

Semantic Concept Detection in Video Using Hybrid Model of CNN and SVM Classifiers

Nita S. Patil

*Datta Meghe College of Engineering
Airoli, Navi Mumbai, India*

nsp.cm.dmce@gmail.com

Sudhir D. Sawarkar

*Datta Meghe College of Engineering
Airoli, Navi Mumbai, India*

sudhir_sawarkar@yahoo.com

Abstract

In today's era of digitization and fast internet, many video are uploaded on websites, a mechanism is required to access this video accurately and efficiently. Semantic concept detection achieve this task accurately and is used in many application like multimedia annotation, video summarization, annotation, indexing and retrieval. Video retrieval based on semantic concept is efficient and challenging research area. Semantic concept detection bridges the semantic gap between low level extraction of features from key-frame or shot of video and high level interpretation of the same as semantics. Semantic Concept detection automatically assigns labels to video from predefined vocabulary. This task is considered as supervised machine learning problem. Support vector machine (SVM) emerged as default classifier choice for this task. But recently Deep Convolutional Neural Network (CNN) has shown exceptional performance in this area. CNN requires large dataset for training. In this paper, we present framework for semantic concept detection using hybrid model of SVM and CNN. Global features like color moment, HSV histogram, wavelet transform, grey level co-occurrence matrix and edge orientation histogram are selected as low level features extracted from annotated groundtruth video dataset of TRECVID. In second pipeline, deep features are extracted using pretrained CNN. Dataset is partitioned in three segments to deal with data imbalance issue. Two classifiers are separately trained on all segments and fusion of scores is performed to detect the concepts in test dataset. The system performance is evaluated using Mean Average Precision for multi-label dataset. The performance of the proposed framework using hybrid model of SVM and CNN is comparable to existing approaches.

Keywords: Semantic Concept Detection, SVM, CNN, Multi-label Classification, Deep Features, Imbalanced Dataset.

1. INTRODUCTION

The semantic concept detection system detects the concepts presents in key-frame of the video based on features and assigns automatic labels to video based on the predefined concept list. Human can assign labels based on the visual appearance with the experience. But automatic semantic detection systems performs mapping of the low level features to high level concepts using machine learning techniques. Such systems are useful for many applications like semantic indexing, annotations and video summarization and retrieval.

Semantic concept detection system works on bridging the semantic gap by performing mapping of low level features to high-level video semantics. Extensive research in this field has improved the efficiency of semantic concept detection systems but it is still a challenging problem due to the large variations in low level features of semantic concepts and inter concept similarities.

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 years, due to the technological advances in computing power deep learning techniques specially Convolutional Neural Network (CNN) has shown promising improvement in efficiency in various field. CNN has the powerful ability of feature extraction and classification on large amount of data and hence widely adopted in concept detection systems.

Systems with unbalanced dataset have less relevant examples as compared to irrelevant examples. This limits classifier accuracy during training phase and the created classifier model may give misclassification. Researchers used methods to deal with unbalanced dataset problem mostly based on over sampling positive examples and down sampling negative samples which may lead to over fitting of the classifier.

This paper proposes framework for effective feature extraction and classification dealing with imbalanced dataset problem. Section 2 covers related work. Section 3 discusses basic semantic concept detection system. Section 4 focuses on the methodology and concept selection for segments as well as the generation of concept training data. Section 5 presents the results of the experimental evaluation of the concept detector system. Finally, Section 6 concludes this paper by discussing the key aspects of the presented concept detection system.

2. RELATED WORK

Feature extraction and feature selection is fundamental and important step in concept detection task. In this section we discuss deep learning framework and semantic concept detection using traditional handcrafted features and deep learning methods.

2.1 Deep Learning Frameworks

Recently Deep Convolutional Neural Networks (DCNN) has shown promising results in various fields because of its ability to extract high level features effectively from video and images.

Deep learning frameworks like Cafe [1] ,Theano [2], Cuda-convnet are adopted in various video processing applications.

Deep convolutional networks proposed by Krizhevsky et al. [3] implemented on Imagnet dataset in the ILSVRC-2010 and ILSVRC-2012 competitions set benchmark in deep learning. Krizhevsky et al. achieved the best results and reduced the top-5 test error by 10.9% compared with the second winner.

A Deep Convolutional Activation Feature (Decaf) [4] was used to extract the features from an unlabeled dataset. Decaf learns the features with high generalization and representation to extract the semantic information using simple linear classifiers such as Support Vector Machine (SVM) and logistic Regression (LR).

Jia et al. [1] proposed Cafe, a Convolutional Architecture for Fast Feature Embedding which contains modifiable deep learning algorithms, but also has several pretrained reference models. Another model for object detection is (R-CNN) [5], which extracts features from region proposals to detect semantic concepts from large datasets.

2.2 Low Level Feature Based Classifiers

Video includes information from various modalities like visual, audio, text. Each modality has information that is useful to improve detection accuracy and their joint processing can boost performance and help uncover relationships that are otherwise unavailable. Conventionally low-level visual features of color, texture, shape and edge are acquired from entire key-frame globally or from region, grid or keypoints of key-frame. Traditional global features of color, texture and shape extracted from segmented video are used by many researchers [6],[7],[8]. Concept detection based on features extracted around keypoints show good detection results using

descriptors like SIFT, SURF, ORB [9][10][11][12]. Duy-Dinh Le et al. [13] evaluated the performance of global features and local features for the semantic concept detection task on Trecvid dataset from 2005 to 2009. They discussed the performance of the individual global features like color moments, color histogram, edge orientation histogram, and local binary patterns on grid of varying size, various color spaces including HSV, RGB, Luv, and YCrCb and variation in bin sizes. The local feature used is SIFT with BOW. They also considered late fusion of all features by averaging probability scores of SVM classifier. They concluded that global features are compact and effective in feature representation as compared to the computational complexity of local features.

2.3 Deep Feature Based Classifier

Karpathy et al. [14] proposed multiple approaches for extending the connectivity of CNN to take advantage of the spatio-temporal information. Results show that CNN can generate strong improvement over handcrafted features. However, the multiple frame models showed a modest improvement compared to the single-frame model. However, these techniques process sample frames selected randomly from full length video. Such random selection of samples may not take into account all useful motion and spatial information.

Next, Simonyan et al. [15] proposed a two stream CNN network for video classification. One network analyzes the spatial information while the second analyzes the optical flow field. Their approach generates significant improvement over the single frame model.

Tran et al. [16] presented the first single stream CNN that incorporate both spatial and temporal information at once. Their approach takes multiple frames as input and examines them with 3D spatio-temporal convolutional filters. They handle the problem of activity recognition and performed evaluation on the UCF101 dataset. They outperformed all previous work. In addition, their technique is fast as it does not require optical flow estimation.

Xiaolong Wang et al. [17] used unsupervised learning methods for visual feature representation. They tracked first and last frame of the moving object or object part in patch using KFC tracker forming two parts and third part from Siamese Triplet Network extracting features from these paths and loss function. They extracted SURF interest points and use Improved Dense Trajectories (IDT) to obtain motion of each SURF point and using thresholding technique to obtain motion surf points. In next step extracted best bounding box to contain maximum moving surf points. Given various patches from these triplets, they extracted feature representation using Siamese Triplet Network and CNN from the last layer by forward propagation and ranking loss function. They obtained boosted result using model ensemble of Alexnet and transfer learning.

In 2013 [18] team from NTT Media Intelligence Laboratories worked on dense SIFT features around local points and used effective feature encoding Fisher Vector for concept detection task. To reduce dimension team used PCA and GMM to generate the codebooks. They used Support Vector Machines (SVM) classifiers with linear kernels for the first run. Ensemble Learning with sparse encoding for fusion of sub classifiers showed better performance than ensemble learning with Clara, the fisher vector run classified by the large linear classification method LIBLINEAR and Deep CNN.

Ha et al. [19] proposed technique using two correlation-based approaches Feature Correlation Maximum Spanning Tree (FC-MST) and Negative-based Sampling (NS), using CNN. First, FC-MST is applied to select the most relevant low-level features, obtained from multiple modalities which decided input layer of the CNN. To solve the problem of imbalanced dataset which result in poor performance of the classifiers batch sampling was used. Negative frame are ranked using MCA [20] and added in all batches. They got better results on 81 semantic concept from NUSWIDE image dataset [21].

Zhongwen Xu et al. [22] introduced latent concept descriptors for video event detection by extracting features from vgg pool 5 layer and descriptors from all frames in video are encoded to

generate video representations and descriptor are encoded using VLAD at the last convolutional layer with spatial pooling.

Podlesnaya A. et al. [23] develop video indexing and retrieval system based on semantic features extracted from GoogLeNet architecture and provided shot summarization using temporal pooling. Video retrieval was created using structured query for keyword based retrieval and indexing for database was prepared using graph based approach with the WorldNet lexical database.

Nitin J. Janwe et al. [24] asymmetrically trained two CNN model to deal with unbalanced dataset problem. They also proposed to combine foreground driven object to background semantic concept.

2.4 Combining Low-level and Deep Features

Foteini Markatopoulou et al.[25] used local descriptors like ORB, Opponent ORB, RGB-ORB, SIFT, Opponent SIFT, RGB-SIFT, SURF, Opponent SURF, RGB-SURF. These Local descriptors are aggregated into global image representations by employing feature encoding techniques such as Fisher Vector (FV) and VLAD after PCA. Also deep feature are extracted from hidden layers fc7 which resulted into a 4096-element vector. They combined features based on Deep Convolutional Networks (Deep CNN) with above local descriptors, both within a cascade and in two other late-fusion approaches.

Bahjat Safadi et al. [26] considered various types audio, image and motion descriptors for evaluation and variants of classifiers and their fusion. Classical descriptors extracted from key-frame includes color histogram, Gabor transform, quaternionic wavelets, a variety of interest points descriptors (SIFT, color SIFT, SURF), local edge patterns, saliency moments, and spectral profiles for audio description. Semantic descriptors which are scores computed on the current data using classifier trained on other data are also considered in fusion step. They used KNN and SVM as base classifiers and in late fusion step MAP weighted average of the scores produced by the two classifiers is combined.

3. SEMANTIC VIDEO CONCEPT DETECTION

Semantic Concept Detection assigns one or multiple label to video or part of video based on visual properties of it using some machine learning algorithm. General framework for semantic concept detection is shown in figure1 and include following broad steps 1) Preprocessing 2) Feature Extraction 3) Classifier Training 4) Concept Detection. The details of the above steps are described below.

3.1 Preprocessing

In preprocessing step video is divided into significant scene. Scene is semantically co-related and temporally adjacent shots depicting its story. Scene is segmented into shots.

3.1.1 Shot Segmentation

Shots are sequence of consecutive frames captured by uninterrupted camera. Frames of scene contain strong co-relation between them. Therefore, video is segmented into shots and used as basic unit in concept detection. In few datasets, annotations are available at shot level. Shot boundary detection is of two types a cut, where the transition between two consecutive shots is abrupt, and gradual transitions where visual content changes slowly. Dissolve, wipe, fade-in, fade-out are types of gradual transition. Gradual transition is more difficult to detect than cut.

3.1.2 Key-frame Extraction

In video multiple shots are present having many frames so it is necessary to represent shot with significant key-frame which depicts summary of shot. Frames of shot may be quite similar with little variation among them. Therefore frames which are representative of entire shot and shows summary of shot are selected as key-frames. Generally first, middle or last frame of shot are considered as key-frames. Other approaches are also available which selects key-frame

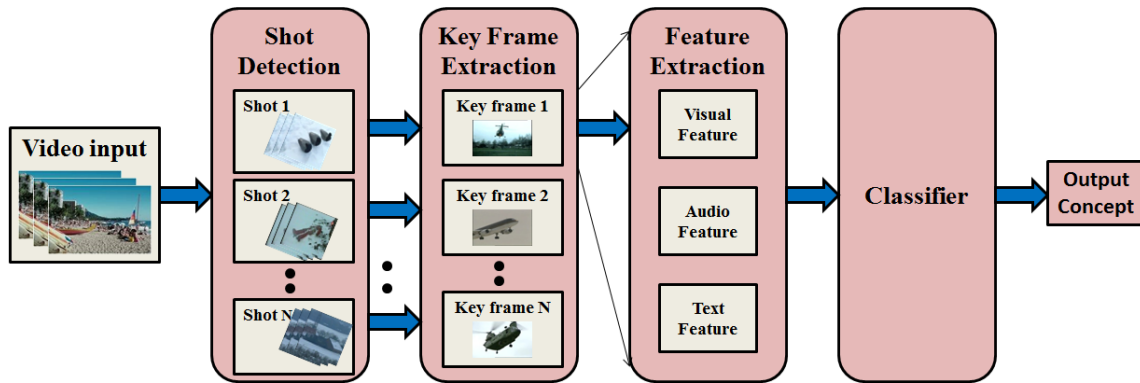


FIGURE 1: Block diagram of Semantic Concept Detection in Video.

considering all frames based on sequential or clustering based methods. In sequential method variation between visual features of frames are calculated and whenever considerable change is observed it is selected as key-frame. In cluster based approaches, frames are clustered based on variation of content of frames of shots and frame nearest to cluster center is selected as key-frame.

3.2 Feature Extraction

The aim of feature extraction is to derive a compact representation for the video shot or key-frame. In video concept detection system annotations are provided along with key-frames in standard dataset. Two kinds of features can be extracted from videos: key frame-based features and shot-based features. From the key-frame, color, texture, shape and object features are derived. In addition, some mid-level features from face detection can also be captured, such as the number of faces.

In the set of shot-based features, audio features and shot-based visual features are extracted. The extracted audio feature set is composed of volume-related, energy-related, and spectrum-flux-related features as well as the average zero crossing rates. For shot-based visual features, the grass ratio is calculated as a mid-level feature, which is useful for detecting sports-related semantic concepts such as soccer players and sports. Furthermore, a set of motion intensity estimation features are extracted, such as the center-to-corner pixel change ratio. Both of the categories of features target at representing / summarizing the video.

Key-frame based features are commonly used in concept detection system. Color features, texture features, and shape features are derived from key-frame along the spatial scale i.e., global level, region level, key-point level, and at temporal level. These features can be used independently or fused together or with other modality. Fusion can also be done at classifier level, where their kernel functions can be fused to improve performance.

3.3 Classifier Training

Semantic concept detection is basically a multi-label classification problem since multiple labels can be assigned to key-frame or video shot. Key-frames extracted from the video shots are given as input to classifier which predicts the set of concepts present in key-frames with score of each concept. Multi-label classification problem is converted into binary classification also known as Binary Relevance (BR) transformation. In this each classifier is trained independently for each concept class.

3.4 Concept Detection

Classifier training creates trained models. New test video or image passed through this model gives confidence score of presence or absence of concept from predefined list of concept list.

4. METHODOLOGY

Figure 2 shows the framework of proposed method using Support Vector Machine (SVM) and Convolutional Neural Network(CNN).

4.1 Shot Detection Methods

Video shots need to be identified in order to get detect concept. For shot boundary detection approach used by Janwe et al. [14] is followed and hierarchical clustering algorithm is used for key-frame extraction. Once key-frame are obtained, automatic or manual annotations method is used to generate groundtruth data for video. Figure 3 shows shots and key-frames obtained from video.

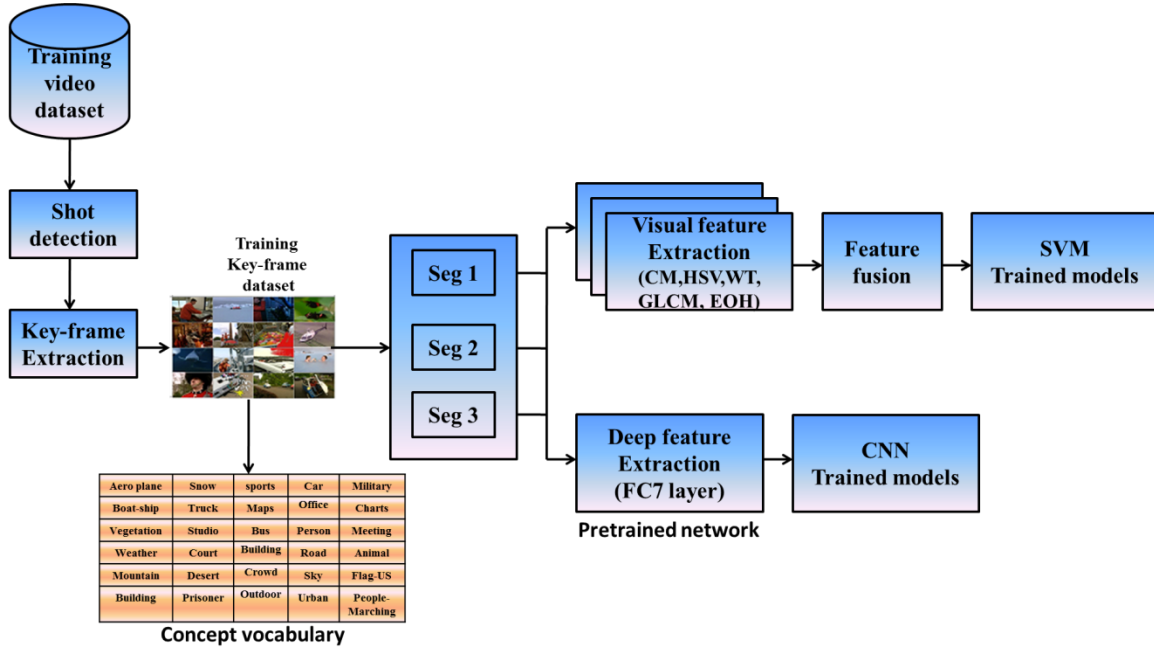


FIGURE 2: Block diagram of Proposed Framework for Semantic Concept Detection Using SVM and CNN.



FIGURE 3: Sample of shots and key-frames obtained from video.

4.2 Unbalanced Dataset Issue

In unbalanced dataset the number of samples belonging to one concept class is significantly

more than those belonging to other class. The predictive classifier model developed using such imbalanced dataset can be biased and inaccurate. Over-Sampling can be used to increase the number of instances in the minority class by randomly replicating them in order to increase the number of samples of the minority class in the sample. Here we have partitioned dataset into three segments based on frequency of the samples in dataset. Concepts having low frequency are kept in segment one, moderate frequency between 0.1 to 0.5 in segment two and more than 0.5 are in segment three.

4.3 Feature Extraction

Low level features of color and texture are extracted from each key-frame of TRECVID dataset for which Groundtruth labels are provided by NIST. In this experiment color moment and HSV histogram are used as color features and wavelet transform as textures derived globally from entire key-frame.

4.3.1 Color Feature

Color acts as a discriminative feature for understanding image or key-frame content. Color feature is independent of image size and orientation. For example blue color is prominent in beach or sky concept whereas brown color is dominating in desert or sunset concepts. There are primarily two methods available based on considering distribution of color and color histogram.

4.3.1.1 Color Moments (CM)

Color moments are invariant to scaling and rotation. The color distribution in an image can be interpreted as a probability distribution. The moments of this distribution can then be used as features to identify that image based on color. In this work two lower order moments Mean and Standard Deviation for each channel in HSV space have been used since most of the color distribution information is contained in the low-order moments. This results in a six dimensional feature vector. Equation 1 and 2 are the formulae to calculate mean and standard deviation.

$$E_i = \frac{1}{N} \sum_{i=0}^N I_{ij} \quad (1)$$

$$\sigma_i = \sqrt{\frac{1}{N} \left(\sum_{j=1}^N (I_{ij} - E_i)^2 \right)} \quad (2)$$

4.3.1.2 HSV Histogram (HSV)

Hue and Saturation define the chromaticity. Hue is a color element and represents a dominant color. Saturation expresses the degree to which white light dilutes a pure color. The HSV model is motivated by the human visual system as it better describes a color image than the RGB model. HSV histogram is distribution of colors of the key-frame in HSV color space. HSV color space is quantized into 40 bins then histogram is generated.

4.3.2 Texture Features

Texture is an important visual feature used in domain- specific applications. It can give us information about the content of an image efficiently. It is a repeated pattern of information or arrangement of the structure with regular intervals. It quantifies the properties such as smoothness, coarseness and regularity in an image. The texture feature used in our system are wavelet transform and grey level co-occurrence matrix.

4.3.2.1 Wavelet Transform (WT)

The wavelet transform is one of the current popular feature extraction methods used in texture classification. The wavelet transform is able to de-correlate the data and provides orientation sensitive information which is vital in texture analysis. It uses wavelet decomposition to significantly reduce the computational complexity and enhance the classification rate.

4.3.2.2 Grey Level Co-occurrence Matrix(GLCM)

GLCM is a statistical method of characterizing texture that considers the structural relationship of pixels in an image. It is constructed by counting the number of occurrences of a pairs of pixel with same values in an image, at a given displacement and angle. Correlation measures the probability of occurrence of grey levels among neighborhood pixels and contrast measures the local variations in GLCM. Homogeneity measures how close are the distribution of GLCM elements to the GLCM diagonal. Energy provides the sum of squared elements in the GLCM. The dimension of GLCM feature vector is 16.

4.3.3 Shape Features

Shape is an important basic visual features used to describe image content. A commonly used shape feature is edge.

4.3.3.1 Edge Orientation Histogram (EOH)

The edge orientation histogram descriptor consists of local edge distribution in the image. The EOH counts the number of pixels in each direction. A popular method for edge detection is canny

Feature	Feature Description	Dimension
Color Moments (CM) (Normalized)	Low order Moments (mean and standard deviation for RGB components) $2 \times 3 = 6$	6
HSV Color Histogram (HSV) (Normalized)	Each of h, s and v channel is quantized to $8 \times 2 \times 2$ bins respectively	32
Wavelet Transform (WT) (Normalized)	Mean and standard deviation of four coefficient $2 \times 20 = 40$	40
Gray level co-occurrence matrix (GLCM)	Co-occurrence matrix Contrast, Correlation, Energy,	16
Edge orientation Histogram (EOH) (Normalized)	Histogram of the 5 types of edges. $5 \times 1 = 5$	05

TABLE 1: Global Features Color and Texture Used In Experiment.

edge detector which is used in our experiment. Edge pixels in vertical, horizontal, two diagonals directions and one non-directional are counted. Table1 summarizes the feature set used in our experiments. The dimension of the concatenated feature vector is 99. SVM creates model accurately when features are in same range. Features are normalized using min-max normalization.

4.3.3. Deep Features

We use deep Convolutional Neural Network (CNN) trained on the ImageNet dataset to extract features from video key-frames. A 4096-dimensional feature vector is extracted from the key-frame of each video shot by using the CNN. The first to fifth layers are convolutional layers, in which the first, second, and fifth layers have max-pooling strategy. The sixth and seventh layers are fully connected. The parameters of the CNN are trained on the ImageNET ILSVRC dataset with 1,000 object categories. Finally, from each key-frame 4096-dimensional feature are extracted to train an SVM for each concept in the Video Concept Detection. Deep features are extracted from fc7 layer of trained model and used as deep features.

4.4 Classifier Design

Classifier stage is built with SVM and CNN as two classifier in hybrid combination.

4.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is basically a classifier technique that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. The output of a SVM can be a classification map that contains class labels for each pixel (or object), or a probability map that contains probability estimates for each pixel (or object) to belong to the assigned class. In this method, the one-versus-all approach is used for the multiclass SVM soft output. The one versus-all approach builds K SVMs (where K is the number of concept classes), each of which is able to separate one class from all the others. SVM is robust and effective. In this experiment, LIBSVM-3.20 package with Matlab implementation is used and One-Versus-All (OVA) technique for multi-class classification is adopted. For Concept detection task RBF kernel is used. SVM Parameter optimization is important to get better result. C (the cost parameter) and γ (the width of the RBF function) need to be tuned. Grid search and cross validation techniques are used to optimize these parameters.

4.5 Dataset Description

TRECVID's 2007 Video dataset and ground-truth dataset is used to conduct experiments. The National Institute of Standards and Technology (NIST) is responsible for the annual Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVID). Every year, it provides a test collection of video datasets along with a task list. It focuses its efforts to promote progress in video analysis and retrieval. It also provides ground-truth for researchers. Fig. 4 shows sample key-frames of concepts used in this experiment. TRECVID 2007 dataset consists of 219 video clips separated into two groups, the development set and testing set. The development set consists of 110 videos and the test set is consists of 109 video clips. Dataset has 36 concepts covering various categories like object, events, scene and activities like walking running. In this experimentation, the development dataset is partitioned into two parts, Partition I and Partition-II.



FIGURE 4: Sample key-frames of various concepts in TRECVID dataset (segment 1).

Dataset	Dataset Name	Partitions	No. of Videos	No. of key-frames
TRECVID Development Dataset	Partition I	Validation Dataset	80	4213
		Training Dataset		16542
	Partition II	Test Dataset	30	12615

TABLE 2: Details of Partitions used from TRECVID Development Dataset.

Partition I contains 80 videos used for training and 30 videos for testing. Training dataset consists of 16542 randomly chosen positive key-frames to perform classifier training. Test dataset consists of 12615 randomly chosen positive key-frames from Partition-II. Table 2 shows number key-frame used in two partitions.

Figure 5 shows number of positive key-frames available for 36 concepts in training dataset. As observed from figure 5 the TRECVID training dataset is highly imbalanced having few positive examples for concepts like US flag as compared to very high frequency of other concepts. Approaches like oversampling for imbalanced dataset are not sufficient in this task as classifier overfits. Classifiers are biased towards classes having more number of positive examples and tend to predict majority classes. Therefore we have partitioned training dataset as described in section 4.2. Table 3 shows distribution of training and testing key-frames used in three segments.

4.6 Classification Design

Two types of the classifiers are used in the experimentation. Training dataset partitioned into three segments is referred as seg1, seg2 and seg3.

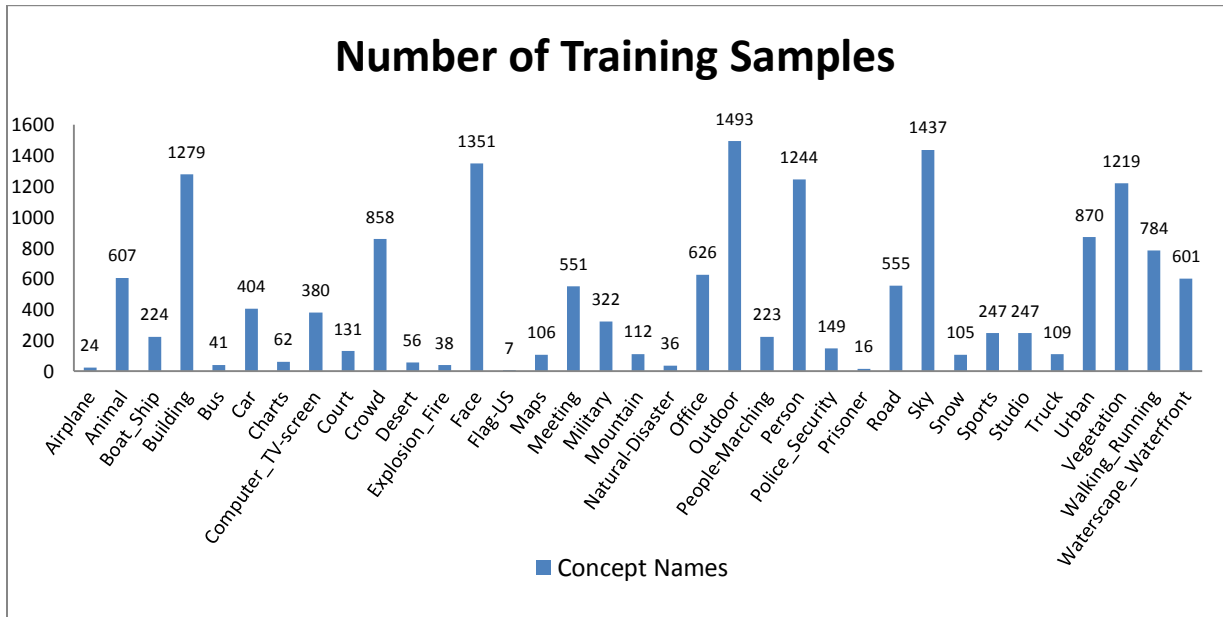


FIGURE 5: Representation of number of positive key-frame in TRECVID dataset.

Segment No. / Key-frames	No. of Training Key-Frames	No. of Testing Key-frames
Segment 1	871	311
Segment 2	3955	1176
Segment 3	11716	10528
Total	16542	12015

TABLE 3: Details of the Number of key -frames of segments of TRECVID development dataset.

Segment 1			Segment 2			Segment 3		
Sr. No	Concept Name	No. of Key-frames	Sr. No	Concept Name	No. of Key-frames	Sr. No	Concept Name	No. of Key-frames
1	Airplane	24	1	Animal	607	1	Building	1279
2	Bus	41	2	Boat_Ship	224	2	Crowd	858
3	Charts	62	3	Car	404	3	Face	1351
4	Court	131	4	Computer_TV-screen	380	4	Office	626
5	Desert	56	5	Meeting	551	5	Outdoor	1493
6	Explosion_Fire	38	6	Military	322	6	Person	1244
7	Flag-US	7	7	People-Marching	223	7	Road	555
8	Maps	106	8	Police_Security	149	8	Sky	1437
9	Mountain	112	9	Sports	247	9	Urban	870
10	Natural-Disaster	36	10	Studio	247	10	Vegetation	1219
11	Prisoner	16	11	Waterscape_Waterfront	601	11	Walking_Running	784
12	Snow	105	Total Key-frames		3955	Total Key-frames		11716
13	Truck	109						
14	Weather	28						
Total Key-frames		871						

TABLE 4: Concepts in all three segments and number of training key-frames for each concept.

4.6.1 SVM

The classifier adopted is SVM. Global features like color moment, HSV histogram and wavelet transform are extracted from training key-frame and combined to form single feature vector. SVM is trained for each segment separately creating three SVM models. The stepwise procedure for building SVM classifier is as follows [24]:

- Normalization: Training and test dataset feature vectors are normalized in the range (0.0 to 1.0).
- Kernel function selection: The choice of kernel function like RBF or linear depend upon feature selection.
- Parameter tuning: In non-linear features, best parameters for C and g are to be obtained for RBF kernel.
- Training: Values of C and g obtained after cross-validation are used to train the entire training set.
- Testing: Model obtained after training, are used for predicting a class for the test dataset.

4.6.2 CNN Classifier

In second pipeline deep features using pretrained Alexnet model from FC7 layer are used to train three SVM models which are called as CNN models.

4.6.3 Hybrid Model

To detect Semantic concept, shot detection is performed on test video and key-frames obtained from test video are passed through two pipelines to get global and deep features. Trained classifiers predict the probability scores for each concept. Each SVM and CNN classifier gives scores for corresponding segment. The combined score from three SVM are called score1 and CNN are called score2. Scores of CNN are range normalized to [0,1]. Scores obtained from two

pipelined are linearly fused to get final scores. These final merged scores are used to predict the concepts. The performance of the fused classifier is evaluated over the test dataset. Figure 6 shows approach for semantic concept detection for test video using hybrid approach.

4.7 Evaluation Measures

Mean Average Precision (MAP) is used as the performance evaluation of the proposed concept detection method by measuring top d ranked concepts sorted in descending order. Here we have considered top 5, top 10 and top N ranked concepts. In top N, N is considered to the maximum number of annotations of key-frame in training dataset.

Computing MAP: The ground-truth key-frames consists of multi-label data. For each key-frame the labels available for each key-frame called as concept set (Y_i) and the count of the concepts are stored in (N_i) are computed and stored. Let D be a multi-label test dataset, consisting of $|D|$ multi-label test examples (x_i, Y_i), $i = 1 \dots |D|$, $Y_i \subseteq L$, where L is a concept vocabulary set for a dataset. When concept prediction scores for all the L concepts are obtained, the following procedure is adopted to compute Top-d MAP [24]:

1. Rank the prediction scores and concepts in descending order for a test sample x_i .
2. Pick-up top d_i scores from a ranked list, let P_i is concept list from top d concepts. The intersection M_i between Y_i and P_i are the detected concepts. Let N_i is the label density for x_i , and intersection between Y_i and P_i is M_i , then M_i are correctly predicted concepts by a classifier. The average precision (AP) for a test sample is computed by (3),

$$AP_i = \frac{|Y_i \cap P_i|}{|P_i|} = \frac{M_i}{N_i} \tag{3}$$

And the Top-d MAP for a classifier, H , on dataset D , is obtained by computing the mean of APs by (4).

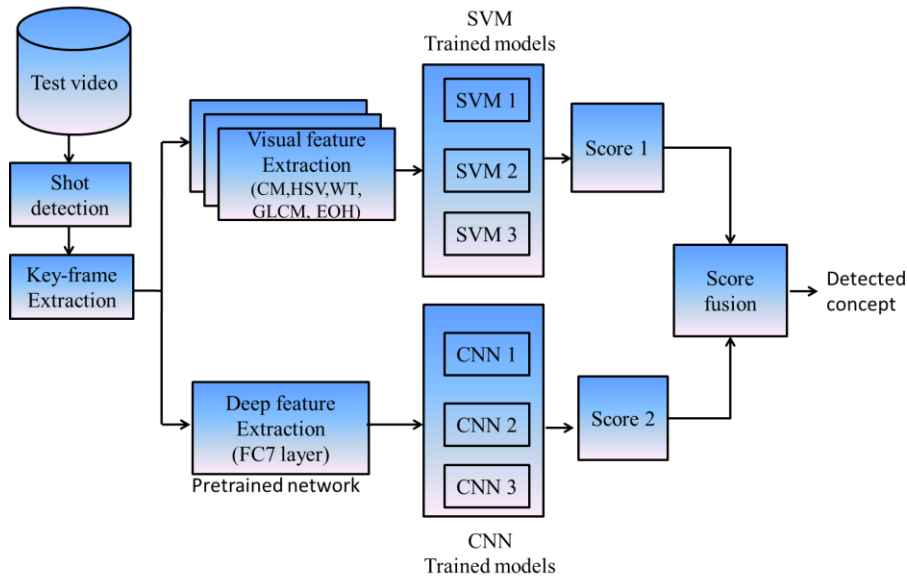


FIGURE 6: Testing phase of Semantic Concept Detection Using SVM and CNN.

$$MAP(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap P_i|}{|P_i|} \tag{4}$$

5. EXPERIMENTAL RESULT AND ANALYSIS

Experiments are performed on Intel i7 processor, 64 GB memory and NVIDIA 1050 GTX GPU. The proposed method is implemented using Matlab R-2017 version. The performance of proposed system is evaluated on the basis on MAP for multilabel dataset. The Libsvm tool is used to train the SVM classifier based on global features. MAP for Top 10 ranked concepts is 0.53.

In the experiment based on CNN, a pretrained Alexnet architecture is used for obtaining deep features. These features are used to train SVM classifier. A multiclass SVM classifier using a fast linear solver is used to build classifier on deep features. Map of 0.56 is obtained for CNN model. MAP weighted average fusion of scores from both the model is performed and probability for each concept class is ranked. Based on the fused score MAP of 0.58 is obtained for hybrid model. Table 5 shows MAP 5, 10 and n of SVM, CNN and hybrid fusion of scores of these classifiers. Table 6 depicts concepts available in groundtruth dataset and correctly detected concepts for test key-frames of few concept classes using hybrid model. As the dataset is partitioned into segments accurate classifier models are created with better discriminating capability. Concepts with less number of samples are recognized by model helping in improving efficiency.

As shown in table 6 out of 8 labels of first key-frame 6 concepts are correctly predicted. similarly, model is able to detect concepts with less annotations and concepts classes having less samples in dataset It is observed that CNN classifiers gives better precision than SVM classifier method and hybrid fusion method. The performance of the semantic concept detector using CNN is better than the other detectors as the automatically extracted deep features are more powerful as compare to low level features used to learn SVM classifier. MAP obtained for hybrid model is less than CNN classifier for top n concepts

Classifier Models	Map 5	Map 10	Map n
SVM	0.42	0.53	0.37
CNN	0.33	0.56	0.48
Hybrid Model	0.40	0.58	0.43

TABLE 5: MAP for Three Classifier Models.

The proposed work is compared with the existing methods presented in [18], [25], [27] and [28] as shown in table 7. The MAP of proposed system is better than these methods.

Sr. No.	Method	Publication year	MAP
1	Proposed Hybrid classifier using CNN and SVM	-	0.58
2	NTT-MD-DUT [18]	2014	0.048 (inf. MAP)
3	Cascading CNN [25]	2015	0.28
4	Deep multi task learning [27]	2016	0.25
5	Tokiyo_tech TRECVID [28]	2014	0.28 (inf .MAP)

TABLE 7: Performance comparison of proposed method with other methods.

CNN has proven to be good classifier in audio classification on large dataset and for acoustic event detection task [29]. In the present work, classifier fusion based on global feature and deep features is demonstrated. The present work can be extended for evaluation of CNN for multimodality fusion with audio features and /or text features for semantic concept detection. Similarly it can be checked for fusion of various machine learning approaches which might improve detection precision. The work can be extended to domain changes of target concept and heterogeneous training datasets where classifier model can be created using two different training datasets to integrate information from different dataset to build robust classifier.








Key-Frame No	Key-Frame Name	Test Key-frame	Concepts in Ground-Truth Dataset	Correctly Detected Concepts by hybrid model.
12214	Shot_102_144_RKF		Crowd, Military, Outdoor, Person, Police_Security, Sky, Snow, Waterscape_ Waterfront	Military, Outdoor, Person, Sky, Snow, Waterscape_ Waterfront
No. Of concepts			8	6
12209	Shot_6_38_RKF		Building, Bus, Car, Outdoor, Road, Sky, Vegetation'	Building, Bus, Car, Outdoor, Sky, Vegetation
No. Of concepts			7	6
11915	Shot_56_21_RKF		Face, Outdoor, Person	Face, Outdoor
No. Of concepts			3	2
207	Shot26_113_RKF		Boat_Ship, Face, Outdoor, Sky, Snow, Waterscape_ Waterfront	Outdoor, Sky, Snow, Waterscape_ Waterfront
No. Of concepts			6	4
1066	Shot4_255_RKF		Building, Face, Meeting, Office, Person, Studio	Building, Face, Meeting, Outdoor
No. Of concepts			6	4
764	Shot63_25_RKF		Building, Car, Outdoor, Sky, Urban	Building, Car, Outdoor, Sky
No. Of concepts			5	4
12108	Shot86_114_NRKF_1		Face, Outdoor, Person, Vegetation	Face, Outdoor, Person
No. Of concepts			4	3

TABLE 6: Detection of Semantic Concepts using Hybrid Model.

6. CONCLUSION

Video concept detection systems detects one or multiple concepts present in the shots or key-frame of video and automatically assign labels to unseen video which provides facilities to automatically indexing of multimedia data. Visual concept detection bridges the semantic gap between low level data representation and high level interpretation of the same by human visual system. Selection of compact and effective low level feature is important. Also because of imbalanced dataset problem, accurate classifier models cannot be created resulting into less accuracy.

In this paper, framework for multi-label semantic concept detection is proposed using classifiers trained on global and deep features. The imbalanced dataset issue is solved by partitioning dataset into three segments. The framework is evaluated on TRECVID dataset using mean average precision as predictive measure for multilabel dataset. Hybrid model of CNN and SVM performed better than individual classifier for top 10 concepts whereas CNN worked better for top n ranked concepts.

7. REFERENCES

- [1] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. "Caffe: Convolutional architecture for fast feature embedding," arXiv:1408.5093, 2014.
- [2] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard and Y. Bengio. "Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning,". NIPS Workshop, pp. 1-10, 2012.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks," In NIPS, 2012
- [4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell. "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," CoRR, abs/1310.1531, vol. 32, 2013.
- [5] R. Girshick, J. Donahue, T. Darrell, U. C. Berkeley, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation Tech report," 2012.
- [6] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel. "Learning automatic concept detectors from online video," In *Comput. Vis. Image Underst.*, vol. 114, no. 4, pp. 429–438, 2010.
- [7] B. Safadi, N. Derbas, A. Hamadi, M. Budnik, P. Mulhem, and G. Qu. "LIG at TRECVID 2014 : Semantic Indexing of the semantic indexing," 2014.
- [8] U. Niaz, B. Merialdo, C. Tanase, M. Eskevich, B. Huet, and S. Antipolis. "EURECOM at TrecVid 2015 : Semantic Indexing and Video Hyperlinking Tasks," 2015.
- [9] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and I. Patras. "A Study on the Use of a Binary Local Descriptor and Color Extensions of Local Descriptors for Video Concept Detection," vol. 8935, pp. 282–293, 2015.
- [10] L. Feng and B. Bhanu. "Semantic Concept Co-Occurrence Patterns for Image Annotation and Retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 785–799, 2016.
- [11] S. T. Strat, A. Benoit, P. Lambert, and A. Caplier. "Retina-Enhanced SURF Descriptors for Semantic Concept Detection in Videos," 2012.
- [12] F. Markatopoulou, V. Mezaris, N. Pittaras, and I. Patras. "Local Features and a Two-Layer Stacking Architecture for Semantic Concept Detection in Video," *IEEE Trans. Emerg. Top. Comput.*, vol. 3, no. 2, pp. 193–204, 2015.

- [13] D. Le. "A Comprehensive Study of Feature Representations for Semantic Concept Detection," Fifth IEEE International Conference on Semantic Computing, pp. 235–238, 2011.
- [14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale Video Classification with Convolutional Neural Networks," 2013.
- [15] K. Simonyan and A. Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos," pp. 1–9, 2014.
- [16] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," *CoRR*, vol. abs/1412.0, 2015.
- [17] X. Wang and A. Gupta, "Unsupervised Learning of Visual Representations using Videos." In ICCV, 2015.
- [18] Y. Sun, K. Sudo, Y. Taniguchi, H. Li, Y. Guan, and L. Liu. "TRECVID 2013 Semantic Video Concept Detection by NTT-MD-DUT," In Sun2013TrecVid2s, 2013.
- [19] H. Ha, Y. Yang, and S. Pouyanfar. "Correlation-based Deep Learning for Multimedia Semantic Concept Detection." In IEEE International Symposium on Multimedia (ISM08), pp. 316–321, Dec 2008.
- [20] H. Tian and S.-C. Chen. "MCA-NN: Multiple Correspondence Analysis Based Neural Network for Disaster Information Detection," In IEEE Third Int. Conf. Multimed. Big Data, pp. 268–275, 2017.
- [21] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-wide: A real-world web image database from national university of singapore," *ACM Int. Conf. Image Video Retr.*, p. 48, 2009.
- [22] Z. Xu, Y. Yang, and A. G. Hauptmann. "A Discriminative CNN Video Representation for Event Detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
- [23] A. Podlesnaya and S. Podlesnyy. "Deep Learning Based Semantic Video Indexing and Retrieval," no. 2214.
- [24] N. J. Janwe and K. K. Bhoyar. "Multi-label semantic concept detection in videos using fusion of asymmetrically trained deep convolutional neural networks and foreground driven concept co-occurrence matrix," In *Appl. Intell.*, vol. 48, no. 8, pp. 2047–2066, 2018.
- [25] F. Markatopoulou, V. Mezaris, and I. Patras. "Cascade of classifiers based on binary, non-binary and deep convolutional network descriptors for video concept detection.," In IEEE Int. Conf. on Image Processing (ICIP 2015), Canada, 2015.
- [26] B. Safadi, N. Derbas, A. Hamadi, M. Budnik, P. Mulhem, and G. Qu, "LIG at TRECVID 2014 : Semantic Indexing LIG at TRECVID 2014 : Semantic Indexing," no. June 2015, 2014.
- [27] F. Markatopoulou, "Deep Multi-task Learning with Label Correlation Constraint for Video Concept Detection," pp. 501–505.
- [28] N. Inoue, Z. Liang, M. Lin, Z. Xuefeng, K. Ueki. "TokyoTech-Waseda at TRECVID 2014", 2014
- [29] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. "CNN architectures for large-scale audioclassification," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017