Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Multimodal Concept Detection in Broadcast Media: KavTan

Medeni Soysal · K. Berker Loğoğlu · Mashar Tekin · Ersin Esen · Ahmet Saracoğlu · Banu Oskay Acar · Ezgi Can Ozan · Tuğrul K. Ateş · Hakan Sevimli · Müge Sevinç · İlkay Atıl · Savaş Özkan · Mehmet Ali Arabacı · Seda Tankız · A. Aydın Alatan · Tolga Çiloğlu

Received: date / Accepted: date

Abstract Concept detection stands as an important problem for efficient indexing and retrieval of large video archives. In this work, for detection of diverse and distinct concepts a semantic analysis system (*KavTan*) that combines multi-modal information sources in a framework is proposed. In the proposed system, there are generalized audio concept detection and audio keyword detection modules that use audio data and generalized visual concept detection, video text detection, visual human detection, nudity detection, blood detection, flag detection and skin detection modules that use visual data. Key high-level concepts are detected by using output of the low-level concept detection modules. Performance of the proposed system in detection of both high and low-level concepts on a significant amount of unconstrained TV data is reported. It has been observed that for most of the concepts high performance can be achieved with this approach.

Keywords Concept Detection · Broadcast Video Indexing · Semantic Indexing

1 Introduction

Immense increase in the number of multimedia content from television and Internet with the ever increasing storage and Internet technologies reveals management, i.e. organizing and indexing, of such content as a problem. Automatic semantic classification of multimedia content into predefined concepts, which will eventually make the data much easier to analyze and search, stands as one of the major research areas in recent years.

The *KavTan* Semantic Concept Detection System is developed for Radio Television Supreme Council of Turkey (RTÜK). RTÜK monitors more than 300 TV chan-

Medeni Soysal
TUBITAK - UZAY, METU Campus, ANKARA
Tel.: +90-312-2101310
Fax: +90-312-2101315
E-mail: medeni.soysal@tubitak.gov.tr

Table 1 Audio-Visual Semantic Concepts that are analyzed in the KavTan System

High Level Concepts	Low Level Concepts	
Violence	*Audio Keyword*	Flag
Nudity	*Video Text*	Crowd
Terrorist Organisation	Explosion Sound	Air
Human Presence	Explosion Image	Cloud
Nature	Gunshot Sound	Grass
	Blood	Animal Sound
	Fire	Soil
	Scream	Skyline
	Crying	Artificial Edge
	Water Image	Speech
	Water Sound	Music
	Human Skin	Silence
	Face	

nels and maintains storage of the most recent 12 months of this huge broadcast data. Since monitoring such a huge data is a vastly challenging problem, utilizing an automated system for semantic analysis and search in this data is an absolute requirement. According to the prioritization of the corporate requirements, five key high-level concepts, namely, violence, nudity, terrorist organization, human presence and nature, are selected as the primary goals of the *KavTan* System. To define these concepts in an objective manner, an audio-visual semantic lexicon containing lower-level concepts is established. The lexicon contains 14 visual and 9 auditory elements, along with two supplementary tools, namely, *overlay video text detection* and *audio keyword spotting*, which are utilized exclusively for the *terrorist organization* concept. A complete list of high-level and low-level concepts (along with supplementary tools) is given in Table 1.

Automatic semantic classification of multimedia content is studied by a wide variety of independent research groups and on a wide variety of semantic concepts. This is observed in TRECVID, which is an annual benchmarking activity, where research groups measure the performance of their algorithms on common datasets using an open, metrics-based approach. In TRECVID 2011, 346 concepts were selected for the semantic indexing task, showing a significant expansion over the 130 concepts selected in TRECVID 2010. In 2010 and 2011, the semantic classification task was called ”Semantic Indexing (SIN)”, which is actually a redefinition of the ”high level feature extraction” task that has been used for representing the same goal for 20 concepts in TRECVID 2009 [45].

There are quite a lot of different approaches for recognition of visual concepts. One of the most successful attempts in the literature is the work of MediaMill group [47, 48]. In their system, multiple frames are selected by temporal sampling of video shots, which are then used to extract features that are to be used in the classification step. The features for each frame are obtained from image pyramids, and converted by using the predefined codebooks, before being classified with Support Vector Machines (SVM). The features that are used for classification are SIFT [30], OpponentSIFT [42] and RGB-SIFT [42] descriptors. The codebooks that are used to con-

vert the features are generated for each type of feature from the training set and are created by using k-Means method. Then, the conceptual classification of a video shot is made from the distribution of the codes from all the features. Similarly, Chang et al.[5] proposes a "bag of words" based system that uses SVM as classifier [22]. The points where the features will be extracted are chosen by Difference of Gaussians and Hessian-Affine methods. Then, from these points SIFT descriptors are computed and those vectors are used as features. From the features obtained from the training set, visual codebook is generated by k-means clustering. The classification is performed according to histograms of this visual dictionary. Peng et al. [36] evaluated the methods described above, to avoid the problems in training of the classifier because of unbalanced distribution of the labels in the training data, they used the training method that they named ABU-SVM. In this method, depending on the penalty function value, negative samples are reduced, while positive samples are increased by copying. The penalty function is calculated based on the classification results, which are obtained by using multiple SVMs that are trained with multiple training sets and their corresponding labels. In a similar research, in TRECVID 2007, IBM and Fudan universities used Sparse Sampling SVM (USVM) [1].

Audio-based semantic classification constitutes an important part of the KavTan System. Majority of the studies in the literature that deal with this problem utilizes the *fingerprint*-based approach. Fingerprints are low-level descriptions that are extracted from short-duration audio segments. Although, these segments are typically fixed-length sliding windows [26, 41], there is another strand of research that utilize variable-duration segments with homogeneous content for descriptor extraction [10, 18, 32]. Mel Frequency Cepstral coefficients (MFCCs) [2, 8, 10, 41], Perceptual Linear Prediction (PLP) coefficients, Spectral Centroid (SC) [50], Spectral Rolloff (SRO) [37, 50] and Zero Crossing Rate (ZCR) [37] are common examples to these low-level descriptions. Biatov et al. [4], takes an alternative approach and use model selection criterion as a confidence metric to retrieve audio events, based on 14 different environmental sounds.

In addition to the generalized concept detection methods, the KavTan System utilizes specialized modules for some of its subtasks. Nudity detection is one of these subtasks. The aim of the nudity detection is to find video shots that contain significantly nude human body instances. Bag-of-Visual-Words (BOVW) method has been used for adult image classification purposes in multiple independent studies [13, 27–29], and successful results were reported. In addition, motion information in video, has also been reported as useful in detecting pornographic content [20, 57].

Visual human detection is another subtask of the KavTan System, whose feasible solution demands methods with specialized structures. In the context of this paper, human presence is objectively defined as existence of at least one of the most prominent human features, i.e. human face and human body. The face detection method of Viola and Jones [52], and human body detection method of Dalal and Triggs [12] are two of the successful methods in the literature.

Detection of predefined flags that consist of distinctive colors and symbols in video frames under a wide range of photometric and geometric transformations stands as a significant challenge by itself in the computer vision area. In the context of the KavTan System, flag detection is considered as another subtask, which requires cus-

tomized treatment, in contrast to the generalized visual concept detection. Significant research regarding this problem have been conducted recently and successful results in the domain of flag and logo detection has been presented. Crandall and Luo proposed a robust algorithm [11] designed to detect a class of compound color objects using a single model image. Their algorithm, which is based on the Color Edge Co-occurrence Histograms, additionally employs perceptual color naming to handle color variation, and pre-screening to limit the search scope. Huttenlocher [19] used Hausdorff distances extracted from edge pixel maps, which can be used to determine the degree of resemblance between two objects that are superimposed on one another. The Hausdorff distance allows tolerance to some geometric distortions. Fergus [14] presented an unsupervised scale-invariant learning method to learn and recognize object class models from unlabeled and unsegmented cluttered scenes. In this method, objects are represented in a probabilistic manner used for all aspects of the object like shape, appearance, occlusion and relative scale. Cheng [7] detected color objects from a cluttered scene based on statistical and spatial color similarity by using Color Region Adjacent Graphs (RAGs) and six one-dimensional histograms corresponding to the RGB and HIS color spaces. Similarity measures of the objects are calculated by histogram intersection (HI) strategy. On the other hand, Phan [38, 39] proposed a logo and trademark retrieval system by using Color Edge Gradient Co-occurrence Histograms (CEGCHs) which is an extended version of Color Edge Co-occurrence Histograms (CECHs). CECGH produces more accurate representation of edges in color images by using edge gradient information. Color quantization scheme is selected based in the Hue-Saturation-Value (HSV) color space. Finally, Wenjing [21] used Color Edge Co-occurrence Histograms (CECHs) to match images by measuring similarities between their CECH histograms. For this purpose, a newly proposed Gaussian weighted histogram intersection (GWHI) is used to measure the similarity between two histograms.

Blood is one of the low-level concepts that are hierarchically under the *Violence* concept in the semantic lexicon (Table 1). Detection of this concept in still images, due to its highly local and irregular shaped appearance characteristics of blood, requires specialized treatment. In this context, despite numerous simple methods have been proposed in the literature that involve pixel or key image matching [6, 9, 34], in the preliminary experiments none of these performed acceptably in unconstrained broadcast videos.

This paper presents a full-fledged semantic concept detection system that operates in the domain of standart, unconstrained broadcast videos. The scope of this system consists of five high-level concepts and 23 lexical low-level concepts (excluding Video Text and Audio Keyword detection supplementary tools). In order to perform detection of this significant set of audio-visual concepts, a generalized framework is designed, which automates otherwise ad-hoc processes like low-level feature selection and classifier parameter optimization. For a short list of concepts, where generalization is not feasible, i.e. nudity, human presence, flag and blood, specialized methods are implemented based on robust predecessors in the literature. The details of the complete list of methods that are utilized in the KavTan System are presented in Section 2. Generalized visual and audio concept detection frameworks are explained in Sections 2.1 and 2.2, respectively. Specialized methods nudity, human presence,

flag and blood are detailed in Sections 2.3, 2.4, 2.5 and 2.6, respectively. High-level decision fusion mechanism that is explained in Section 2.7 concludes the description of the KavTan system. In Section 3, results of the experiments, which are performed on a significant amount of unconstrained TV broadcast data, and evaluated by using objectively labeled ground-truth data, are presented. The presentation is concluded by Section 4, in which experimental evidence is combined with some remarks in order to reach some important conclusions and identify possible research directions.

2 Proposed System

The KavTan System is designed to perform detection of audio-visual concepts in large video archives. In order to achieve this goal, a hybrid framework is designed, which utilizes automated processes that perform low-level feature selection and classifier parameter optimization, along with specialized concept detection methods. Overall structure of the proposed concept detection system is given in Fig. 1. This section is organized to introduce parts of the system in an intuitive format. In Section 2.1, generalized visual concept detection module is introduced, along with a brief overview of basic tools like shot detection and keyframe generation. The generalized audio concept detection module is explained in Section 2.2. Specialized concept detection modules, namely, nudity, human presence, flag and blood detectors are detailed in the sections that follow (Section 2.3, 2.4, 2.5 and 2.6 respectively).

2.1 Generalized Visual Concept Detection

Generalized visual concept detection module of the *KavTan* System builds upon the earlier work of Saracoğlu et al. [43], which utilizes visual codebooks and an SVM-based classifier structure. In the proposed system, features that belong to various spatial categories are extracted from samples that represent video data. The set of extracted features include numerous dense color and texture descriptors that are defined by the MPEG-7 standart [31], along with sparse interest point-based descriptors.

Main aim of the proposed generalized visual concept detection system is to enable systematic training of experts for the wide range of visual concepts listed in Table 1. The structure of the system is designed to serve this purpose of using various attributes that represent visual information from different approaches in combination. Features are quantized using codebooks to provide a robust summary of visual information in the form of visual codewords. Concept specific clustering is achieved by using combinations of codewords from a wide range of visual codebooks, each of which represent a different aspect of the inherent visual characteristic. Four main components of this universal framework for training and recognition of different visual concepts are introduced in Sections 2.1.1-2.1.4.

2.1.1 Spatio-Temporal Sampling

As the first step of visual concept recognition, video is divided into spatial and temporal parts. Temporal sampling of video can be summarized in two steps as shot

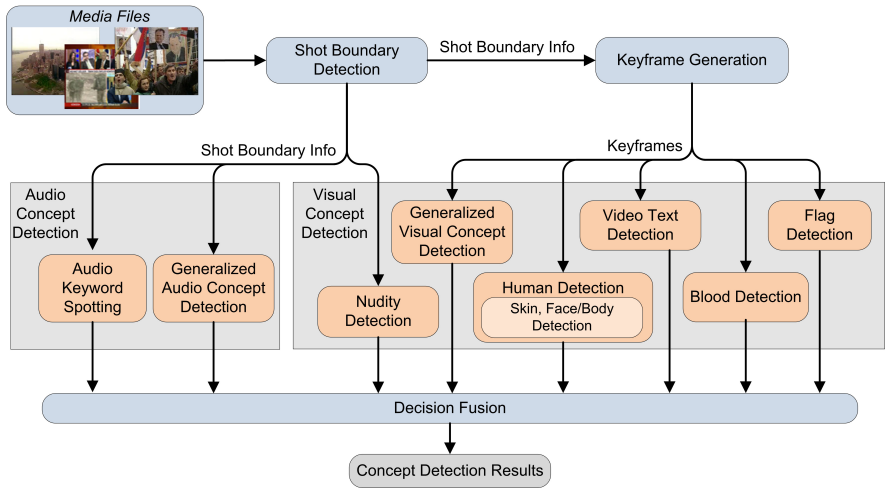


Fig. 1 Outline of the KavTan Concept Detection System

(a) (b) (c)

Fig. 2 Spatial Sampling: (a) Full frame (b) 2x2 grid (c) 3x3 grid

detection and selection of keyframes from these shots. This stage need to be carried out prior to feature extraction system as a prerequisite. Temporal shot boundary information of the detected shots, along with the keyframes representing the shot is provided as input to the next step. Keyframes that represent a shot are obtained by uniformly sampling that shot throughout its duration. The frequency of the samples is left as an implementation parameter.

Spatial segmentation is performed by splitting keyframes obtained from shots into regular-size grids for feature extraction. Three types of grid structures are used in the proposed system, which divide a frame into 1x1 (full frame), 2x2 and 3x3 cells as shown in Fig. **??**. Descriptors are extracted separately from each cell of a grid, providing a semi-local descriptors. By using smaller blocks of a frame, in addition to using the entire key frame, local information is embedded into the descriptors.

2.1.2 Visual Feature Extraction

Temporal and spatial sampling process is followed by the extraction of low-level descriptors from each representative video sample. Low-level visual descriptions that

are extracted from the selected video frames belong to one of the three spatial categories, which are global, semi-global and sparse. Global descriptors are obtained by extracting dense color and texture features from an entire video frame as a whole. These dense features are mainly based on the MPEG-7 Standard and include color features like color structure and color layout, as well as texture features like homogeneous texture and edge histogram. These descriptors are briefly explained in Table 2.

Table 2 Visual Feature Descriptors that are used in Generalized Visual Concept Detection

Visual Feature	Description
Color Correlogram [17]	Global color structure represented by a 166 dimensional HSV color space auto-correlation
Color Moments [49]	First three moments of colors that are calculated from a 5x5 grid
Wavelet Texture [54]	Haar distributions among 12 sub-bands are extracted from a 3x3 grid
Co-occurence Texture [16]	Entropy, energy, and homogeneity of texture in a normalized 96 dimensional vector
Dominant Color [31]	Dominant colors and their statistics such as variance, represented in an efficient, small and intuitional way
Scalable Color [31]	Color Histogram computed in the HSV color space and encoded by Haar transform
Color Structure [31]	Color content (like color histogram) and the structure of this content, described by using a structure element
Color Layout [31]	Spatial distribution of colors in an image, encoded using Discrete Cosine Transform
Homogeneous Texture [31]	Region texture represented by mean energy and energy deviation from a set of frequency channels that are modeled by Gabor functions
Edge Histogram [31]	Global and local edge distribution into 5-bin histograms as horizontal, vertical, 45°, 135° and non-directional
SIFT [30]	Scale Invariant Feature Histogram. Extraction of location, scale and rotation invariant distinctive feature vectors
SURF [3]	Speeded-up Robust Features. 2D Haar wavelet responses computed using integral images

Dense visual features are chosen in a way enabling the representation of visual concepts from different perspectives. Global descriptions, despite the fact that they represent the visual characteristic of a frame using different tools, may only represent this characteristic in the level of the entire frame. In order to bridge this gap between semantics and low-level descriptions, we use semi-global descriptions. Semi-global descriptors, in fact, utilize the same representative feature projections, but this time in a sub-frame level. This is achieved by representing each video frame by dense color and texture descriptors that are extracted from cells of 2x2 and 3x3 grids. In total for a single keyframe and for each single descriptor feature, 14 descriptor vectors, 13 of which are semi-global and one global are extracted.

The third category of low-level features are sparse features, which represent local interesting characteristics inside video frames. SIFT [30] and SURF [3] descriptors, which are well-known and robust representatives of this category are utilized in the proposed system. These descriptors are also briefly explained in Table 2.

2.1.3 Codebook Generation and Classification

Low-level features that are extracted from frames that are sampled from videos vary significantly between different instances of a concept. In order to provide a more robust and efficient representation for training phase, a low-level feature need to be projected into a less complex space, in which determining similar characteristics is feasible. In order to achieve this goal, low-level features extracted from a video shot are mapped into histograms of visual codewords before the classification phase. This mapping is performed using predefined visual codebooks. As a result of this conversion, for each kind of feature a single feature vector that represents the entire shot is obtained, leading to a significant reduction in total feature and computational complexity. The codebooks to be used in the conversion are prepared separately for each type of feature and each concept. These codebooks are generated from the training set feature vectors by k-means clustering method. At the conversion step, all the vectors extracted for a single feature type are paired up with the closest cluster center and histogram of the distribution of features in the cluster space is calculated.

The output of the feature conversion/projection step consists of histograms of codewords, which are accumulated for keyframes of each video shot and for a single low-level feature. These histograms are then utilized in training and classification of nu-Support Vector Regression (nu-SVR) based SVMs [44]. At the end of the classification phase, for each histogram of codewords that constitutes the visual representation of a shot, a confidence value in the range of [0, 1] is obtained. This decision value is calculated using the kernel-based model of Equation 1, whose weights are determined in the training phase.

$$y(\mathbf{x}) = \sum_{k} w_k l_k K(\mathbf{x}, \mathbf{x_k}) + b \tag{1}$$

In Equation 1, confidence value, input histogram vector and training data support vectors are represented by y, $\mathbf{x}$ and $\mathbf{x_k}$, respectively. $K(\mathbf{x},\mathbf{y}) = \exp(-Cd(\mathbf{x},\mathbf{y}))$ is the RBF function, which is used as the SVM kernel. The weight and the class of the support vector $\mathbf{x_k}$ are represented by w_k and l_k, respectively. Lastly, b represents the bias of the decision function.

2.1.4 Generalized Visual Decision Fusion

The confidence values for a concept, for each low-level feature are calculated by using trained models and input histogram vectors as explained in Section 2.1.3. The confidence values obtained by using various visual features of a concept are grouped to reach a more sophisticated collective decision. This decision fusion process flow is illustrated in Fig. 3. As illustrated in Fig. 3, confidence values obtained from each feature are combined in a single vector, which is then classified by a second SVM, in order to produce a final confidence value for a concept.

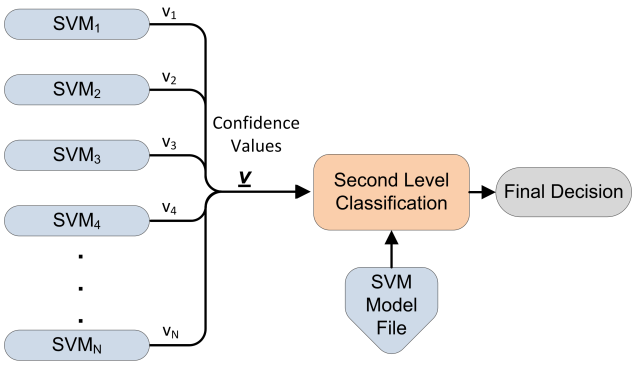


Fig. 3 Generalized Visual Decision Fusion

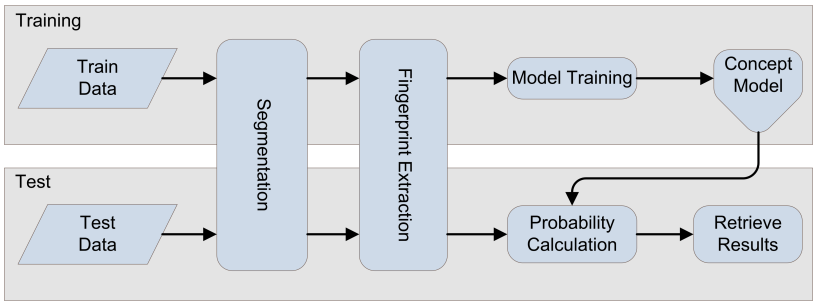


Fig. 4 Generalized Audio Detection Flowchart

2.2 Generalized Audio Concept Detection

Audio contains information that is important for extracting semantic information from multimedia content. For example, explosion sound, weapon sound and scream might imply violence content, whereas crowd, whistle and applause sounds might imply sports programs or protest marches.

The proposed method is applied in four stages. First, the audio data is segmented into homogeneous temporal parts [56]. Next, for each segment, fingerprints that represent the character of the sound are generated. This is followed by the generation of GMMs that are generated from the audio fingerprints obtained from the training data. In the last step, likelihood values of each concept is calculated for test videos using the trained concept models. The outline of the proposed system is given in Fig. 4.

2.2.1 Segmentation

Audio segmentation is used many areas such as auditory-based audio classification, archive management and speaker tracking applications. In an audio stream, multiple concepts can exist at the same time segment. To capture significant auditory

Table 3 Audio descriptors that are utilized in the KavTan System

Descriptor Name	Abbreviation
Mel Frequency Cepstral Coefficients	MFCC
Perceptual Linear Prediction Coefficients	PLP
Spectral Band Power	SBP
Specrtral Roloff	SRO
Spectral Flow Direction	SFD
Harmonicity	HRM
Zero Crossing Rate	ZCR
Harmonic Specral Deviation	HSD
Spectral Flux	SFX
Spectral Flatness	SFL
Spectral Centroid	SC

changes is important for concept recognition since they may represent the transition between the concepts. In this context, unsupervised energy-based segmentation method is used to segment audio data into homogeneous small segments where concept doesn't change [56]. After segmenting long audio streams, by using the feature vectors obtained from consecutive segments, parts that contain silence or background noise are eliminated.

Eliminating conceptually meaningless segments is important for improving the precision of the system. Silence is defined as the regions that have sound energy below the level of minimum energy that ear can hear, as well as regions surrounded by relatively higher-energy regions and considered as background in meaning and sound level. Since broadcast audio contain various records from different sources, recording energy is inconsistent thus determining a dynamic or fixed threshold value for activity/non-activity detection does not result successfully. Therefore, a method that evaluates the energies of successive segments relative to each other is suggested [56].

2.2.2 Fingerprinting

In this study, temporal and spectral descriptors are used to distinguish the different characteristics of audio concepts. For each 10 milliseconds, the short-time attributes that are listed in Table 3 are extracted from 25 millisecond sliding windows.

2.2.3 Model Training

Audio concepts are modeled based on the training data using Gaussian Mixture Models. GMM mixture numbers are optimized based on training data modeling performance and vary between 2 and 64. Confidence values are calculated for each concept seperately using Equation 2.

$$confidence = \log\left(\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right) \tag{2}$$

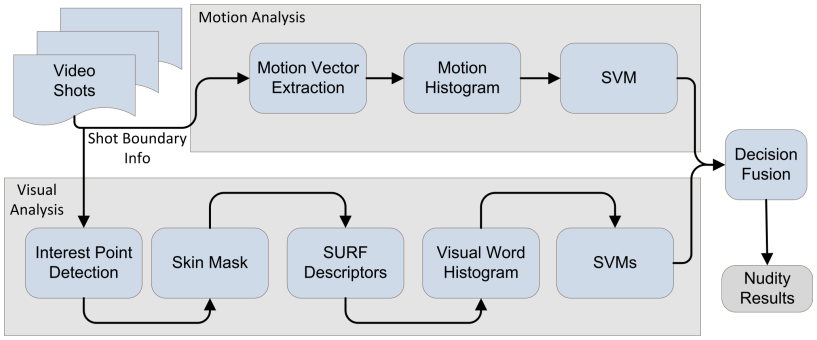


Fig. 5 Nudity Detection Flowchart

2.3 Nudity Detection

The nudity detection module in the proposed system utilizes both stationary visual statistics and motion statistics in the context of nude body detection. In the proposed approach, Bag-of-Visual-Words and motion histogram methods are utilized for visual and motion classification, respectively. The results of the visual and motion classification methods are combined by using probability models for discriminating data that contains significant nudity from normal data, and thus yielding a final classification result.

2.3.1 System Description

The algorithm flow for nudity detection is given in Fig. 5. The system has three main parts, namely, visual classification, motion classification and decision fusion. Stationary visual attributes-based classification utilizes Bag-of-Visual-Words method in order to represent underlying concept-based characteristics. The aim is to discriminate video shots as belonging to one of the two categories “normal” or “nudity”.

The first step of the visual classification is the extraction of interest points. Due to its extensive use in the literature and reported success in capturing low-level local blob-shaped details, Speeded-up Robust Features (SURF) method [3] is utilized in detecting interest points. Detected interest points are then filtered by using a skin filter [23], due to the strong assumption that most of the nudity related information is present on the skin areas and their immediate vicinity. SURF descriptors are then extracted from interest points, which reside on areas that are labeled as skin by this filter. In the next step, SURF descriptors from skin areas are mapped to visual codewords according to a predefined visual codebook. All visual words that are extracted from keyframes of a shot are combined to give a visual word histogram, which is interpreted as the shot’s fingerprint. As an implementation detail, in this module, we represent video shots with keyframes that are taken with 0.6 second intervals. In the final step, fingerprints that are obtained for each shot are classified by using SVMs as “normal” or “nudity”.

The second information source for nudity detection is selected as motion statistics of video shots. For this purpose, motion classification method proposed by Jansohn et al. [20] is utilized. In order to obtain motion histograms, first of all, the frame is divided into 3 rows and 4 columns. Next, motion vectors that are already calculated by the video encoder (MPEG-4) are extracted from the video data and 12 independent motion histograms are created as shown in the Fig. 6. Eventually, these histograms are normalized and concatenated into a single histogram, which will be used for normal/nudity classification by an SVM.

Independently obtained decision results for visual and motion classification are combined by using the decision fusion function given in Equations 3 and 4. SVMs that are used for visual and motion classification produce confidences in addition to the classification labels. In the training phase, we create probability models for positive (nudity) and negative (normal) data sets by using confidence values of shots. Through these models, expected confidence values for nudity and normal shots are computed. In the classification phase, we use the decision label and decision confidence values of the classification SVMs to compute the likelihood $p(nudity|c_v, c_m)$ of a shot containing nudity. c_v and c_m represent the confidence values of visual and motion SVM classifiers, respectively. Similarly, $p(normal|c_v, c_m)$ is the likelihood that the shot does not contain nudity concept. The ratio of $p(nudity|c_v, c_m)$ and $p(normal|c_v, c_m)$ terms, i.e. r_{nudity}, is compared with the threshold λ_{nudity} to reach the final normal/nudity decision. Value of the λ_{nudity} parameter is determined empirically by analyzing the precision-recall relationship during the training experiments.

$$\frac{p(nudity|c_v, c_m)}{p(normal|c_v, c_m)} = r_{nudity} \tag{3}$$

$$Decision = \begin{cases} nudity, & r_{nudity} > \lambda_{nudity} \\ normal, & r_{nudity} \leq \lambda_{nudity} \end{cases} \tag{4}$$

The final confidence value for nudity (c_{nudity}) on each shot decision is calculated by the Equation 5. This function is obtained by empirical analysis of system results and does not affect decision label, instead, it merely operates as a normalizer on confidence values.

$$c_{nudity} = \frac{1}{1 + \frac{1000}{r_{nudity}}} \tag{5}$$

2.3.2 Model Creation

Approximately two hours of video during training, validation and testing steps of the model creation phase. It should be noted that this dataset does not have any overlap with the dataset that is used for performance measurements in Section 3. Table 4 shows the fragmentation of the dataset into training, validation and test subsets. This fragmentation is performed with randomization, during which, each shot has 60% chance of belonging to training, 20% to validation and 20% to test.

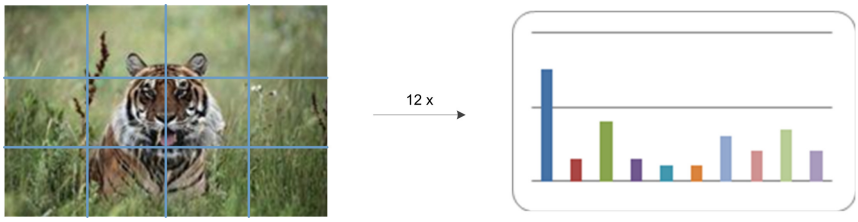


Fig. 6 Motion histogram generation from video frame grid cells

Table 4 Training and Test Dataset Statistics

	Duration	# of Shots
Training	1 hour 17 min. 52 sec.	4669
Validation	25 min. 57 sec.	1556
Test	25 min. 57 sec.	1557
Total	**2 hours 9 min. 48 sec.**	**7782**

This results in a dataset fragmentation of approximately 60/20/20, corresponding to train/validation/test data sizes.

The content of the data set includes videos from television broadcasts and short adult videos collected from the Internet. Television broadcasts are mainly used as negative (normal) data, while Internet-based videos contain only adult content. Positive to negative ratio in the data is deliberately kept as 1:2 to prevent imbalanced training.

Training of visual classification starts with creating visual codebook. In order to create our codebook, SURF descriptors are extracted from training shots as described in Section 2.3.1. Extracted SURF descriptors (more than 7 million) are then clustered by using standard k-Means clustering with 512 cluster centers. The number of cluster centers are determined empirically through performance analysis of results obtained by using 256, 512, 1024 and 2048 cluster centers.

SVM kernel selection during training phase has a significant effect on the final performance of the system. Different kernels yield different classification performances depending on the domain of the problem. In the context of nudity detection, a comparison between performances of RBF, CHI-RBF, Chi square and Histogram Intersection kernels were performed, which resulted in the selection of the CHI-RBF kernel.

Training of visual classification system by using the training data set is evaluated on the validation data set for preliminary performance analysis in terms of accuracy. In our experiments, after learning and parameter optimizations, visual classification system achieved 99% classification performance on training data set and 89.88% classification performance on validation data set.

In the training of motion classification system, different motion vector normalization methods and different SVM kernels were investigated. In experimental comparisons, Euclidean normalization method performed best in normalization of 12 motion

Table 5 Expected Confidence Values for Nudity/Normal Training Data

	Normal	Nudity
Visual	0.103	0.732
Motion	0.208	0.592

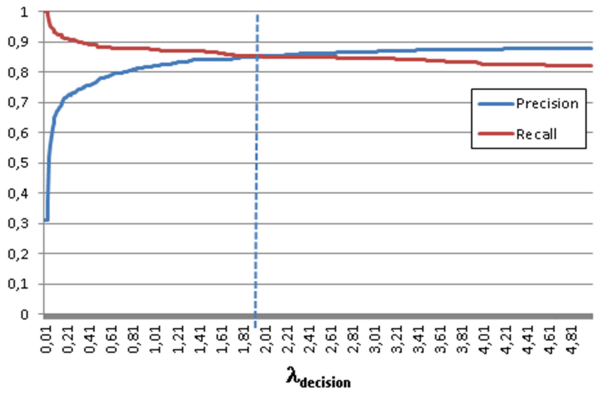


Fig. 7 Precision-Recall curve with respect to the change in $\lambda_{decision}$ parameter

histograms that are obtained for each shot. Similar to the stationary visual features, for the classification of motion histograms SVM with CHI-RBF kernel performed best during the training experiments. The best performance values achieved were 87.21% for the training data set and 81.7% for the validation data set.

Shot decisions (Equation 4) and their confidence values (c_v, c_m) are used to build the probability models, i.e. $p(nudity|c_v, c_m)$ and $p(normal|c_v, c_m)$ for decision fusion. Table 5 shows the expected nudity confidence values for training samples that belong to normal and nudity classes.

Classification of new shots are performed by using the SVMs and the probability models that are trained as explained previously. A video shot is first evaluated using visual and motion classifiers, and final decision is made by using Equations 3 and 4. Calculation of the threshold λ_{nudity} that is used in the final decision step is illustrated on precision recall curves of training phase that are given in Fig. 7.

Final tests during the model creation phase is performed by using the test data set, which constitutes 20% of total training data. The combined nudity model achieved 85.3% classification performance on this dataset.

2.4 Visual Human Detection

Visual human detection module is composed of two submodules, namely, multi-view face detection and whole-body detection. Output of these two submodules are fused empirically for efficient visual human detection. The general flow of the visual human detection module is given in Fig. 8.

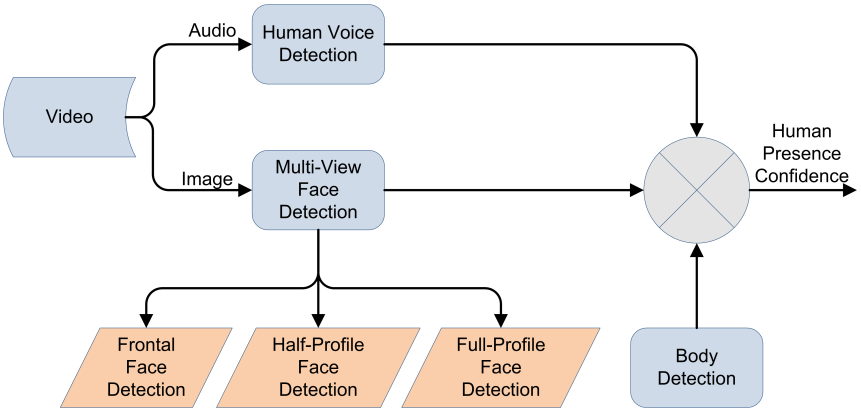


Fig. 8 Visual Human Detection Flowchart

(a) (b) (c) (d)

Fig. 9 (a,b) Face Detection Results, (c,d) Body Detection Results

2.4.1 Multi-View Face Detection

Face detection is one of the most popular applications of computer vision and lies in the foundation of many human-computer interaction systems. It forms an important building block of visual human detection and face recognition systems. There are two main categories of challenges for achieving robust face detection. The first one is coping with variations in shape, size, pose and orientation of faces. Dealing with environmental factors like brightness variations, illumination changes and occlusions is considered as the second category.

In the KavTan System, face detection algorithms are based on the method proposed by Viola and Jones [52, 53]. For achieving robustness against pose variations, face images are separated into different classes depending on pose and and Adaboost-based classifiers are trained for each class. Training of these classifiers are performed using images from the FERET data set [40]. Representative results of the proposed multi-view face detection submodule are given in Fig. 9.

2.4.2 Human (Body) Detection

Utilization of human body detection methods are widespread in a variety of application domains, such as security, crowd analysis, vehicle automatic pedestrian detection and content-based video analysis. In the literature, various approaches are proposed

to achieve this goal. Viola and Jones proposed an Adaboost based system, which integrates image intensity information with motion information [25]. This system works even in low resolutions, and have demonstrated its ability to detect human bodies with a pixel size of at least 20 by 15. Mikolajczyk and Schmid proposes a system that is based on probabilistic combination of different body parts detectors [33]. In their system, parts-based training is performed by using Adaboost classifiers. Arguably, Dalal and Triggs's algorithm [12], which is based on Histogram of Oriented Gradients (HOG) stands out as the most popular among all body detection approaches. In this approach, a locally normalized histogram of gradient orientations is calculated within overlapping search windows of size 64x128. These histograms are then classified by using a linear SVM, with an emphasis on speed and efficiency. During the implementation of the human body (whole body) detection submodule, Dalal and Triggs' approach is adhered due to speed and efficiency considerations. Representative detection results of the implemented method are given in Fig. 9.

2.5 Flag Detection

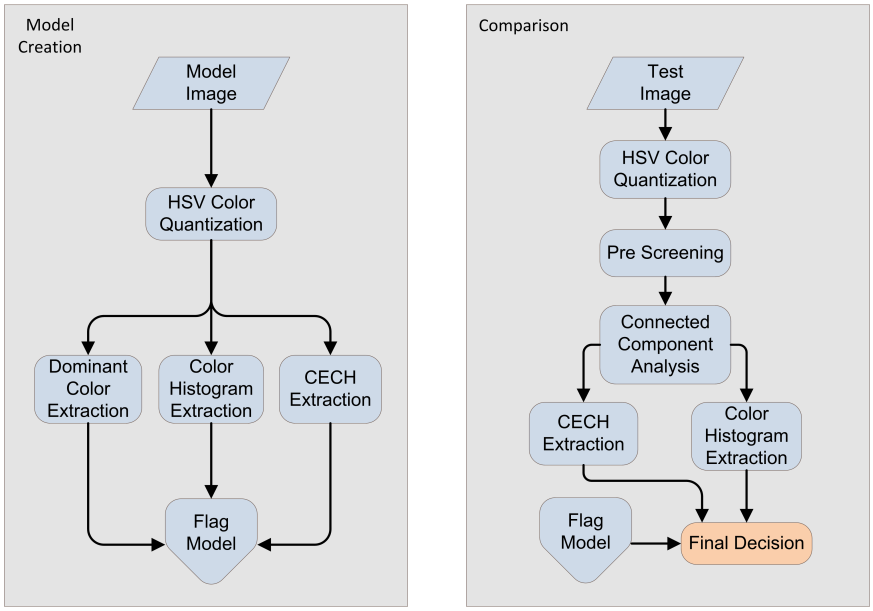
In the context of the KavTan System, the goal of the flag detection module is limited to detection of flags that symbolize a specific terrorist group. A robust algorithm for detecting compound color objects (flags, logos etc.) in unconstrained images is developed and utilized for the specific goals of the KavTan System.

2.5.1 System Description

Compound color objects are objects, which can be characterized by a specific set of colors that are arranged in a unique spatial layout. Unconstrained, general purpose object detection is considered as the holy grail of computer vision with an infinite number of potential applications. Despite years of research, there has been little success in creating a single general-purpose algorithm that can detect arbitrary objects in unconstrained images. Instead, proposed algorithms are typically customized for the specific object under consideration.

Detection of compound color objects like flags lies on the high-difficulty side of the problem, since their appearance may vary drastically from scene to scene due to the non-rigid nature of the base material. For instance, a flag is subject to self-occlusion and non-affine distortion, depending on wind conditions. Since orientation of images and objects are not always known, the detector must be invariant to rotation. It must also be robust to color variations arising from illumination changes and inherent color differences among various instances of the object [11].

In the KavTan System, a Color Edge Co-Occurrence Histogram (CECH) [11] based method is applied to the specific flag detection problem. For this purpose, a simple color histogram (10 bins) and Color Edge Co-Occurrence Histograms are used to describe images. A linear regression-based similarity measure is utilized for performing histogram comparisons. In the proposed system, color quantization and pre-screening steps are performed in HSV color space. The proposed method is illustrated in Fig. 10.



(a) (b)

Fig. 10 Flag detection flowchart: (a) Feature extraction from a model image (b) Object detection in a search image

2.5.2 Model Creation and Comparison

The algorithm starts to work with a model image M and an input search image I which may contain zero, one, or multiple target objects. Firstly, a single model image is used for color quantization and extracted some features like dominant color, color histogram and color edge co-occurrence histogram. Fig. 10(a) shows the feature extraction process from the model image. Next, a search image is examined for the possible objects by comparing the search image with the model image. For this purpose, the colors of the search image are quantized and a pre-screening routine is used by using model image dominant colors. Following the pre-screening step, possible object regions are labeled by using connected component analysis for color histogram and CECHs extraction. Finally, the resulting color histograms and CECHs for the possible object regions are compared with the model image histograms to reach a decision about the object presence. Fig. 10(b) shows the object detection flowchart for the search image.

2.6 Blood Detection

The aim of the blood detection module is to robustly detect shots that contain images or visual illusions of blood in TV broadcast videos. There are many approaches in the literature, which use simplistic methods, such as pixel-matching or key image com-

Fig. 11 Representative flag (Terrorist Organization) detection results under various attacks. (a) Moderate brightness change (b) Extreme brightness change, (c) Multiple instances with moderate photometric and geometric transformations, (d) Extreme non-rigid geometric transformation

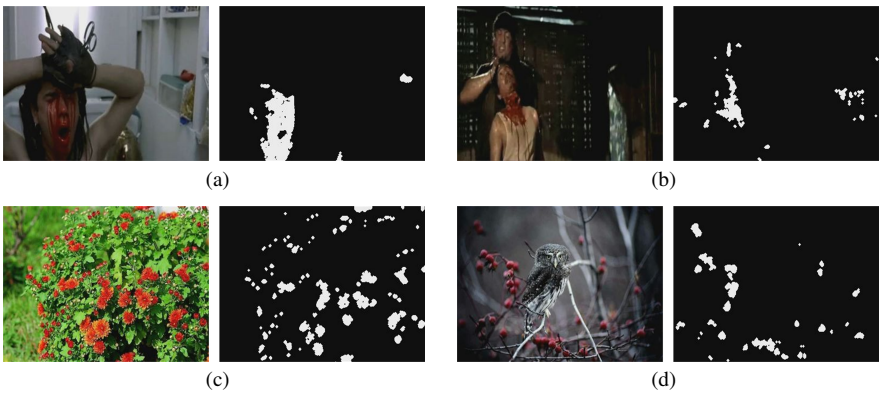


Fig. 12 Pixel Color-based Blood Detection Results: (a,b) True Detections, (c,d) False Positives

parison [6, 9, 34] for blood detection. These algorithms are vulnerable against false positives in unconstrained TV content (Fig. 12). In this study we use a combination of methods that are based on color and shape descriptors.

The proposed algorithm, which is illustrated visually in Fig. 13, can be summarized as follows: (1) Apply color enhancement in order to render the method robust against illumination changes and blood color variations; (2) Segment the im-

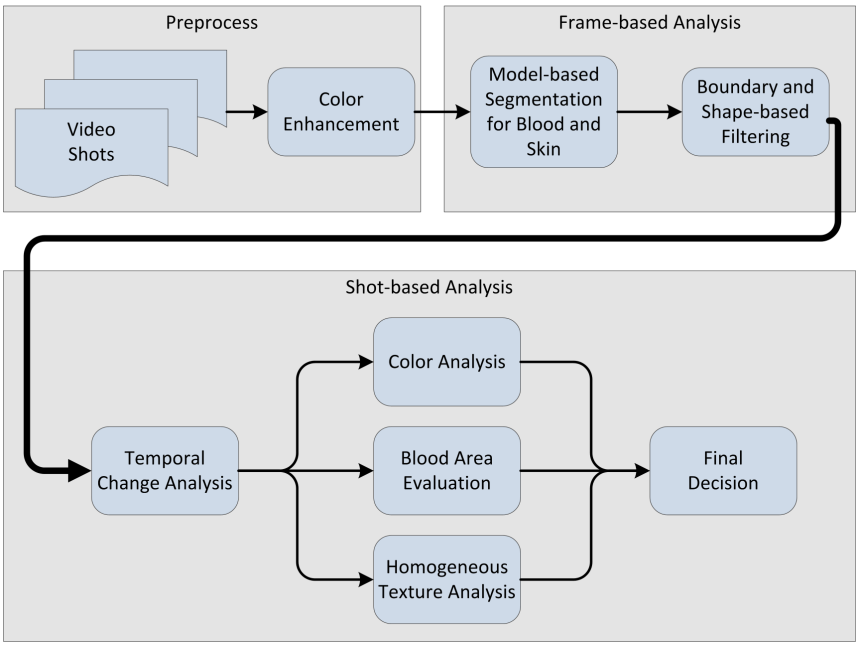


Fig. 13 Blood Detection Flowchart

age and determine candidate blood regions by using statistical color distribution in scenes that contain blood; (3) Extract region and boundary descriptors from candidate frames, and filter them using region-based shape information; (4) Analyze the temporal change in candidate regions and filter regions that do not comply with the expected characteristics; (6) Calculate multi-modal probabilities for blood existence in video shots; (7) Consolidate final video shot-based confidence values.

2.6.1 Preprocess

Color enhancement algorithm (Fig. 13) enhances color images during conversion from RGB to HSV space by using a non-linear transfer function, taking pixel neighborhood into account. During this process, in order to maintain the color balance, only the V (luminance value) component is enhanced in terms of luminance and contrast. Both luminance and contrast enhancement are performed through dynamic range compression using specially designed nonlinear transfer functions defined in [51] and [15], respectively.

2.6.2 Frame-based Analysis

A central task in visual learning is the construction of statistical models of image appearance from pixel data. For this purpose, the color distribution statistics of blood and normal pixels were extracted in the form of color histograms, and modeled with

Fig. 14 Frame analysis result for blood detection: (a) Original image (b) Input image (after color enhancement) (c) Color based image segmentation (d) Input image masked with segmented binary image (e) Boundary analysis (f) Skin detection result

Gaussian mixture models (GMM). Training set consists of 2435 images, which are utilized for modeling 16 mixture models for each class. Hue-Saturation-Value (HSV) is empirically selected as the most distinctive color space for blood concept.

Skin detection module has been added to the blood detection since most of the bloody scenes include human existence. In this work, statistical color modeling is used for skin detection [24]. Skin regions affect the confidence value calculation of the scenes (Section 2.6.3).

After color based image segmentation, boundary and shape descriptors are extracted from the segmented regions. The regions which have low convexity defects ratio and/or small size are filtered out since blood regions generally have high convexity defect ratios. Frame-based analysis steps are illustrated visually in Fig. 14.

2.6.3 Shot-based Analysis

Candidate blood regions, which are the output of the frame-based analysis step are examined in the temporal domain. In this step, temporally static regions like TV logos or objects, which do not change for a long time, are eliminated.

At this point, all candidate regions that remain are considered as positive detections, and qualify for confidence calculations. Confidence values are calculated based on three independent feature categories, namely, color, texture and region size, and converted into a shot confidence (c_{shot}) according to the formula in Equation 6. $w_{texture}$ and $w_{color-region}$ represent the weights of texture and color-region size combined confidence values, respectively. Texture-based confidence value ($c_{texture}$) is obtained by using an SVM model (Equation 1), trained on MPEG-7 Homogeneous

Texture [31] descriptors of blood and normal regions.

$$c_{shot} = w_{texture} \cdot c_{texture} + w_{color-region} \cdot c_{color-region} \tag{6}$$

On the other hand, $c_{color-region}$, which is the color and region size-based combined confidence, is calculated by using Equation 7.

$$c_{color-region} = w_{color} \cdot c_{color} + w_{region} \cdot c_{region} + b_{color} \tag{7}$$

In this formula, w_{color} represents the weighting factor for the color-based confidence, c_{color}. Each keyframe in a video shot is assigned into one of the three categories, as frames that consist of both blood and skin regions, only blood regions and neither blood nor skin regions. c_{color} is calculated by assigning the quantized values 1.0, 0.5 and 0 to each of these categories, and averaging the sum on each video shot. b_{color} is a color-based bias value for a shot, whose value (in $[0, 0.65]$) depends on the existence of blood and skin in any of its frames. Finally, w_{region} represents the weighting factor for the region size-based (c_{region}) confidence, which is calculated by using the piecewise function given in Equation 8, which penalizes regions with size out of an empirically defined range (15-50% of frame area). μ_{Area} in the formula represents the ratio of the blood region area to the video frame size.

$$c_{region} = \begin{cases} 1 - \exp(-30 \cdot \mu_{Area}), & \mu_{Area} < 0.15 \\ 1, & 0.15 \leq \mu_{Area} \leq 0.5 \\ \exp((0.5 - \mu_{Area})/0.08) & \mu_{Area} > 0.5 \end{cases} \tag{8}$$

The proposed algorithm has been tested on a video dataset that consists of 9943 shots. 369 of these 9943 shots contain blood images. The results showed that the proposed algorithm can eliminate most of the regions whose color and region are similar to the blood such as channel logos, objects or background. On the other hand, significantly dark scenes and scenes with very small blood regions remain as problematic cases.

2.7 High-Level Decision Fusion

Typically, each low-level concept detector produces a confidence value between $[0, 1]$ for each video shot. These confidence values from relevant low-level detectors are then merged into a single high-level concept confidence value by using rule-based methods. This decision fusion process for reaching each of the high-level concepts from low-level lexical elements is explained in this section.

High level decision fusion process aims to generate a single confidence value in the range $[0, 1]$ for each of the 5 high level concepts by using low level concept detection results. For each high-level concept, one or more of the low-level concept detection modules, which also produce condence values in the range $[0, 1]$ for each video shot, are utilized. Decision fusion methods vary from one high-level concept to the other, since each concept is related to low-level concepts in its own unique way. Details of the high-level decision fusion process, which is tailored for each high-level concept according to training performance results is given in Sections 2.7.1 through 2.7.5.

2.7.1 Nudity

For the nudity detection high-level concept, results of the specialized module defined in Section 2.3 is directly utilized. The module produces confidence values in the range $[0,1]$ for each video shot.

2.7.2 Violence

For detection of the violence concept, Generalized Visual Concept Detection (GVCD), Generalized Audio Concept Detection (GACD) and blood detection modules are utilized. By the GACD module, scream, gunshot and explosion sound are identified and a confidence value for each one is produced. The output of the GACD module is selected as the maximum of these confidence values. A similar procedure is followed in the Generalized Visual Concept Detection module. Explosion image and fire detection results are obtained by this module and the maximum of these confidence values is chosen as the final output. Specialized blood detection module directly contributes with its own confidence value, which is also in the range $[0,1]$.

Final confidence value for violence concept ($c_{violence}$) is calculated using Equation 9, as the weighted maximum of the outputs of these three modules. The weights w_{GVCD}, w_{GACD} and w_{Blood} are determined based on the empirical reliability of GVCD, GACD and blood detection results on the training data set.

$$c_{violence} = \max(w_{GVCD} \cdot c_{GVCD} \ , \ w_{GACD} \cdot c_{GACD} \ , \ w_{Blood} \cdot c_{Blood}) \tag{9}$$

2.7.3 Terrorist Organization

For the terrorist organization concept, audio keyword spotting (AKS), overlay video text detection (VTD) and flag detection modules are used. Special keywords are searched using VTD and AKS modules and a confidence value in the range $[0,1]$ is calculated by each module. Specialized flag detection module directly contributes with its own confidence value in the range $[0,1]$.

Final confidence value for the terrorist organization concept (c_{TO}) is calculated using Equation 10, as the weighted maximum of the outputs of these three modules. The weights w_{VTD}, w_{AKS} and w_{Flag} are determined based on the empirical reliability of VTD, AKS and flag detection results on the training data set.

$$c_{TO} = \max(w_{VTD} \cdot c_{VTD} \ , \ w_{AKS} \cdot c_{AKS} \ , \ w_{Flag} \cdot c_{Flag}) \tag{10}$$

2.7.4 Nature

Two different methods are tested for reaching the nature concept. In the first method, cloud, grass, air, water, soil and skyline concepts are identified by Generalised Visual Concept Detection (GVCD) module, and the maximum of the confidence values is chosen as the final result.

In the second method, the nature concept is considered as a low-level concept is directly modeled by the GVCD module. During the modeling phase of this method,

training datasets for all of the nature-related low-level visual concepts, i.e. cloud, grass, air, water, soil and skyline, are used in combination. In the experiments, the second method outperformed the first one with a clear margin. Therefore, the performance results that are obtained by this method are reported in Table 8.

2.7.5 Human Presence

Human Presence is another concept that utilizes a specialized module, namely visual human detection. In addition to visual human detection module, Generalized Audio Concept Detection (GACR) is used for human voice detection. As a heuristic specific to this concept, presence of human voice does not guarantee the presence of a human in a video shot. Commercials, documentaries and news videos, which are narrated by a background speaker, are typical examples of this phenomenon.

Due to the aforementioned reasons, human voice detection result is considered as a secondary degree clue, or a modifier for visual human detection output. In accordance with this fact, if the result obtained from visual human detection module is very low or very high, then GACR module has no effect on the final result. In these cases, the results of the visual human detection module alone forms the human presence confidence values. In other cases, where the visual human detection result is in a predefined range, GACR module contributes to the final confidence through an empirically-determined weight, along with the visual human detection.

3 Experimental Results

The main purpose of the experiments presented in this section is to objectively assess the semantic classification performance of the system that is defined in Section 2 for five high-level semantic concepts, namely, violence, nudity, terrorist organization, human presence and nature. In order to perform labeling of test data in a normative and unambiguous manner, these concepts are defined in terms of a semantic lexicon containing lower-level concepts. According to this definition, in order for some video content to be labeled as having one of the high-level concepts, it should contain one or more of the related elements from the lower-level concepts, or in other words, audio-visual semantic lexicon (e.g. presence of nudity requires the presence of human skin). This audio-visual semantic lexicon contains 14 visual and 9 auditory elements. Visual elements of the lexicon are explosion, blood, fire, human skin, face, flag, crowd, water, air, cloud, grass, soil, skyline and artificial edge. On the other hand, auditory lexicon elements are speech, music, silence, explosion sound, gunshot sound, animal sound, scream, crying and water sound. In addition to these elements, *overlay video text detection* and *audio keyword spotting* tools are utilized to provide supplementary clues (used with a predefined keyword list and only for the terrorist organization concept). Operators responsible for labeling test data for high-level concepts and lexical elements are provided with normative semantic definitions beforehand to minimize subjectivity in the labeling process.

We tested the proposed system mainly on unconstrained video data that is recorded from broadcasts of Turkish national television (TV) networks. For those concepts that

Table 6 Frequency of Concepts in Common TV Broadcast

Concept Appearance Frequency Category	Low		Medium		High	
Concept Semantic Level	Low	High	Low	High	Low	High
Minimum Number of Positive Shots	100	250	250	500	1000	2000
Minumum Negative / Positive Shot Ratio	8	8	4	4	1	1

are rare in common national TV broadcast, test data is complemented using video collected from the Internet. The amount of test data to be used for each concept and lexicon element is determined based on an objective criterion. Each of the 28 semantic concepts (5 high-level and 23 low-level), along with concept-related video text and audio keyword are first assigned to one of the three frequency categories (low, medium or high) according to their frequency of appearance in common national TV broadcast. Depending on a concepts frequency category and semantic level (i.e. high-level or low-level), the constraints on the amount of its test data varies. The guidelines for calculating the constraints on the amount of test data are detailed in Table 6. These guidelines are used to approximately determine the meaningful amount of positive and negative data in the test set. According to the guidelines in Table 6, the annotated test data distribution utilized during the experiments is detailed in Table 7. Each shot in a test video is annotated for all of the relevant concepts as positive, negative or ambiguous. Only the positive and negative labels for each concept are included in the performance computations. On the average, 912 positive shots, 4725 negative shots and 6 hours of video are utilized for testing the system on a single concept. At this point, it is important to emphasize the fact that there is no overlap whatsoever in terms of content between the test and the training datasets, in compliance with the prerequisites of a fair evaluation.

The performance of the proposed system is evaluated using *Mean Average Precision* (mAP) metric [46], which is a widely accepted criterion for measuring retrieval performance in large databases [35, 46]. This metric is defined in Equation 11.

$$mAP = (\frac{1}{N_p}) \sum_{k=i}^{N_p} \frac{i}{o_i}, \tag{11}$$

where N_p is the number of shots in the test data that are annotated as consisting of the concept (positive labels) and o_i is the order of the i^{th} positively labeled shot in the returned result list.

During our experiments, in addition to the original, we derived another useful definition in order to better assess the performance of the system from perspectives of various users. This modified definition is given in Equation 12.

$$mAP_k = (\frac{1}{N_{pk}}) \sum_{i=1}^{N_{rk}} \frac{i}{o_i} \tag{12}$$

where $N_{pk} = min(k, N_p)$ is the number of positively labeled shots that fits into the first k retrieved results, N_{rk} is the number of positive shots returned in first k results and o_i is the order of the i^{th} positively labeled shot in the returned result list.

Table 7 Concept-based Distribution of the Test Dataset

	Semantic Level	Appearance Frequency	# Positive Labels	# Negative Labels	Total Test Data Duration (hr:min)
Violence	High	Low	1206	10957	11:15
Nudity	High	Low	271	13984	11:10
Terrorist Organisation	High	Low	829	8271	13:50
Human Presence	High	High	4283	3982	09:00
Nature	High	Medium	1413	5422	08:00
Video Text	Low	Low	387	4295	07:00
Explosion Image	Low	Low	110	3556	04:00
Blood	Low	Low	369	9574	08:40
Fire	Low	Low	427	3161	03:34
Human Skin	Low	High	1359	2224	04:00
Face	Low	High	3773	4518	09:00
Flag	Low	Low	212	2100	04:35
Crowd	Low	Low	365	4029	04:00
Water Image	Low	Low	334	3628	05:45
Air	Low	Medium	715	4294	05:00
Cloud	Low	Medium	337	2722	04:45
Grass	Low	Medium	563	4490	06:00
Soil	Low	Low	385	3276	04:45
Skyline	Low	Low	300	4141	05:00
Artificial Edge	Low	High	1780	5038	08:15
Audio Keyword	Low	Low	125	2585	03:20
Speech	Low	High	4049	4417	07:00
Music	Low	High	1735	1492	03:50
Silence	Low	Low	186	5771	04:44
Explosion Sound	Low	Low	412	5068	04:44
Gunshot Sound	Low	Low	141	4528	03:44
Animal Sound	Low	Low	204	2584	04:15
Scream	Low	Low	350	4386	04:00
Crying	Low	Low	246	3476	03:40
Water Sound	Low	Low	503	3781	05:00

Using the metric in Equation 12, experiences of users with different levels of focus can be simulated. To exemplify, a user with limited time and interest would only analyze the retrieved results superficially by typically looking at the first 20 retrieved results. The success of the system from this users perspective can be captured in mAP_{20} compared to the original mAP, which requires considering thousands of data for performance computation. In our experimental results, we present the performance of the system in four detail levels, namely, mAP_{20}, mAP_{100}, mAP_{1000} and original mAP (denoted as mAP_{∞}).

The retrieval performance for five high-level semantic concepts are given in Table 8. Performances measured for each of the 14 low-level visual concepts of the semantic lexicon, along with the visual supplementary tool *overlay video text detection* that is utilized for the *terrorist organization* high-level concept are presented in Table 9. Lastly, retrieval performances for the 9 low-level audio concepts and the *audio keyword spotting* tool that is utilized for the *terrorist organization* concept are given in Table 10. In each of these tables, namely, Tables 8, 9 and 10), in addition

Table 8 High-level Concept Retrieval Performances

	mAP_{20}	mAP_{100}	mAP_{1000}	mAP_{∞}	Computation Time (min)
Violence	0.875	0.771	0.364	0.459	72.2
Nudity	0.503	0.405	0.321	0.357	16.8
Terrorist Organisation	1.000	1.000	0.395	0.558	14.9
Human Presence	1.000	1.000	0.888	0.826	34.3
Nature	0.463	0.462	0.292	0.445	79.5

Table 9 Low-level Visual Concept Retrieval Performances

	mAP_{20}	mAP_{100}	mAP_{1000}	mAP_{∞}	Computation Time (min)
Video Text	1.000	0.876	0.543	0.597	7.4
Explosion Image	0.000	0.002	0.008	0.032	20.1
Blood	0.050	0.016	0.013	0.057	12.8
Fire	0.452	0.398	0.263	0.339	63.3
Human Skin	0.590	0.618	0.594	0.677	34.8
Face	1.000	1.000	0.898	0.834	20.3
Flag	0.912	0.438	0.345	0.373	9.5
Crowd	0.717	0.620	0.491	0.536	7.5
Water Image	0.412	0.296	0.258	0.312	66.2
Air	0.850	0.633	0.347	0.462	73.7
Cloud	0.533	0.194	0.143	0.213	71.8
Grass	0.473	0.363	0.268	0.372	60.8
Soil	0.398	0.326	0.249	0.320	70.3
Skyline	0.516	0.352	0.302	0.344	75.1
Artificial Edge	0.766	0.521	0.323	0.467	74.1

Table 10 Low-level Audio Concept Retrieval Performances

	mAP_{20}	mAP_{100}	mAP_{1000}	mAP_{∞}	Computation Time (min)
Audio Keyword	0.495	0.233	0.240	0.270	10.4
Speech	1.000	0.933	0.902	0.880	7.2
Music	0.912	0.899	0.755	0.790	7.0
Silence	0.933	0.821	0.863	0.863	6.7
Explosion Sound	0.750	0.621	0.560	0.593	7.0
Gunshot Sound	0.364	0.259	0.373	0.377	7.1
Animal Sound	0.689	0.355	0.407	0.427	8.9
Scream	0.149	0.103	0.144	0.205	7.5
Crying	0.483	0.250	0.195	0.235	7.5
Water Sound	0.209	0.137	0.141	0.238	7.6

to the retrieval performances, the approximate computation times for each concept detector are also included (last column in each table). These computation times (in minutes) are measured on a typical one hour benchmark video containing 1900 shots and include all of the operations performed on a video, i.e. shot boundary detection, feature extraction, low-level classifications, decision fusion and high-level decision. Average retrieval performances of the system in terms of mAP_{20}, mAP_{100}, mAP_{1000} and mAP_{∞} for the high and low-level concepts are given in Table 11. The definitions of mAP_{∞} and mAP_k are given in Equations 11 and 12, respectively.

Table 11 Average Concept Retrieval Performances

	mAP_{20}	mAP_{100}	mAP_{1000}	mAP_{∞}
High-level Concepts	0.768	0.728	0.452	0.529
Low-level Concepts	0.602	0.469	0.405	0.449
Overall	**0.616**	**0.497**	**0.401**	**0.452**

4 Conclusion

We presented a framework for application of classification methods to high-level semantic classification in large-scale TV archives. The performance of the presented framework was assessed objectively over a wide range of semantic concepts (5 high-level, 14 visual, 9 auditory, 2 supplementary) Performance results were obtained by using a significant amount of precisely labeled ground truth data.

The concepts that were targeted in this study were selected according to the needs and requests of RTÜK, a governmental agency that is responsible for evaluating the content of the TV and radio broadcast, in terms of adherence to guidelines that are defined by laws. Most of the concepts in this work are either not existent or defined differently in the literature. Due to this fact, finding a meaningful baseline for performance comparison was not possible. However, a target success level is devised from the average performances obtained by similar methods in the literature. The most extensive content-based classification benchmark in the literature, TRECVID [35], which accepts contributions from organizations all around the world is used for this purpose. Although, most of the concepts in our study are not part of the target concept lists of TRECVID, the performances of significant organizations on similar high-level concept classification problems provides a valuable baseline for our performances. In this study, we evaluated the algorithm performances using Mean Average Precision (mAP) metric [46]. TRECVID measures performance levels in the presence of incomplete and conflicting judgments of participating organizations and therefore, uses an estimated version of the original mAP metric called *Inferred Average Precision (infAP)* [55]. Considering the performances obtained in TRECVID over the last five years in terms of infAP, it can be seen that there is a large spread among the scores of the target concepts. In addition, the performances vary significantly between the utilized datasets. In the light of the average performance results obtained, and by evaluating the requirements of the end users of the system, in our studies we set a relatively high average target performance level of 0.4 in terms of the mAP metric [46].

In total, 30 semantic concepts were indexed automatically in 60 hours of TV broadcast. This test set consist of 27369 positive and 141750 negative labels to be used as ground truth annotations. Our system achieved an average performance result of 0.452 in terms of mAP. From the high-level concepts perspective, in accordance with the TRECVID results [35], performance varies greatly between the well-defined and relatively ambiguous concepts. Human presence and terrorist organization (specific) are concepts that can be almost completely defined by objective low-level semantic tokens like face, silhouette, speech, flag, overlay text and audio keywords.

These concepts are detected with a remarkable performance in our experiments. On the other hand, violence, nudity and nature are concepts that require a much more richer low-level lexicon in order to cover the possible variations. Therefore, despite the fact that typical instances of violence, nudity and nature concepts are successfully retrieved at the top of the result list (mAP_{20}), marginal examples that could not be placed at the desired positions cause a degradation in the overall performance (mAP) of these concepts.

Low-level visual concepts behavior vary greatly between the concepts. The most significant reason for this variation is spatio-temporal extent of typical concept instances. Concepts like human skin, crowd and skyline typically appear in larger spatial proportions of video frames and within longer video shots. On the contrary, concepts like blood, fire and explosion tend to be occupy a limited part of frames, which are part of short video shots. This difference arises from the nature of these concepts and can be asserted as the main reason for the performance variations among visual low-level concepts. On the other hand, the performance of the proposed system in detection of low-level auditory concepts also have important implications. The top performing concepts, namely, speech, music and silence have an abundance of natural training data and therefore, despite their generic definition, detected with a remarkable success in the majority of the test dataset. Scream and crying training data, however, contains mostly artificial data that is a produced by actors as a result of the role-playing process. These fabricated concepts possess a significant amount of variation and discrepancy in terms of the auditory characteristics, and when combined with relatively lower frequency of appearance in the TV broadcast, results in a lower detection performance. The remaining auditory concepts, namely, explosion, gunshot and animal sound possess a moderate difficulty in terms of training data acquisition, which is compensated by the discriminative characteristics of the concepts.

In view of the experimental results, it can be stated that high-level semantic concepts that have generic definitions profit remarkably from the enrichment of the low-level audio-visual lexicon supporting them. This can be seen from the results presented in Table 8. However, low-level visual lexicon, should mainly consist of well defined concepts that are dominant in terms of spatial and temporal extent. For those visual concepts that appear in shorter time extent and smaller spatial extent, specially developed algorithms should necessarily replace the generalized frameworks. Using specialized algorithms that are expensive both in terms of computation time and memory, may be limited to special instances, where generalized clues point out their existence. On the contrary to the visual case, auditory concept detection, does not experience a difficulty in cases where the data is limited to a short time extent. Finally, in our opinion, detection performance of auditory concepts' are highly dependent on the subjectivity level of the concept. Auditory concepts like crying and scream, vary greatly from one source to the other and therefore, depend mostly on one's interpretation of the underlying emotions.

References

1. Akbani R, Kwek S, Japkowicz N (2004) Applying support vector machines to imbalanced datasets. Proc ECML pp 39–50
2. Barrington L, Chan A, Turnbull D, Lanckriet G (2007) Audio information retrieval using semantic similarity. In: Proc. ICASSP, Ieee, vol 2, pp II–725
3. Bay H, Ess a, Tuytelaars T, Van Gool L (2008) Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding 110(3):346–359
4. Biatov K, Hesseler W, Koehler J (2008) Audio data retrieval and recognition using model selection criterion. In: Proc. ICSPCS, IEEE, pp 1–5
5. Chang S, He J, Jiang Y, Khoury E, Ngo C, Yanagawa A, Zavesky E (2008) Columbia university at trecvid2008: high-level feature extraction and interactive video search. In: Proc. TRECVID
6. Changkaew P, Kongkachandra R (2010) Automatic movie rating using visual and linguistic information. In: Proc. ICIIC, IEEE, pp 12–16
7. Cheng J, Drue S, Hartmann G, Thiem J (2000) Efficient detection and extraction of color objects from complex scenes. In: Proc. ICPR, IEEE, vol 1, pp 668–671
8. Chu S, Narayanan S, Kuo C (2009) Environmental sound recognition with time–frequency audio features. Audio, Speech, and Language Processing, IEEE Transactions on 17(6):1142–1158
9. Clarin C, Dionisio J, Echavez M, Naval P (2005) Dove: Detection of movie violence using motion intensity analysis on skin and blood. PCSC 6:150–156
10. Clavel C, Ehrette T, Richard G (2005) Events detection for an audio-based surveillance system. In: Proc. ICME, IEEE, pp 1306–1309
11. Crandall D, Luo J (2004) Robust color object detection using spatial-color joint probability functions. In: Proc. CVPR, IEEE, vol 1, pp I–379
12. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proc. CVPR, Ieee, vol 1, pp 886–893
13. Deselaers T, Pimenidis L, Ney H (2008) Bag-of-visual-words models for adult image classification and filtering. In: Proc. ICPR, IEEE, pp 1–4
14. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-invariant learning. In: Proc. CVPR, IEEE, vol 2, pp II–264
15. Ghimire D, Lee J (2010) Color image enhancement in hsv space using nonlinear transfer function and neighborhood dependent approach with preserving details. In: Proc. PSIVT, IEEE, pp 422–426
16. Gotlieb CC, Kreyszig HE (1990) Texture descriptors based on co-occurrence matrices. Computer Vision, Graphics, and Image Processing 51(1):70 – 86
17. Huang J, Kumar S, Mitra M, Zhu WJ, Zabih R (1997) Image indexing using color correlograms. In: Proc. CVPR, pp 762 –768
18. Huang R, Hansen J (2006) Advances in unsupervised audio classification and segmentation for the broadcast news and ngsw corpora. Audio, Speech, and Language Processing, IEEE Transactions on 14(3):907–919
19. Huttenlocher D, Klanderman G, Rucklidge W (1993) Comparing images using the hausdorff distance. PAMI 15(9):850–863
20. Jansohn C, Ulges A, Breuel T (2009) Detecting pornographic video content by combining image features with motion information. In: Proc. MM, ACM, pp

601–604
21. Jia W, Zhang H, He X, Wu Q (2006) Image matching using colour edge cooccurrence histograms. In: Proc. SMC, IEEE, vol 3, pp 2413–2419
22. Jiang Y, Ngo C, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proc. CIVR, ACM, pp 494–501
23. Jones M, Rehg J (1999) Statistical color models with application to skin detection. In: Proc. CVPR, IEEE, vol 1
24. Jones M, Rehg J (2002) Statistical color models with application to skin detection. IJCV 46(1):81–96
25. Jones M, Viola P, Jones M, Snow D (2003) Detecting pedestrians using patterns of motion and appearance. In: Proc. ICCV, Citeseer
26. Lin C, Chen S, Truong T, Chang Y (2005) Audio classification and categorization based on wavelets and support vector machine. Speech and Audio Processing, IEEE Transactions on 13(5):644–651
27. Liu Y, Xie H (2009) Constructing surf visual-words for pornographic images detection. In: Proc. ICCIT, IEEE, pp 404–407
28. Lopes A, de Avila S, Peixoto A, Oliveira R, Araújo A (2009) A bag-of-features approach based on hue-sift descriptor for nude detection. In: Proc. ESPC, Citeseer
29. Lopes A, de Avila S, Peixoto A, Oliveira R, de M Coelho M, de A Araujo A (2009) Nude detection in video using bag-of-visual-features. In: Proc. SIBGRAPI, IEEE, pp 224–231
30. Lowe D (2004) Distinctive image features from scale-invariant keypoints. IJCV 60(2):91–110
31. Manjunath B, Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface, vol 1. John Wiley & Sons Inc
32. Mesaros A, Heittola T, Eronen A, Virtanen T (2010) Acoustic event detection in real life recordings. In: Proc. ESPC, pp 1267–1271
33. Mikolajczyk K, Schmid C, Zisserman A (2004) Human detection based on a probabilistic assembly of robust part detectors. Proc ECCV pp 69–82
34. Nam J, Alghoniemy M, Tewfik A (1998) Audio-visual content-based violent scene characterization. In: Proc. ICIP, IEEE, vol 1, pp 353–357
35. Over P, Awad G, Fiscus J, Antonishek B, Qu G (2011) TRECVID 2011 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In: Proc. TRECVID
36. Peng Y, Yang Z, Yi J, Cao L, Li H, Yao J (2008) Peking university at trecvid 2008: High level feature extraction. In: Proc. TRECVID, vol 3
37. Petridis S, Giannakopoulos T, Perantonis S (2010) A multi-class method for detecting audio events in news broadcasts. Artificial Intelligence: Theories, Models and Applications pp 399–404
38. Phan R, Androutsos D (2010) Content-based retrieval of logo and trademarks in unconstrained color image databases using color edge gradient co-occurrence histograms. Computer Vision and Image Understanding 114(1):66–84
39. Phan R, Chia J, Androutsos D (2008) Colour logo and trademark detection in unconstrained images using colour edge gradient co-occurrence histograms. In: Proc. CCECE 2008, IEEE, pp 000,531–000,534

40. Phillips P, Moon H, Rizvi S, Rauss P (2000) The feret evaluation methodology for face-recognition algorithms. PAMI 22(10):1090–1104
41. Portelo J, Bugalho M, Trancoso I, Neto J, Abad A, Serralheiro A (2009) Non-speech audio event detection. In: Proc. ICASSP, IEEE, pp 1973–1976
42. van de Sande KEa, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. PAMI 32(9):1582–96
43. Saracoglu A, Tekin M, Esen E, Soysal M, Logoglu K, Ates T, Sevinç A, Sevimli H, Acar B, Zubari U, et al (2010) Generalized visual concept detection. In: Proc. SIU, IEEE, pp 621–624
44. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. Neural Computation 12(5):1207–1245
45. Smeaton AF, Over P, Kraaij W (2009) High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In: Divakaran A (ed) Multimedia Content Analysis, Theory and Applications, Springer Verlag, Berlin, pp 151–174
46. Snoek C, Worring M, Koelma D, aWM Smeulders (2007) A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. IEEE Transactions on Multimedia 9(2):280–292
47. Snoek CGM, van de Sande KEA, de Rooij O, Huurnink B, Gavves E, Odijk D, de Rijke M, Gevers T, Worring M, Koelma DC, Smeulders AWM (2010) The mediamill trecvid 2010 semantic video search engine. In: Proc. TRECVID
48. Snoek G, et al (2006) The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. PAMI 28(10):1678–1689
49. Stricker MA, Orengo M (1995) Similarity of color images. In: Proc. SPIE, pp 381–392
50. Sundaram S, Narayanan S (2008) Audio retrieval by latent perceptual indexing. In: ICASSP, IEEE, pp 49–52
51. Tao L, Asari V (2004) An integrated neighborhood dependent approach for nonlinear enhancement of color images. In: Proc. ITCC, IEEE, vol 2, pp 138–139
52. Viola M, Jones M, Viola P (2003) Fast multi-view face detection. In: Proc. CVPR, Citeseer
53. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proc. CVPR, IEEE, vol 1, pp I–511
54. Wu P, Manjunanth B, Newsam S, Shin H (1999) In: Proc. CBAIVL, pp 3 –7
55. Yilmaz E, Aslam JA (2006) Estimating average precision with incomplete and imperfect judgments. In: Proc. CIKM, ACM, pp 102–111
56. Zubari Ü, Ozan E, Acar B, Ciloglu T, Esen E, Ateş T, Önür D (2010) Speech detection on broadcast audio. In: EUSIPCO
57. Zuo H, Wu O, Hu W, Xu B (2008) Recognition of blue movies by fusion of audio and video. In: Proc. ICME, IEEE, pp 37–40