CONTENT BASED VIDEO CLASSIFICATION

CS4090 Project

Final Report

R.Surya teja(B140600CS) ,Varsha A(B140163CS) ,Vishnupriya Santhosh(B140237CS)
Guided By: Pournami P.N.

May 1, 2018

Abstract

With the development of multimedia data types and available bandwidth there is huge demand of video retrieval systems, as users shift from text based retrieval systems to content based retrieval systems. Selection of extracted features play an important role in content based video retrieval regardless of video attributes being under consideration. In this approach, we are using a content based video processing in which a given video strip is being processed and classified into predefined classes. Here we have discussed methods for video processing including video segmentation, frame detection, shot boundary detection for extracting the key frames of the video strip. Feature extraction methods like Histogram of Oriented Gradients(HOG) and Motion History Images(MHI) used in combination are used to extract features of interest and to be used as inputs of the Support Vector Classifier.

First, the proposed method extracts the Key Frames of the video using the Two Pass Block based Adaptive Threshold Algorithm by frame difference of successive frames of a video. Further the MHI Templates of the corresponding videos are created. The histogram of oriented gradients (HOG) of the MHI is then computed for the Feature Extraction. Finally, Support Vector Machine (SVM) is trained using the Train Set Feature vectors and the Test set is applied on the classifier to check for accuracy.

Keywords: *Motion History Images, Histogram of Oriented Gradients, Support Vector Machine*

1 Introduction

Video is rapidly becoming one of the most popular multimedia due to its high information and entertainment capability. It also consists of audio, video and text together. The importance of content based video classification has swelled out. The significance and recognition of Content Based Video Retrieval has promoted many researches in this arena. In order to make efficient use of video databases, it is crucial to explore efficient ways to index their content based on its features. Human action recognition in video has a huge demand for vision-based applications such as security surveillance, home-care/healthcare monitoring, human-computer interaction, video retrieval and sport/exercise analysis.In this paper, we study the human action recognition problem based on motion features directly extracted from video. Motion History images are used for the purpose of detecting motion in the videos and the motion history templates generated can be subjected to feature extraction techniques like Histogram of Oriented Gradients. The histogram of oriented gradients is a feature descriptor used for the purpose of object detection. The combination of Motion History Images and Histogram of Oriented Gradients are used for Feature Extraction. The extracted features can be used in a Support Vector Machine for the classification.

2 Problem Statement

A given video input can be classified for recognizing the human gestures and an action label can be associated with it. The video is segmented into its constituent frames and the Key Frames representing the video are extracted using the Two Pass Block based Adaptive Threshold Algorithm. These frames are used to generate a Motion History Template for the particular video. The template is further analyzed for features using the method of Histogram of Oriented Gradients and the Feature Vectors generated are used to learn the Support Vector Classifier.

3 Literature Survey

Combination of Text,Images and Audio is called a Video.Content Based Video Retrieval System can be done by different methods of which one is text-based that is done using subtitles and the other one is meta-based method which uses the video.[1] And the third one, Audio-based methods in which different speech recognition extraction techniques and speech as keywords are used. Finally, Content-based method which is integration of all methods mentioned above.From the works of M. Gitte [2] and K. Sanghavi [3] the pre-processing stage of any video retrieval dealt with video segmentation. The complete video is first converted into scenes, then shots and finally shots are converted into various frames. Shot transition detection is utilized to separate up a video into basic temporal units referred to as shots; a shot could be a series of reticulated consecutive footage taken contiguously by one camera and representing never-ending action in time and space[3][4]. It is an elementary step for automatic classification and content-based video retrieval or summarization applications which offer a proficient access to very large video repositories. There are two types of transitions

1. *Abrupt Transitions*: These are unexpected transitions from one shot to another, i. e. one frame belongs to the first shot, and the subsequent frame belongs to the second shot. They are conjointly known as merely cuts.

2. *Gradual Transitions*: Here the two shots are joined using chromatic, spatial or spatial chromatic effects that progressively swap one shot by another. These are soft transitions are classified as wipes, dissolves, fades.

Existing general-purpose CBIR systems[5] roughly fall into two categories namely;

1 *Global Feature Based*: they extract features from the whole image not from certain regions in it; these features are referred to as Global features.

2 *Region Based*: These systems extract features from images at the object-level to overcome the deficiencies of Global Feature based systems. If the decomposition is ideal,it applies image segmentation to decompose an image into regions, which correspond to objects.

FEATURE EXTRACTION

Feature extraction[5] is defined as, extracting compact but semantically valuable information from images. Similar images should have similar information,which is the signature for that images. Images representation needs to consider features which are most useful for the representation of the contents of images and which approaches can effectively code the attributes of the images. Some of the features are texture, shape and color.

MOTION HISTORY IMAGE

Motion History Image[14]is a motion feature for detecting motions in a video. It is a view-based temporal template method.Which is simple but powerful in representing movements. It is widely used by various research groups for action detection, motion analysis and other related applications[13]. Approaches based on matching of a template and then convert an image sequence into a either Motion History Image or Motion Energy Image. Motion History Image and Motion Energy Image are known as static shape pattern.Which is further compare to pre-stored action prototypes during recognition.

HISTOGRAM OF ORIENTED GRADIENTS

Histogram of oriented gradients[13] is a feature descriptor used to find objects in image processing and computer vision. This technique count the no of occurrences of gradient orientation in localized portions of an image detection window, or region of interest (ROI). Histogram of oriented gradients ensures Local object appearance.HOG characterized the shape well compared to edge directions or distribution of local intensity gradient.

Machine learning techniques[5] are helps to derive high level semantic features from the image database.There are two types of machine learning techniques i.e. supervised machine learning technique and unsupervised machine learning technique.

1 *Supervised Machine Learning Techniques*

Neural networks, Decision trees, and Support Vector Machines (SVMs) are some of the supervised machine learning techniques, which learn the high level concepts from low-level image features. The supervised machine learning techniques perform the classification process with the help of the already categorized training data. For the training data, the input (low level image features) and the desired output is already known. Hence, given a query image, the low level features are extracted and it is given as input to any one of the machine learning algorithms which is already trained with the training data. The machine learning algorithm predicts the category of the query image which is nothing but the semantic concept of the query image. Hence instead of finding similarity between the query image and all the images in database, it is found between the query image and only the images belonging to the query image category. Also when the entire database is searched, the retrieval result contains images of various categories.

- Neural Network Neural networks are also useful in concept learning[5]. The low level features of the segmented regions of the training set images are fed into the neural network classifiers, to establish the link between the low level image features and high level semantics.
- Support Vector Machine Support Vector Machines (SVMs) are supervised learning methods used for image classification. It views the given image database as two sets of vectors in an n dimensional space and constructs a separating hy-

per plane that maximizes the margin between the images relevant to query and the images not relevant to the query.

- k-Nearest Neighbors This classification method has been used successfully in image processing and pattern recognition. Based on the training result, KNN is applied for the query data images. KNN helps to classify the input data; also it fixes the code book which means the training result can be self-adapted.

2 *Unsupervised Machine Learning Techniques*
Unsupervised learning refers to the problem of trying to find hidden structure in the unlabelled data. It has no measurements of outcome, to guide the learning process. Image clustering is a typical unsupervised learning technique. It groups the sets of image data in such a way, that the similarity within a cluster should be maximized, and the similarity between different clusters must be minimized.

CLASSIFICATION USING SUPPORT VECTOR MACHINE

In video classification, After feature extraction, the next step is video content modeling.Support vector machines mainly used for pattern classification. It is built by mapping the input patterns into a higher dimensional feature space using a nonlinear transformation (kernel function), and then optimal hyperplanes are built in the feature space as decision surfaces between classes. Nonlinear transformation of input patterns should be such that the pattern classes are linearly separable in the feature space[15]. The goal of a support vector machine is to find a particular hyperplane for which the margin of separation is maximized. Where margin of separation is the separation between the hyperplane and the closest data point.

According to Cover's theorem,nonlinearly separable patterns in a multidimensional space, when transformed into a new feature space are likely to be linearly separable with high probability, provided the transformation is nonlinear, and the dimension of the feature space is high enough. Training and Test data are constitute by splitting the total number of clips of each video genre into half. Small subset of the training data makes support vectors . A Gaussian kernel based support vector machine was constructed for each class by using the one against the rest approach.

During testing phase, given a pattern can result a measure of the distance of the pattern vector from the hyper plane constructed as a decision boundary between the particular class and rest of the classes. If it will result a positive value means that the pattern belongs to the target. Based on the outputs from all the support vector machines for a given pattern vector, the class label corresponding to the model giving the highest positive value can be assigned to the pattern vector.Two different approaches can be used to make decision at video clip level:

1 For each model, Average the values of all pattern vectors belonging to a video clip, the highest average positive value is used to assign the class label for that clip.

2 Assign the class label to the clip based on the highest number of positive outputs per model.

4 Work Done

The input video is first passed in as the input and all constituent frames are extracted from it. The Block based Adaptive threshold method is applied on these frames for the Shot Boundary Detection and the Key Frames representing each of the shots are extracted. The MHI algorithm is run on the Key Frames to generate Motion History Templates. Further, Histogram of Oriented Gradients is run on each of the templates to generate Feature vectors. These Feature vectors are used as inputs for the Support Vector Classifier for performing classification.

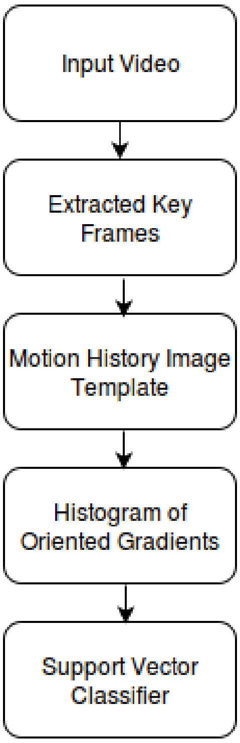

Input Video
Extracted Key Frames
Motion History Image Template
Histogram of Oriented Gradients
Support Vector Classifier

Figure 1: Flow Chart

4.1 Design

4.1.1 Video Parsing

It consists of temporal segmentation of the video contents into smaller units. Video parsing methods extract structural information from the video by detecting temporal boundaries and identifying significant segments, called shots [8].

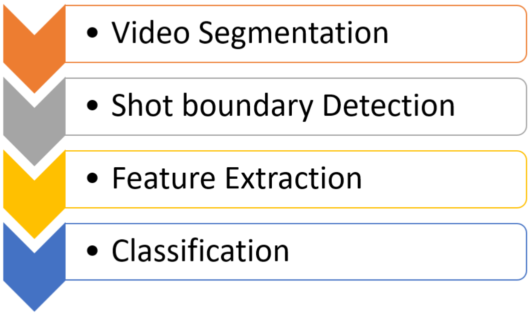


Figure 2: Design

- Shot: A sequence taken by a single camera
- Scene: single dramatic event taken by a small number of related cameras.
- Frame: A still image

Firstly the video is divided into a number of frames i.e these frames when clubbed together yields the entire movie. These frames are then analyzed to get the shots.

Converting to Single Intensity Images
Each pixel of the frame consists of three values i.e RGB values. The given formula is used to convert the RGB values of a pixel into a single intensity. [Equation 1]

$$IntensityVal = \left(\left(\left(\frac{RedVal}{64} \right) * 4 + \frac{GreenVal}{64} \right) * 4 + \frac{BlueVal}{64} \right) \tag{1}$$

4.1.2 Key Frame Extraction

A key-frame is the still image extracted from the video data that best represents the contents of a shot in an abstract manner.This is done using the Shot Boundary Detection Algorithm which is explained below.

Shot Boundary Detection Algorithm

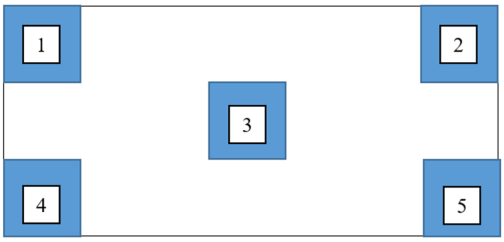


Figure 3: Block division

Shot detection is the process of detecting boundaries between two consecutive shots, so that a sequence of frames belonging to a shot will be grouped together. This partitioning is typically achieved by consecutive measurement of inter-frame variations and learning their variances. Histogram based technique is used for shot boundary detection.The gray level or color histograms of two consecutive frames are computed and if the bin-wise difference between the two histograms is above a threshold, a shot boundary is said to be found[9][11].

`Two Pass Block Based Adaptive Threshold Algorithm`

Finds Consecutive frame differences, using block-based color histogram. We have used a combined step of the statistics- based and histogram-based methods for the shot boundary detection.

`FIRST PASS:`

i Each of the given frame is segmented into 5 blocks i.e 4 corners and the middle. Blocks of size 60x60 are segmented at Top Right, Top Left, Bottom Right, Bottom Left and the Middle [Figure 3].

ii These blocks are further utilized to make a 64-bin color histogram.

iii Compute the accumulated histogram based dissimilarity between the f_m and f_n frames as stated below

$$S_d(f_m,\ f_n) = \sum_{i=1}^{r} B_i * S_p(f_m,\ f_n, i) \tag{2}$$

where, B_i is the predetermined weighting factor for block and $S_p(f_m, f_n, i)$ is the partial match, which has been obtained either by histogram matching for each block and r is total number of blocks (i.e. r = 5)

i	**Bi**
1	1
2	2
3	4
4	2
5	1

Table 1: Block weights

iv Compute similarity difference for all consecutive frames in the video data.

`SECOND PASS:`

i Use the sequence of similarity measures

ii A sliding window of predetermined size 9 over the samples is chosen

iii Find the middle sample, a shot boundary detected if two conditions satisfied

1 The middle sample is the maximum in the window

2 The middle sample satisfies the condition as given in equation 3

$$m_\tau > max\left\{\left(\mu_{left}+3\sqrt{\sigma_{left}}\right),\left(\mu_{right}+3\sqrt{\sigma_{right}}\right)\right\}) \tag{3}$$

Where μ And σ are the mean and standard deviation of the previous left and next right samples of the middle sample within the window.

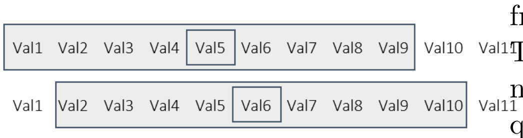


Figure 4: Sliding Window

4.1.3 Feature Extraction

Feature extraction is an approach to diminish the dimensionality of an available set of features by performing inter-feature transformations in order to obtain a new dimensionally reduced representation without largely sacrificing relevant information from the original set.

Motion History Images

The motion history image (MHI) is a static image template helps in understanding the motion location and path as it progresses. In MHI, the temporal motion information is collapsed into a single image template where intensity is a function of recency of motion. Thus, the MHI pixel intensity is a function of the motion history at that location, where brighter values correspond to a more recent motion. Using MHI, moving parts of a video sequence can be engraved with a single image, from where one can predict the motion flow as well as the moving parts of the video action.

The MHI

$$H_\tau(x, y, t)$$

can be computed from an update function

$$\Psi(x, y, t):$$

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \Psi(x,y,t)=1 \\ max(0, H_\tau(x,y,t-1)-\delta & \text{otherwise} \end{cases} \tag{4}$$

Here, *(x,y)* and *t* show the position and time, Ψ(x,y,t) signals objects presence (or motion) in the current video image,the duration τ decides the temporal extent of the movement(e.g., in terms of frames), and δ is the decay parameter. This update function is called for every new video frame analyzed in the sequence. The result of this computation is a scalar-valued image where more recently moving pixels are brighter and vice-versa.

Histogram of Oriented Gradients

In the HOG feature descriptor, the distribution (histograms) of directions of gradients (oriented gradients) are used as features. Gradients (x and y derivatives) of an image are useful because the magnitude of gradients is large around edges and corners (regions of abrupt intensity changes) and we know that edges and corners pack in a lot more information about object shape than at regions.

`Algorithm`

1 Global normalization applied on the image

2 The gradient images in both x-direction and y-direction are computed

3 The histograms of the gradients are computed

4 Normalization is done across the blocks

5 All individual histograms are flattening into a single feature vector

`Algorithmic Overview`

In the first stage, a global image normalization equalization is applied on the image to reduce the influence of illumination effects. This is an optional step.

The first order image gradients are computed in the second stage. An image gradient is a directional change in the intensity or color in an image.

Information from the images are extracted using these image gradients. Gradient images are created from the original images to serve this purpose.For each pixel in the gradient image, the intensity change of the corresponding point in the original image in a given direction is measured. To get the full range of direction, gradient images in both x-direction and y-direction are computed.

The third stage produces an encoding which is sensitively takes care of the local image content while remains resistant to small changes in pose or appearance. Each cell will accumulate a local 1 dimensional histogram of gradients or edge orientations all over the pixels of the cell. The basic orientation histogram is obtained by combining the 1 dimensional cell-level histograms. The gradient angle range is divided into a fixed number of predetermined bins using the orientation histogram. The gradient magnitudes of the pixels in the cell are used to vote into the orientation histogram.

Normalization is computed in the fourth stage. It takes local groups of cells and contrast normalizes their overall responses before passing to next stage.

In the final step, HOG descriptors from all the blocks of a dense overlapping grid of the blocks covering the detection window are collected and combined to form a feature vector for use in the window classifier.

4.1.4 Support Vector Classifier

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. Here the Training Data set is used to learn the Classifier and the predetermined classes are specified. The Test Data Set is input to the classifier to test the classification accuracy. The output of the Classifier is visualized using a confusion matrix which compares the predicted label and true label of the tested videos.

4.2 Implementation

Dataset

Implementation is done using Python as the programming language on the Human action KTH Data set. Data set contains videos, which belong to different categories like walking, running, boxing, hand waving and hand clapping. The video Resolution is 160x120. 160 videos out of the Data set were taken as the Training set, 180 as the Test set and rest for the Validation set.Each video in the Data set is processed using below mentioned steps to find the feature vector.

4.2.1 Video Parsing

Using the openCV module in Python, an input video clip was separated into its constituent frames. These frames may contain redundant information in the form of repeated occurrences of the same frame. Hence we extract Key frames which are representative frames of each shot for the further processing to reduce complexity.

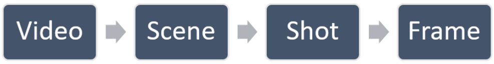


Figure 5: Video Segmentation

4.2.2 Shot Boundary Detection

For the Shot Boundary Detection we are using the Two-Pass Block Based Adaptive Threshold Algorithm which uses the Dugad Model for the detection. Initially, every pixel of each of the Extracted Frames is converted into a single Intensity pixel value.[Equation 1].

In the *First Pass* of the algorithm, the frame differences of the consecutive frames were calculated using the color histogram. The block weighted differences were written into a text file.6

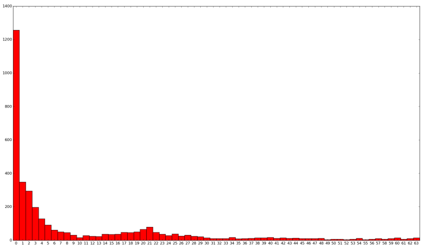


Figure 6: Color Histogram

In the *Second Pass* of the algorithm finds the representative key frames of the video, using a Sliding Window of predefined size 9. This uses the Dugad Model to identify whether a particular frame in the window satisfies the Dugad equation. 3

4.2.3 Feature Extraction

In this stage, for learning the Support Vector Classifier we are extracting certain features from the Extracted Key Frames obtained in the above stage.

Motion History Images
We have obtained the Key Frames of the input video which holds for its richest contents. Motion History Images are generated for the extracted Key Frames which are used to develop a Motion History Template for each of the videos. These templates are used for further pattern matching and comparisons for classifications. The Motion History Templates for each genre of the videos are shown in Figures

Histogram of Oriented Gradients
The Histogram of Oriented Gradients were computed for each of the Motion History Template of the input video and the Feature Vectors obtained were written into a text file. These Feature Vectors are used as inputs for the Support Vector Classifier.

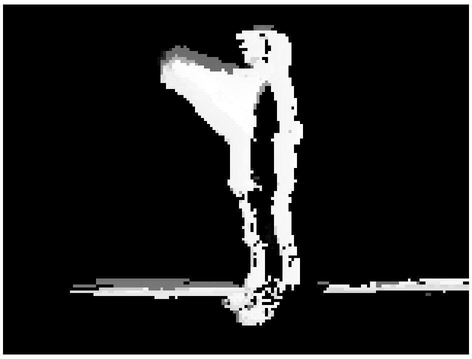


Figure 7: Boxing MHI Template

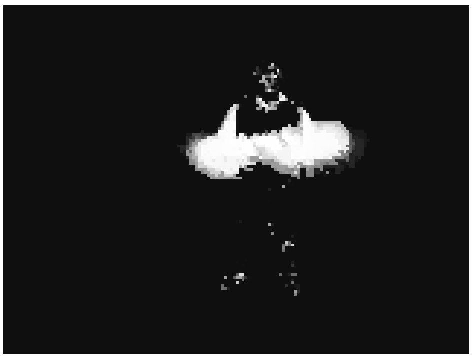


Figure 8: Handclapping MHI Template

We have used the scikit-learn tool,matplotlib,PIL,and openCV for our implementation of Feature Extraction.

4.2.4 Classification using SVM

The Features extracted from the Train Set of Videos are used to form the Train set of features, which are used to learn the Support Vector Classifier. The Classifier defines the set of decision boundaries based on this Training examples. Further the Train Set of data is applied on the Classifier to check the accuracy of the classifier. The accuracy is determined using the Confusion Matrix, which compares the True label and the Predicted label of the input videos.

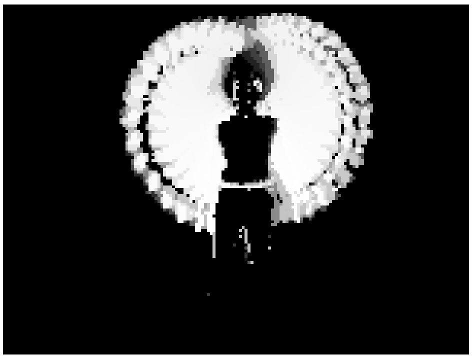


Figure 9: Handwaving MHI Template

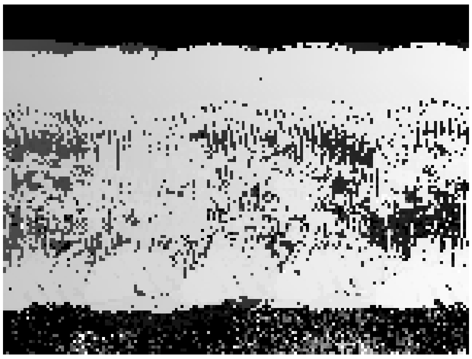


Figure 11: Walking MHI Template

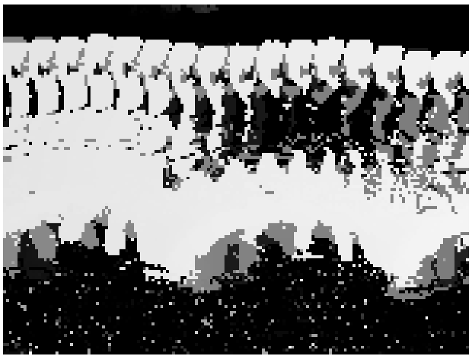


Figure 10: Running MHI Template

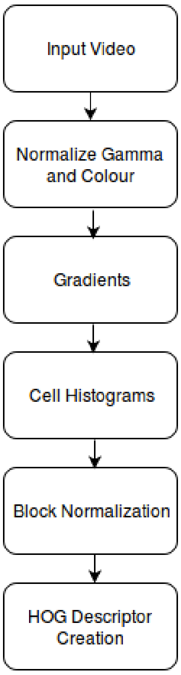
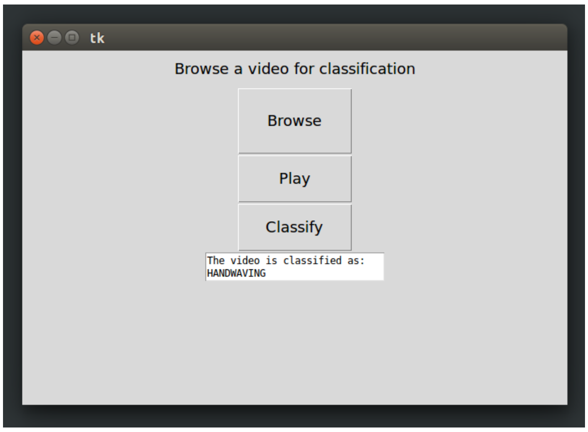


Figure 12: HOG

4.2.5 Graphical User Interface

Tkinter is the standard GUI library for Python. The Graphical User Interface implemented here prompts the user to select a video and input it to the classifier. There are options for Playing the video and to classify the video. The video is passed into the classifier and the genre label associated with the video is obtained as the output. The labelled class name is then displayed.



Browse a video for classification

```
The video is classified as:
HANDWAVING
```

Figure 13: Graphical User Interface

4.3 Results and Analysis

The efficiency of the classifier is analyzed using the generation of the Confusion Matrix. The diagonal of the matrix represents the correct genre classification. In this project, we have used a combined Feature Extraction Technique with Motion History Images and Histogram of Oriented Gradients. The method was found to have correct classification of **86.67%**. Further for testing the category of any video, a Graphical User Interface was implemented which takes up an input video of the users choice and displays the action label associated with it.

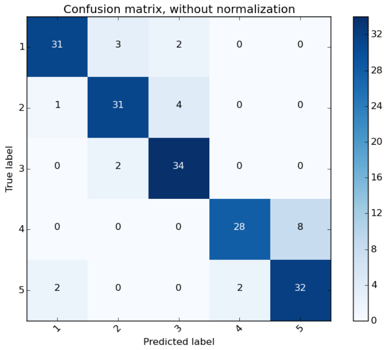


Figure 14: Confusion Matrix

Video Category	Percentage
Boxing	86
Hand Clapping	86
Hand Waving	94
Running	77
Walking	88

Table 2: Percentage of Correct Classification

5 Future Work and Conclusions

The fusion of Motion History Images and Histogram of Oriented Gradients on the Extracted Frames obtained by the Two pass Block based adaptive threshold algorithm was found to give better results than the ones obtained by applying the techniques individually. The motion or movement aspect of the video was captured using the Motion History Images and further feature detection techniques were done using the HOG. Other Feature Extraction Techniques like Optical Flow may also be applied on the MHI Templates to obtain good results.

References

[1] P. Chivadshetti, K. Sadafale, and K. Thakare, Content based video retrieval using integrated feature extraction and personalization of results, *2015 Int. Conf. Inf. Process.*, pp. 170175, 2015.

[2] S. A. Gitte Madhav ,Bawaskar Harshal ,Sethi Sourabh, Content Based Video Retrieval Systems, *Int. J. UbiComp*, vol. 3, no. 2, pp. 1330, 2012.

[3] M. S. Kainjan Sanghavi, Dr. Rajeev Mathur, International Journal of Modern Trends in Engineering and Research, *Int. J. Mod. Trends Eng. Res.*, no. 2349, pp. 645652, 2015.

[4] I. Chugh, R. Gupta, R. Kumar, and P. Sahay, Techniques for key frame extraction: Shot segmentation and feature trajectory computation, *Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. 2016*, pp. 463466, 2016.

[5] S. Tunga, A Comparative Study of Content Based Image Retrieval Trends and Approaches, *Int. J. Image Process.*, vol. 9, no. 3, pp. 127155, 2015.

[6] E. P. Ijjina and K. M. Chalavadi, Human action recognition using genetic algorithms and convolutional neural networks, *Pattern Recognit.*, vol. 59, pp. 199212, 2016.

[7] S. Sadanand and J. J. Corso, Action bank: A high-level representation of activity in video, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. May, pp. 12341241, 2012.

[8] V. D. Amit Fegade, Survey on content based video retrieval, *Int. J. Appl. Eng. Res.*, vol. 9, no. 24, pp. 2805528078, 2014.

[9] G. Rathod and D. Nikam, An Algorithm for Shot Boundary Detection and Key Frame Extraction Using Histogram Difference, *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 8, pp. 155163, 2013.

[10] A. V Kumthekar and P. J. K. Patil, Key frame extraction using color histogram method, *Int. J. Sci. Res. Eng. Technol.*, vol. 2, no. 4, pp. 207214, 2013.

[11] Suresh, Vakkalanka, et al. "Content-based video classification using support vector machines." *International conference on neural information processing.* Springer, Berlin, Heidelberg, 2004.

[12] A. F. Bobick and J. W. Davis, *The recognition of human movement using temporal templates*, In Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 3, (2001), pp. 257-26 7

[13] Huang, Chin-Pan, et al. *"Human action recognition using histogram of oriented gradient of motion history image."*Instrumentation, Measurement, Computer, Communication and Control, 2011 First International Conference on. IEEE, 2011.

[14] Meng, Hongying, et al. *"Motion history histograms for human action recognition."*Embedded Computer Vision. Springer, London, 2009. 139-162.

[15] Latah, Majd. *"Human action recognition using support vector machines and 3D convolutional neural networks."* International Journal of Advances in Intelligent Informatics 3.1 (2017): 47-55.