Electronic Letters on Computer Vision and Image Analysis 16(3):14-29; 2017

Semantic Video Concept Detection using Novel Mixed-Hybrid-Fusion Approach for Multi-Label Data

Nitin J. Janwe* and Kishor K. Bhoyar+

* *Department of Computer Technology, Rajiv Gandhi College of Engineering, Babupeth, Chandrapur, India*

+ *Department of Information Technology, Yeshwantrao Chavan College of Engineering, Hingna Road, Nagpur, India*

Received 17th May 2016; accepted 18th Jul 2017

Abstract

The performance of the semantic concept detection method depends on, the selection of the low-level visual features used to represent key-frames of a shot and the selection of the feature-fusion method used. This paper proposes a set of low-level visual features of considerably smaller size and also proposes novel ‘hybrid-fusion’ and ‘mixed-hybrid-fusion’, approaches which are formulated by combining early and late-fusion strategies proposed in the literature. In the initially proposed hybrid-fusion approach, the features from the same feature group are combined using early-fusion before classifier training; and the concept probability scores from multiple classifiers are merged using late-fusion approach to get final detection scores. A feature group is defined as the features from the same feature family such as color moment. The hybrid-fusion approach is refined and the “mixed-hybrid-fusion” approach is proposed to further improve detection rate. This paper presents a novel video concept detection system for multi-label data using a proposed mixed-hybrid-fusion approach. Support Vector Machine (SVM) is used to build classifiers that produce concept probabilities for a test frame. The proposed approaches are evaluated on multi-label TRECVID2007 development dataset. Experimental results show that, the proposed mixed-hybrid-fusion approach performs better than other proposed hybrid-fusion approach and outperforms all conventional early-fusion and late-fusion approaches by large margins with respect to feature set dimensionality and Mean Average Precision (MAP) values.

Key Words: Semantic Video Concept Detection, High-Level Feature Extraction, Semantic Gap, Video Retrieval, Support Vector Machine, Hybrid-Fusion, Mixed-Hybrid-Fusion, Multi-Label Classification.

1 Introduction

Recent technological development in the field of multimedia and particularly video storage, compression techniques and networking are resulting into huge amounts of rich video archives. It has been a common strategy to develop automatic analysis techniques for deriving metadata from videos which describe the summarization that facilitates browsing, search, retrieval, delivery and manipulation of video data in an efficient manner.

The content of a video segment is also called high-level features or semantic concept for describing, indexing and searching video information. The semantic concepts could be a car, bus, road, vehicle, tree, forest, mountains, person or an animal for a particular segment of a video. The objective of concept detection or high

Correspondence to: <nitinj_janwe@yahoo.com>

Recommended for acceptance by Angel Sappa
https://doi.org/10.5565/rev/elcvia.927
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

level feature extraction is to build mapping functions from the low-level features to the high-level concepts with some machine learning techniques [1]. The state-of-the-art concept detection system consists of low-level feature extraction, feature fusion, and classifier training. Thus, the kind of low-level features and fusion methods chosen and classifier models adopted have critical impact on the performance of concept detection.

1.1 Semantic Video Concept Detection

The goal of semantic video concept detection is to detect semantic concepts of a video segment on its visual appearance. Human beings interpret the semantic meaning for a video segment based on visual appearance. But automatic semantic detection techniques express the semantics on the basis of low-level features extracted from the video segment. There is a difference in semantics of these two representations. This is called 'semantic gap'. The main challenge is to understand the video content by bridging the semantic gap between the video signals and the visual content interpretation. And to minimize the semantic gap, early efforts focused on methods exploiting simple handcrafted decision rules which maps a set of low-level visual features to a single high-level concept. Vailaya et al. [2] worked on concepts detectors for cityscape, landscape, mountains and forests. However, such dedicated approach to concept detection becomes expensive when a large-scale concepts need to be detected. Therefore, bridging of semantic gap is not possible by designing dedicated detector for each concept. Some generic approaches for large-scale concept detection have come into existence as an alternative to dedicated methods. These approaches [3, 4] exploit the observation that, if the low-level features of a video segment are to be mapped to a large number of high-level semantic concepts, it requires too many decision rules. Therefore, these rules must be derived using some type of machine learning mechanism. Many efficient concept detection schemes exist today based on machine learning approach, which allow access to multimedia as well as video data at the semantic level.

1.2 Typical Concept Detection System

The pipeline of a typical semantic video concept detection system is shown in Fig. 1. The four important stages of state-of-the-art system are as follows-

1. Stage-I: Video segmentation or Shot boundary detection.
2. Stage-II: Key-frame/s extraction.
3. Stage-III: Low-level feature extraction for key-frame/s and classifier training.
4. Stage-IV: Score-fusion to compute final concept detection scores.

1.2.1 Shot Boundary Detection (Video Segmentation)

In order to detect the semantic concepts precisely from video, video shots need to be identified perfectly. The automatic shot boundary detection and video segmentation is a well understood problem and highly robust methods exist [5]. Shot exhibits strong content correlations between frames hence shots are considered to be

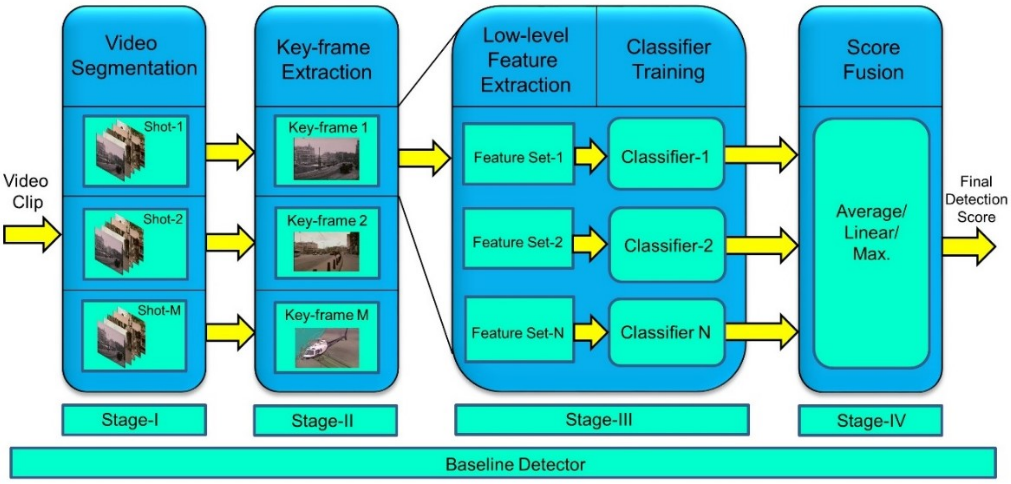


Fig. 1: The pipeline of a typical semantic video concept detection system

the basic units in concept detection. Generally shot boundaries are of two types, a cut, where the transition between two consecutive shots is abrupt and gradual transitions where boundary is stretched over multiple frames. The examples are dissolve, fade-in and fade-out etc. The shot boundary detection methods usually extract visual features from each frame and the similarities are measured and detect shot boundaries between frames that are dissimilar.

1.2.2 Key-Frame Extraction

In video processing applications, many times, the shots are often represented by a frame, called a key-frame, which is supposed to be a representative frame for a shot. There are great similarities among the frames from the same shot; therefore certain frames that best reflect the shot contents are chosen as key-frames. Mostly, the middle frame of a shot is taken as a key-frame, assuming that middle segment contains key contents, but many more other techniques do exist to get a key-frame. It is not necessary that a shot is always represented by a single frame; in some cases; however, multiple key-frames are required to represent a shot. The approaches like unsupervised clustering can be used, where frames in a shot are clustered depending on the variation in shot content and then choose frame closest to the cluster centre as a key frame. Each cluster is represented by a unique key-frame. So a single shot can have multiple key-frames. The choice of a key-frame may also depend on the object or the event one is looking for. Whichever frame that best represents the object or the event can be chosen as a key-frame.

1.2.3 Low-level Feature Extraction & Classifier Training

1.2.3.1 Low-Level Feature Extraction

The aim of feature extraction is to derive a compact representation for the video shot. In video concept detection system, a shot is represented by a key-frame(s). Such a key- frame can then be described using text features, audio features, visual features and their combinations. Here, it is attempted to summarize the most common visual features, as used in many concept detection methods. As mentioned earlier, the major bottleneck for automatic concept detection system is bridging the semantic gap between low-level feature representations that is extracted from video and high-level human interpretation of the video data. Hence, visual features need to represent the wide diversity in appearance of semantic concepts. If the viewpoint, lighting and other conditions are varied in the scene recording will deliver different data, whereas the semantics has not been changed. These variations induce the so-called sensory gap, which is the lack of correspondence between a concept in the world and the information in a digital recording of that concept. Therefore, visual features are needed to be minimally affected by the sensory gap [6], while still being able to distinguish concepts with different semantics. Invariant visual features are needed, such that the feature is tolerant to the accidental visual changes caused by the sensory gap.

Visual features are of three types, i.e., color features, texture features, and shape features; and they are computed along the spatial scale i.e., global level, region level, key-point level, and at temporal level. These features when extracted can be used independently or they can be fused to achieve more detector accuracy. Fusion can also be done at classifier level, where their kernel functions can be fused [7, 8] to improve performance.

1.2.3.2 Multi-label Data Classifier

Automatic video concept detection in segmented video is an inherently machine learning multi-label classification problem. In multi-label classification, the examples are associated with a set of labels. In a multi-label classification system, given the input feature space $\mathcal{X} \in \mathbb{R}^d$ and the output label space $\mathcal{Y} = \{0,1, \dots . n\}$, where n is no. of labels in the label set, a mapping function h: $\mathcal{X} \rightarrow \mathcal{Y}$ can be used to predict the corresponding label vector $y \in \mathcal{Y}$ for each input data instance $x \in \mathcal{X}$. That means the input feature vector of an instance of a key-frame of shot is mapped to a vector of labels. Multi-label learning focuses on identifying a good mapping function h from the training data. Many feature extraction techniques do exist to choose from, and a variety of supervised machine learning techniques to learn the mapping between. In supervised machine learning, in the first phase, the machine has to be trained i.e. classifier by supplying a set of optimal input feature vectors, and in the second phase, the classifier assigns a probability *p(C*$_j$*|* $\mathcal{X}_i$*)* to each input feature vector for every semantic concept. In automatic video concept detection methods, the two main factors which play a crucial role in the performance of a classifier are the extracted features and the supervised machine learning model.

1.2.3.2.1 Supervised Learning

Here, general methods are discussed that may exploit multimedia features used to train a machine to find the concept of a video shot. A better overview of machine learning is given in [9]. The supervised learning paradigm is most suitable for concept detection problems because the number of concepts in predefined concept list is fixed and known. The number of classes, the classifier will be trained for; will be equal to the number of concepts in a list. The objective of supervised learning is to optimize for a certain learning task and with limited amount of training data. This measure quantifies the performance of a classifier when classifying test patterns are not used during training. Poor generalization ability is commonly attributed to the over-fitting [10], It also attributes to curse of dimensionality, where the number of training examples used are two small compared to the number of features used. Therefore it is expected that, a supervised learning method should maintain a balance between the invariant features to use, and at the same time void over-optimization of parameters. Moreover, for concept detection, ideally, it must learn from a limited number of examples, it must handle imbalance in the number of positive versus negative training examples. Support Vector Machine framework [11] has become the default choice in most concept detection schemes because it proved to be the most effective machine learning technique for concept detection. In the experimentation, support vector machine is used to build concept classifiers.

1.2.3.2.2 Support Vector Machine (SVM)

The SVM framework, searches for an optimal hyper-plane which separates an *n*-dimensional feature space into two distinguished classes: one class represents the concept under consideration and second represents rest of the concepts, i.e. $yi = \pm 1$. A hyper-plane is considered optimal when the distance to the closest training examples is maximized for both classes. This distance is called the margin. It is parameterized by the support vectors, $\lambda i > 0$, which are obtained by optimizing:

$$\min_{\lambda}\left(\lambda^T \Lambda K \Lambda \lambda + C\sum_{z} \xi i\right) \tag{1}$$

during training under the constraints: $y_i g(x_i) \geq 1-\xi_i$, $i=1,2,\ldots.,z$, where Λ is a diagonal matrix containing the labels y_i, C is a parameter used to balance training error and to model complexity, z is the total number of shots in the training set, when the data is not perfectly separable, slack variables are introduced and is represented by $\xi_{i.}$, and for all training pairs, K is the matrix which stores the values of the kernel function $K(x_i,x')$. It is of interest to note the significance 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 obtained. Once the support vectors are known, it is straightforward to define a decision function for an unseen test sample x'.

1.2.4 Score-Fusion

Score fusion is a feature-fusion technique where scores resulting out of classifiers are combined using some strategy and final detection scores for each concept are computed. It is discussed in detail in next section.

In the section II, the early and late feature fusion approaches used in concept detection methods and the proposed *hybrid-fusion* and *mixed-hybrid-fusion* approaches are presented. This paper focuses on video concept detection methods over the benchmark dataset, using proposed hybrid-fusion and mixed-hybrid-fusion approaches based on variety of low-level visual features and SVM classifier. Section III presents the discussion about the low-level visual features used for training concept detector. Section IV, discusses the procedure to extract high-level features i.e. concept labels and brief description about the dataset selection is given. In the section V, detailed experimental results are presented. Section VI presents the conclusion.

2 Feature Fusion

Naturally, robust concept detection can be achieved by fusing many features extracted from video data. Selection of a set of features is very important as far as the concept detection accuracy is concerned. Some form of independence of features is required to make feature fusion to be effective. To achieve independence, following two general approaches are identified in the literature. The first approach relies on the so called unimodal features, where the features are extracted from a single modality, e.g., the audio stream, only. The second approach relies on multimodal features, where features are extracted from multiple modalities, for example, the speech transcript and the visual content. After feature combination, both unimodal and multimodal feature fusion methods rely on supervised learning to classify semantic concepts. Most unimodal feature fusion approaches rely on visual information. As different visual features describe different characteristics of a key-frame; color, texture, shape and motion can be considered statistically independent from a conceptual point of view. In this section, the classical early and late-fusion [12] schemes and proposed hybrid-fusion and mixed-hybrid-fusion schemes are presented.

2.1 Early-Fusion (EF) and Late-Fusion/Score Fusion

In EF, all visual features are combined into one larger feature vector and the concept detector is trained using this vector. Fig. 2(a) shows schematic diagram of *EF* approach where the fusion of the feature vectors takes place before training.

In *LF*, all individual detection scores for each concept from separated classifiers are combined using any of the merging strategy like linear, max or average and final score is obtained. The detailed scheme is shown in Fig. 2(b).

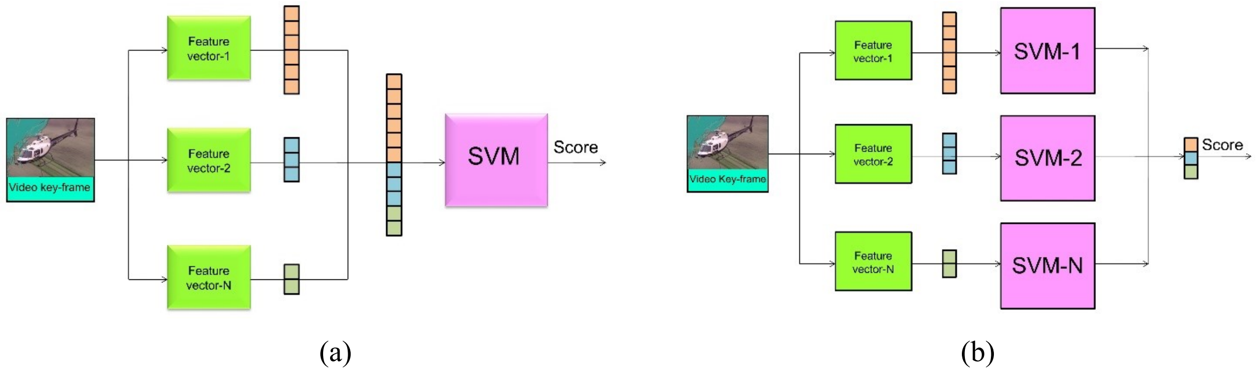


Fig. 2: (a) EF and (b) LF approaches with multiple visual features and SVM classifier/s

2.2 Proposed Hybrid-Fusion (HF) & Mixed-Hybrid-Fusion (MHF)

The *EF* and *LF* strategies have their own inherent merits and demerits. As the *EF* approach combines all feature vectors into one large vector, the training time increases, this is the biggest disadvantage while advantage is the number of classifiers required is only one. With *LF* approach, the number of classifiers required are equal to number of feature used, whereas the size of individual vectors is smaller.

In proposed *HF* approach, an attempt has been made to balance out the advantages and disadvantages of both the fusion methods by combining them. Therefore if these approaches are properly combined in some way, the concept detection performance can be increased. Here, the idea is to apply *EF* to combine the same group features (e.g. all color moment features like 2x2, 3x3 block features or all HSV histogram block features) into single large vector before classifier training *(EF)* and combining all individual detection scores of separated classifiers *(LF)* after training to get the final detection score. If all the feature groups are combined in this fashion, it is to call *hybrid-fusion (HF)* and is shown in Fig. 3(a). But, it is found that, combining each feature group using EF is not always fruitful. Sometimes, non- combining *(LF)* is beneficial in terms of performance.

Therefore, the MAP value for each feature group is computed for EF and LF methods using validation dataset, the method resulting into higher MAP is chosen as the fusion method for that feature group. Accordingly feature groups are fused using EF or LF methods. This scheme is called *mixed-hybrid-fusion (MHF).* Following description explains the process of EF/LF scheme selection.

Let $MAP_{EF}(\,)$ and $MAP_{LF}(\,)$ are the functions to compute MAPs for *EF* and *LF* schemes respectively and *d* is the difference between the two MAPs for an individual feature group *fg*.

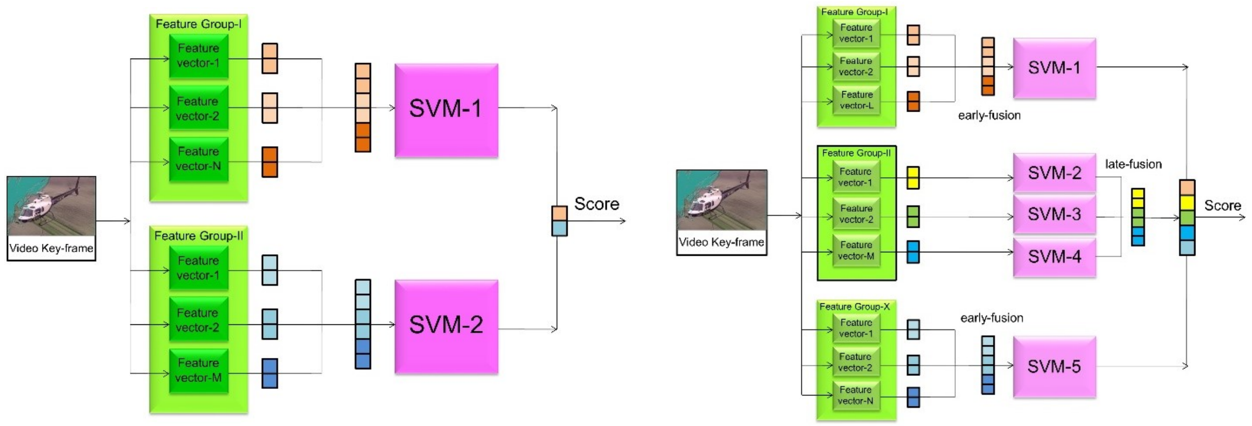
Let for a feature group *fg*, *x* and *y* are the values of MAP for *EF* and *LF* schemes respectively and $x = EF_{MAP}(fg)$ and $y = LF_{MAP}(fg)$ then,

$$d = x - y\ = (EF_{MAP}(fg) \text{ - } LF_{MAP}(fg)) \tag{2}$$

Therefore, the selection of *EF* or *LF* strategy for a feature group *fg* is done by the equation (3).

$$MAP(fg) = \begin{cases} x & (EF),\ d \geq 0 \\ y & (LF),\ d < 0 \end{cases} \tag{3}$$

From equation (3), *EF* is selected if difference *d* is positive or equal to zero, else *LF* is selected. The detailed scheme is explained diagrammatically in Fig. 3(b).



(a) (b)

Fig. 3: Proposed (a) HF and (b) MHF approaches with SVM classifiers

3 Low-Level Features

In the experimentation, five low-level visual features of type color, texture and shape structure are extracted for each key-frame of the ground-truth data provided by NIST. The first two are *color moments* texture features, taken on 2×2 and 3×3 block level, resulting into two feature vectors of dimension 24 and 54 respectively. Since these features belong to the same group type i.e. color moments, they are clubbed into Group-I. Next two features are *edge histogram* structural features taken on a global level 1×1 and 2×2 block level, resulting into vectors of dimension 4 and 16 respectively. They are grouped into Group-II. Next feature is *GLCM* texture consists of *contrast*, *correlation*, *energy* and *homogeneity*, taken on a global level resulting into a vector of size 16 and is placed in Group-III. This way, the total dimension of a feature set is 114-D. Table 1 lists the low-level features and their dimensions in detail.

Feature Name	Group	Description	Dimension
2×2 Color moment	Group-I	Based on 2 by 2 grid division of images: 1) Std. Deviation & 2) Mean for RGB components For a block: 2×3=6-D For 2×2 blocks: 4×6=24-D **Total dimension: 24-D**	24-D
3×3 Color moment		Based on 3 by 3 grid division of images For 3×3 blocks: 9×6=54-D	54-D
1×1 Edge histogram	Group-II	Edge histogram: For "horizontal" direction:2-D For "vertical" direction:2-D **Total dimension: 4-D**	4-D
2×2 Edge histogram		Based on 2 by 2 grid division of images: For 2×2 blocks: 4×4=16-D **Total dimension: 16-D**	16-D
1×1 GLCM texture	Group-III	Co-occurrence matrix texture features for gray images: For 4 filters feature extracted: 1) Contrast 2) Correlation 3) Energy 4) Homogeneity Total features: 4×4=16-D **Total Dimension: 16-D**	16-D

Table 1: Low-level visual features used for concept detection

4 Concept Detection/High-Level Feature Extraction for Multi-Label Data

The most important step in video concept detection is building classifier.

4.1 Building a Classifier using SVM

The multi-label video annotation task is posed into binary classification problem. SVM [13, 14, 15] is used as the baseline. As described in section 3, five low-level visual features are used; all of these features were utilized to build SVM classifiers. In EF approach, a single large feature vector is formed by combining all five feature vectors and a SVM classifier is trained. In LF, SVM classifiers are trained individually over each of the five feature spaces which results into five classifiers. In the proposed HF approach, the feature vectors are merged under the same group and the classifiers are trained resulting into 3 classifiers; and in MHF approach the number of classifiers required will vary and will depend on whether the EF or LF is used to fuse a feature group. The SVMs are implemented using LIBSVM (Version 3.18) [16].

The stepwise procedure for building SVM classifier is as follows:

a. Scaling: conduct simple scaling of the training and test dataset feature vectors.

b. Selecting proper kernel function: e.g. RBF or linear kernel function.

c. Parameter tuning: use cross-validation to find the best parameters C and g.

d. Training: use the best C and g to train the whole training set.

e. Testing: predicting a class for the test sample.

SVMs work well when features are roughly in the same range. Here, the features are normalized using statistical normalization. For *M* feature vectors $\{x_1, x_2, \ldots, x_M\}$ in which x_i is an *N*-dimensional feature vector $[x_{i1}, x_{i2}, \ldots.., x_{iN}]^T$, the mean vector ($\mu$) and the standard deviation vector (σ) are to be computed. The *mean* of a vector is defined as the average of a set of data elements in a vector. The mean vector (μ) is comprised of *M* mean values computed for each of *M* feature vectors and is computed by equation (4) as follows,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4}$$

where *N* is number of data elements in a vector x_i.

The *standard deviation* is the measure of dispersion of a set of data from its mean. It measures the absolute variability of a distribution; the higher the dispersion or variability, the greater is the standard deviation and greater will be the magnitude of the deviation of the value from their mean, the standard deviation is computed by equation (5) as follows,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{5}$$

Equation (5), computes the square of the difference $(x_i - \mu)^2$ between a data element and a mean of a feature vector for all the elements of a vector and then computes the average, this is called variance. The standard deviation is the square root of the variance. The process is repeated for all *M* feature vectors. This way, a standard deviation vector (σ) of size *M* is computed.

The features are then normalized through the equation (6),

$$x^* = \frac{(x - \mu)}{\sigma} \tag{6}$$

where x^* is the normalized feature. The division operation is applied to each component of the feature vector. Here, the features are normalized using statistical normalization which shifts the feature distribution to zero mean (i.e., $\mu = 0$) and unified standard deviation (i.e., $\sigma = 1$).

Sr.No.	Concept	Sr. No.	Concept	Sr. No.	Concept
01	Airplane	13	Face	25	Prisoner
02	Animal	14	Flag-US	26	Road
03	Boat_ship	15	Maps	27	Sky
04	Building	16	Meeting	28	Snow
05	Bus	17	Military	29	Sports
06	Car	18	Mountain	30	Studio
07	Charts	19	Natural-Disaster	31	Truck
08	Computer_ TV-screen	20	Office	32	Urban
09	Court	21	Outdoor	33	Vegetation
10	Crowd	22	People-Marching	34	Walking-Running
11	Desert	23	Person	35	Waterscape-Waterfront
12	Explosion_Fire	24	Police-Security	36	Weather

Table 2: Concept list in TRECVID development dataset

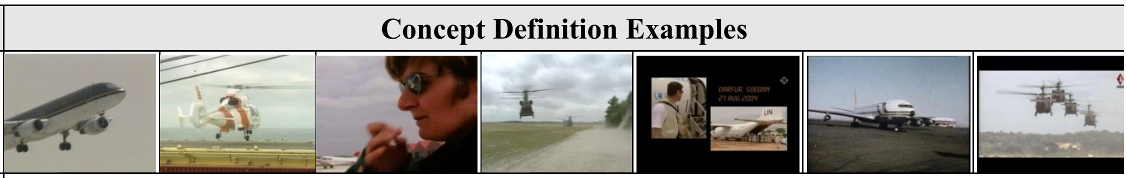
Concept	**Concept Definition Examples**
	
Airplane	Segment contains a shot of an airplane
Boat_Ship	Segment contains a shot of a boat or ship
Building	Segment contains a shot of an exterior of a building
Car	Segment contains a shot of a car
Crowd	Segment contains a shot depicting a crowd
Face	Segment contains a shot depicting a face
Road	Segment contains a shot depicting a road
Sports	Segment contains a shot depicting any sport in action
	SPREKEND NEDERLAND PRODUCTIE NEW YORK ANNE GEENEN REGIE ASSISTENTIE JOSE SCHRAAUWERS
Snow	Segment contains a shot depicting snow
Walking_ Running	Segment contains a shot depicting a person walking or running

Table 3: Concept definition examples from the TRECVID development dataset

4.2 Dataset Selection

Since 2001, the National Institute of Standards and Technology (NIST) [17] has been sponsoring the annual Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVID) [18]. TRECVID provides a large-scale test collection of video datasets every year, along with a task list and focuses its efforts to promote progress in video analysis and retrieval. It also provides ground-truth for data like a list of shots and a list of key-frames for a given TRECVID datasets for genuine researchers. Many researchers [19] [20] [21] and research teams present their high quality research contributions in yearly organized TRECVID conferences and workshops.

The TRECVID dataset is composed of 219 video clips separated into two groups, the development set and testing set. The development set consists of 110 videos while the test set is composed of 109 video clips. The videos in development dataset have partitioned into 19140 shots and 664850 key-frames. There are 36 defined concepts in the dataset. The concept list is given in Table 2. The 36 concepts are manually annotated over these key-frames. NIST has prepared a ground-truth-data for the above dataset for genuine researchers. The ground-truth consists of video shots and their representative key-frame/s for video clips. It is to be noted that, as per key-frame extraction method used by NIST, a shot in a video clip may have one or more positive and/or negative key-frames. For a concept, positive key-frame is defined as a frame containing a said concept as a visual content. The ground-truth dataset consists of both positive as well as negative examples. Table 3 lists some of the concepts and concept defining key-frames in the dataset.

For the experimentation, the ground-truth data for the development dataset is used. As shown in Table 4, the dataset is partitioned into two parts, Partition-I and Partition-II. Partition-I is further divided into Validation set and Training set and Partition-II is Test-set. Validation/Selection dataset consists of 5398 randomly chosen positive key-frames from Partition-I to perform cross validation to find out optimal parameters C and g for RBF kernel function in SVM. The Validation/Selection dataset is also used in *mixed-hybrid-fusion* approach to compute the MAP for each feature group for selecting one between EF and LF. Training dataset consists of 17114 randomly chosen positive key-frames to perform classifier training. Test dataset consists of 9352 randomly chosen positive key-frames from Partition-II and is used to test classifier performance. Fig. 4 illustrates the distribution of positive examples for each individual 36 concepts in the Training dataset.

4.3 Parameter Selection

Although the only parameters of the SVM are C and the kernel function K (•), it is well known that the influence of these parameters on concept detection performance is significant. Since the RBF Kernel is used, two parameters: C (the cost parameter) and γ (the width of the RBF function) need to be tuned. Since Libsvm-3.18 toolset is used to implement SVM, the data unbalance problem is handled through imposing penalty weights on respective classes at the time of training classifiers. In practical implementation, the penalty weight for a particular class is the ratio $\frac{Nmax^+}{N^+}$, where $N_{max}{}^+$ is the maximum number of positive training examples of any class and N^+ is the number of positive training examples for a respective class.

Dataset	Dataset Name	Partitions	# of Videos	# of Key-Frames
TRECVID Development Dataset	Partition-I	Validation/ Selection Dataset	90	5398
		Training Dataset		17114
	Partition-II	Testing Dataset	20	9352

Table 4: Partition details of TRECVID development dataset

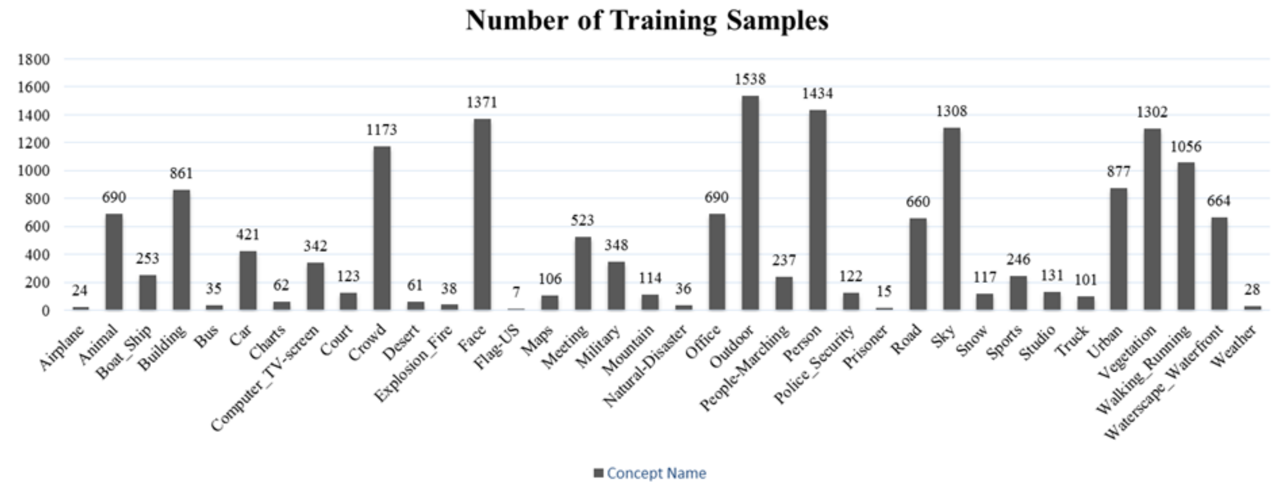


Fig. 4: Number of positive frames in TRECVID training dataset

4.4 Applying Score Fusion

After separated classifiers for each visual feature or feature group are learned, the LF approach is applied to combine all detection scores for each concept as shown in Fig. 2(b). The three fusion strategies are Linear, Average, and Max and their details are as follows-

- Linear: Performs a grid search in fusion parameter space to select the optimal weights.
- Average: The scores resulting from each classifier are simply averaged to generate the fused score.
- Max: For each concept, the best performance is selected.

Its results were considered as the fused results. In the experimentation average fusion strategy is used.

4.5 Computing Average Precision (AP) and Mean Average Precision (MAP) for Multi-label Data

The ground-truth dataset consists of the key-frames manually annotated with multi-label data. Therefore when dealing with multi-label frames, it is very important to know the way the *Average Precision* (AP) and *Mean Average Precision* (MAP) are computed as the performance is evaluated by these measures, which are the official performance metric in TRECVID evaluations. Some processing has been done over the ground-truth test dataset. The label set (concept set), $Y_{i,}$ and label count or *label density*, N_i, for each test sample, x_i, are computed. Let D be a multi-label test dataset, consisting of $|D|$ multi-label test examples (x_i, Y_i), $i = 1\ldots |D|$, $Y_i \subseteq L$, where L is a label set for a dataset. When the detection score (probability score) for all the 36 concepts for a test example are combined in the score-fusion phase, following procedure is followed to compute AP and MAP:

1. Rank the final scores of probabilities in descending order for all 36 concepts for a test example x_i.
2. If N_i is the *label density* for x_i, then top N_i scores from a ranked list (top N_i predicted concepts), P_i, and concepts in Y_i from a test sample are considered and their intersection is found out. The result of the intersection operation between sets Y_i and P_i is the number of concepts, $M_{i,}$ that are correctly predicted by a classifier.
3. The average precision AP for a test sample is computed by equation (7),

$$AP_i = \frac{|Y_i \cap P_i|}{|P_i|} = \frac{M_i}{N_i} \tag{7}$$

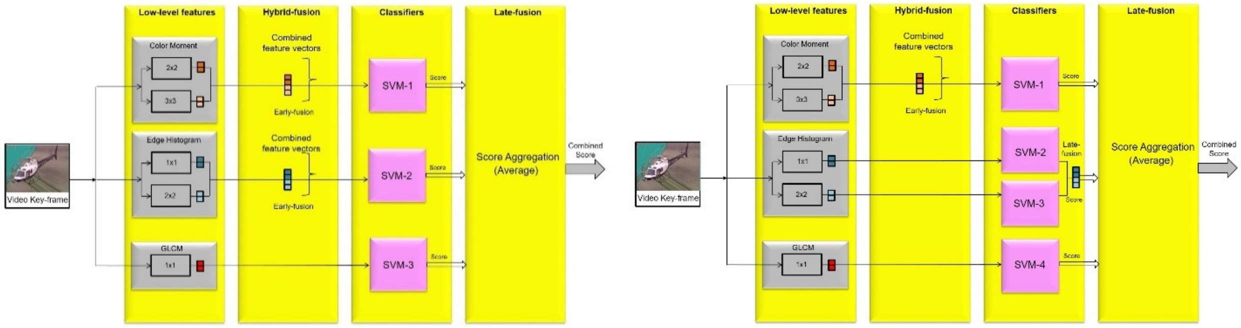
And the MAP for a classifier, H, on dataset D, is obtained by computing the mean of APs by equation (8),

$$MAP\ (H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap P_i|}{|P_i|} \tag{8}$$

5 Experimental Results

5.1 Experimental Evaluation

To demonstrate the effectiveness of the proposed feature fusion approaches in improving the video concept detection rate, the performance of concept detection using proposed HF and MHF approaches are compared with the performance of the existing EF and LF methods. In the experimentation, video concept detectors using multi-class SVM are implemented and compared for four approaches: VCD_EF (video concept detection using early-fusion), VCD_LF (video concept detection using late-fusion), VCD_HF (video concept detection using proposed hybrid-fusion approach) and VCD_MHF (video concept detection using proposed mixed-hybrid-fusion approach). The task is to detect the presence of 36 predetermined benchmark concepts in test dataset. Fig. 5(a) and Fig. 5(b) shows the detailed schematic diagram of video concept detection using HF and MHF approaches respectively.



(a) (b)

Fig. 5: Video concept detection using proposed (a) HF approach and (b) MHF approach

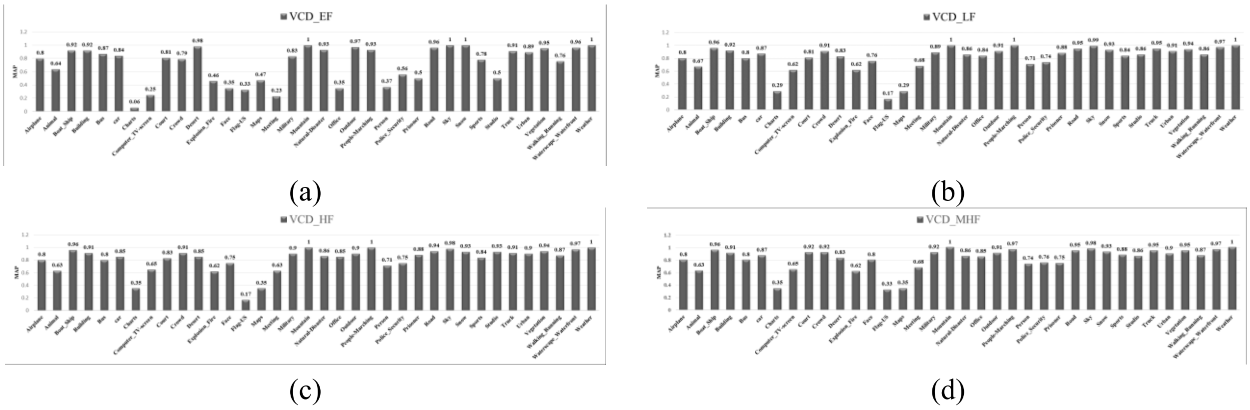
The performance is evaluated by AP and MAP. The decision table required to implement MHF approach is given by Table 5, it also presents the experimental evaluation results (MAPs) using Selection dataset for EF and LF schemes.

Feature Group	EF (MAP)	LF (MAP)	Decision? EF/LF
Color-Moments	0.43	0.40	EF
Edge-Histogram	0.42	0.45	LF

Table 5: Decision table for individual feature groups

5.2 Performance Evaluation & Results Comparison

The performance of all the above methods are evaluated on the basis of MAP values. Fig. 6(a), Fig. 6(b), Fig. 6(c) and Fig. 6(d) shows the results for all the 36 individual defined concepts using VCD_EF, VCD_LF, VCD_HF and VCD_MHF methods respectively. Fig. 7 presents the combined comparison of existing approaches with proposed HF and MHF approaches in terms of APs. It is observed that, the APs obtained with HF and MHF approaches are lot better than the existing EF and LF approaches. From Fig. 7, it is seen that, for a concept like *Charts*, the detection rate is a bit worst using EF (0.06) and LF (0.29) approaches, than using proposed methods. There is a significant improvement using the proposed HF and MHF (0.35) approaches. For concept Court the detection rate is 0.81 using EF and LF while it is 0.83 and 0.92 using HF and MHF



(a) (b)

(c) (d)

Fig. 6: APs of 36 concepts of TRECVID dataset obtained using (a) VCD_EF (b) VCD_LF (c) proposed VCD_HF and (d) proposed VCD_MHF

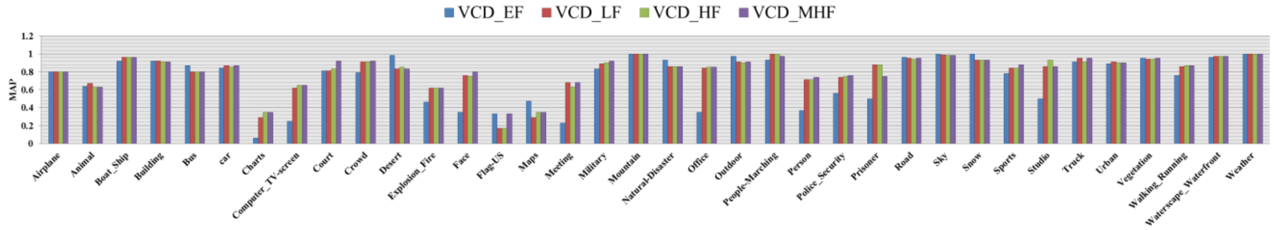


Fig. 7: Comparison of APs of 36 concepts, obtained using VCD_EF, VCD_LF, VCD_HF and VCD_MHF

Key-Frame No.	**Test key-frame**	**Concepts in Ground-Truth Data**	**Correctly Detected Concepts**			
			EF	**LF**	**HF**	**MHF**
540		Building-Crowd-Outdoor-People_Marching-Urban-Walking_Running	Building-Outdoor-Urban-Walking_Running	Crowd-Outdoor-Walking_Running	Building-Crowd-Outdoor-Urban-Walking_Running	Building-Crowd-Outdoor-Urban-Walking_Running
# of concepts		**6**	**4**	**3**	**5**	**5**
1569		Face-Person	--	--	Face-Person	Face-Person
# of concepts		**2**	**0**	**0**	**2**	**2**
2375		Building-Face-Outdoor-Person-Sky-Waterscape_Waterfront	Outdoor-Sky-Waterscape_Waterfront	Face-Outdoor-Sky-Waterscape_Waterfront	Face-Outdoor-Person-Sky-Waterscape_Waterfront	Face-Outdoor-Person-Sky-Waterscape_Waterfront
# of concepts		**6**	**3**	**4**	**5**	**5**
2547		Face-Outdoor-Person-Sky-Vegetation	Outdoor-Sky-Vegetation	Outdoor-Sky-Vegetation	Face-Outdoor-Sky-Vegetation	Face-Outdoor-Person-Sky-Vegetation
# of concepts		**5**	**3**	**3**	**4**	**5**
9095		Face-Office-Person-Walking_Running	Walking_Running	Person	Face-Person-Walking_Running	Face-Person-Walking_Running
# of concepts		**4**	**1**	**1**	**3**	**3**

Table 6: Result comparison of some of the sample test frames using proposed MHF and HF approaches and EF and LF methods

respectively. For Police-Security, the detection rate is 0.56 and 0.74 for EF and LF and is 0.75 and 0.76 for HF and MHF respectively. This shows that the concept detection rate for some concepts greatly improves using proposed HF and MHF approaches over EF & LF. Table 6 presents the experimental results for some of the test samples, showing the comparison of correctly detected concept count for the EF, LF and proposed HF and MHF approaches.

For a test sample key-frame no. 540, the count of concepts in the ground-truth is 6, namely *Building*, *Crowd*, *Outdoor*, *People-Marching*, *Urban* and *Walking-Running*. Out of these concepts, EF has detected 4, namely *Building*, *Outdoor*, *Urban*, and *Walking-Running* while LF has detected 3 i.e. *Crowd*, *Outdoor* and *Walking-Running* and the proposed HF and MHF detected 5 concepts namely; *Building*, *Crowd*, *Outdoor*, *Urban* and *Walking-Running*. For other test samples too, the proposed methods exhibit better performance than EF and

Sr.No.	Fusion Method	MAP
1	EF	0.33
2	LF	0.41
3	HF	0.49
4	MHF	0.52

Table 7: Performance comparison of proposed methods with EF and LF

Sr. No	Method	Database used	Features used	Dimension of feature vector	Classifier used	MAP
1	**Proposed method using Mixed-Hybrid-Fusion**	TRECVID2007	1) 2×2 color moment 2) 3×3 color moment 3) 1×1 edge histogram 4) 2×2 edge histogram 5) 1×1 GLCM texture	24-D 54-D 4-D 16-D 16-D Total: 114-D	Multi-class Support Vector Machine	**0.52**
2	TRECV0705 Model [22]	TRECVID2007 & Partial TRECVID2005	1) AutoCorrelogram 2) 3×3 color moment 3) 5×5 color moment 4) 7×7 color moment 5) Co-occurrence texture 6) Edge distribution histogram 7) Face 8) HSV color histogram 9) Wavelet PWT & TWT texture	144-D 81-D 225-D 441-D 16-D 75-D 7-D 64-D 128-D Total: 1181-D	Support vector machine	0.286
3	Multi-Label LGC [23]	TRECVID2006	1) 5×5 block-wise Color moment in Lab color space	Each block is described by 9-D features. Total: 225-D	Graph-based Semi-supervised learning	0.329
4	Multi-Label GRF [23]				Graph-based Semi-supervised learning	0.346
5	SGAL_noCorr [24]	NUS-WIDE-Lite Dataset	1) 5×5 block-wise color moments 2) edge direction histogram 3) wavelet texture	225-D 73-D 128-D Total: 426-D	Sparse-graph-based Semi-supervised learning	0.279

Table 8: Performance comparison of video concept detection using proposed MHF with state-of-the-art other existing methods

LF methods. The results thus obtained with all the above methods are compared and given in Table 7. It is observed that, the MAP for the proposed MHF and HF approaches are 0.52 and 0.49 respectively, which are much better than 0.33 and 0.41 for EF and LF respectively. MHF exhibits substantial improvement of approximately 25% over LF and it is also observed that, the performance of MHF outperforms all. Fig. 8 shows the performance comparison of proposed approaches with conventional EF and LF methods. The performance of the proposed MHF approach is also compared with the state-of-the-art other existing video concept detection methods given by Zha at el. [22] & [23] and Tang et al. [24] as shown in Table 8. The proposed MHF approach gives the best performance amongst all other approaches.

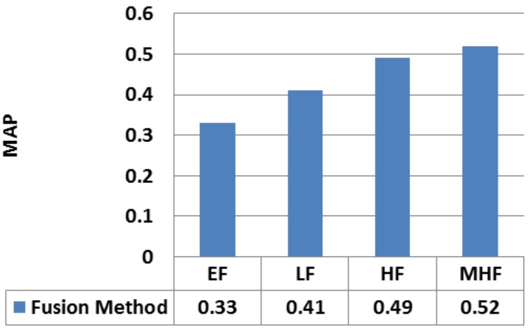


Fig. 8: Performance comparison of proposed MHF and HF with EF & LF methods

6 Conclusion

In video semantic concept detection methods, *semantic gap* directly controls the concept detection rate. Lower the semantic gap, higher the concept detection accuracy. The selection of low-level visual features and their dimensions to represent key-frame/s of a video shot and the selection of the feature-fusion methods are the two important factors to control the semantic gap. This paper has presented a work on these two important aspects and proposed 1) a set of low-level visual features of considerably smaller size (114) as compared to others and 2) novel feature fusion approaches namely, *hybrid-fusion* and *mixed-hybrid-fusion* with an aim to minimize semantic gap and to improve performance of video concept detection. Multi-class SVM is used to build classifiers. Extensive experimentation conducted on the multi-label ground-truth data for TRECVID development dataset have demonstrated that by combining EF and LF approaches in typical fashion which resulted into HF and MHF approaches, can substantially improve concept detection rate. In the experiments, video concept detectors are built using proposed HF and MHF approaches and their detection rate is compared with EF and LF methods. Experimental results show that, the proposed mixed-hybrid-fusion approach, MHF (MAP=0.52) performs better than our other proposed hybrid-fusion approach, HF (MAP=0.49) and outperforms conventional early-fusion, EF (MAP=0.33) and late-fusion, LF (MAP=0.41) approaches by large margins in terms of concept detection rate. The MHF approach is compared with other state-of-the-art methods in the category and exhibits enhanced performance over the methods.

References

[1] Cees G. M. Snoek and Marcel Worring, “Concept-Based Video Retrieval,” *Foundations and Trends in Information Retrieval archieve*, vol. 2, no. 4, pp. 215-322, 2008. DOI: 10.1561/1500000014

[2] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. J. Zhang, “Image classification for content-based indexing,” *IEEE Transactions on Image Processing*, vol.10, pp.117-130, 2001. DOI:10.1109/83.892448

[3] A. Amir, M. Berg, S. F. Chang, W. Hsu, G. Iyengar, C. Y. Lin, M. R. Naphade, A. P. Natsav, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, “IBM Research TRECVID-2003 video retrieval system,” *In Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[4] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Koelma, F. J. Seinstra and A.W. M. Smeulders, “The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing,” *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 1678-1689, 2006. DOI: 10.1109/TPAMI.2006.212

[5] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, "A formal study of shot boundary detection,"*IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 168–186, 2007. DOI:10.1109/TCSVT.2006.888023

[6] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1349–1380, 2000. DOI:10.1109/34.895972

[7] Stephane Ayache, Georges Quenot and Jerome Gensel, "Classifier Fusion for SVM-Based Multimedia Semantic Indexing," *Proceedings of the 29th European Conference on IR Research*, ECIR 2007, Rome, Italy, pp. 494-504, April 2-5, 2007. DOI:10.1007/978-3-540-71496-5_44.

[8] B. L. Tseng, C. Y. Lin, M. Naphade, A. Natsev, and J. R. Smith, "Normalized Classifier Fusion for Semantic Visual Concept Detection," *In Proceedings of IEEE ICIP*,2003. DOI:10.1109/ICIP.2003.1246735

[9] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, 2000. DOI:10.1109/34.824819

[10] H. R. Naphide and T. S. Huang, "A Probabilistic Framework for Semantic video Indexing, Filtering and Retrieval," *IEEE Transactions on Multimedia*, vol. 3, pp. 141–151, 2001. DOI:10.1109/6046.909601

[11] V. N. Vapnik, "The Nature of Statistical Learning Theory," *New York, USA:Springer-Verlag*, *2nd ed.,* 2000.

[12] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis*," In proceedings of ACM Multimedia*, pp. 399-402, 2005. DOI:10.1145/1101149.1101236

[13] A. Yanagawa, S. F. Chang, L. Kennedy, and W. Hsu, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," *Columbia University ADVENT Technical Report* # 222-2006-8, March 20, 2007.

[14] L. Duan, Ivor W. Tsang, Dong Xu, Stephen J. Maybank, "Domain Transfer SVM for Video Concept Detection," IEEE, 978-1-4244-3991-1/09/, 2009. DOI:10.1109/CVPR.2009.5206747

[15] Xinxing Xu, Dong Xu, "Video Concept Detection Using Support Vector Machine with Augmented Features," *Proceedings of the 2010 Fourth Pacific-Rim Symposium on Image and Video Technology*, pp. 381-385, 2010. DOI:10.1109/PSIVT.2010.70

[16] Chih Chung Chang and Chih Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*,27:1--27:27,2011. DOI:10.1145/1961189.1961199

[17] NIST: http://www.nist.gov

[18] TRECVID: http://www-nlpir.nist.gov/projects/trecvid/

[19] J. Liu, Y. Aytar, B. Orhan, J. han and M. Shah, "University of Central Florida at TRECVID 2007 Semantic Video Classification and Automatic Search," *In Proceedings of TRECVID Workshop*, 2007.

[20] Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi, Haojie Li, Yue Guan, Lijuan Liu, "TRECVid 2013 Semantic Video Concept Detection by NTT-MD-DUT," *TRECVID'13*,Nov.26-28, 2013, Gaithersburg, Maryland, USA.

[21] T.Mei, X. S.Hua, W.Lai, L.Yang, "MSRA-USTC-SJTU AT TRECVID2007:HIGH-LEVEL FEATURE EXTRACTION AND SEARCH," *In TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[22] Zheng Jun Zha, Yuan Liu, Tao Mei, Xian Sheng Hua, "Video Concept Detection using Support Vector Machines-TRECVID 2007 Evaluations," *Tech. Report Microsoft Research, Asia*, MSR-TR-2008-10.

[23] Zheng Jun Zha, Tao Mei, Jingdong Wang, Zengfu Wang, and Xian Sheng Hua, "Graph-based Semi-Supervised Learning with Multiple Labels," *In proceedings of Journal of Visual Communication and Image Representation,* vol 20, issue 2, pp. 97-103, Feb. 2009. DOI:10.1016/j.jvcir.2008.11.009

[24] Jinhui Tang, Zheng-Jun Zha, Dacheng Tao, and Tat-Seng Chua, "Semantic-Gap-Oriented Active Learning for Multilabel Image Annotation," *In IEEE Transactions on Image Processing*, vol. 21, no. 4, April 2012. DOI: 10.1109/TIP.2011.2180916