision-based human action recognition image sequences with action labels. Robust solutions to this problem have applications in domains such as visual surveillance, video retrieval and human–computer interaction. The task is challenging due to variations in motion performance, recording settings and inter-personal differences. Maxime Devanne, Hazem Wannaous et al [3] 3D human action recognition is an important current challenge at the heart of many research areas lying to the modeling of thespatio-temporal information. In this paper, we propose representing human, actions using spatio-temporal motion trajectories. Xiaowei Gu et al [4] employed handcrafted and deep learning techniques for HAR.

II. PROPOSED METHODOLOGY

To resolve the above-mentioned problems, we proposed a hybrid methodology which initially enhances the frames to extract the moving objects and later classifies the regions based on feature vectors. The pre-processing step is very important to resolve the problems related to contrast and noise; therefore, we are giving a good weight to this step. Overall, the proposed method is divided into four primary steps:

(a) frame acquisition and enhancement,
(b) segmentation of moving region,
(c) feature extraction and fusion, and
(d) feature selection and action recognition.

Also, in the proposed method, the classification of humans is done with other objects such as vehicles. Our major contributions are enumerated below:

a) Implementation of a contrast stretching technique to make foreground objects (human) maximally differentiable compared to the background.
b) Implementing velocity estimation to identify the motion regions which are later segmented using fusion of uniform distribution-based method and expectation-maximization (EM) segmentation.
c) Utilizing serial-based fusion technique which integrates HOG and texture features with LBP features.
d) Implementation of a joint entropy-PCA-based feature selection, based on maximal score. The selected features are later classified using a KNN for action recognition.
e) A detailed comparison of proposed action classification method with existing algorithms

The selected datasets include KTH, MSR Action, CASIA, INRIA, Weizmann, UIUC, and Muhavi. The proposed method is verified with KNN which acts as a base classifier. The performance of our proposed algorithm is based on multiple measures which include recall rate, false positive rate, false negative rate, accuracy, and precision.

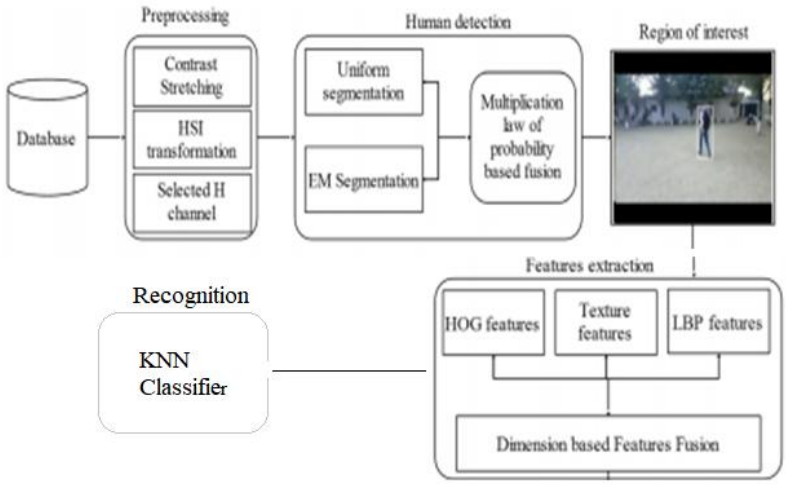


FIG 2 BLOCK DIAGRAM

2.1 PREPROCESSING

Pre-processing is a common name for operations with input video frames at the lowest level of observations. The input frame is captured by the given video sequence, which is originally RGB format. The major aim of pre-processing is an improvement of the frame data that enhances some foreground features for further processing.

2.1.1 FRAME ENHANCEMENT

For the videos, the processing is done frame-by-frame because each processed frame provides us with different results having single or multiple moving objects. For this specific case, each frame can have a different number of moving objects. In the designed algorithm, firstly, image frames are enhanced and then transformed into hue saturation-intensity (HSI) color space. In the first step, contrast enhancement is implemented for each RGB color channel, utilizing histogram equalization detailed with the Algorithm 1.

Algorithm 1. Histogram equalization of gray channels

Step 1: For each color channel with L gray levels and with pixels' intensity value k^1,

$$hist[k^1] = hist[k^1] + 1, \text{ when } i=0 \text{ to } L-1$$

Step 2: The cumulative frequency of histogram H_{cf} is given as:

$$H_{cf}[k^1] = h_{cf}[k^1 - 1] + hist[k^1]$$

Step 3: The equalized histogram is generated by H_{cf} and total number of pixels 'N' in the frame .

$$H_{eq}[k^1] = \lfloor \frac{L * h_{cf}[k^1] - N^2}{N^2} \rfloor$$

Step 4: For each k^1, replace previous values with the new mapping gray value $H_{eq}[k^1]$.

2.2 FRAME SEGMENTATION

In this article, optical flow is used to identify motion of pixels in each frame sequence. After velocity estimation, a segmentation technique is implemented, named as uniform segmentation, which is improved with EMsegmentation. The purpose of segmentation is to collect common features of an image such as color and texture. The fundamental problem faced was "how to exact foreground information with prominent variations in the contrast?" To deal with this problem, the proposed segmentation method worked significantly well. The detailed description of each section presented below

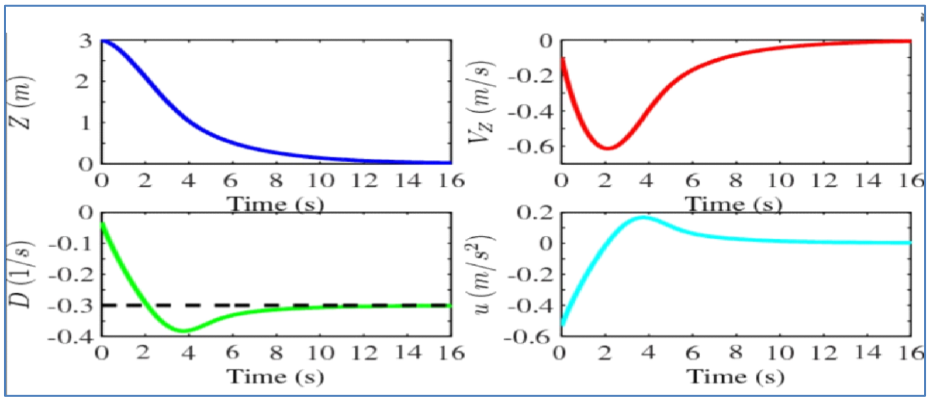


FIG 3 OPTICAL FLOW BASED VELOCITY ESTIMATION

2.2.1 VELOCITY ESTIMATION

To calculate the velocity estimation of motion regions in the video sequences is still an existing research topic in the area of pattern recognition. To estimate the flow of moving objects in the video sequences, we

implemented an optical flow algorithm. The optical flow algorithm identifies the active pixels in the video sequences from time t to time t+1. It provides active pixel information in all three directions as horizontal, vertical, and time. The detailed description of optical flow algorithm is presented in the Algorithm 3, where h and v are motion vectors, hav and vav are the average of four neighbors, ζx, ζy are the displacement functions in x and y direction, ζt is the function of time, γ is the smoothing parameter, P is parallel flow, and D is normal flow. The motion information is segmented by uniform segmentation and EM segmentationand then fuse both segmented frames by implementing multiplication law of probability.

Algorithm 3. Velocity estimation using optical flow

Step 1: $i = 0$

Step 2: Initialize $h^i v^i$ randomly

Step 3: Repeat until Convergence

{

$$h = h_{av} - \zeta_x \frac{P}{D}$$

and v is calculated using equation:

$$v = v_{av} - \zeta_y \frac{P}{D}$$

Where;

$$P = \zeta_x h_{av} + \zeta_y v_{av} + \zeta_t$$

and D is calculated as:

$$D = \gamma + \zeta_x^2 + \zeta_y^2$$

}

Step 4: Return

2.2.2 EM SEGMENTATION

Human detection under different conditions of visual surveillance is a key problem which requires prudent decisions. The proposed technique deals with moving object detection and classification by utilizing consecutive frame subtraction. In the real-time, video frames may contain multiple moving objects, e.g., humans and vehicles, and the proposed hybrid strategy classifies the moving regions with maximum accuracy. The central concept revolves around the detection of the motion vector from the optical flow which is embedded into the video sequence using segmentation of moving regions. For the detection of motion regions, we implement a hybrid technique, which is a combination of uniform distribution and EM segmentation. The implementation of EM segmentation is given as follows:

The EM segmentation is an unsupervised clustering method and utilized for density estimation of the data points. The EM consists of two steps: (1) expectation and (2) maximization. Supposedly, we have a set of observations; in our research, ϕ^H frame is utilized as a input with $\xi_i = \varsigma_i^H$, for i=1 with the ith pixel's value in ϕHchannel. The data are represented as (1×D) matrix where dimension D represent hue pixels in the frame. To calculate the Knumber of mixture densities, the following equation is used:

$$p\left(\varsigma_i, |\phi_j\right) = \sum_{j=1}^{k} \alpha_j p_j\left(\varsigma_{i,}; m_{j,}, \sigma_j\right)$$

where αj is a mixing parameter $\sum_{j=1}^{k} \alpha_j = 1$ for eachGaussian mixture model ϕj =(∂j,mj), where ∂j, mj are the mean and standard deviations of mixtures. The variance is fixed to 1. A K-dimensional binary random variable z is introduced with all zero entries except the Kth entry zj =zj1, zj2, …, zjk. The value of zj satisfies the condition zj∈ [0,1]. The joint distribution p(ξ,z) is defined in terms of marginal distribution p(z) and conditional distribution p(ξ|z) given byp(z)p(ξ|z).

$$\sum_z p(z)p(\xi|z) = \sum_{j=1}^{k} \alpha_j N(\xi|m_{j,}, \partial_j)$$

Let, g(α1,α2,…,α(k− 1);m1,m2,...,mk;σ1,σ2,…,σk) be a vector of estimated parameters. E-Step: Calculate the post probability with heuristic initialized means, fixed variances, and randomly selected alpha. Evaluating the responsibilities:

$$\beta_{ij}^{(u)} = \frac{\alpha_i p(\xi; m_j^{(u)}, \partial_j^{(u)}}{\sum_{j=1}^{k} \alpha_j p\left(\varsigma_i; m_j^{(u)}, \sigma_j^{(u)}\right)}$$

2.2.3 UNIFORM SEGMENTATION

The uniform distribution based segmentation technique utilized for accurate detection of multiple humans in a given scenario. This technique is also well performed in low-resolution sequences and high variation. The uniform segmentation work based on mean and variances of motion regions. The idea behind uniform segmentation is that equally utilized each motion pixel and create a border based on their mean and change in variances. The mean and variances of uniform distribution are calculated as follows:

$$\mu = \int_q^r \phi f(\phi) d\phi, \frac{1}{(r-q)} \int_q^r \phi d\phi$$

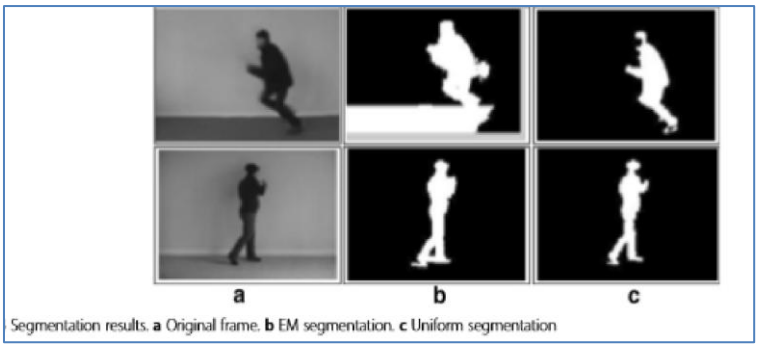
$$= \frac{1}{r-q} \left[\frac{\phi^2}{2}\right]_q^r$$

$$\mu = \frac{r+q}{2}$$

where r and q denote the maximum and minimum motion pixels of the processed frame. After this, both segmented frames are fused by implementing of multiplication law of probability. The fused frame has more information as compared to the individual frame. The basic goal of frame fusion is to improve the detection results in terms of graphical and tabular.

2.2.4 FRAMES FUSION

After foreground segmentation, both segmented frames are fused to get a new foreground which embeds more information compared to a single segmented frame. The main goal of image fusion is to integrate the common information of two images into one image, which contains more information and is easier for human and machine perception compared to individual image [47]. In this article, the fusion of the two segmented frames is done based on the multiplicative law of probability. The fusion using multiplication law is described in the following.



Segmentation results. **a** Original frame. **b** EM segmentation. **c** Uniform segmentation

FIG 4 SEGMENTATION RESULTS

Let m1 denotes the number of ξ(ϕ) pixels, and m2 denotes the number of ϕEM pixels, where ξ(ϕ) is the uniform segmentation frame and ϕEM is the EM segmented frame. Let m3 denotes the matching pixels between ξ(ϕ) andϕEM. Then the fusion of both segmented frames is defined as:

$$\widetilde{\phi_{fusion}} = P\left(\xi(\phi), \widetilde{\phi_{EM}}\right) = \frac{m_3}{n}, \frac{m_3}{n} \times \frac{m_1}{m_1},$$

$$= \frac{m_1}{n} \times \frac{m_3}{m_1},$$

$$\phi_{fusion} \sim = \xi(\phi) \times \mathrm{P}(\xi(\phi)\phi_{EM} \sim)$$

2.3 IMAGE REPRESENTATION

In computer vision, feature extraction is a major step for a description of the input query. For this purpose, many feature extraction techniques are implemented as discussed in literature review. In this article, we extract three types of features as HOG and texture features with LBP features. The texture features are also represented as GLCM (gray-level covariance matrix) features. The HOG features are originally introduced by Dalal et al. which produce shapebased features. The HOG features are implemented in four steps: (1) gradient

computation using magnitude and orientation, (2) cell and blocks creation, (3) votes, and (4) normalize the block size. The block size is 16 × 16, and the size of the input query is fixed at 128 × 64. Hence, the size of the HOG feature vector is 1 × 3780. Secondly, extract local binary patterns (LBP) of detected regions having feature vector size (1 × 59). Whereas, LBP features which are originally introduced to represent human silhouette are more evident and also resolve the problem of contrast of bright objects against a dark background.

The LBP features are calculated as follows:

$\varphi_{u;v} = \sum_{\varphi=1}^{\varrho-1} s(\Phi\Gamma\Phi\lambda)$;where $s(\Gamma)$

$= \begin{cases} 1 & \Gamma \geq 0 \\ 0 & \Gamma < 0 \end{cases}$

$\varrho = 8$, which are the total number of neighboring pixels, $\Phi\lambda$ is the value of the pixels at (u, v), and $\Phi\Omega$ is the value of pixels in the Ωth location on the circle of the radius R around $\Phi\lambda$. The size of LBP feature vectors is 1 × 59 that are further fused with HOG features based on their vector size

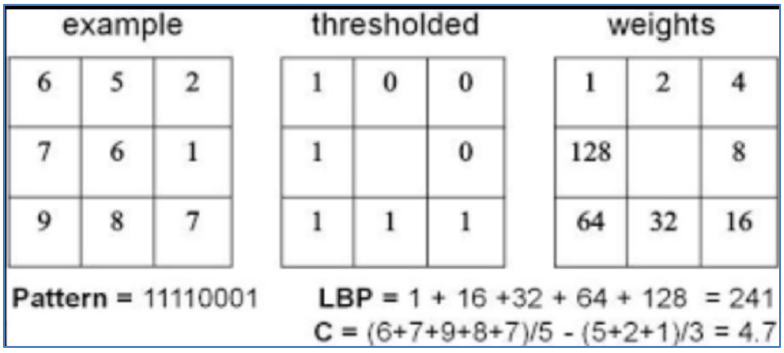


FIG 5 LOCAL BINARY PATTERN

2.4 ACTION CLASSIFICATION

When an image representation is available for an observed frame or sequence, human action recognition becomes a classification problem. An action label or distribution over labels is given for each frame or sequence. This Section discusses approaches that classify image representations into actions without explicitly modeling variations in time. Nearest neighbor**classification k-Nearest neighbor (NN)** classifiers use the distance between the image representation of an observed sequence and those in a training set. The most common label among the k closest training sequences is chosen as the classification. For a large training set, such comparisons can be computationally expensive. Alternatively, for each class, an action prototype can be calculated by taking the mean of all corresponding sequences. The ability to cope with variations in spatial and temporal performance, viewpoint and image appearance depends on the training set, the type of image representation and the distance metric. NN classification can be either performed at the frame level, or for whole sequences. In the latter case, issues with different frame lengths need to be resolved, for example by using majority voting over all frames in a sequence. 1-NN with Euclidean distance are used by Blank et al. for global features and Batra et al. for histograms of codewords. Euclidean distance might not be the most suitable choice given the type of image representation. Bobick and Davis use Hu moments of different orders of magnitude. Mahalanobis distance is used to take into account the variance of each dimension. Rodriguez et al. describe a method to generate spatio-temporal templates that effectively capture the intra-class variance into a single prototype. Several authors have used NN classification in combination with dimensionality reduction. Wang and Suter either use the minimum mean frame-wise distance in an embedded space, or a frame-order preserving variant. Turaga et al. focus on parametric and non-parametric manifold density functions and describe distance functions for Grassmann and Stiefel manifold embeddings. Tran et al. and Poppe and Poel use a learned discriminative distance metric in the NN classification. It has been observed that many actions can be represented by key poses or prototypes. Sullivan and Carlsson recognize forehand and backhand tennis strokes by matching edge representations to labeled key poses. Wang et al. also use edge representations but learn action clusters in an unsupervised fashion. We manually provide action class labels after the clustering. Weinland et al. learn a set of action key poses as 3D voxel representations. These methods use only a single frame for action classification. As many poses are only weakly informative for the action class, considering a sequence of poses over time is likely to reduce ambiguities. Weinland and Boyer use the minimum distance of

each key pose to the frames in the sequences. The set of key poses is discriminatively selected. Lin et al. store prototypes in a tree to allow for efficient matching.

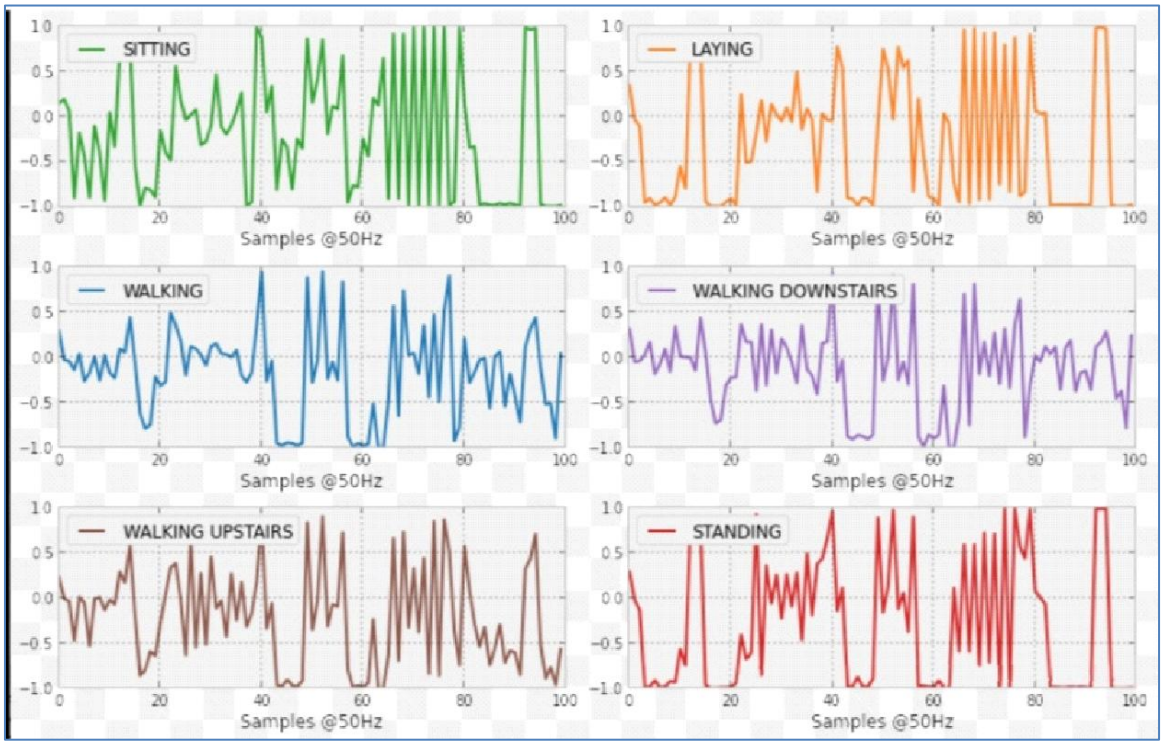


FIG 6 KNN RESULTS FOR 6 ACTIONS

III. DATASETS

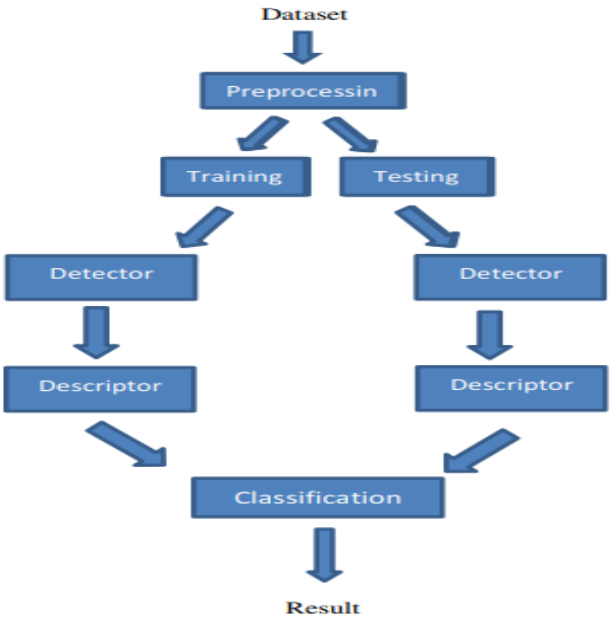


FIG 7 PROCESS OF DATASET

The use of publicly available datasets allows for the comparison of different approaches and gives insight into the (in)abilities of respective methods. We proposed here the most widely used sets i.e. KTH human motion dataset. The KTH human motion dataset contains six actions (walking, jogging, running, boxing, hand waving and hand clapping), performed several times by 25 different actors. Four different scenarios are used: outdoors (s1), outdoors with zooming (s2), outdoors with different clothing (s3) and indoors (s4). There is

considerable variation in the performance and duration, and somewhat in the viewpoint. The backgrounds are relatively static. Apart from the zooming scenario, there is only slight camera movement. Currently the database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25 ps frame rate. The sequences were downsampled to the spatial resolution of 160 x 120 pixels and have a length of four second in average.

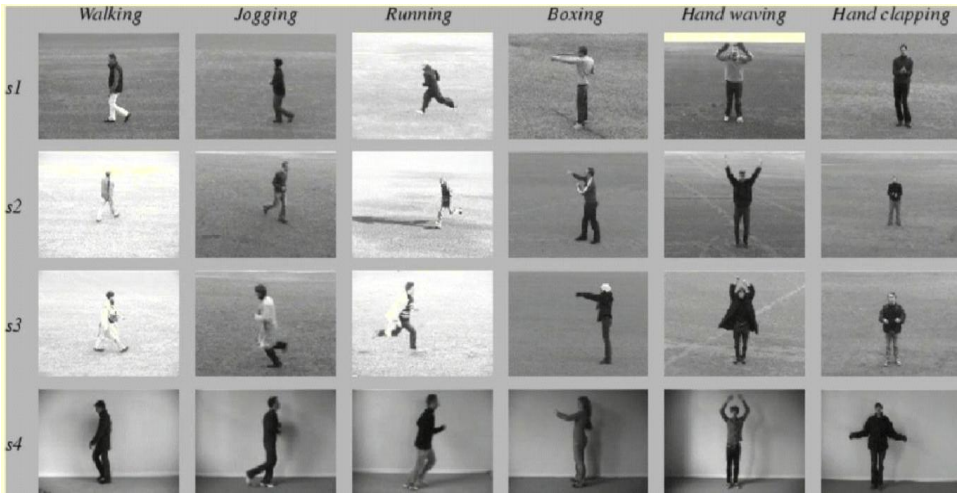


FIG 8 KTH DATASET - 6 ACTIONS IN 4 SCENARIOS

3.1 EXPERIMENTAL APPROACH

Among the sequences perfomed by 25 persons, sequences of 8 persons is grouped as a training set, another set of sequences of 8 persons is grouped into a validation set and the remaining set of sequences by 9 persons is taken as a test set. There are 25x6x4 =600 videos for each combination of 25 subjects(persons), 6 actions and 4 scenarios. Each video contains about four subsequences used as a sequence in our experiments. The subdivision of each file into sequenes in terms of start_frame and end_frame.

The different scenarios are considered as d1,d2,d3,d4

Where, d1 - Static homogeneous background
d2 - Static homogeneous background + Scale variations
d3 - Static homogeneous background+ Different clothes
d4 - Static homogeneous background + Lighting variations

In our experiments, we used the following subdivision of sequences with respect to the subjecti.e among the 25 persons.

Training : person11, 12, 13, 14, 15, 16, 17, 18
Validation : person19, 20, 21, 23, 24, 25, 01, 04
Test : person22, 02, 03, 05, 06, 07, 08, 09, 10

The validation set was used to optimize the parameters of each method. The recognition results were obtained on the test set.

IV. RESULTS AND DISCUSSION

In this work, we detect and classify human action by extract the corner and blob features with suitable learning algorithm, and we could observe that KNN is overcome other classifiers and also the KNN performance is faster than others and the big challenge we face here is classify the similar classes, where the different in accuracy of classification of KNN between similar and non-similar almost 12% higher than other classifiers like SVM. We know to improve the classification process it depends on the characteristics of the data that enter it so there is no classification method that works perfectly with the problems, as mentioned there is a lot of challenge in this subject and we still try to increase the performance. In our experiment, a given video sample is divided into 50 frames. The following is the sequence of 50 frames of a cycling video.

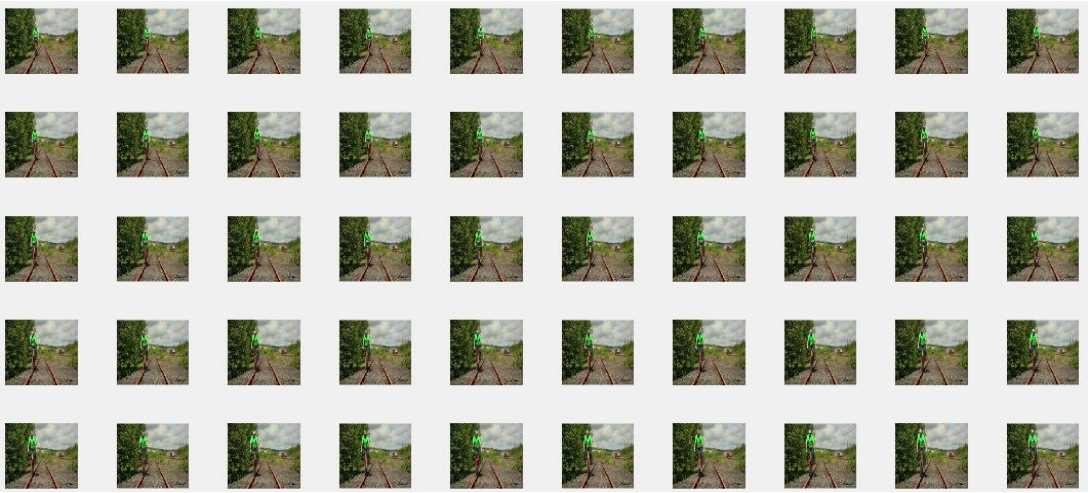


FIG 9. 50 FRAMES OF CYCLING VIDEO

The following is the pixel values of each frame of video sequences.

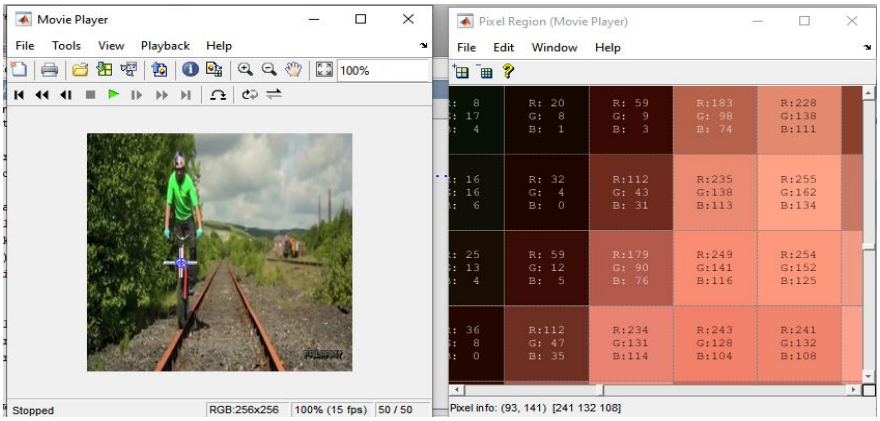


FIG 10 PIXEL REGION

The frames are tested against the KTH dataset using the Confusion matrix of the KNN classifier.

FIG 11 Result of classification of similar action (running and walking)

V. CONCLUSION

In this work, we detect and classify human action by extract the corner and blob features with suitable learning algorithm, and we could observe that KNN is overcome other classifiers and also the KNN performance is faster than others and the big challenge we face here is classify the similar classes, where the different in accuracy of classification of KNN between similar and non-similar almost 12% higher than other classifiers like SVM. We know to improve the classification process it depends on the characteristics of the data that enter it so there is no classification method that works perfectly with the problems, as mentioned there is a lot of challenge in this subject and we still try to increase the performance.

REFERENCES:

[1]. J. Aggarwal and Q. Cai. Human motion analysis: A review. CVIU, 73(3):428–440, 1999.

[2]. S. Belongie, C. Fowlkes, F. Chung, and J. Malik. Spectral partitioning with indefinite kernels using the nystromexten- ¨ sion. In Proc. ECCV, volume 2352 of LNCS, page III:531 ff. Springer, 2002.

[3]. M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using view-based representation. IJCV, 26(1):63–84, 1998.

[4]. Mohiuddin Ahmad, Seong-Whan Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (7) (2008) 2237–2252.

[5]. C. Chen, K. Liu, and N. Kehtarnavaz, ''Real-time human action recognition based on depth motion maps,'' J. Real-Time Image Process., vol. 12, no. 1, pp. 155–163, Aug. 2013.

[6]. A. Farooq, F. Farooq, and A. V. Le, ''Human action recognition via depth maps body parts of action,'' KSII Trans. Internet Inf. Syst., vol. 12, no. 5, pp. 2327–2347, 2018.

[7]. Qluwatoyinpopoola and kejun wang 2012 Video based abnormal human behaviour recognition IEEE.

[8]. Jesper BirksøBækdah 2016 Human Action Recognition using Bag of Features IEEE.

[9]. Ruoyun Gao 2009 Dynamic Feature Description in Human Action Recognition Dynamic Feature Description in Human Action Recognition Leiden Institute of Advanced Computer Science.

[10]. Ronald Poppe and Mannes Poel 2008 Discriminative human action recognition using pairwise CSP classifiers, in (FGR'08) September 2008.

[11]. Zhang Y and Liu Z 2007 Irregular behaviour recognition based on treading track inProc.Int.Conf.WaveletAnal.PatternRecog.

[12]. Wiliem A, Madasu V, Boles W and Yarlagadda P 2008 Detecting uncommon trajectorie in Proc. Digital Image Computing: Tech. and Applications.

[13]. Zhong H, Shi J and VisontaiM,"Detectingunusualactivityinvideo," in Proc. 2004 IEEE Comput. Vis. Pattern Recog., 2008.

[14]. Hara K, Omori T and Ueno R 2002 Detection of unusual human behaviour in intelligent house," in Proc. 12th IEEE Workshop Neural Netw. Signal Process., 2002.

[15]. Y. Hung, C. Chiang, S. J. Hsu, and C. Chan, "Abnormality detection for improving elder's daily life independent," in Proc. 8th Int. Conf. Smart Homes Health Telematics, Jun.22–24, 2010.

[16]. Lee C K, Ho M F, Wen W S and Huang C L 2006 Abnormal event detection in video using N-cut clustering Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.

[17]. Haralick R 1983 Ridges and Valleys on Digital Images Computer Vision, Graphics, and Image Processing.