Video Concept Detection Using SVM and CNN

[1]Jeevan J.Deshmukh, [2]Nita S.Patil, [3]Dr.Sudhir D.Sawarkar
[1]Student, [2]Assistant Professor, [3]Professor
Datta Meghe College of Engineering,Navi Mumbai,India

Abstract—In fast growing digital world, with very high speed internet videos are uploaded on web. It becomes need of system to access videos expeditiously and accurately. Concept detection achieves this task accurately and is used in many applications like multimedia annotation, video summarization, annotation, video indexing and retrieval. The execution of the approach lean on the choice of the low-level visible features employed to show the key-frames of a shot and the preference of method used for extracting the feature. The syntactic differences among low-level features abstracted from video and human analysis of the video data are linked by Concept Detection System. In this proposed work, a set of low-level visible features are of greatly smaller size and also proposes effective union of Support Vector Machine(SVM) and Convolutional Neural Networks (CNNs) to improve concept detection, where the existing CNN toolkits can abstract frame level static descriptors. To deal with the dataset imbalance problem, dataset is partitioned into segments and this approach is extended by making a fusion of CNN and SVM to further improve concept detection. To increase efficiency and to get the result within lesser time the existing systems lags and so using the video reader to extract the frames. Frame undergoes Hu_moments and HSV histogram to produce feature vector for classification. This paper makes two contributions first, the two classifiers SVM and CNN are separately trained on data set and this enriches efficient result. The accuracy of each classifier is individually calculated. Second, the fusion of two classifiers is performed to efficiently detect the concepts in test dataset. After the fusion of two classifiers, the accuracy is calculated. The proposed framework using fusion of SVM and CNN gives effective video concept detection. Accuracy is used as measure to evaluate the system performance for UCF 101dataset. The fusion of CNN and SVM classifiers provides better results in comparison with individual classifier. The proposed framework is validated on standard UCF 101 dataset using accuracy as predictive measure.

IndexTerms—Support vector machine; Video Concept Detection; Convolutional Neural Network; Key Frame Extraction; Feature Extraction.

I. INTRODUCTION

Videos are becoming trendy for entertainment from several years. The Development of the video browsing and indexing application is growing faster as the end user feel necessity for better control over the video data. The Methods like video indexing, video browsing and video retrieval are taken into more consideration for Content-based video analysis due to rapid growth of video data. In video indexing, retrieval and annotations etc, for these video based applications automation of identifying semantic concepts for video samples is the prominent thirst research area. Image processing is an approach to process digital images by applying few techniques to procure enhanced images and valuable information. Useful knowledge can be fetched from these images by employing classifying schemes. Image processing is a part processing of signals where image is taken as input, and yield might be image or features of image. Initially, segmentation of video in shots is performed and later extraction of key-frame for individual shot is carried out separately. In conquering the semantic gap, the content based schemes faces difficulties for retrieval of semantics depending on color, texture as low level features. The latest schemes perform semantic search using concept detectors such as bike, car and airplane for fetching of semantics from low level features. The goal of concept detection system is to define video shots with many concepts, selected from list stored in database. In concept detection system, first video is divided into shots and from each shot key-frame is extracted which uniquely identifies the shot. The key-frame is used to extract the different features of shot. With the help of extracted features training of classifiers is done individually for every concept. After training of classifiers, it labels the video shot and also gives the score for each concept. Markatopoulou, Foteini [1] discussed the architecture for video concept detection that has enhanced the computational complexity as matched with typical state-of-the-art late fusion architectures. Figure 1 shows the basic architecture of concept detection system.

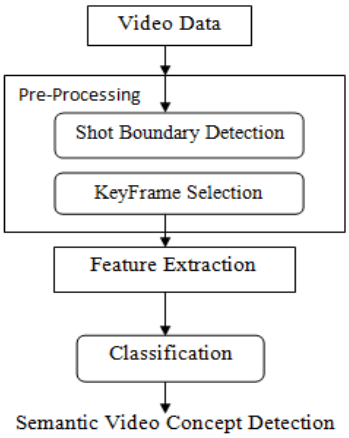

Fig.1: Architecture of concept detection system

Tong, Wenjing, et al. [2] proposed a novel video shot boundary detection method where CNN model is used to generate frames' TAGs. It is efficient to find out both CT and GT boundaries and also combines TAGs of one shot for implementation of video annotation on that shot. Karpathy, Andrej, et al. [4] used CNN for large-scale video classification_where CNN architectures are efficient to learn persuasive features from weakly-labeled data that is the best feature based methods in execution. Krizhevsky, Alex, Ilya, et al. [3] proposed deep Convolutional neural network and presented that deep-CNN is efficient to accomplish best output on highly challenging data sets using completely supervised learning. Xu, Zhongwen, et al. [5] is first to use CNN descriptors for video representation and proposed that CNN descriptors are generated more accurately with help of suppressed concept descriptors. D. Ciresa et al. [6] proposed that the DNN is generic image classifier with raw pixel intensities as inputs, without ad-hoc post-processing. C. Snoek et al.[7][8] discussed to develop mapping functions from the low-level features to the high-level concepts with some machine learning techniques for concept detection or high level feature extraction. N. Janwe et al. [9] proposed that in concept detection method the concept detection rate is directly controlled by semantic gap. The semantic gap is controlled by considering set of low level visual feature of very smaller size and selecting the feature-fusion methods like hybrid-fusion to improve performance of concept detection.

The concept probabilities for a test frame are produced by classifiers like SVM. The modern concept detection system consists of low-level feature extraction, classifier training and weight fusion. Earlier researchers focused on improving accuracy of the concept detection system using global and local features obtained from key-frame or shot of the video and various machine learning algorithm. In recent times, due to the technological advances in computing power deep learning techniques specially Convolutional Neural Network (CNN) has shown promising improvement in efficiency in various fields. CNN has the powerful ability of feature extraction and classification on large amount of data and hence widely adopted in concept detection systems. The Proposed method is to incorporate SVM and CNN for the challenging video concept detection problem due to its known ability to classify feature vector and gives efficient output. To increase efficiency and to get the result within lesser time the existing systems lags and by using the video reader to extract the frames. Frame undergoes Hu_moments and HSV histogram to produce feature vector for classification. The video frames undergo different layers the CNNs for descriptor extraction. Convolutional Neural Network a category of Neural Networks in which feature are extracted by weights of the convolution layer and the fully connected layer is used for classification.

The Proposed method follows four steps. First step is key Frame extraction from video. The Next step is feature extraction using CNN and HSV model to produce feature vector for classification. The Next Step is Classication using SVM and CNN. Next is fusion of classifiers. The rest of the paper is described below. Section II discusses the proposed method. In Section III experimental results are presented. Section IV contains conclusion and future work.

II. PROPOSED METHOD

The figure 2 shows the proposed system methodology. It shows the flow of system where video is taken as input, it undergoes different steps and processes and in output step with help of SVM and CNN concept of video is concluded.

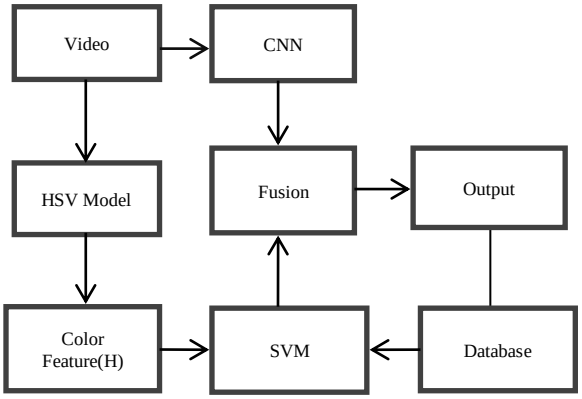

Fig.2: Proposed System methodology

A. Key Frame extraction

The key-frame is the frame for depicting outstanding content and data of the shot. The key-frame can be visual index as it enables simple browsing and navigation through its organization. Video consist of Number of frames. The group of Frames will form a shot and set of shots will form a scene. The same shot contains many similar frames; therefore positive frames that perfectly revert the shot contents are selected as key-frames. Some times; shot is represented by many key-frames as per requirement. The selection of a key-frame may also base on the object or the event end user required. Whichever frame that perfectly represents the object or the event can be chosen as a key-frame. Firstly, video is divided into shots. From the shot key frame is extracted. Extracted key-frame further processed in system as well as features are extracted from key-frame. In proposed system first I-frame is consider as key-frame in video. Features extracted from key-frame are given to classifier for generating the result.

B. Feature extraction and feature vector generation

a) HSV Model

The main part of video indexing and retrieval is extraction of features as per result of video structural analysis. This system concentrates on visible features of key-frames, objects and motions useful for video indexing and retrieval. In HSV model, Hue is standard of the wavelength appeared in dominant color collected by sight while saturation is height of the size of white light mixed in hue. In Mathematical field Image is function of two dimensions which is in continuity with intensity of light in the field. In order to process image digitally through computer it must be presented with discrete values. Hue_moments are group of 7 numbers that are calculated with help of central moments that are proportional to image transformation. It is proved that first 6 moments are proportional to translation, scale, reflection and rotation where 7^{th} moments flag changes for image reflection. The Hue represents the color type. The generated Hue_moments and HSV histogram is given to SVM for classification. The Key-frame extracted from shot is given to HSV. From input frame hu_moment and HSV histogram is calculated then harlick method is applied to convert image into grayscale and harlick texture feature vector is generated. From image HSV histogram is computed then frame matrix is formed and given to classifier to predict output. In CNN, image in matrix form is taken as input to Convolutional layer. Features are extracted at Convolutional layer and pooling layer by applying different filters. Generated feature map is passed to fully connected layer to predict the class.

III. CLASSIFICATION OF FEATURE VECTOR

a) Support Vector Machine(SVM) Classifier

Support vector Machine is selective classifier specified by separating hyper-plane. The proper labeled training data is given and algorithm outputs with optimal hyper-plane which categorizes into two parts. In two dimensional plane cases, the hyper-plane divides space into two sets where each group lay in either side.

The optimal hyper-plane is searched by the SVM and that hyper-plane forms two distinguished classes of *n*-dimensional feature space. One class shows the concept under consideration and second class means rest of the concepts, i.e. $yi = \pm 1$. A hyper-plane is said to be optimal when the distance to the nearby training examples is maximized for both classes. This distance is called the margin. The margin is specified by the support vectors, $\lambda i > 0$, which are obtained by optimizing:

$$\text{Min } \lambda \left(\lambda^T \Lambda K \Lambda \lambda + C \sum_z \xi i \right)$$

during training under the constraints: $yig(xi) \geq 1-\xi i$, $i=1,2,\ldots,z$, where Λ is a diagonal matrix containing the labels *yi*, *C* is specified to balance training error and to model complexity, *z* is the overall shots in the training set, when the data is not ideally separable, slack variables are received and is represented by ξi., and for all training pairs, *K* is the matrix which stores the values of the kernel function $K(xi,x')$. It is of interest to note the implication of this kernel function $K(\cdot)$, as it maps the distance between feature vectors into a higher dimensional space in which the hyper-plane separator and its support vectors are gained. Once the support vectors are admitted, it becomes easy to specify a decision function for an unseen test sample x'. Hue features were utilized to build SVM classifiers. The simple scaling of the training is planned to build SVM classifier and test dataset feature vectors. The appropriate kernel function is used like RBF or linear kernel function. Cross-validation is used to find the correct parameters for C and gamma. Use the appropriate values of C and gamma (γ) to train the entire training set and after training class is predicted for the test sample. In SVM, global features Hue_moments and HSV histogram extracted from key-frame are given to SVM classifier to predict the output. In proposed framework, RBF kernel is used with optimized values of C and γ to correctly predict the output.

b) Convolutional Neural Network(CNN)

Convolutional Neural Networks are used in image recognition, Image classification, Object Detection.CNN classifier takes image as input, process the image and classify image into certain category. Computer takes input image as array of pixels and that depends on resolution of image. Input image looks like h*w*d where h=height, w=width, d=dimension. Height and width are considered as special dimensions for feature extraction. To Train and Test the input image it undergoes series of Convolutional layers with kernels, pooling layer and fully connected layer then object is classified by applying Softmax function with probabilistic values ranges between 0 to 1. Figure 3 shows the basic architecture of CNN. At output layer, classifier predicts the class of image.

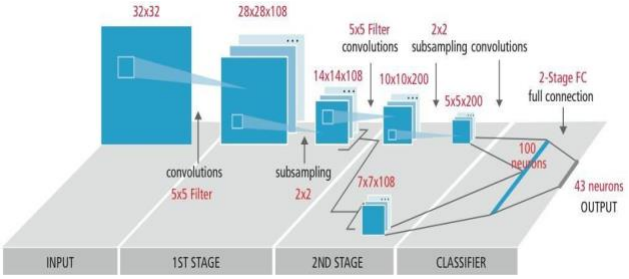

Fig.3: Basic Architecture of CNN

b.1) Convolution Layer

The very first layer is Convolutional Layer where features are extracted from input image. In this layer feature map is generated by convolving filter with the input. In mathematical form convolution takes two inputs as image matrix and filter. Convolution of image using different filters process activities like edge detection, blur and sharpen image. Sometimes filter does not fit with input image so pad the matrix with zero so that filter fits and also drop the part of image matrix where filter does not fit is a valid padding which maintains only valid part of matrix. CNN is type of Neural Networks where features are extracted by using weights of convolution layer.

b.2) Pooling Layer

The pooling layer performs reduction of dimensionality size of image matrix. The number of parameters is decreased when image is too large in pooling layer. Max pooling and average pooling are two different types of pooling used in CNN. In Max pooling largest element is taken from feature map.

b.3) fully Connected layer

In FC layer feature map matrix from pooling layer is flattened into vector and feed into fully connected layer such as in neural network. Here, features are combined together to create a model and then activation function like Softmax is used to classify the output into desired class. In The fully connected layer the output matrix from pooling layer is converted into vector and classification function being used for classifying it into appropriate class which is determined during the training process.

The CNN architecture consists of 7 layers. CNN is trained during training phase with set key frames and in test phase CNN gives desired prediction probability. In fully connected layer every neuron in the previous layer is connected to every neuron on the next layer. The output from the Convolutional and pooling layers represent high-level features of the input image.

IV. EXPERIMENTAL RESULT

Experiment is implemented on python environment. The proposed frame work tested on standard UCF-101 dataset. In system, for training 537 videos and for testing 205 videos of different classes are used. Table 3, 4 and 5 shows the confusion matrix of SVM, CNN and fusion of SVM and CNN respectively. In experiment five classes are used. Table 3 presents output confusion matrix of five classes for SVM classifier with obtained accuracy is 0.47.

Data Set:

Table 1: Detail partition of UCF-101 data set

Dataset	Partition	Name of Dataset	No. of Key-Frames
UCF-101 Dataset	Partition-1	Training Dataset	**537**
	Partition-2	Testing Dataset	**205**

Table 2: different classes used in dataset

Class Name	No. of Training Videos	No. Of Testing videos
The Archery	104	41
Baseball Pitch	107	43
Cricket Bowling	103	36
Cricket Shot	118	49
Kayaking	105	36

Table 3: Confusion Matrix for SVM

Actual Class	Predicted Class				
	Archery	***Baseballpitch***	***CricketBowling***	***Cricketshot***	***Kayaking***
Archery	20	9	0	5	7
Baseballpitch	16	18	5	2	2
CricketBowling	5	0	12	14	5
Cricketshot	8	9	4	26	2
Kayaking	12	0	0	3	21

Table 4 presents the output confusion matrix for classifier CNN with obtained accuracy is 0.476

Table 4: Confusion Matrix for CNN

Actual Class	Predicted Class				
	Archery	***Baseballpitch***	***CricketBowling***	***Cricketshot***	***Kayaking***
Archery	2	9	16	2	12
Baseballpitch	1	31	1	1	9
CricketBowling	0	0	48	1	0
Cricketshot	0	0	31	4	1
Kayaking	4	3	6	10	13

Table 5: Confusion Matrix for fusion of SVM and CNN

Actual Class	Predicted Class				
	Archery	***Baseballpitch***	***CricketBowling***	***Cricketshot***	***Kayaking***
Archery	21	8	0	5	7
Baseballpitch	10	33	0	0	0
CricketBowling	0	0	35	1	0
Cricketshot	7	9	4	27	2
Kayaking	6	0	0	2	28

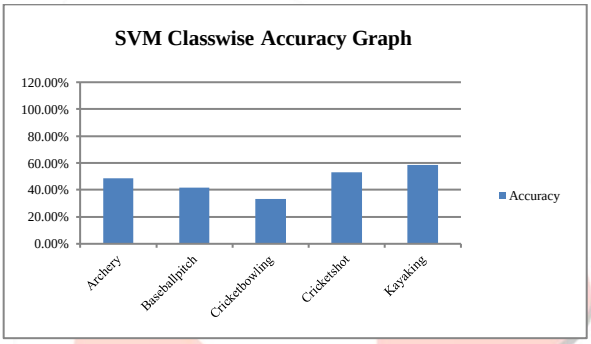

Fig.4 SVM class wise accuracy graph

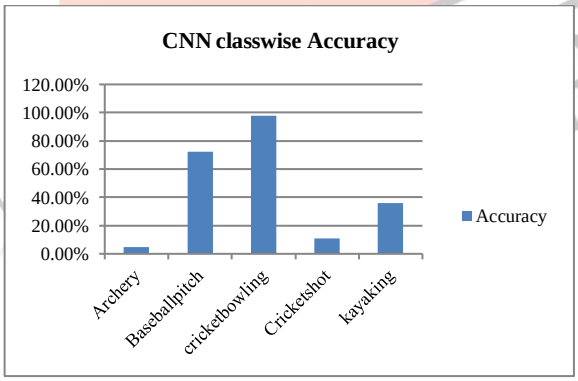

Fig.5 CNN class wise accuracy graph

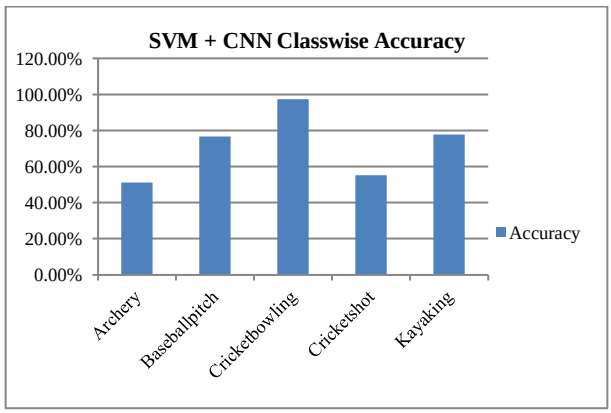

Fig.6 class wise accuracy after fusion of SVM and CNN

V. CONCLUSIONS

In this work, video concept detection technique is presented by using Support Vector machine (SVM) and Convolutional Neural Network (CNN). A set of low-level visible features are of greatly smaller size and also proposes effective union of SVM and CNNs to improve concept detection, where the existing CNN toolkits can abstract frame level static descriptors. Initially SVM is developed using global features like Hue_moments and HSV histogram extracted from key-frame. CNN is developed using extracted key-frames from videos. The two classifiers SVM and CNN are separately trained on data set and this enriches efficient result. The accuracy of each classifier is individually calculated. The fusion of two classifiers is performed to efficiently detect the concepts in test dataset. After the fusion of two classifiers, the accuracy is calculated. The proposed framework using fusion of SVM and CNN gives effective video concept detection. The proposed framework is validated on standard UCF 101 dataset using accuracy as predictive measure. The fusion of CNN and SVM classifiers provides better results in comparison with individual classifier.

REFERENCES

[1] Markatopoulou, Foteini, Vasileios Mezaris, and Ioannis Patras. "Cascade of classifiers based on binary, non-binary and deep convolutional network descriptors for video concept detection." *2015* IEEE International Conference on Image Processing (ICIP). IEEE, 2015.

[2] Tong, Wenjing, et al. "CNN-based shot boundary detection and video annotation." 2015 IEEE international symposium on broadband multimedia systems and broadcasting. IEEE, 2015.

[3] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[4] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.

[5] Xu, Zhongwen, Yi Yang, and Alex G. Hauptmann. "A discriminative CNN video representation for event detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[6] Ciresan, Dan, et al. "Deep neural networks segment neuronal membranes in electron microscopy images." Advances in neural information processing systems. 2012.

[7] Snoek, Cees GM, and Marcel Worring. "Concept-based video retrieval." Foundations and Trends® in Information Retrieval2.4 (2009): 215-322.

[8] Snoek, C. G. M., et al. "MediaMill at TRECVID 2013: Searching concepts, objects, instances and events in video." NIST TRECVID Workshop. 2013.

[9] Janwe, Nitin J., and Kishor K. Bhoyar. "Neural network based multi-label semantic video concept detection using novel mixed-hybrid-fusion approach." Proceedings of the 2nd International Conference on Communication and Information Processing. ACM, 2016.