Foundations and Trends® in
Information Retrieval
Vol. 2, No. 4 (2008) 215–322

DOI: 10.1561/1500000014

now
the essence of knowledge

Concept-Based Video Retrieval

By Cees G. M. Snoek and Marcel Worring

Contents

Foundations and Trends® in
Information Retrieval
Vol. 2, No. 4 (2008) 215–322

DOI: 10.1561/1500000014

now
the essence of knowledge

Concept-Based Video Retrieval

Cees G. M. Snoek[1] and Marcel Worring[2]

[1] *University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands, cgmsnoek@uva.nl*
[2] *University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands, worring@uva.nl*

Abstract

In this paper, we review 300 references on video retrieval, indicating when text-only solutions are unsatisfactory and showing the promising alternatives which are in majority concept-based. Therefore, central to our discussion is the notion of a semantic concept: an objective linguistic description of an observable entity. Specifically, we present our view on how its automated detection, selection under uncertainty, and interactive usage might solve the major scientific problem for video retrieval: the semantic gap. To bridge the gap, we lay down the anatomy of a concept-based video search engine. We present a component-wise decomposition of such an interdisciplinary multimedia system, covering influences from information retrieval, computer vision, machine learning, and human–computer interaction. For each of the components we review state-of-the-art solutions in the literature, each having different characteristics and merits. Because of these differences, we cannot understand the progress in video retrieval without serious evaluation

efforts such as carried out in the NIST TRECVID benchmark. We discuss its data, tasks, results, and the many derived community initiatives in creating annotations and baselines for repeatable experiments. We conclude with our perspective on future challenges and opportunities.

1

Introduction

1.1 How to Retrieve Video Content?

This question is highly relevant in a world that is adapting swiftly to visual communication. Online services like YouTube and Tudou show that video is no longer the domain of broadcast television only. Video has become the medium of choice for many people communicating via Internet and their mobile phones. Digital video is leading to an abundance of narrowcast repositories, with content as diverse as Al Jazeera news, concerts of the Royal Philharmonic Orchestra, and the baby panda at your local zoo, to name just three examples. A nation's broadcast archive can be expected to contain petabytes of video data, requiring careful treatment for future preservation and disclosure. Personal video archives are likely to be much smaller, but due to the amount of effort involved, the willingness to archive for the purpose of future retrieval will be lower. At all stages and for all target groups, effective and efficient video retrieval facilities will be necessary, not only for the public broadcasters, but also for any private broadcaster and narrowcaster-to-be.

User needs determine both the effectiveness and efficiency of video search engines. To understand what are the user needs for video retrieval, we draw inspiration from the video production process. According to Jain and Hampapur [106], the purpose for which a video is created is either entertainment, information, communication, or data analysis. For all these purposes, the user needs and demands vary substantially. A consumer who wants to be entertained, for example, will be satisfied if a complete movie is accessible from an archive through a mobile phone. In contrast, a cultural anthropologist studying fashion trends of the eighties, a lawyer evaluating copyright infringement, or an athlete assessing her performance during training sessions might be more interested in retrieving specific video segments, without going through an entire video collection. For accessing complete video documents, reasonable effective commercial applications exist, YouTube and Netflix being good examples. Video search applications for consumers and professionals targeting at retrieval of specific segments, however, are still in a nascent stage [112]. Users requiring access to video segments are hardly served by present-day video retrieval applications.

In this paper, we review video search solutions that target at retrieval of specific segments. Since humans perceive video as a complex interplay of cognitive concepts, the all-important step forward in such video retrieval approaches will be to provide access at the semantic level. This is achievable by labeling all combinations of people, objects, settings, and events appearing in the audiovisual content. Labeling things has been the topic of scientific endeavor since Aristotle revealed his "Categories." Following in this tradition are Linnaeus (biology), Werner (geology), Mendeleev (chemistry), and the Human Genome Project (genetics) [263]. In our information age, Google labels the world's textual information. Labeling video content is a grand challenge of our time as humans use approximately half of their cognitive capacity to achieve such tasks [177]. Two types of semantic labeling solutions have emerged: (i) the first approach relies on human labor, where labels are assigned manually after audiovisual inspection; (ii) the second approach is machine-driven with automatic assignment of labels to video segments.

1.2 Human-Driven Labeling

Manual labeling of (broadcast) video has traditionally been the realm of professionals. In cultural heritage institutions, for example, library experts label archival videos for future disclosure using controlled vocabularies [56, 148]. Because expert labeling [50, 155] is tedious and costly, it typically results in a brief description of a complete video only. In contrast to expert labor, Web 2.0 [172] has launched social tagging, a recent trend to let amateur consumers label, mostly personal, visual content on web sites like YouTube, Flickr, and Facebook. Alternatively, the manual concept-based labeling process can be transformed into a computer game [253] or a tool facilitating volunteer-based labeling [198]. Since the labels were never meant to meet professional standards, amateur labels are known to be ambiguous, overly personalized, and limited [69, 73, 149]. Moreover, unlabeled video segments remain notoriously difficult to find. Manual labeling, whether by experts or amateurs, is geared toward one specific type of use and, therefore, inadequate to cater for alternative video retrieval needs, especially those user needs targeting at retrieval of video segments [204].

1.3 Machine-Driven Labeling

Machine-driven labeling aims to derive meaningful descriptors from video data. These descriptors are the basis for searching large video collections. Many academic prototypes, such as Medusa [22], Informedia classic [255], and Olive [51], and most commercial video search engines such as Baidu, Blinkx, and Truveo, provide access to video based on text, as this is still the easiest way for a user to describe an information need. The labels of these search engines are based on the filename, surrounding text, social tags, closed captions, or a speech transcript. Text-based video search using speech transcripts has proven itself especially effective for segment-level retrieval from (English) broadcast news, interviews, political speeches, and video blogs featuring talking heads. However, a video search method based on just speech transcripts results in disappointing retrieval performance, when the audiovisual content is neither mentioned, nor properly reflected in the associated text. In addition, when the videos originate from non-English speaking

countries, such as China, and the Netherlands, querying the content becomes much harder as robust automatic speech recognition results and their accurate machine translations are difficult to achieve.

It might seem that video retrieval is the trivial extension of text retrieval, but it is in fact often more complex. Most of the data is of sensory origin (image, sound, video) and hence techniques from digital signal processing and computer vision are required to extract relevant descriptions. In addition to the important and valuable text data derived from audio analysis, much information is captured in the visual stream. Hence, a vast body of research in machine-driven video labeling has investigated the role of visual content, with or without text. Analyzing the content of visual data using computers has a long history [195], dating back to the 1960s. Some initial successes prompted researchers in the 1970s to predict that the problem of understanding visual material would soon be solved completely. However, the research in the 1980s showed that these predictions were far too optimistic. Even now, understanding visual data is a major challenge. In the 1990s a new field emerged, namely content-based image retrieval, where the aim is to develop methods for searching in large image archives.

Research in content-based retrieval has resulted in a wide variety of image and video search systems [17, 32, 34, 61, 67, 72, 114, 143, 180, 197, 207, 214, 258, 266]. A common denominator in these prototypes is their dependence on low-level visual labels such as color, texture, shape, and spatiotemporal features. Most of those early systems are based on query-by-example, where users query an archive based on images rather than the visual feature values. They do so by sketches, or by providing example images using a browser interface. Query-by-example can be fruitful when users search for the same object under slightly varying circumstances and when the target images are available indeed. If proper example images are unavailable, content-based image retrieval techniques are not effective at all. Moreover, users often do not understand similarity expressed in low-level visual features. They expect semantic similarity. This expected semantic similarity, is exactly the major problem video retrieval is facing.

The source of the problem lies in the *semantic gap*. We slightly adapt the definition by Smeulders et al. [213] and define it as: "The lack of

correspondence between the low-level features that machines extract from video and the high-level conceptual interpretations a human gives to the data in a given situation." The existence of the gap has various causes. One reason is that different users interpret the same video data in a different way. This is especially true when the user is making subjective interpretations of the video data related to feelings or emotions, for example, by describing a scene as *romantic* or *hilarious* [76]. In this paper, those subjective interpretations are not considered. However, also for objective interpretations, like whether a *windmill* is present in a video, developing automatic methods is still difficult. The main difficulties are due to the large variations in appearance of visual data corresponding to one semantic concept. Windmills, for example, come in different models, shapes, and colors. These causes are inherent to the problem. Hence, the aim of video retrieval must be to bridge the semantic gap.

1.4 Aims, Scope, and Organization

In this paper, we review state-of-the-art video retrieval methods that challenge the semantic gap. In addition, we also address the important issue of evaluation. In particular, we emphasize *concept-based video retrieval*. A recent breakthrough in the field, which facilitates searching in video at a segment-level by means of large sets of automatically detected (visual) concepts, like a *telephone*, a *flamingo*, a *kitchen*, or one of the concepts in Figure 1.1. Evidence is accumulating that when large sets of concept detectors are available at retrieval time, such an approach to video search is effective [36, 219, 227]. In fact, by using a simulation study, Hauptmann et al. [85] show that even when the individual detectors have modest performance, several thousand detectors are likely to be sufficient for video search in the broadcast news domain to approach standard WWW search quality. Hence, when using concept detectors for video retrieval, we might be able to reduce the semantic gap for the user.

In contrast to other reviews on video retrieval, which emphasize either content-based analysis [221], machine learning [158], text and image retrieval [283], search strategies [118], interactive browsing

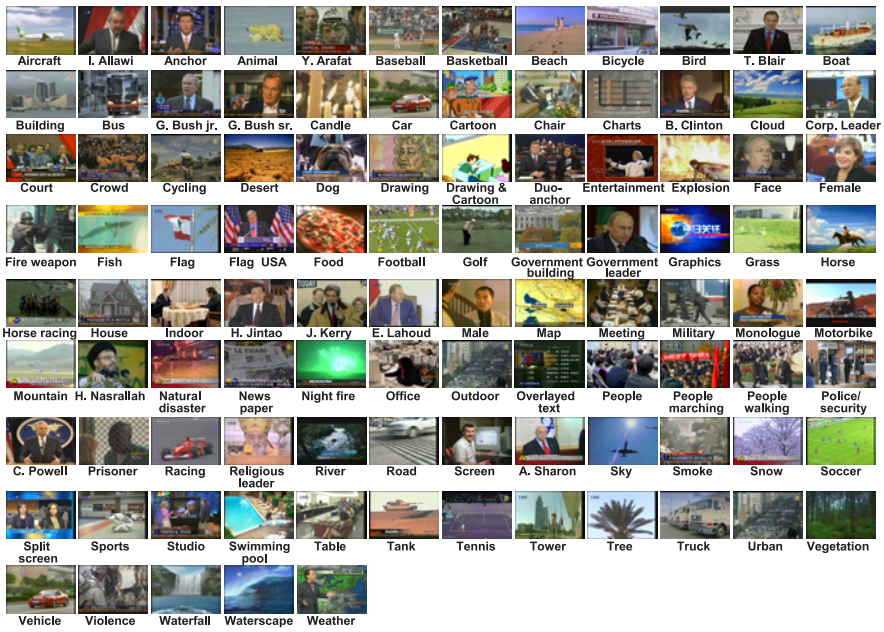


Fig. 1.1 Visual impression of 101 typical semantic concepts [230] for which automatic detection, retrieval, and evaluation results on video data are described in this paper.

models [86], or challenges for the future [129], the aim of this paper is to cover the semantic gap completely: from low-level features that can be extracted from the video content to high-level interpretation of video segments by an interacting user. Although these concepts have a visual nature, concept-based video retrieval is different from still-image concept detection as the concepts are often detected and retrieved using an interdisciplinary approach combining text, audio, and visual information derived from a temporal video sequence. Our review on concept-based video retrieval, therefore, covers influences from information retrieval, computer vision, machine learning, and human–computer interaction. Because of this interdisciplinary nature, it is impossible for us to provide a complete list of references. In particular, we have not attempted to provide an accurate historical attribution of ideas. Instead, we give preference to peer-reviewed journal papers, over earlier published conference papers, and workshop papers, where possible. Throughout the review we assume a basic familiarity with computer

science and information retrieval, but not necessarily the specific topic of (visual) video retrieval. For in depth, technical details on the fundamentals underlying many concept-based video retrieval methods, the interested reader is referred to recent books [16, 18, 19, 74, 128, 147, 199], review papers [48, 103, 140, 213, 241, 261], special issues [75, 270], and online proceedings [171], that provide further entry points into the literature on specific topics.

We organize the paper by laying down the anatomy of a concept-based video search engine. We present a component-wise decomposition of such an interdisciplinary multimedia system in our aim to bridge the semantic gap. The components exploit a common architecture, with a standardized input–output model, to allow for semantic integration.

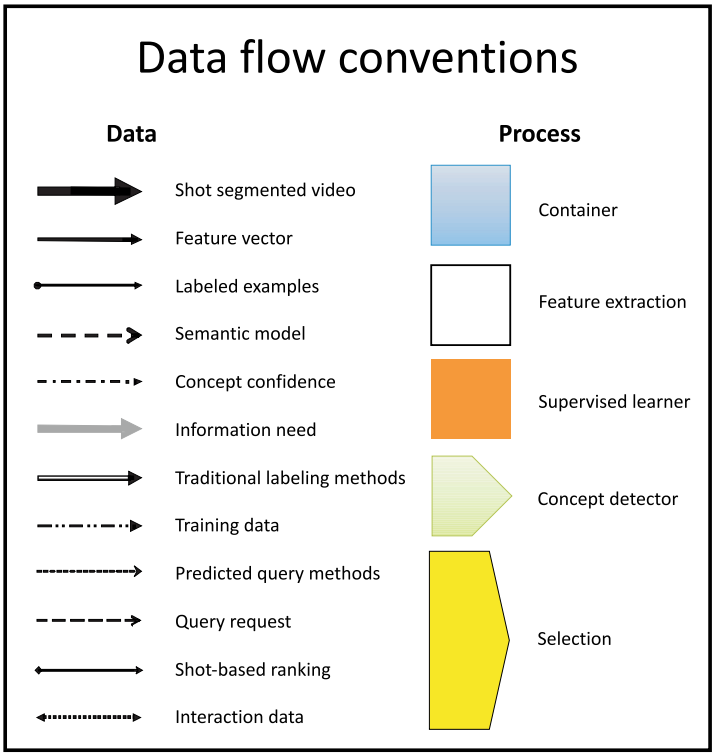


Fig. 1.2 Data flow conventions as used in this paper. Different arrows indicate difference in data flows.

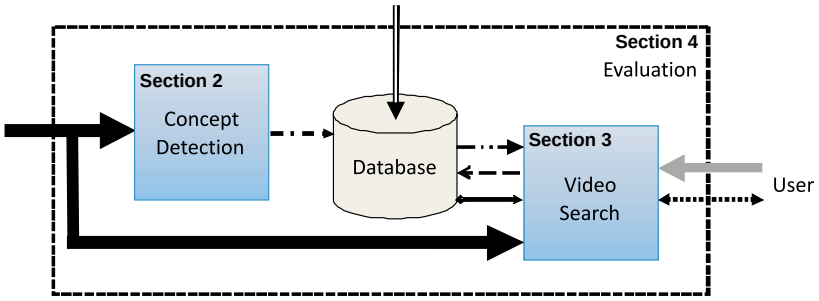


Fig. 1.3 We organize our review by laying down the anatomy of a concept-based video search engine, building upon the conventions introduced in Figure 1.2. First, we detail generic concept detection in Section 2. Then, we highlight how concept detectors can be leveraged for video search in combination with traditional labeling methods and an interacting user in Section 3. We present an in depth discussion on evaluating concept-based video retrieval systems and components in Section 4.

The graphical conventions to describe the system architecture are indicated in Figure 1.2. We will use the graphical conventions throughout this paper. Based on these conventions, we follow the video data as they flow through the computational process, as sketched in Figure 1.3. We start in Section 2, where we present a general scheme for generic concept detection. We cover the common concept detection solutions from the literature and discuss how they interconnect for large-scale detection of semantic concepts. The availability of a large set of concept detectors opens up novel opportunities for video retrieval. In Section 3, we detail how uncertain concept detectors can be leveraged for video retrieval at query time and how concept detectors can be combined with more traditional labeling methods. Moreover, we discuss novel visualizations for video retrieval and we highlight how to improve concept-based video retrieval results further by relying on interacting users. In Section 4, we turn our attention to evaluation of concept-based video search engines and their most important components. We introduce the de facto benchmark standard, its most important tasks, and evaluation protocols. In addition, we highlight the many community efforts in providing manual annotations and concept-based video retrieval baselines against which scientific progress in the field is measured. We conclude the review with our perspective on the challenges and opportunities for concept-based video search engines of the future.

2

Detecting Semantic Concepts in Video

2.1 Introduction

In a quest to narrow the semantic gap, early research emphasized concepts with a small intra-class and large inter-class variability in appearance. This research yielded a variety of dedicated methods exploiting simple handcrafted decision rules mapping restricted sets of low-level visual features, such as color, texture, or shapes, to a single high-level concept. Typical concepts and their detectors are *news anchor person* by Zhang et al. [298], *sunsets* by Smith and Chang [214], *indoor* and *outdoor* by Szummer and Picard [233], and the work on distinguishing *cityscape*, *landscape*, *mountains*, and *forests* by Vailaya et al. [243, 244]. These dedicated detectors have become icons for detection of concepts in video. However, such a dedicated approach to concept detection is limited when considering the plethora of concepts waiting to be detected. It is simply impossible to bridge the semantic gap by designing a dedicated tailor-made solution for each concept one can think of. As an adequate alternative for dedicated methods, generic approaches for large-scale concept detection have emerged [1, 6, 157, 224]. In contrast to specific methods, these approaches exploit the observation that mapping low-level (multimedia) features to a large

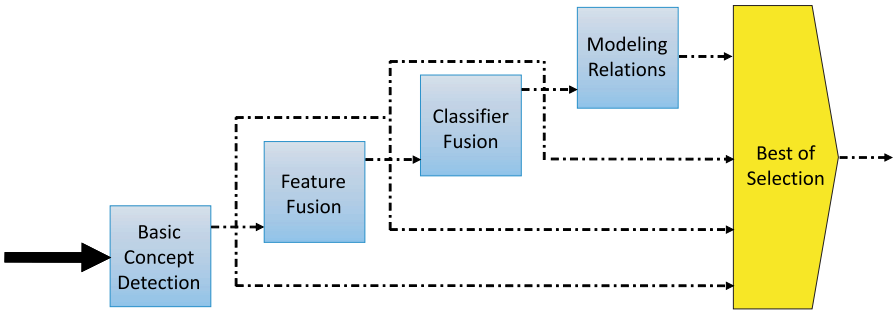


Fig. 2.1 General scheme for generic concept detection in video. Detail of Figure 1.3, using the conventions of Figure 1.2. The scheme serves as the blueprint for the organization of this section.

number of high-level semantic concepts requires too many decision rules. Hence, ideally, these rules need to be distilled automatically using machine learning. The machine learning approach to concept detection has led to powerful schemes, which allow access to video data at the semantic level.

In this section, we will detail the generic concept detection paradigm. We start with describing the basics of concept detection in Section 2.2, briefly introducing (visual) features, and supervised learning. After the basics, we highlight several possible extensions to improve concept detection. These extensions cover fusion of multiple features in Section 2.3, fusion of multiple classifiers in Section 2.4, and modeling of relationships between multiple concept detectors in Section 2.5. Naturally, the best possible analysis approach for generic detection is concept-dependent, therefore, we discuss automatic selection of most appropriate analysis stages in Section 2.6. We end the section with a discussion in Section 2.7. The organization of this section is detailed in the graphical overview scheme in Figure 2.1, which simultaneously serves as the general scheme for a generic large-scale concept detection system.

2.2 Basic Concept Detection

To cater for retrieval of automatically detected semantic concepts at a fine granularity, video sequences have to be divided into workable segments first. The most natural candidates for this segmentation are

video shots, which consist of one or more related frames that represent a continuous (camera) action in time and space [49]. The automatic segmentation of video into shots has a long history and is by now a well understood problem [23, 297] for which highly robust automatic methods exist [292]. All methods rely on comparison of successive frames with some fixed or dynamic threshold on either pixel, region, or frame level. For ease of analysis and computation, a segmented video shot is often represented by a single frame, the so-called key frame. Typically, the central frame of a shot is taken as the key frame, but many more advanced methods exist [23]. Once the video is segmented, we are ready to detect concepts.

Concept detection in segmented video is commonly viewed as a machine learning problem. Given an n-dimensional feature vector x_i, part of a shot (or key frame) i, the aim is to obtain a measure, which indicates whether semantic concept ω_j is present in shot i. We may choose from various feature extraction methods to obtain x_i, and from a variety of supervised machine learning approaches to learn the relation between ω_j and x_i. The supervised machine learning process is composed of two phases: training and testing. In the first phase, the optimal configuration of features is learned from the training data. In the second phase, the classifier assigns a probability $p(\omega_j \,|\, x_i)$ to each input feature vector for every semantic concept. The basic architecture of a generic concept detector is illustrated in Figure 2.2. For the

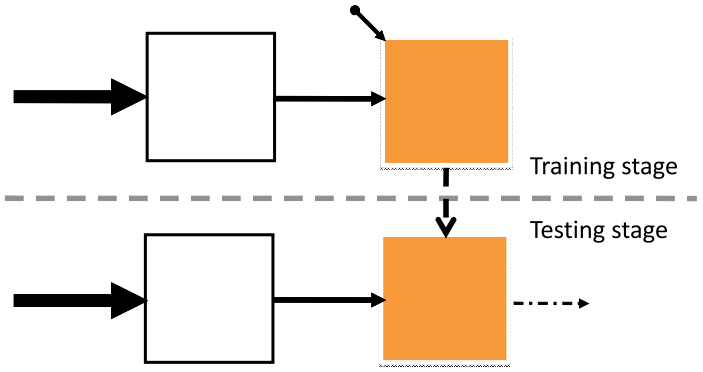


Fig. 2.2 General scheme for basic concept detection, using the conventions of Figure 1.2.

moment, the two main things to consider are the features to extract and the supervised machine learning scheme.

2.2.1 Feature Extraction

The aim of feature extraction is to derive a compact, yet descriptive, representation of the pattern of interest. When considering video, the common approach is to take a camera shot as the pattern of interest. Such a pattern can then be described using text features, audio features, visual features, and their combinations. Since the sole unique requirement for a video is the availability of a sequence of image frames, it is clear that the dominant information in a video is encapsulated in the visual stream. Here, we focus on summarizing the most common visual features, as used in many concept detection methods. The interested reader is referred to the following publications for text features [147, 199] and audio features [140, 261].

As highlighted in the Introduction, the major problem for automated concept detection is bridging the semantic gap between low-level feature representations that can be extracted from video and high-level human interpretation of the data. Hence, visual features need to model the wide diversity in appearance of semantic concepts. There are, however, also variations in appearance which are not due to the richness of the semantics. Varying the viewpoint, lighting and other circumstantial conditions in the recording of a scene will deliver different data, whereas the semantics has not changed. These variations induce the so-called *sensory gap* [213], which is the lack of correspondence between a concept in the world and the information in a digital recording of that concept. Hence, we need visual features minimally affected by the sensory gap, while still being able to distinguish concepts with different semantics. Some form of *invariance* is required [213], such that the feature is tolerant to the accidental visual transformations caused by the sensory gap. To put it simply, an invariant visual feature is a computable visual property that is insensitive to changes in the content, for example, caused by changing the illumination color, illumination intensity, rotation, scale, translation, or viewpoint. Features become more robust when invariance increases, but they loose discriminatory

power. Hence, effective visual features strike a balance between invariance and discriminatory power, but the ideal mix often depends on the application.

Visual features are many [48, 128, 213, 241], but we provide a brief summary only. To structure our discussion on features, we consider them along two dimensions. Namely, their type, i.e., color features, texture features, and shape features; and the spatial scale on which they are computed, i.e., global level, region level, keypoint levels, and their temporal extensions.

Color: A key frame is an array of pixels, where each pixel has a color. There are numerous color spaces to choose from in which to compute the numerical color values per pixel, including standard RGB, the intuitive HSV space, the perceptually uniform $L^*a^*b^*$ space, or the invariant sets of color spaces in [63, 66]. The first thing to realize is that the color of a concept is actually a color spectrum, indicating how much a certain wavelength is present, where white light contains an equal amount of all wavelengths. This is the basis for defining HSV. To be precise the three components of HSV are as follows: H(ue) is the dominant wavelength in the color spectrum. It is what you typically mean when you say the object is red, yellow, blue, etc. H is orientation-invariant with respect to the illumination intensity and camera direction, but not for illumination color [66]. S(aturation) is a measure for the amount of white in the spectrum. It defines the purity of a color, distinguishing, for example, signal-red from pink. Finally, the V(olume) is a measure for the brightness or intensity of the color. This causes the difference between a dark and a light color if they have the same H and S values. $L^*a^*b^*$ is another color space that is used often. The L^* is similar to the V in HSV. The a^* and b^* are similar to H and V, and are both invariant to intensity. An important characteristic of the $L^*a^*b^*$ space is that the distance between colors in the color space is approximately equal to the human perceived difference in the colors. As such, the space is intuitive for (interacting) users.

Texture: Where color is a property that can be measured at every pixel in the visual data, texture is a measure which considers local patterns. Texture is of importance in classifying different materials like

the line-like pattern in a brick wall, or the dot-like pattern of sand. Again, there is a variety of methods to choose from [202] with varying characteristics in terms of invariance, especially with respect to scale, orientation, and intensity invariance. Many can be traced back to statistics on co-occurrence of grey values in a local neighborhood, or filters emphasizing patterns in certain directions [189], like Gabor filters [21, 104] and wavelets [145]. In general, there will be different colors and/or textures in a key frame. This means that there will be many frame locations where there is a significant change in visual data, in particular a change in color or texture. These changes form (partial) lines called edges.

Shape: Alternatively, a key frame can be segmented by grouping pixels in the image based on a homogeneity criterion on color, texture, or both [32, 54, 89], or by connecting edge lines [65]. Once we have such a segmentation, we may describe the segmented region using shape features. A distinction can be made between shape features that describe a region or features describing the shape contour [123, 250]. A segmented key frame region is typically summarized by easy to compute shape properties such as the area, centroid, and orientation of the region. These the so-called moments are useful because we can define functions upon them which yield features invariant under image translation, scale, rotation, viewpoint, and illumination [94, 120, 152]. We observe more variation in methods for describing the contour of a region, which include Fourier descriptors, wavelets, principal components, and polygon approximations using curvature scale space (see [250] for detailed descriptions). A clear drawback of both contour-based and region-based shape features is their dependence on a good visual segmentation, an ill-defined and notoriously difficult problem in computer vision. In general, visual segmentation leads to a decomposition of the key frame, where each region in the decomposition has a uniform visual property. This process is known as weak segmentation [213] as it is purely based on the visual data itself, not on the interpretation of the data. The latter is known as strong segmentation, which is delineating the contour of a semantic concept in the key frame. As a consequence of weak segmentation, shape features are only applicable for those applications

where the background can be controlled easily, like the recognition of trademarks [105], or recognizing shapes of marine life species in line drawings [182].

Global: For representation of visual features, we often use a global measure, summarizing the information in the entire key frame. Most commonly, these measures are in the form of histograms [232], for example, simply counting how many pixels have a certain color value or edge orientation. It was shown in [64] that the complete range of image statistics in natural textures can be well-modeled with an integrated Weibull probability distribution. This modeling step allows us to reduce a complete histogram to just two Weibull distribution parameters. Although histograms are invariant to translation and rotation, and robust to changes in viewing angle, scale, and occlusion, they loose all information on spatial configurations of the pixels. If a peak occurs in the histogram at the color red, for example, the pixels can be scattered all around the key frame, or it can be one big red region. Color coherence vectors [178], color correlograms [95], and Markov stationary features [130] are alternative representations which also consider spatial neighborhood information. As the above global measures are summaries of the data, they are irreversible. Once we compute the feature vector, we cannot reconstruct the original key frame.

Region: Descriptors that suffer less from the irreversibility of global measures rely on segmentation or partitioning. A partitioning uses a fixed set of rectangles, or one fixed rectangle in the middle of the frame, and a further partitioning of the remaining space in a fixed number of equal parts (see e.g., [32, 161, 247, 257]). Rather than relying on predefined regions, Shotton et al. [205] suggest a data-driven segmentation based on simple functions of pixels in image patches. For the moment, however, strong (or precise) segmentation in an unconstrained setting seems impossible. Some methods, therefore, focus on the weak segmentation, describing the resulting regions individually, or move on to the detection of keypoints.

Keypoint: Keypoints were originally used as an efficient means to capture the essence of a scene by detecting the information-rich pixels in an image, such as those representing spots, edges, corners, and

junctions [241]. While effective for many computer vision applications, keypoints have become especially useful for concept detection by two recently proposed consecutive operations on such points. The first operation is the Scale-Invariant Feature Transform (SIFT) descriptor by Lowe [139], which measures the region around a keypoint and describes each region using an edge orientation histogram. Its' superior performance is commonly attributed to its invariant properties with respect to intensity, color, scale, and rotation [207, 245]. The second operation is a vector quantization procedure (see e.g., [207]), which clusters the edge orientation histograms into a so called codebook. In this compact codebook, each cluster represents a visual codeword. The frequency distribution of the codewords is used as feature vector for a key frame. Utilizing keypoints in combination with codebooks for concept detection is a heavily researched subject in computer vision, and improvements are reported regularly on many components, including: various keypoint sampling strategies [124], computational efficient keypoint descriptors [13], the addition of color-invariant descriptors [28, 245], and better codebook construction [154, 246].

Temporal: All of the above features are important for video data. As a video is a set of temporally ordered frames, its representation clearly shares many of the representations considered above for key frames. However, the addition of a time component also adds many new aspects. In particular, we can analyze the motion pattern to find the camera motion [237], track regions or points of interest through the sequence and describe their tracks [206], or derive general measures of the activity in the scene [185]. We can also consider the motion of individual objects segmented from the video data. For segmentation we have, in addition, to color- and texture-based grouping of pixels, motion-based grouping which groups pixels if they have the same optic flow [170], i.e., move in the same direction with the same speed. In principle, these additional information sources can enhance the interpretation of the video, but due to their expensive computational costs, the use of temporal feature extraction is not common place in concept-based video retrieval, yet.

We have made a first step in limiting the size of the semantic gap, by representing video data in terms of various multimedia features.

In particular, we have considered features of the visual class which are tolerant to the influence of the sensory gap. Selecting, implementing, and extracting visual features can be a laborious effort. A good starting point for a set of practical and easy to use visual features is defined in the MPEG-7 standard [146], which includes features for HSV color histograms, edge histograms, and motion activity, and so on. Once the features are extracted and each shot or key frame is represented in the form of a feature vector x_i, we are ready to learn concepts.

2.2.2 Supervised Learning

In this section, we will consider general techniques that may exploit multimedia features to find the conceptual label of a piece of video content. The techniques required to perform this task are commonly known as machine learning. A good overview of learning methods is presented in [18, 103]. Here, we focus specifically on methods that adhere to the supervised learning paradigm, i.e., learning a concept detector based on training examples.

The goal of supervised learning is to optimize for a certain learning task and a limited amount of training data, a model with the best possible generalization performance. This measures quantifies the performance of a supervised learner when classifying test patterns not used during training. Poor generalization ability is commonly attributed to the *curse of dimensionality*, where the number of features used is too large relative to the number of training examples [103], and *over-fitting*, which indicates that the classifier parameters are too intensively optimized on the training data [103]. Hence, a supervised learner needs to strike a balance between the number of (invariant) features to use, while simultaneously avoiding over-optimization of parameters. Moreover, for the purpose of 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, and it should account for unknown or erroneously detected data. In such heavy demands, the support vector machine framework [27, 249] has proven to be a solid choice. Indeed, it has become the default choice in most concept detection schemes.

Support vector machine: In the support vector machine framework, an optimal hyperplane is searched that separates an n-dimensional feature space into two different classes: one class representing the concept to be detected and one class representing all other concepts, or more formally $y_i = \pm 1$. A hyperplane is considered optimal when the distance to the closest training examples is maximized for both classes. This distance is called the margin. The margin is parameterized by the support vectors, $\lambda_i > 0$, which are obtained during training by optimizing:

$$\min_{\lambda} \left(\lambda^T \Lambda K \Lambda \lambda + C \sum_{z} \xi_i \right), \tag{2.1}$$

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 that allows to balance training error and model complexity, z is the number of shots in the training set, ξ_i are slack variables that are introduced when the data is not perfectly separable, and the matrix K stores the values of the kernel function $K(x_i, x')$ for all training pairs. 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 hyperplane 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'.

Ideally, one would have a posterior probability, $p(\omega_j \,|\, x')$, that given an input feature vector x' returns a probability for a particular concept ω_j. But, the model dependent output of a support vector machine, $g(x')$, is not a probability. A popular and stable method for distance conversion was proposed by Platt [184]. This solution exploits the empirical observation that class-conditional densities between the margins are exponential. Therefore, the author suggests a sigmoid model. The output of this model results in the following posterior probability:

$$p(\omega_j \,|\, x') = \frac{1}{1 + \exp(\alpha_j g(x') + \beta_j)}, \tag{2.2}$$

where the parameters α_j and β_j are maximum likelihood estimates based on the training set. To obtain the posterior probabilities in a more efficient way, Lin et al. [134] proposed a computational improvement over Platt's algorithm. The support vector machine thus obtained

allows for concept detection with probabilistic output $p(\omega_j \,|\, x', q)$, where q are parameters of the support vector machine yet to be optimized.

Parameter optimization: Although the only parameters of the support vector machine are C and the kernel function $K(\cdot)$, it is well known that the influence of these parameters on concept detection performance is significant [156, 224]. In most cases, a kernel based on a Gaussian radial basis function is a safe choice. However, it was recently shown by Zhang et al. [299] that for concept detection approaches that quantize keypoint descriptors into codebooks, as explained in Section 2.2.1, the earth movers distance [196] and χ^2 kernel are to be preferred. In general, we obtain good parameter settings for a support vector machine, by using an iterative search on a large number of values for both C and $K(\cdot)$. From all parameters q, we simply select the combination that yields the best performance, yielding q^*.

To prevent over-fitting, parameters are typically evaluated in a cross-validation setting, where the training set is randomly split into X folds and each of these X folds is used as a hold-out set once. In order to minimize the effect of the random sampling, the standard procedure is to repeat the cross-validation several times. Different sampling strategies have been proposed [257, 286], which take the number of available positive annotations into account, thereby limiting the influence of unbalanced data. To prevent bias in the random sampling, it was argued by Van Gemert et al. [248] that shots from the same video should not be separated during parameter optimization. Hence, they propose a cross-validation method that separates folds on a video-level rather than shot-level. In all cases, the result of the parameter search over q is the improved model $p(\omega_j \,|\, x', q^*)$, contracted to $p^*(\omega_j \,|\, x')$.

We have made a second step in limiting the size of the semantic gap, by detecting concepts in video using (invariant) features in combination with supervised machine learning. In particular, we have considered supervised machine learning using a support vector machine. Similar to feature extraction, implementing machine learning algorithms can be laborious. Here also, stable software is available, e.g., [33, 111]. Now that we have introduced the basics of concept detection, we are ready to extend it into schemes that are more powerful.

2.3 Feature Fusion

Naturally, the many features one can extract from video data can be fused to yield more robust concept detection. A key question to ask before fusing features is what features to include in the first place. For feature fusion to be effective, here interpreted as leading to increased concept detector accuracy, some form of independence of features is required [103]. We identify two general approaches in the literature to achieve independence. 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. The general scheme for feature fusion is illustrated in Figure 2.3 and detailed next.

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. Many approaches in the literature follow a visual feature fusion approach [1, 6, 71, 240, 259, 267]. In [240], for example, Tseng et al. extract a varied number of visual

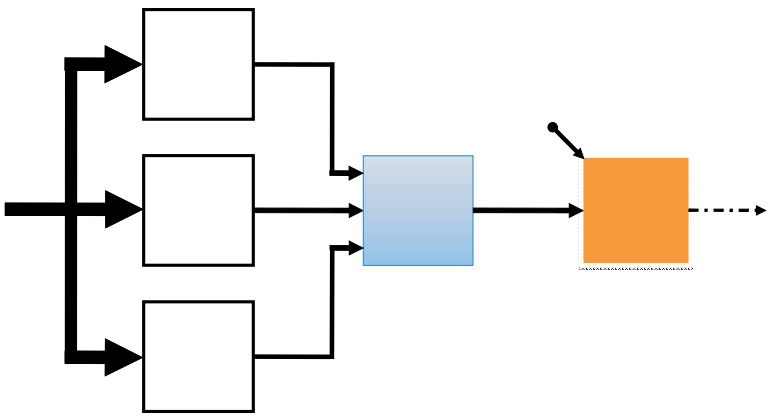


Fig. 2.3 General scheme for feature fusion. Output of included features is combined into a common feature representation before a concept is learned, special case of Figure 2.2.

features, including color histograms, edge orientation histograms, wavelet textures, and motion vector histograms. In [267], the authors use six features from the MPEG-7 standard, covering variation both in color and texture. Apart from achieving independence by varying the feature properties, (visual) feature fusion may also achieve independence by varying the spatial scale on which the features are computed. Here, we distinguish between visual features that operate on global, region, and keypoint levels. This feature fusion approach is often combined with variation in feature properties [1, 6, 71, 161, 219, 240, 259]. In [6], for example, the authors consider various frame and region features, covering color, texture, shape, and motion. In [219], the fusion approach covers features at global, region, and keypoint levels, and properties related to color, texture, and shape. While fusion approaches relying completely on visual features are known to be effective, they cannot cover all information relevant for detection of a concept.

Multimodal feature fusion methods compute features from different information streams. Although for some concepts in video, the audio and visual features may be correlated on a semantic level, the individual features are often considered data-independent of each other. Again, the visual modality is dominant in such approaches. Multimodal extensions focus on fusing visual features with textual features [9, 144, 224, 228, 230], with auditory features [1, 6, 39], or with both [79, 226]. A specific advantage for multimodal feature fusion is that it yields a truly multimedia feature representation for concept detection. In [224, 228] for example, the authors segment key frames from the visual modality into regions matching common color-texture patches. This results in a vector, which expresses a key frame in terms of pixel-percentages corresponding to visual patches. In the textual modality, the authors learn a list of concept-specific keywords. Based on this list, they construct a word frequency histogram from shot-based speech transcripts. This approach and many of its alternatives [1, 6, 9, 79, 144, 226, 230], yield a separate vector for each modality analyzed.

Once independence of features is assured, by unimodal or multimodal diversity, the extracted features need to be combined into a single representation. Of course, independence between features alone is not sufficient for an effective fusion. The individual features need

to perform reasonably in isolation too. A general advantage of feature fusion is the requirement of a single learning phase only, although learning time might increase substantially for large vectors. Moreover, combining features in an effective representation might be problematic as the features are often defined on different layout schemes and the features use different numerical ranges. And, most importantly, because the total number of features rapidly becomes too large relative to the number of training examples available, they might suffer from the curse of dimensionality. To counter these three problems we consider solutions related to feature synchronization, normalization, and transformation.

Synchronization: When combining different feature extraction results, difficulties arise with respect to their synchronization and alignment. Text features derived from a speech transcript are often defined on a time scale. In contrast, shape features are typically defined on the frame level. In addition, motion features are often defined on a shot segmentation. While feature synchronization seems an important issue for detection of concepts in video, it is often ignored in the literature. In the papers where synchronization is addressed, it is mostly used to align the video content to text obtained from: speech transcripts [51, 255]; video optical character recognition [11, 200, 255, 277]; closed captions [10, 22, 200, 220]; or websites [92, 277, 278, 279]. For synchronization, a pragmatic and widely adopted solution is to convert all features to a common segmentation scheme. As the common scheme researchers typically choose camera shots or its key frame equivalent [221]. Since all features are now defined within the same temporal representation, there is no need for synchronization anymore. To combine several synchronized feature extraction results, researchers often employ an *ad hoc* solution using simple concatenation of feature vectors.

Normalization: Although effective concept detection results can be achieved using vector concatenation only [240], it does not relieve us from the problem of how to handle diversity in numerical ranges resulting from different feature extraction algorithms. As different modalities have different characteristics, the problem is especially apparent for feature fusion methods exploiting diverse multimodal features. Moreover,

it is well known that the Euclidean distance, and hence a support vector machine with radial basis function kernel, is sensitive to the magnitude of the feature values it receives as input. Consequently, it will favor the feature vectors with large ranges. Therefore, normalizing features with respect to their dynamic range to a value between zero and one is often a good choice [1, 228, 286]. Popular normalization techniques include range normalization, which shifts feature values between zero and one based on the minimum and maximum values; Gaussian normalization, which shifts feature values based on the mean and variance; and rank normalization, which shifts feature values based on their mutual ordering [1, 228, 286].

Transformation: Once we obtain synchronized and normalized feature vectors, it is tempting to just concatenate them all together. Unfortunately, this simplistic approach will invoke the curse of dimensionality, eventually leading to deteriorated concept detection performance. In order to reduce the dimensionality, many schemes for combining features focus on feature transformations [9, 30, 39, 71, 144, 259]. In [39], for example, the authors propose to reduce the dimensionality of visual and audio features separately using Fisher's linear discriminant, and to concatenate the projected feature vectors afterwards. Rather than treating all input features equally, it has also been proposed to use feature-specific transformations, like specific kernels [9, 30, 71, 259], to combine various features into a single fused vector. Magalhães and Rüger [144] propose to transform modality-specific features to an optimal intermediate representation, inferred using the minimum description length principle. It balances the trade-off between feature space dimensionality and data representation. For sparse textual feature representations, the authors suggest to compress features. In contrast, for dense visual features, the authors suggest to expand the feature space. Once each modality is represented optimally, the authors rely on vector concatenation for multimodal feature fusion.

2.4 Classifier Fusion

Rather than emphasizing the fusion of features, one can also opt for fusion of classifiers. Kittler et al. [121] identify two reasons for fusing

classifiers: efficiency and accuracy. One can increase efficiency by using a simple detector for relatively easy concepts, and using more advanced schemes, covering more features and classifiers, for difficult concepts. To obtain a more accurate concept detector than can be expected from a single classifier alone, again some form of independence is required. We identify three common approaches in the concept-based video retrieval literature to achieve independence for classifier combination: using separate features, separate classifiers, and separate sets of labeled examples. As soon as concept detection probabilities are available, classifier combination methods rely on a combine function which provides the final concept detection result. The general scheme for combining classifiers is illustrated in Figure 2.4 and detailed next.

The classical method to obtain separate classifiers is to learn the same concept detector on several unimodal [109, 240, 257, 267, 286] or multimodal features [135, 228, 272]. In [135], for example, the authors learn a concept detector for *weather news* from the visual modality by representing a key frame as a histogram computed on block-segmented channels of the RGB color space, in combination with a support vector machine. In the textual modality, the authors construct a normalized word frequency histogram from the speech transcript. Again a support vector machine is used to learn the concept detector for *weather news*.

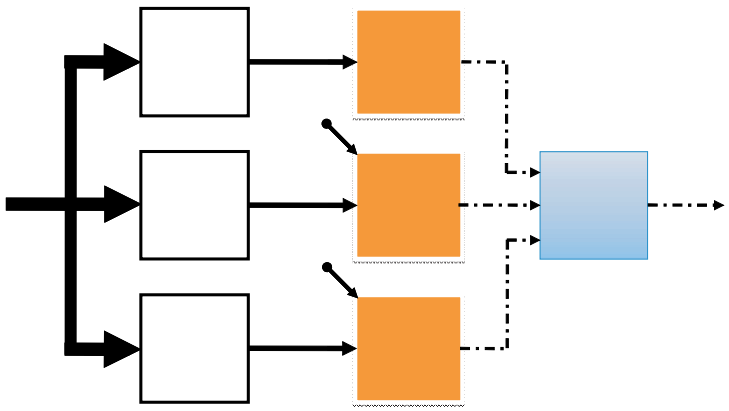


Fig. 2.4 General scheme for classifier fusion. Output of feature extraction is used to learn separate probabilities for a single concept. After combination a final probability is obtained for the concept, special case of Figure 2.2.

However, by treating multiple modalities separately from the start of the analysis, their interdependencies are left unexplored. To counter this problem, Wu et al. [272] start from a large (multimodal) feature pool. In contrast to feature fusion approaches, they do not learn a single classifier directly from this pool. Instead, they suggest to extract several diverse feature sets from this pool first, which are as independent as possible. For this purpose, they propose a three-stage approach. First, to remove noise and simultaneously reduce the feature dimensionality, they apply principal component analysis. Second, to obtain estimates of independent feature components they apply independent component analysis. Third, to reduce the number of components, and hence the number of features, the authors suggest a clustering approach that maximizes cluster size on a pre-defined range by cross-validation. Each resulting cluster is considered as an optimal independent feature set and forms the input for a support vector machine. Separation of features assures independence from the start, but it is not a prerequisite for classifier combinations.

Apart from variation resulting from diversity in the input features, we may also vary the classification algorithms. This can be helpful as different classifiers, like decision trees and naive bayes, may cover different regions in feature space. Hence, their combined coverage of feature space could result in improved concept detection performance. In addition to variety in the classifiers, we may also keep the classification algorithm fixed, but instead vary its parameters. Neural networks, for example, behave differently for each initialization due to the randomness in its training stage [103]. As variety in the classifiers increases the computation costs at training time considerably, exploiting variety by separating classifiers is not widespread in the concept detection literature. A few notable exceptions are the methods reported in [1, 6, 109, 161, 219, 257]. The methods by Adam et al., Amir et al., Snoek et al., [1, 6, 219] rely on support vector machines. They explore synergy with other classifiers such as Gaussian mixture models [1], neural networks [6], Fisher's linear discriminant [219], and logistic regression [219]. In [109, 161, 257], the authors rely on a support vector machine also, but they vary its kernel functions. These include linear kernel, histogram intersection, earth movers distance, radial basis

function, Laplacian, and χ^2 variants. Clearly, we may also employ multiple classifiers with multiple parameters, but this will further increase the computation load.

As labeled concept examples are a scarce resource, it is not surprising that the third, and last, method to achieve independence for classifier combination, using separate labeled examples at training time, is hardly explored in the concept detection literature. In [225], the authors suggest to use a bagging approach, which randomly re-samples the training data, and trains a support vector machine for each sample instance. In total, 200 detectors for the concept *news subject monologue* were evaluated, each based on a different sample of the original training set. In contrast to [225], who treat all examples equally, [257] suggests to sample labeled examples in proportion to the errors made by previous detectors. In addition, the authors suggest taking only a small sample of the negative examples into account as these typically outnumber the positive examples significantly. The proposed framework is inspired by the RankBoost algorithm and scales well to a large set of semantic concepts.

In designing an appropriate classifier combination scheme the first question to ask is what to combine? The choices include classifier rankings, binary classifier results, and classifier confidence scores. Solutions using classifier rankings and binary classifier results exist, for general approaches see [88, 121, 103], but most concept detection methods rely on classifier confidence scores. When the confidence score is not a probability already, the confidence scores are normalized to a range between zero and one [240, 267], as shown in Section 2.3, to assure an appropriate combination. Once we know what to combine, the next question to ask is how to combine? We identify two common approaches in the literature: *supervised* and *unsupervised*.

Supervised classifier combination: Supervised classifier combination schemes first unite individual concept probabilities into a common representation. Similar to feature fusion, researchers typically rely on vector concatenation [135, 272, 228]. Subsequently the vector forms the input for a supervised learner; again mostly a support vector machine is used nowadays. After parameter optimization, this yields a

combined, and ideally a more robust, classification result. An advantage of supervised combine functions is that they can exploit nonlinear relationships between individual classifiers. An obvious disadvantage of supervised combine functions is their dependence on labeled examples. Since supervised combine functions build upon individual classifiers, which are also dependent on labeled examples, the labeled examples for the supervised combine function need to be drawn from an independent set to prevent over-fitting. This additional demand for training data can be expensive if only few labeled examples are available.

Unsupervised classifier combination: To avoid the problems introduced by the additional learning stage of supervised combine functions, unsupervised combine functions rely on arithmetic operations. These range from basic operations like taking the maximum, the minimum, the average, or the geometric mean of individual classifier probabilities [109, 219, 225, 240, 257, 286]; to more advanced operations like the one suggested in [267], which builds upon Dempster–Shafer theory. From all unsupervised combine functions on offer, the ones based on robust statistics, such as the geometric mean [219, 257], are most resilient to outliers in the classifier combination.

2.5 Modeling Relations

Until now, we have only considered the case where a detector modeled a single semantic concept. In this section, we move on to the situation where multiple different detectors are taken into account simultaneously. When multiple detectors for different concepts are available, their semantic relationship can be exploited for analysis too. Naphade and Huang [157] were probably among the first to make this observation. They suggest to model several concepts simultaneously. To illustrate the main idea, once we detect with a certain probability that a shot contains a *car*, and we also detect presence of *road*, we might decide to increase our overall probability in both of them. An important additional purpose for modeling relations between concepts is its support for inference. For example, once an algorithm detects, with a certain probability, that a video segment contains both *sand* and *sky*, chances for concepts like *desert* increase, while the likelihood of detecting a *polar*

bear decreases substantially. For both improved accuracy and support for inference, the notion of co-occurrence is crucial. The co-occurrence between concepts can be learned from training data. We consider two general approaches in the literature to exploit co-occurrence of concept detectors: *learning spatial models* and *learning temporal models*. In contrast to methods that model relationships between concept detectors bottom-up, based on machine learning, one can also exploit a top-down approach by incorporating world-knowledge captured in ontologies. All approaches follow the general scheme for modeling relationships between concept detectors as illustrated in Figure 2.5 and are detailed next.

Learning spatial models: When we talk about learning spatial models we refer to methods that exploit simultaneous occurrence of multiple concepts in the same video frame. The common approach to learn spatial co-occurrence models between concepts starts with multiple individual concept detection results. Subsequently, these concept detection results are exploited simultaneously, in combination with labeled

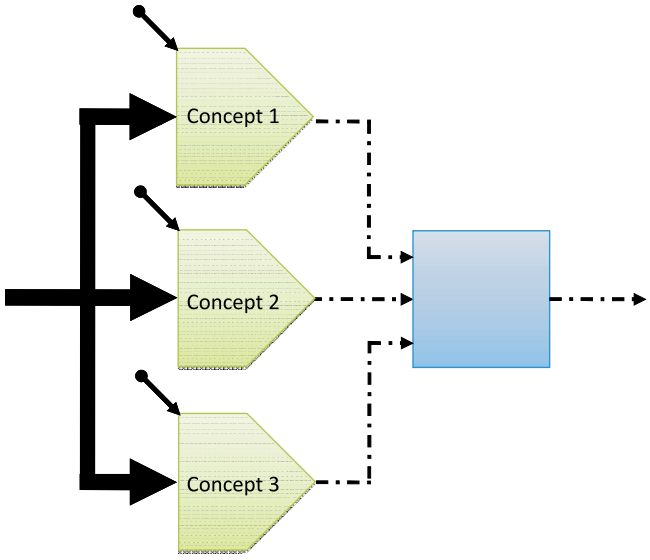


Fig. 2.5 General scheme for modeling relations between concept detectors. Outputs of multiple concept detectors are combined to improve existing detectors or to infer new concepts.

examples, to learn spatial co-occurrence models aiding in achieving increased accuracy or support for inference. We distinguish explicit and implicit frameworks.

Approaches that model spatial co-occurrence between concepts explicitly rely on directed [2, 157, 186, 280] or undirected [107, 186, 280] graphical models. Graphical models represent several random variables, in our case the individual concept detectors, and their linkage, in our case the probabilistic relationships between concept detectors, into a common framework. To allow for tractable computation of the joint distribution over all random variables, most graphical models represent individual concept detection results as binary variables [157, 160, 186, 280]. Even by using a binary representation, however, an apparent drawback of graphical models remains their computational complexity, especially when the number of available concept detectors increases. In order to alleviate the computation further, these methods typically exploit numeric approximation tools [160, 186, 280].

Although graphical models are currently the most popular approaches to model spatial co-occurrence, other methods have appeared which exploit relations between concept detectors more implicitly. As these implicit methods suffer less from computational complexity, the individual concept detection results are typically represented as probabilities. In [2, 6, 9, 31, 224], for example, the authors concatenate multiple concept detection probabilities into a single vector, which serves as the input for a support vector machine. In this approach, it is assumed that the support vector machine exploits the spatial co-occurrence between concepts at training time, while simultaneously handling the inherent uncertain individual concept detection result. Alternatively, Liu et al. [138] rely on association rules derived from the labeled training data to uncover hidden concept relations. They exploit association rules with sufficient support and confidence to adjust individual concept detection probabilities, leading to improved concept detection accuracy. Another data-driven approach is proposed by Weng et al. [264]. They employ a greedy algorithm, which partitions the training data into several hierarchical concept-specific binary trees.

Because errors in individual concept detectors propagate, a problem for both explicit and implicit modeling approaches is their dependence on the individual concept detection stage. Moreover, each stage also requires a separate set of labeled examples to prevent over-fitting. It was shown by Qi et al. [186] that by modeling individual detectors and their correlation in a single graphical model formulation, training data can be leveraged more efficiently without error propagation.

Learning temporal models: In principle, the above methods can be extended to include temporal modeling of concept detectors, but so far only few have considered the temporal dynamics of concept detection probabilities throughout a video. For learning temporal relations between concept detectors, it is needed to consider models that explicitly take the dynamic aspects of video into account. The hidden Markov model [188], which assigns a pattern to a class based on a sequential model of state and transition probabilities, is often considered a suitable tool [55, 186, 274]. In [55], for example, a hidden Markov model is used to model the temporal development of concept detection probabilities within a single shot to arrive at detectors for dynamic concepts like *riot* and *handshaking*. Based on the estimated probabilities for static concepts like *airplane*, *sky*, and *vegetation*, for example, the authors are able to model dynamic concepts like *airplane takeoff* and *airplane landing*.

Many alternatives to the hidden Markov models exist [275], and include the work reported in [138, 220, 264, 274]. In [220], the temporal relations between concept detectors are modeled using time intervals. In this approach, each detector is defined on an interval. The authors represent the relation between various concept detectors using an adaptation of Allen's interval algebra [5]. By choosing one detector as reference interval, the relation between all other detectors can be expressed in terms of a limited set of binary Allen relations. This binary pattern can subsequently be used for supervised machine learning of temporal concepts, like *goals* in soccer matches. Other alternatives, such as [138, 264, 274], rely on techniques originating from data mining. By studying frequent temporal patterns in annotated training data, it was found by Xie et al. [274] that good predictors for *highway* are concepts

outdoors, *road*, and *sky* in the current shot, in combination with presence of *car* in the previous shot. In both [138] and [264], the learned temporal relationships were also exploited for concept detection.

Whether temporal models are learned on top of semantic concepts using hidden Markov models, support vector machines, or data mining techniques, the result will always be dependent on the number and the performance of the individual detectors used in the initial stage. As highlighted above, the framework suggested by Qi et al. [186] relieves this limitation by modeling low-level features and interaction between concepts simultaneously. Interestingly, their approach is also capable to model temporal relationships, albeit requiring many computation cycles.

Including knowledge: A key issue when including knowledge in the form of ontologies is how concepts, their properties, and relations should be represented. The general problem is that concepts in common ontologies, like Cyc [126], WordNet [59], and ConceptNet [136], have a linguistic foundation, whereas video has an inherently multimedia nature. It is the reason why ontologies have initially been used mainly to support manual annotation of visual content [101, 201]. Recently, approaches have appeared in the literature making the connection from linguistic terms to the visual content of video explicit. We distinguish between two general approaches to incorporate linguistic knowledge in a visual representation, namely: approaches that make the connection from low-level (visual) features to linguistic terms, and approaches that connect by means of concept detectors.

An early approach that aimed to connect existing ontologies with low-level visual descriptors is the work by Hoogs et al. [91], who manually extended WordNet with a restricted set of visual features related to properties of regions, scenes, geometry, and motion. Others start from scratch and build their own multimedia ontology by manually linking linguistic terms with dedicated visual features. Due to the amount of effort involved in constructing such multimedia ontologies, they currently exist for narrow domains only, like medical video [58], soccer video [15], and Formula One video [47]. To prevent dependence on dedicated and specific visual feature sets many efforts have concentrated on

embedding standard visual features, like the MPEG-7 visual descriptors, into existing ontologies like WordNet [14, 90] and DOLCE [183]. Once visual features are connected to an ontology, it offers possibilities for reasoning and classification. In [90], for example, the authors argue that a visual ontology is suited for annotation. They achieve such an ontology by first detecting visual region properties in video frames. Subsequently, they match the properties with those in the ontology. Finally, the list of possible concept annotations is reduced to those that match the detected visual properties. While promising and convincing results have been reported with methods connecting low-level visual descriptors to linguistic terms, the problem of the semantic gap remains a critical obstacle.

The second general approach to include linguistic knowledge in a visual representation, which prevents the fact that low-level visual features need to be connected to linguistic terms, links terms directly to concept detectors. In [217], for example, the authors connect 100 of concept detectors to WordNet synsets manually based on similarity between concept definitions. The largest manual effort, so far, in linking concept detector definitions to an existing ontology, is the connection between Cyc and the concept definitions provided by the large scale concept ontology for multimedia [163]. Similar to the approaches linking ontologies to low-level features, approaches using concept detectors connected to ontologies can also be exploited for reasoning and classification. A problem here is that detectors are uncertain whereas ontologies use symbolic facts. This is the reason why the main application of ontologies in video concept detection, so far, is to improve overall detection accuracy [273, 296]. In [273], for example, two factors are defined which increase concept detector probability when an hierarchical relation between detectors is present, such as *trees* and *vegetation*; or decrease probability when the ontology indicates that concepts cannot co-exist simultaneously, e.g., *classroom* and *truck*. A further improvement of this idea is presented in [296], where the authors construct an ontology in a data-driven manner by taking into account hierarchical relations and pairwise correlations from labeled examples. These are subsequently exploited to refine learned concept detector probabilities. In effect, this approach resembles a graphical model.

2.6 Best of Selection

From the discussion above, it becomes clear that many methods exist for generic concept detection in video. Since all approaches have their own strengths and limitations, it is advisable to use them in concert when the purpose is large-scale concept detection. Then the problem becomes a matter of selection: what feature extraction methods should be relied upon in the feature fusing stage?; how many classifiers should be combined?; which combine function should be used?; how to model co-occurrence between concept detectors?; can ontologies help? To tackle these questions, several complex architectures have appeared in recent years that integrate various methods for effective large-scale concept detection. We identify two common schemes for such complex architectures: *schemes optimizing components* and *schemes optimizing the process flow*.

Schemes optimizing components assume that the performance of concept detection improves in each analysis stage. Moreover, these schemes also assume that each concept needs the same analysis flow. Therefore, such architectures typically combine several analysis stages in succession [162]. These architectures start with (multimodal) feature extraction, followed by consecutive fusion stages on features and classifiers. Many architectures consider only feature and classifier fusion [35, 109, 231, 286], others also cover analysis stages for modeling relationships between concepts. A pioneering and successful example of such an architecture is the one by IBM Research [1, 6, 30, 156]. A strong feature of the system is that in each stage of the pipeline it selects the best of multiple hypotheses based on validation set performance. Other architectures, like the Video Diver by Tsinghua University [31, 257, 293] or the pipeline by Microsoft Research Asia [151], have emphasized diversity of (visual) features in combination with machine learning parameter optimization. While proven effective, a clear drawback of architectures optimizing components is that their founding assumptions can be too rigid, leading to over-fitting and therefore suboptimal results.

To overcome the limitation of schemes that optimize components, many have proposed to take into account the fact that the best analysis

results typically vary per concept. In [245], for example, van de Sande et al. show that for light intensity changes the usefulness of invariance is concept-dependent. In addition, Naphade et al. [161] show that the most informative key frame region is concept-dependent too. An experimental comparison between feature fusion and classifier fusion indicates that the latter tends to give better performance in general; however, when the former performs better, the results are more significant [228]. Similar observations hold for modeling relations using co-occurrence and temporal models [264, 280], and quite likely for the inclusion of ontologies too. Indeed, by taking analysis results on a per-concept basis into account the performance of concept detection architectures can be optimized further. To realize such an architecture, schemes are needed which optimize the processing flow. In [224, 230], this is achieved by incorporating a process that selects the best among multiple analysis stages. In [230], the authors select the best component from a basic concept detection stage, a feature fusion stage, and a classifier fusion stage, whereas in [224], the authors also cover a stage that models relationships between detectors. Selection is typically based on maximum performance of individual processes on validation data, but in principle, it can be decided based on supervised learning also if sufficient labeled examples are available.

2.7 Discussion

In the past years, we have witnessed a transition from specific to generic concept detection methods, and from approaches emphasizing multimedia analysis to visual-only schemes. In all cases, the influence of supervised machine learning has been prevalent. Indeed, the machine learning paradigm has proven to be quite successful in terms of generic detection, currently yielding lexicons containing more than 100 concepts. Concept detection performance, however, is still far from perfect; the state-of-the-art typically obtains reasonable precision, but low recall. Hence, robustness of concept detection methods needs further improvement. More discussion on supervised machine learning aspects of evaluating concept detection appears in Section 4.6.

Analysis paths left to explore to improve the robustness are many, including the important question as to how to leverage the value of (non-speech) auditory information for generic concept detection. Because it seems that the fields of computer vision and machine learning have picked up the topic of concept-based video retrieval, we anticipate that the biggest leap in concept detection robustness is to be expected from methods emphasizing visual pattern recognition. In particular, we would like to highlight the role of temporal visual analysis. Due to the high computational cost involved, currently only few methods analyze video data beyond the key frame even though it is known to increase robustness [223, 231]. Apart from improving the robustness, temporal visual analysis could aid in tracking and selecting features over time, better definition and detection of temporal concepts, and potentially it allows modeling dynamic concept interactions more effectively. Moreover, temporal analysis could aid in localizing concepts in the video frame.

For all sketched solution paths, however, one major bottleneck needs to be taken into account: *efficiency*. Increasingly, large-scale concept detection approaches rely on high-performance computing [203, 224, 257] for both visual feature extraction and parameter optimization of supervised learning schemes. We deem it unlikely, however, that brute-force alone is sufficient to solve the generic concept detection problem. Hence, we need to spend more time on clever algorithms that maintain a robust level of performance while being efficient in their execution. The sketched solution paths should eventually lead to large sets of robust concept detectors.

3

Using Concept Detectors for Video Search

3.1 Introduction

As shown in the previous section, research in automatic detection of semantic concepts in video has now reached the point where over a hundred, and soon more than thousand, concept detectors can be learned in a generic fashion, albeit with mixed performance. The shot-based probability values associated with large sets of concept detectors offer novel opportunities for video retrieval. In [209], Smeaton even states: "This appears to be the road map for future work in this area." Despite the potential of concept-based video retrieval, however, automatic methods will not solve all search problems. Thus, eventually user involvement is essential.

In this section, we focus predominantly on video retrieval using concept detectors, but we will not ignore traditional methods for video search. To that end, we first discuss basic query methods for video retrieval in Section 3.2. Then we will elaborate in Section 3.3 on video retrieval approaches that automatically predict the appropriate query method given a user query. Naturally, several queries might be relevant. Therefore, we will address combination strategies

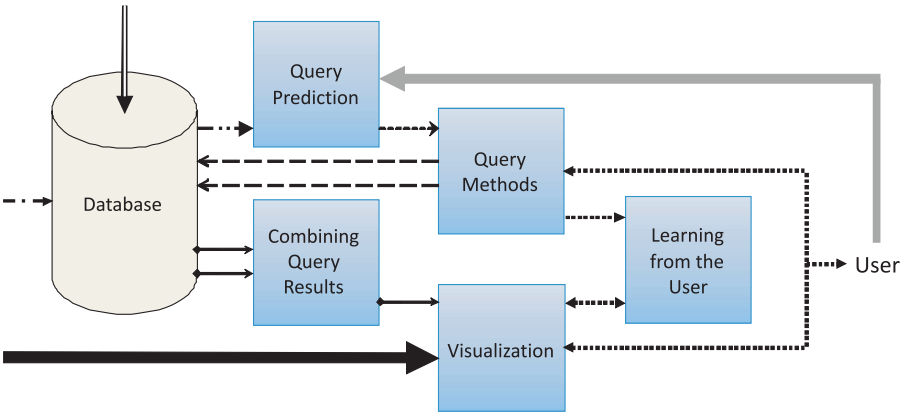


Fig. 3.1 General scheme for the retrieval frontend of a concept-based video search engine. Detail of Figure 1.3, building upon the conventions introduced in Figure 1.2. The scheme serves as the blueprint for the organization of this section.

in Section 3.4. We present several advanced visualization techniques that aid the user by displaying and browsing video retrieval results in Section 3.5. User involvement is the topic of Section 3.6, where we highlight how active learning and relevance feedback can aid in an interactive video search scenario. We end the section with a discussion in Section 3.7. A schematic overview of this section is presented in Figure 3.1, which simultaneously serves as the general scheme for a retrieval frontend of a concept-based video search engine.

3.2 Query Methods

Similar to traditional document retrieval, the aim in video retrieval is to retrieve the most relevant documents given a user query. In contrast to document retrieval, however, more variety exists in representing a video document; it can be an entire video, a scene, a shot, or a single frame, for example. Consequently, there is also more variation in posing a query to a video search engine. In order to avoid an excessive amount of formal notation to cater for all possibilities, we prefer a data-driven approach to structure our discussion on query methods. To that end, we make a distinction between systems using the common query-by-keyword approach extending upon text retrieval, methods starting

from image retrieval resulting in query-by-example, and query methods relying on concept detectors resulting in query-by-concept.

Query-by-keyword: Text retrieval is what many commercial video retrieval systems, especially on the web, are based on. In these systems, the document corresponds to a complete video and the document representation builds on index terms derived from the filename, social tags, the text surrounding the video on a web page, or hyperlink analysis. In order to retrieve the most relevant document given a textual user query, the systems typically rely on variables related to document statistics, like the well-known vector space model [199]. Video document terms may also be derived from closed captions or speech recognition results, when available. They provide the opportunity to retrieve video segments based on keyword matching [22, 51, 255]. When text from closed captions and recognized speech is combined with automated video segmentation techniques, like shot segmentation [292] or story segmentation [45], video retrieval on a finer document granularity becomes possible. Due to the difficulty of automatic story segmentation [45], however, most approaches in the literature rely on a shot-based video document representation. For shot-based video retrieval, Yan and Hauptmann [283] present an extensive analysis of the most commonly used text-based approaches, being variants of the aforementioned vector space model and the Okapi model [194]. They conclude that Okapi BM-25 models tend to outperform the vector space model. In addition, Huurnink and De Rijke [99] quantify the importance of expanding speech recognition results over temporal-adjacent shots as speech and the visual content it is describing are not necessarily aligned. Despite their proven effectiveness, query-by-keyword methods ignore the visual content of the video. Especially, in situations where text is absent, an alternative retrieval scenario is required.

Query-by-example: In contrast to text-based video retrieval, the content-based image retrieval research community has emphasized a visual-only approach using a single image as the document representation. Several methods extend the typical image retrieval paradigm, as highlighted in Section 1.3, to video. Chang et al. [34] extended the query-by-sketch paradigm to video. In their system, the user draws

a sketch where the object is identified with a rough shape; an indication of its color, and a sketch of its motion trajectory. The system then returns video clips resembling this pattern as closely as possible. The query-by-example paradigm is usually extended to video by taking some key frames out of the query clip and matching those to the video collection based on low-level features (see e.g., [57, 236, 255]). This is effective if one is searching for near identical videos e.g., versions of the query video that are converted from one encoding to another. This is clearly important in fighting copyright infringement, but for general search this is not what people are looking for. For those cases where the user has a clip at his disposal, he is more likely to be interested in related videos showing the same type of concepts i.e., sharing the same semantics. Smith et al. [215] were probably the first to use query by semantic visual example. They simply apply all available (visual) concept detectors to the query images, resulting in a vector with concept detection probabilities. This query vector is matched with the vector of concept detection probabilities of shots (represented by key frames) in the archive. For matching, they use various normalization schemes like the ones discussed in Section 2.3. The authors in [190], extended upon this work by developing an elaborate probabilistic framework for query by semantic visual example.

Query-by-concept: Indexing the video collection with a large set of video concepts, as explained in Section 2, allows us to employ a query-by-concept paradigm. In this paradigm, a user manually selects a predefined video concept, after which the system orders the shot-based documents by computed presence of the concept [227]. Detectors with sufficiently high performance allow us to filter out irrelevant results or bring forward correct query results [41]. When the lexicon of available concepts is small, manual selection by users is feasible indeed. When more than hundred concepts are available to choose from, users experience difficulty re-membering and selecting the appropriate detectors for an information need, especially when the quality of the detectors is unknown. Clearly, instead of having the user select the concept, we can employ automatic concept selection mechanisms. These and other query prediction methods are discussed next.

3.3 Query Prediction

Given an information need expressed in short natural language text snippets, and possibly containing some example images or video clips, the task of query prediction is to automatically translate this information need into an appropriate set of query methods. After query parsing, traditional query prediction methods consider a combination of query-by-keyword and query-by-example. For an extensive review on query prediction, including an in-depth survey of query prediction methods for text and image retrieval, we refer to [118]. Here, we focus exclusively on query prediction for video retrieval. In its most simple form, query prediction translates an information need into a single query, e.g., by-keyword, by-example, and by-concept. This query-independent approach, however, is known to be sub-optimal [118]. Learning an optimal prediction for each possible query is unlikely to be a valid alternative either. Therefore, many video search methods parse incoming queries, and classify them into a restricted set of pre-defined query classes. Another query prediction method, that comes within reach once thousands of concept detectors are available at video retrieval time, boils down to selecting the appropriate subset of detectors from the lexicon in response to a user query. Query prediction using query classes and automatic selection of concept detectors are detailed next.

3.3.1 Query Classes

The basic assumption in query class approaches is that similar information needs can be tackled with a similar search strategy, and that incoming queries can be accurately classified. Initial work in this direction constructed the query classes manually [46, 285, 300], resulting in categories tailored to the video domain at hand. In [46, 285], for example, the query classes are optimized for broadcast news video and contain categories like *named person* and *finance*. Alternatively, Zhang et al. [300] consider categories like *music* and *funny* for retrieval of online video clips. These approaches learn the appropriate query method(s) for individual query classes from a labeled training corpus [46, 285], or query history and logged user clicks [300], and often rely considerably on query-by-keyword.

The usage of query classes for predicting the appropriate query method is known to improve performance, but it also has some drawbacks. Clearly, the largest drawback of manual query class construction is their dependence on expert domain knowledge. Hence, scalability to hundreds of query classes and generalizability to multiple domains becomes problematic. To counter these issues, the authors of [119, 282] propose data-driven methods for query class construction. Kennedy et al. [119] rely on clustering to discover query classes. They exploit consistency in search performance, using a ground truth, and consistency in query formulation when training data is unavailable. Alternatively, the approach by Yan and Hauptmann [282] relies on the observation that each query can be described by a mixture of latent query classes. They propose a family of probabilistic methods that automatically discover query classes, capable of handling multiple classes per query, using a statistical latent class model. Indeed, the authors report improved performance over manual query class construction methods.

3.3.2 Automatic Selection of Concept Detectors

An alternative query prediction approach for video retrieval relies on concept detectors only. In such a scenario, query prediction is framed as selecting the appropriate set of concept detectors, given an information need. Based on the different query methods and the information need, we identify three general approaches in the literature for selecting an appropriate concept detector: *text-based selection*; *visual-based selection*; and *ontology-based selection*. The third approach exploits ontological knowledge in trying to capture the original user intent motivating the information need.

Text-based selection: The basic approach for text-based selection of concept detectors is to rely on exact text matching between search query and concept detector description [36, 164, 217, 256]; or speech transcripts that can be associated with concepts [164]. In [217, 256], for example, the authors represent each detector description by a term vector, where individual vector elements correspond to unique normalized words. To match the words in the detector descriptions to the words from the user query, they rely on the vector space model. Based on the

similarity between description vector and query vector, concept detectors are selected automatically. When we consider the query: *find shots of an office setting*, for example, an exact match on an *office* detector would in theory be sufficient to answer the retrieval request.

Although exact text matching might result in a highly accurate detector selection process, it is applicable to a very limited number of queries, especially when only a small set of concept detectors is available for video retrieval. The largest drawback of exact text matching, however, is its ignorance of potentially relevant detectors not covered directly by the query. Continuing the example, a *table* and *computer* detector might be relevant for the office query too. To increase the likelihood of correspondence between search query and the concept detectors available for retrieval, more extensive text-based approaches are presented in [164, 166]. They successfully employ data-driven query expansion approaches, using external (online) text resources, and semantic text annotation for text-based selection of concept detectors. Since data-driven query expansion offers a broader choice of terms to choose from, it is less restrictive, and therefore potentially more reliable than exact text matching. However, data-driven query expansion might also introduce noisy terms, which could result in inappropriate concept detector selection.

Visual-based selection: Concept detectors may also be selected by using visual query images or clips. Although it is hard to expect that general users will prefer to provide a number of image examples rather than explicitly specifying the semantic concepts they need using text, visual-based selection of concept detectors might prove a valuable additional strategy when other selection strategies fail.

Rather than using all detectors, as suggested in [215], one may also select the detectors with the highest posterior probabilities as most relevant. It was, however, argued in [217] that the most robust detectors, e.g., *person* and *outdoor*, are often the least informative for retrieval. Hence, for effective retrieval it seems better to avoid frequently occurring but non-discriminative detectors in favor of less frequent but more discriminative detectors. Metrics to take such concept frequency into account explicitly for concept detector selection were proposed in

[132, 256]. Li et al. [132] embed all detected concept probabilities for both the video shots and query images in a vector space model. The authors base selection of concept detectors on a modified *tf-idf* measure. In addition, they also consider a language model approach to concept detector selection by treating each query term as an independent event. Selection of concept detectors is then based on the probability of generating the query terms given the detector descriptions. Alternatively, Natsev et al. [164] formulate concept detector selection as a discriminative learning problem. Their algorithm maps the query images into the space spanned by concept detector probabilities. Subsequently, they use support vector machines to learn which concept detectors are most relevant.

Ontology-based selection: As an alternative to these data-driven query expansion methods, knowledge-driven approaches to selection of concept detectors have been investigated too. When concept detectors are interconnected within a general-purpose ontology, it becomes possible to exploit (hierarchical) concept relations, for example, to disambiguate between various concept interpretations. In this approach, concept detectors are selected by first parsing the user query. The typical approach is to employ part-of-speech tagging to extract nouns and noun chunks from the query. Subsequently, the extracted terms are matched with the terms in the ontology, e.g., WordNet synsets [59]. For the matching of query nouns and WordNet synsets several word sense disambiguation approaches exist [26]. In the video retrieval literature many ontology-based selection methods [166, 217] rely on Resnik's measure [193], which is a measure of semantic similarity between hierarchically structured concept terms based on information content derived from a text corpus. Alternatively, Natsev et al. [164] employ Lesk's measure [12, 127], which estimates semantic relatedness based on overlapping words from WordNet glosses. It was recently shown in [78] that the common corpus to estimate the information content between terms, i.e., the Brown Corpus, is outdated for present-day video retrieval requests, resulting in many concepts being unaccounted for. Hence, the authors suggest the use of web-based corpora as an alternative to better reflect modern search vocabularies.

A critical issue in ontology-based selection of concept detectors is the linking of concept detectors to the ontology, here taken as WordNet. To allow for scalability one prefers to obtain the link between concept detectors and WordNet synsets automatically. However, automatically mapping a concept detector to an ontology is still a difficult issue. Therefore, many approaches in the literature create the links between concept detectors and WordNet synsets manually [78, 164, 166, 217]. When automatic reasoning methods that automatically link concepts to ontology nodes with high accuracy become available, these might at least partly substitute the manual process. To address the scalability issue to some extent, Wei et al. [262] propose to construct a data-driven ontology using agglomerative hierarchical clustering. On top of the resulting clusters, they suggest to construct a vector space. This approach selects concept detectors based on cosine similarity between query terms and concept name. By doing so, the approach outperforms many traditional ontology-based selection methods.

Whether concept detectors or query classes are used, in both cases accurate query prediction is a hard problem. Moreover, an accurately predicted query method is no guarantee at all for effective video retrieval results. In order to compensate for the limitations of individual query methods, most video search approaches in the literature combine results from several methods.

3.4 Combining Query Results

We consider two general approaches for combining query results in the video retrieval literature: *parallel combination of query results, and sequential combination of query results.*

3.4.1 Parallel Combination

In parallel combination methods, all video retrieval results are taken into account simultaneously. Until recently, most combination approaches relied upon a mixture of query-by-keyword and query-by-example. Good examples are the probabilistic combination models proposed by Westerveld and De Vries [266] and Iyengar et al. [102]. With

the introduction of concept detectors in the shot-based video retrieval process, the dominant combination approach has shifted to a weighted average of individual query results, although other methods exist [150]. Yan and Hauptmann [281] present a theoretical framework for monotonic and linear combination functions in a video retrieval setting. They argue that a linear combination might be sufficient when fusing a small number of query results. Indeed, most work in this direction restricts itself to the combination of only few query results, e.g., [217, 256] who use a pair-wise combination of a couple of concept detectors.

A general problem of weighted averaging methods is the estimation of the individual weight parameters per query result. Ideally, one wants to learn the optimal weights based on validation set performance, but in practice, the amount of available training samples is too small. This is the reason why methods estimating weight parameters based on training data hardly exist, yet. As an alternative to learning, Chang et al. [36] determine the weight based on a text-match score. Other methods includes the work by Wang et al. [256], who fix the weights per query method, and the approach proposed by Wei et al. [262], who determine the weight of individual concept detectors by relating it to the cosine distance with respect to the query issued. However, all these combination methods are likely to be suboptimal, as the optimal weights typically vary per query. Naturally, the query classes introduced in Section 3.3.1, may also contain multiple query methods. When query classes are used, the individual weights have to be estimated for a limited number of query classes only. In addition to these parallel combination approaches, a further improvement of video retrieval might be achieved by inclusion of a sequential combination approach.

3.4.2 Sequential Combination

In contrast to parallel combination methods, where all video retrieval results are taken into account simultaneously, sequential combination methods update video retrieval results in succession. The common approach in sequential combination methods is to rely on variants of pseudo-relevance feedback algorithms [46, 82, 93, 117, 132, 137, 165, 284]. In general, pseudo-relevance feedback algorithms simply assume

a substantial number of video shots in the top of the ranking are relevant. The information associated with these top-ranked pseudo-relevant shots is then used to update the initial retrieval results.

Associations are based on different features. In [46], for example, the authors improve the ranking by using the text features of the top ten retrieved shots for re-ranking. Alternatively, Hsu et al. [93] suggest to exploit the visual consistency in the top-ranked shots, using clustering of visual features, to update the results. Apart from text and image features, one can also opt to update the initial ranking using concept detector probabilities. This approach is followed in [117, 132, 137]. In [117], for example, the authors construct for a number of top-ranked and bottom-ranked shots a feature vector containing 75 concept detector probabilities, which is then used in combination with a support vector machine to update the retrieval results. Clearly, all of these sequential combination methods depend on a reasonable initial ranking. When this initial ranking is unsatisfactory, updating the retrieval scores using pseudo-relevance feedback in combination with any multimedia feature becomes problematic, yielding modest performance.

To account for modest performance in the top-ranked video retrieval results, it was suggested in [284] to sample the bottom-ranked shots instead. They exploit the query images to identify the most dissimilar images in the ranking. Subsequently, they apply a support vector machine to re-rank the retrieval results. It was shown by Natsev et al. [165], however, that random sampling of pseudo-negative examples results in even better performance than sampling bottom-ranked shots. Especially when executed in a bagging scenario, where each bag uses the same set of positive query images and a random set of negative samples.

Traditional pseudo-relevance feedback assumes that most top-ranked results are relevant; the feedback algorithm proposed in [82], however, only requires the top-ranked shots to contain more relevant results than the bottom-ranked shots. The key innovation of the algorithm is to consider shot relevance together with all available concept detectors, even those not considered relevant for the query, as latent variables in an undirected graphical model. The algorithm jointly

optimizes these latent variables in an iterative fashion, which is guaranteed to converge in a few iterations, making it useful in practice.

Both parallel and sequential combination strategies may improve video retrieval. Present-day automated video retrieval solutions, however, have not yet received a robust level of performance. To allow for more robustness it is therefore wise to let a human user assess the relevance of individual shots interactively. In addition, she may take the automatic retrieved results as starting point for further exploration using a visualization of the video collection in the interface.

3.5 Visualization

Before we start delving into advanced visualization solutions let us first consider a typical web-based video search engine. These are almost exclusively based on textual meta data to find complete clips, rather than video segments, and results are displayed in a grid by showing one frame for each video. Navigation is limited to next page and scrolling. See Figure 3.2 for an example of such a standard interface. The solutions we will present below go beyond these basic techniques for visualizing individual video frames and navigating them.

3.5.1 Video Visualization

Let us first consider the presentation of one segment of video. The basic means for visualizing video content is, of course, a generic video player, but for a video search system we need different techniques especially when the video segments are not small coherent clips, but are programs of half an hour or more. Watching the entire video is not an option, therefore, many people have developed techniques to reduce the length of the video and keep only the relevant information. Extensive overviews of methods for summarization are presented in [153] and [239]. Here, we focus on some basic techniques. The most common way of visualizing the content is by a set of static key frames, typically 1–3 per shot. Lee and Smeaton [125] give an overview of dimensions in the design space for key frame-based video browsing systems. They distinguish the layeredness, the spatial versus temporal presentation and the

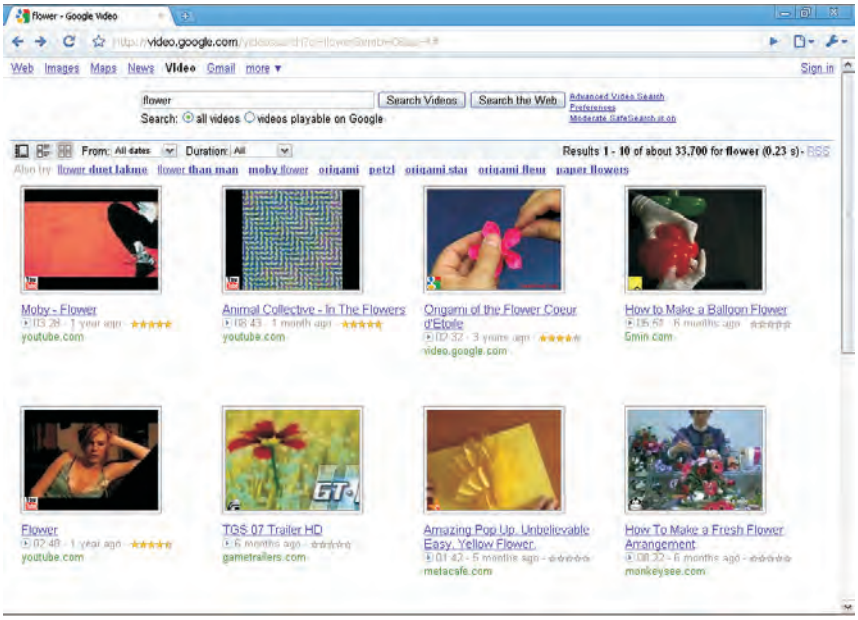
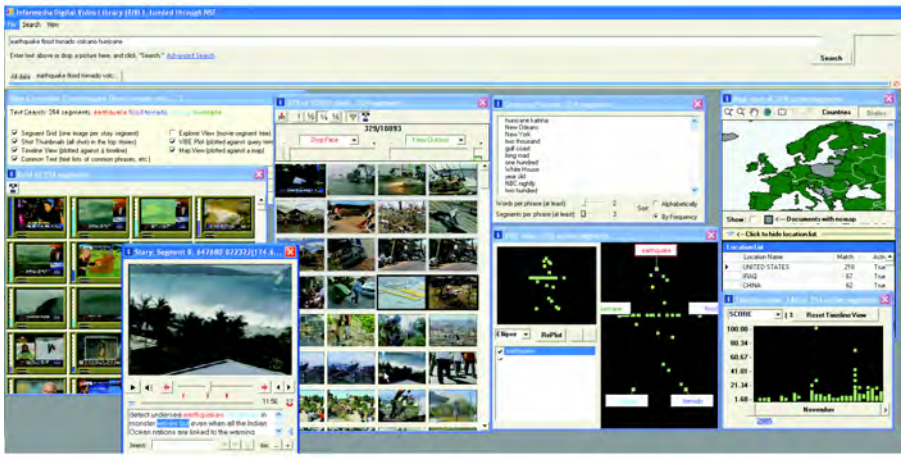


Fig. 3.2 Common visualization of video retrieval results by a video search engine on the Internet. Note the influence of textual meta data, such as the video title, on the search results.

temporal orientation as dimensions. Most of the existing browsers can be mapped to specific values for these dimensions, an advanced example is given in Figure. 3.3a. Simple storyboard methods are actually very effective [43]. Rather than taking the shot as basis, [3] visualize the story units in the video as a collage of key frames. The system in [42] also creates collages of key frames. It exploits named entity extraction on the automatic speech recognition results to map the key frames to the associated geographic location on a map and combine this with various other visualizations to give the user an understanding of the context of the query result, as illustrated in Figure. 3.3b. When moving to specific domains, systems can provide more specific summaries and visualizations of the video content. A good example of this is the video browsing environment presented in [77] for the domain of video lectures. It uses standard mechanisms like showing the timeline, and also shows faces of speakers appearing in the video, the speaker transitions, and slide changes.

(a)



(b)

Fig. 3.3 (a) The Fischlar/K-space interface [268] showing a set of key frames for a number of videos. (b) The Informedia classic interface [255] showing the result as a storyboard: A 2D grid with subsequent key frames taken from the video segment. In addition the geographic and conceptual context for the segment are displayed.

3.5.2 Video Set Visualization

Almost all systems take a user query as the basis for visualization. As the query yields a ranked list, systems present a linear set of key frames to the user. Commonly, the result is paginated and each page is displayed in a 2D grid. In the grid, the linear ranking of results is mapped to left-right top-down reading order. To go through the results, the user just flips through the various pages. Hauptmann et al. [84] follow an extreme approach by presenting pages dynamically to the user in a very rapid fashion. The user only has controls to indicate whether the results are correct or not. Rather than assuming reading order, Zavesky et al. [294, 295] employ the 2D nature of the grid by computing optimal data-driven coordinate axes. Thus in every page of results, clusters of visually related key frames appear. Pecenovic et al. [179] and Nguyen and Worring [168] also use a true 2D display, but they do not rely on a grid. Instead, they place key frames on the screen in such a way that the dissimilarity among the features of the key frames are optimally preserved, using various projection methods. Both organize the data in a hierarchy to facilitate browsing at the overview level, viewing one representative frame per cluster, as well as more detailed viewing of individual clusters.

A disadvantage of the query-based methods is that the user has to go through the result list and has to switch back to the query screen when she cannot find more results that are relevant. To break this iterative query-view approach we should give options to depart from the initial result presented and delve into local browsing opportunities. Rautiainen et al. [191] suggest presenting the timeline of the video on the horizontal axis. They use a full grid, where the vertical dimension shows similar shots for every key frame in the timeline. The CrossBrowser presented in [227] does not use a full grid but shows the ranked results as a vertical linear list, while using the horizontal dimension for showing the timeline of the video the current shot is part of. The browsers presented by De Rooij et al. [52, 53] take this a step further and give up to eight different ranked lists starting at the current shot to explore, including dimensions for visually similar shots, and also for shots that are similar given their concept probabilities. This allows the user to find new results

Fig. 3.4 The ForkBrowser [53] of the MediaMill semantic video search engine [227], which allows the user to browse the video collection along various dimensions exploring different characteristics of the collection.

by browsing along different linear lists, instead of entering new queries. This is illustrated in Figure 3.4. Using such a similarity based approach is good for search, but for explorative browsing a potential disadvantage is that the user might be exploring redundant information. Wu et al. [271] employ visual duplicate detection and speech transcripts to help the user in focusing on novel information only. As an alternative to using linear lists, Luo et al. [142] visualize relations between shots with a hyperbolic tree, allowing for hierarchical browsing. The NN^k networks described in [87] give users unconstrained navigation in a full network of relations between images.

At this point, we should realize that visualizations give the user insight in the video collections and the results of the query, but that the main purpose of visualization is to provide the means for effective user interaction.

3.6 Learning from the User

In content-based image retrieval, the role of interaction has been recognized early as key to successful applications [213]. For video retrieval, the user performs two types of interactions with the system: selection of relevant/irrelevant results and commands to navigate the collection of results. Methods for learning from the navigation paths chosen are few; we focus here on the user selections.

Extensive overviews of generic methods for learning from user interaction are presented in [96, 302]. We make a general distinction between methods based on relevance feedback and active learning. In relevance feedback [197], the system uses the current sets of relevant and irrelevant results to update the model it has built for the particular query. The system then presents the next set of results and the process repeats. Active learning not only optimizes the model, it also computes which elements from the unlabeled data pool are most informative. Hence, the system picks these for presentation to the user.

Chen et al. [38] were among the first to apply active learning in concept-based video retrieval. They select the elements closest to the current boundary between the relevant and non-relevant items as these are the most uncertain elements. Then they employ relevance feedback to improve precision in the top-ranked list. Their results indicate that this approach often leads to a focus on a small part of the information space. They, therefore, propose to perform active learning for different modalities and combine them in the next phase. Goh et al. [68] argue that different semantic concepts require different elements to be judged by the user, and different sampling strategies. The system in [141] takes such an adaptive sampling strategy to obtain a balance between seeking high precision and high recall. The system starts by presenting elements far away from the decision boundary, i.e., elements most likely to be relevant to the query. If the user feedback indicates that those elements are indeed relevant, the system continues sampling this part of the information space. When the number of elements labeled as relevant by the user becomes too small, the system starts selecting elements close to the boundary to update the current model, hoping to increase recall by finding new areas to explore. The interface is depicted in Figure 3.5.

Fig. 3.5 Interface of the VisionGo system [141], which uses active learning to make the interactive search process more effective.

Nguyen et al. [169] also choose elements close to the boundary. They, however, do so in an information space based on distances to a number of selected prototypes. They show that with a reasonable number of prototypes, a distance preserving projection to 2D space can be used as a good approximation to the original space. This has the advantage that the decision boundary and the selected examples can be shown to the user in a manner leading to intuitive feedback. Instead of determining prototypes based on the full feature vectors, Bruno et al. [24] determine different prototypes for each modality, and learning is performed separately on every modality. Hence, the number of prototypes becomes larger, but as the feature vectors are smaller, the classifier can be of lower complexity.

Although interaction was recognized early as a key element, the above interaction methods have only started to unlock the potential that intelligent interaction has for video retrieval. A thorough understanding of the relation between the space spanned by different concepts, the proper sampling strategy, intuitive displays and effective navigation of this space is needed.

3.7 Discussion

For video search solutions targeting at retrieval of specific segments, query-by-concept is closest to what we are used to in traditional online search engines. Whenever the quality and quantity of the detectors is sufficiently high, we consider it the query method of choice for video retrieval. Automatic concept selection methods are getting to the point that, for a given information need in many cases, they will select the same concept a human would [100]. For retrieval tasks precisely covered by one of the concepts in the lexicon, even weak performance of a detector may be sufficient. However, for combination tasks like *find shots of demonstrators burning a flag*, present-day search results often yield nothing but noise. Hence, we need a major boost in detection performance before we can achieve meaningful concept detector combinations. The concept detector selection methods are also not yet capable of employing common sense, associating the information need to, objectively spoken, a not directly related concept. How to teach a system to select the *political leader* concept when searching for *people shaking hands*? Furthermore, how a system should decide on the semantic and visual coverage, while simultaneously taking the quality of available detectors into account is an open problem.

For the moment, interaction will remain part of any practical video search system. So, rather than considering the interaction as a method to counteract any errors of the automatic system, we should incorporate the interaction into the design of the system. To that end, we fortunately see a large body of work in the machine learning community on optimal active learning-based techniques. These should be adapted to deal with the large volumes of data in video and the varying quality of some of the concept detector results. In addition, visualizing the query results in an intuitive way is as important in the design of any video retrieval system. We should develop methods that truly integrate active learning with visualization and do so while conforming to principles from the field of human–computer interaction. Finally, new emerging areas to explore, with many challenges, are collaborative search [4, 210] and mobile video retrieval [98, 210, 276].

4

Evaluation

4.1 Introduction

Evaluation of concept-based video retrieval systems has always been a delicate issue. Due to copyrights and the sheer volume of data involved, video archives for evaluation are fragmented and mostly inaccessible. Therefore, comparison of systems has traditionally been difficult, often even impossible. Consequently, many researchers have evaluated their concept-based video retrieval methodologies on specific video data sets in the past. To make matters worse, as the evaluation requires substantial effort, they often evaluated sub-modules of a complete video retrieval system only. Such a limited evaluation approach is hampering progress, because methodologies cannot be valued on their relative merit with respect to their concept-based video retrieval performance.

To tackle the evaluation problem, many evaluation campaigns have been introduced in recent years. Notable examples are the ETISEO [167] and PETS [60] initiatives, which focus on surveillance video, the AMI project [192], which emphasizes the analysis of meeting videos; VideoCLEF [122], which studies retrieval tasks related to multi-lingual video content; and, the US Government Video Analysis and Content

Extraction program, which ultimately aims for general video content understanding based on visual recognition (see e.g., [113]). While all these initiatives are important to foster video retrieval progress, so far, none of them played a role as significant as the benchmark organized by the American National Institute of Standards and Technology (NIST). NIST extended their successful Text Retrieval Conference (TREC) series [254] in 2001 with a track focusing on automatic segmentation, indexing, and content-based retrieval of digital video, which became an independent evaluation workshop known as TRECVID in 2003 [212]. Due to its widespread acceptance in the field, resulting in large participation of international teams from universities, research institutes, and corporate research labs, the TRECVID benchmark can be regarded as the *de facto* standard to evaluate performance of concept-based video retrieval research. Already the benchmark has made a huge impact on the video retrieval community, resulting in a large number of video retrieval systems and publications that report on the experiments performed within TRECVID, as exemplified by many references in Sections 2 and 3.

In this section, we introduce in Section 4.2, the TRECVID benchmark and discuss its video data, its tasks, and its criteria related to evaluation of concept-based video retrieval systems. In Section 4.3, we elaborate on various manual annotation efforts that have emerged during the past years. We summarize various baseline systems and their components, against which concept-based video retrieval systems can be compared in Section 4.4. We provide an overview of concept detection and retrieval results obtained on various editions of relevant TRECVID benchmark tasks in Section 4.5. We end the section with a discussion on evaluation in Section 4.6.

4.2 TRECVID Benchmark

The aim of the TRECVID benchmark is to promote progress in the field of video retrieval by providing a large video collection, uniform evaluation procedures, and a forum for researchers interested in comparing their results [211, 212]. Similar to its text equivalent, TRECVID is a controlled laboratory-style evaluation [234] that attempts to model real

world information needs. TRECVID evaluates the information retrieval systems that solve these information needs, or significant components of such systems that contribute to the solution, by defining specific benchmark tasks. Researchers may participate in these benchmark tasks by obtaining the video data, optimizing results on the training set, and submitting their test set results to NIST. Finally, NIST performs an independent examination of results using standard information retrieval evaluation measures. We will now detail the relevant TRECVID video data sets, benchmark tasks, and the criteria used for evaluation of concept-based video retrieval systems.

4.2.1 Video Data

Although NIST provided video data sets in 2001 and 2002 also, as part of TREC, we consider these editions of the benchmark initial attempts that aided in formulating the tasks and their evaluation for the later editions. Starting from TRECVID 2003, the video data changed significantly in both quality and quantity. The selection of video material has been a compromise between availability on one hand, and ability to cover retrieval needs relevant to multiple target groups on the other hand. In order to reduce the amount of effort involved in securing large amounts of video data and creation of source-specific training sets, TRECVID works in cycles, meaning that the same video data source is re-used for multiple editions of the benchmark, with the exception of the test data of course. For each video, an automatic shot segmentation result is provided. These shots serve as the basic unit of testing and performance assessment within the benchmark. NIST splits the shot-segmented archive into a training and test set. We provide an overview of the TRECVID video data sets as used since 2003 in Table 4.1, and detail them per cycle below.

Cycle 1, 2003–2004: US broadcast news. The first cycle in the TRECVID workshop series included the 2003 and 2004 editions and relied on a US broadcast news video archive containing ABC World News Tonight and CNN Headline News from the first half of 1998. In addition to the broadcast news data, the archive contained 12 hours of C-SPAN video, mostly covering discussions and public hearings from

Table 4.1 Overview of TRECVID video data set statistics since 2003, for both training and test data for the benchmarks tasks related to concept-based video retrieval. Key frames were not provided in 2008. Pointers to obtain all TRECVID-related data are available from: http://trecvid.nist.gov.

	TRECVID video data set statistics									
	Hours		***Videos***		***Shots***		***Key frames***		***Gigabytes***	
Edition	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
2003	61	56	133	113	34,950	32,318	66,499	57,662	48	45
2004	105	61	221	128	66,076	33,367	122,917	48,818	83	45
2005	86	85	137	140	43,907	45,765	74,337	77,814	62	61
2006	171	166	277	259	89,672	79,484	152,151	146,007	124	127
2007	56	53	110	109	18,120	18,142	21,532	22,084	31	29
2008	109	109	219	219	36,262	35,766	43,616	—	60	60

1999 to 2001. The video archive of the 2004 benchmark extends the data set used in 2003, but without the C-SPAN video data. The training data consists of the archive used in 2003. The 2004 test data covers the period of October until December 1998. For this first TRECVID cycle, LIMSI [62] donated automatic speech recognition results, and CLIPS-IMAG [187] provided the camera shot segmentation and associated key frames.

Cycle 2, 2005–2006: International broadcast news. During the second cycle of the TRECVID workshop, the emphasis shifted from US broadcast news to international broadcast news video. The archive of the 2005 benchmark contains in total 277 Arabic, Chinese, and English news broadcasts, all recorded during November 2004. The video archives from the second cycle come with several automatic speech recognition results and machine translations for the Arabic and Chinese language programs, donated by various sources [174, 175]. It should be noted that both the speech recognition and machine translations yield error prone detection results. What is more, due to the machine translation, the text is unsynchronized with the visual content. Since the speech transcripts are less accurate then the ones used in the first cycle, this archive is more challenging. In addition, this archive also contains a greater variety in production style, i.e., it includes 13 news programs from six different TV channels from three countries. The training data for TRECVID 2005 is taken from the first half of November 2004, the remaining data was used for testing. This entire 2005 archive was

extended with new test data for the 2006 edition of the TRECVID benchmark. The 259 additional news broadcasts are from November to December 2005. For all videos in this cycle, the Fraunhofer (Heinrich Hertz) Institute [181] provided a camera shot segmentation, and Dublin City University created a common set of key frames [208].

Cycle 3, 2007–2009: Dutch broadcast infotainment. For the third TRECVID workshop cycle, the Netherlands Institute for Sound and Vision has provided about 400 hours of broadcast video, containing news magazines, science news, news reports, documentaries, educational programming, and historical (gray scale) video for use within the benchmark editions of 2007 until 2009. Similar to the previous two cycles, NIST extend the entire archive of 2007 with new test data for reuse in TRECVID 2008. The video archives from the third cycle come with output of an automatic Dutch speech recognition system provided by the University of Twente [97], which in turn has been translated into English using machine translation provided by Queen Mary, University of London. Similar to Arabic and Chinese, the quality of Dutch speech recognition is not as robust as its English equivalent, as only limited amounts of Dutch language training data are available. For all videos in this cycle, the Fraunhofer (Heinrich Hertz) Institute [181] again provided a camera shot segmentation, and for the first time no official key frames were provided by NIST to encourage video processing beyond the key frame.

We provide a number of key frame samples from the various video data sets used in the TRECVID benchmark 2003–2008 in Table 4.2.

4.2.2 Benchmark Tasks

Benchmark tasks defined within TRECVID include camera shot segmentation [292], story segmentation [45], content-based copy detection [173], surveillance event detection [173], video summarization [176], concept detection [162], and several search tasks [81]. While all of these tasks are important for the field of video retrieval at large, not all of them are directly relevant for the evaluation of concept-based video retrieval systems. In this section, we restrict ourselves to the TRECVID benchmark tasks relevant for concept-based video retrieval

Table 4.2 Key frame samples from the TRECVID video data sets as specified in Table 4.1.

Edition	**TRECVID video data set samples**
2003	
2004	
2005	
2006	
2007	
2008	

as used since 2003: *concept detection*,[1] *interactive search*, and *automatic search*. These tasks, their participants, and submissions are summarized in Table 4.3, and we detail them per task below.

Concept detection task: The goal in the concept detection task is to detect in the shot-segmented video data the presence of a set of pre-defined semantic concepts. Indeed, the methods detailed in

[1] TRECVID refers to this task as the high-level feature extraction task, to prevent misunderstanding with feature extraction as used in Section 2 we refer to it as the concept detection task.

Table 4.3 Overview of the concept detection, interactive search, and automatic search task as evaluated in the TRECVID benchmark since 2003, including number of allowed runs, total submissions, and number of participating teams per task. The automatic search task was a pilot in 2004, the official version started in 2005.

	TRECVID benchmark task statistics											
	Concept detection task				***Interactive search task***				***Automatic search task***			
Edition	*Concepts*	*Runs*	*Submits*	*Teams*	*Topics*	*Runs*	*Submits*	*Teams*	*Topics*	*Runs*	*Submits*	*Teams*
2003	17	10	60	10	25	10	36	10	—	—	—	—
2004	10	10	82	12	23	10	61	14	23	10	23	5
2005	10	7	110	22	24	7	49	14	24	7	42	11
2006	20	6	125	30	24	6	36	18	24	6	76	17
2007	20	6	163	32	24	6	32	11	24	6	82	17
2008	20	6	200	42	24	6	34	14	48	6	82	17

Section 2 aim for the same goal. The basis for a concept is its textual description as defined by NIST. This description is meant to be clear for both researchers developing concept detectors, as well as for assessors who evaluate the final ranked list of results. For each concept in the set, participants are allowed to submit a list of at most 2000 shots from the test set, ranked according to concept presence likelihood. Since 2003 the popularity of this task has increased substantially, indicating that the concept detection task has evolved into a mature benchmark task within TRECVID.

The set of concepts defined and evaluated within each TRECVID edition contains a mixture of concepts related to people (e.g., *news subject monologue*), objects (e.g., *computer screen*), settings (e.g., *classroom*), and events (e.g., *basket scored*). Although the defined concepts have become more and more complex over the years, consecutive editions of the concept detection task do contain partial overlap in the set of concepts to be detected. This is useful to measure consistency in detection over multiple cycles, as well as generalization capabilities of detectors over multiple video data sets. We summarize a number of concept definitions together with visual examples from each TRECVID edition in Table 4.4.

Interactive search task: The aim of the interactive search task is to retrieve from a video archive, pre-segmented into n unique shots, the best possible answer set in response to a multimedia information need. Given such a video information need, in the form of a search topic, a user is allowed to engage in an interactive session with a video search engine. Based on the results obtained, a user interacts with a video

Table 4.4 Concept definition examples from the TRECVID concept detection task as specified in Table 4.3 using the video data from Table 4.1.

Edition	**TRECVID concept definition examples**
	
2003 Aircraft	Segment contains at least one aircraft of any sort.
	
2004 Beach	Segment contains video of a beach with the water and the shore visible.
	
2005 Mountain	Segment contains video of a mountain or mountain range with slope(s) visible.
	
2006 People-Marching	Shots depicting many people marching as in a parade or a protest.
	
2007 Police-Security	Shots depicting law enforcement or private security agency personnel.
	
2008 Flower	A plant with flowers in bloom; may just be the flower.

search engine; aiming at retrieval of more and more accurate results. To limit the amount of user interaction and to measure the efficiency of a video search system, all individual search topics are bounded by a time limit. The upper limit has been 15 min until TRECVID 2007; starting in 2008 NIST reduced the upper limit to 10 min. For each individual topic, participants are allowed to submit up to a maximum

of 1000 shot-based results, ranked according to the highest possibility of topic presence.

The TRECVID guidelines prohibit optimizing a search system for the topics evaluated. Therefore, NIST reveals the topics only a few weeks before the submission deadline. Every submission must contain results for each search topic using the same interactive video retrieval system variant. Each individual topic result should stem from only one searcher, but within a run, different searchers may solve different topics. In order to avoid bias in the experimentation and comparison among system settings, TRECVID has provided example experimental designs for measuring and comparing the effectiveness of interactive video search engines, including user questionnaires, but, so far, only a few participants follow these suggestions in their experimental design [40, 242, 269]. Whether or not these suggestions are followed, obviously none of the searchers should have any experience with the test video data. Moreover, they should have no experience with the topics beyond the general world knowledge of an educated adult. Since its inception, the interactive search task has attracted a wide interest, resulting in a steady number of participants ranging from 10 in 2003 to 14 in 2008.

Each year NIST provides about 24 search topics for the interactive search task. NIST creates the topics manually by examining a subset of the test set videos. In 2003 and 2004, NIST also examined the video archive query logs provided by the BBC, corresponding to the time period of the US broadcast news cycle. The search topics are provided as an XML file containing for each topic a textual description, and a number of audio-visual examples from the training set and/or the Internet. The search topics express the need for video concerning one or more people (e.g., *find shots of Boris Yeltsin*), objects (e.g., *find shots of one or more soccer goalposts*), settings (e.g., *find shots of the front of the White House in the daytime with the fountain running*), events (e.g., *find shots of a train in motion*) or their combinations (e.g., *find grayscale shots of a street with one or more buildings and one or more people*). Early editions of the interactive search task, covering the cycles on broadcast news, emphasized topics looking for named persons frequently appearing in the news, like *Pope John Paul II*, *Hu Jintao*, and *Tony Blair* [287]. In the third cycle, TRECVID emphasizes complex

Table 4.5 Search topic examples from the TRECVID interactive and automatic search task, as specified in Table 4.3, using the video data from Table 4.1.

Edition	**TRECVID search topic examples**
	
2003	Find shots of Osama Bin Laden.
	
2004	Find shots of a hockey rink with at least one of the nets fully visible from some point of view.
	
2005	Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people.
	
2006	Find shots of a group including at least four people dressed in suits, seated, and with at least one flag.
	
2007	Find shots of a road taken from a moving vehicle through the front windshield.
	
2008	Find shots of a person pushing a child in a stroller or baby carriage.

topics requesting for multiple people and objects in specific settings and specific events, which forces searchers to inspect video sequences rather than key frames only. We summarize a number of search topics together with visual examples from each TRECVID edition in Table 4.5.

Automatic search task: Similar to TRECVID's interactive search task, the goal of the automatic search task is to satisfy a number of video information needs. In contrast to the interactive search task, however, the automatic search task disallows any form of user interaction. Thus, an automatic video retrieval system must solve all search topics autonomously. To quantify the influence of using visual information in addition to text, an additional constraint, specific for this task, is that each submission must also be accompanied by a text-only baseline run. This baseline should use the provided speech recognition results and/or their machine translations. A first pilot of the automatic search task started in 2004. Since this task became an official task in 2005 it has always used the same set of search topics as the interactive search task, but starting in 2008 the number of topics for the automatic search task has been extended to cover 48 topics.

For sake of completeness, it is of interest to note that TRECVID has been organizing an additional search task, namely the manual search task. In this task, a user is allowed to interpret a search topic by formulating a single query to a video search engine, but without any further user interaction afterwards. Due to the declining number of total search runs that participants are allowed to submit in TRECVID and the increasing popularity of the automatic search task, participation in the manual search task had diminished to a level that we no longer consider it relevant.

4.2.3 Evaluation Criteria

TRECVID lets human assessors at NIST inspect the submitted ranked results for both the concept detection task as well as the search tasks. The assessors are instructed to consider shots containing video of the target concept and topic correct, but shots containing videos of physical objects representing the target concept or topic, such as static photos, paintings, toy versions, or models, should be considered wrong. Moreover, it is not required for the concept or topic to be visible for the entire duration of the shot, it is already sufficient if they are visible for some frame sequence within the shot. By establishing a ground truth for the various tasks, NIST is able to compute several

common performance metrics, which in turn may be used to compare the various systems participating in the benchmark task. From all performance metrics on offer, the *average precision* [254] and its recent derivative the *inferred average precision* [290] are considered the most reliable metrics to evaluate relative video retrieval system performance.

Average precision: The average precision is a single-valued measure that is proportional to the area under a recall-precision curve. This value is the average of the precision over all relevant judged shots. Hence, it combines precision and recall into a single performance value. Let $L^k = \{l_1, l_2, \ldots, l_k\}$ be a ranked version of the answer set A. At any given rank k, let $R \cap L^k$ be the number of relevant shots in the top k of L, where R is the total number of relevant shots. Then average precision is defined as

$$\text{average precision} = \frac{1}{R} \sum_{k=1}^{A} \frac{R \cap L^k}{k} \psi(l_k), \tag{4.1}$$

where indicator function $\psi(l_k) = 1$ if $l_k \in R$ and 0 otherwise. In case, a subset of the answer set is evaluated, e.g., the top 1000 ranked shots only, the fraction $1/R$ should be replaced with $1/\min(R, A)$. As the denominator k and the value of $\psi(l_k)$ are dominant in determining average precision, it can be understood that this metric favours highly ranked relevant shots. In addition, the metric is biased by the number of relevant shots available, causing differences in *a priori* random baseline performance.

TRECVID uses a pooled ground truth P, to reduce labour-intensive manual judgments of all submitted runs. They take from each submitted run a fixed number of ranked shots, which is combined into a list of unique shots. Every submission is then evaluated based on the results of assessing this merged subset, i.e., instead of using R in Equation (4.1), P is used, where $P \subset R$. This is a fair comparison for submitted runs, since it assures that for each submitted run at least a fixed number of shots are evaluated at the more important top of the ranked list. However, using a pooled ground truth based on manual judgment comes at a price. In addition to mistakes by relevance assessors that may appear,

using a pooling mechanism for evaluation means that the ground truth of the test data is incomplete. When comparing new results against the pooled ground truth these limitations should be taken into account.

Inferred average precision: Although the pooling process assures that relative system performance can be measured reliably, it still requires substantial amount of human effort. Hence, the number of concepts and topics that can be evaluated is bounded by human labour and its associated cost. An obvious solution would be to restrict the depth of the ranked lists evaluated or to use random sampling strategies. It was shown in [20, 25, 290], however, that evaluation measures, such as average precision, become unstable when judgment sets become less complete. This behaviour is caused by the fact that average precision considers all unjudged shots to be non-relevant. As a result the average precision value is reduced, especially for the best performing systems [290].

To counter the sensitivity of average precision to incomplete judgments, Yilmaz and Aslam [290] start from the observation that average precision can be viewed as the expectation of a random experiment. They show that this expectation can be estimated under the assumption that relevance judgements are a uniform random subset of the set of complete judgements. They propose the inferred average precision, which is a close approximation of average precision. In case, judgments are complete, the two measures are equivalent. If judgments are incomplete, unjudged shots are considered a random sample from the set of complete judgements and estimated accordingly. When used in such a scenario, inferred average precision underestimates the true average precision, but the relative systems ranking is similar to those based on regular average precision.

Overall system quality: As an indicator for overall quality of both concept detection and search systems, TRECVID computes the mean average precision and mean inferred average precision, over all concepts and/or search topics from one run. The introduction of inferred average precision offered TRECVID the opportunity to scale-up in the number of evaluated concepts and search topics by using a reduced random

sample of the usual pools, and hence, to provide a better estimate of overall system performance, by averaging over a larger set of concepts and/or topics. It should be noted, however, that runs can not be evaluated on these single average values alone, and that various runs should at least be be compared via randomization tests also to verify their significance [83].

4.3 Annotation Efforts

The concept-based video retrieval approaches from Sections 2 and 3 rely heavily on supervised machine learning, which in turn depend on labeled examples. Hence, annotated visual examples are a valuable resource. Moreover, when aiming for repeatability of experiments this ground truth needs to be shared among researchers. Manual annotation of visual data, however, is a time and resource consuming process. Furthermore, despite the high costs involved, the accuracy of the annotations is never guaranteed. To limit costs, time, resources, and the amount of possible mistakes, many annotation initiatives have emerged during several editions of the TRECVID benchmark. The annotation effort has resulted in various sets of annotated concepts and search topics, which have become publicly available, as detailed below.

4.3.1 Concept Annotations

Given a limited amount of human labeling effort, the goal of concept labeling is to maximize the annotation result. Consequently, for any annotation effort, there is always the need to strike a balance between the level of spatio-temporal annotation detail, the number of concepts to be annotated, and their individual amount of positive and negative labels. For this reason, many research teams have developed labeling tools and made them available for collaborative annotation [8, 133, 251]. The annotation is mostly performed on a global image level. Some initiatives have also provided annotations on regions inside images [133, 291, 293], typically by providing the image coordinates of bounding boxes. Despite their importance, we are aware of only one effort to annotate dynamic concepts [116]. Obviously, region and temporal annotation of video frames are more expensive in terms

Fig. 4.1 User interface of a concept annotation system [251] for manual labeling of individual key frames on a global level.

of human effort than annotating global frames. A good example of a tool that maximizes human labeling effort is the Efficient Video Annotation system by Volkmer et al. [251]. This annotation tool provides a web-based interface allowing for global annotation of a small number of images per shot, commonly referred to as sub-shot level, see Figure 4.1.

The number of concepts to be annotated has probably the biggest impact on both the quality and the number of concept labels. Since several concepts are defined within TRECVID's concept detection task each year, these concepts often form the basis for the annotation effort [8, 133, 224, 235, 291, 293]. In addition to this task-defined lexicon of concepts, several annotation efforts have provided labeled examples for many more concepts. The most structured and complete annotation effort is the large scale concept ontology for multimedia, commonly known as LSCOM, by IBM,

Carnegie Mellon University, Columbia University, Cyc, and several library scientists [163]. They led a collaborative annotation effort during TRECVID 2005, yielding annotations for a vocabulary of 39 concepts, known as the light-scale concept ontology for multimedia (LSCOM-Lite) [159]. Concepts in this ontology are chosen based on presence in WordNet [59] and extensive analysis of video archive query logs. The concepts are related to program categories, setting, people, objects, activities, events, and graphics. MediaMill [230] extended this lexicon to 101 concepts. We refer to Figure 1.1 in the Introduction for visual examples from both lexicons. LSCOM-Full extended LSCOM-Lite to a total of 449 annotated concepts in broadcast news, which are useful for retrieval, observable by humans, and presumed feasible to detect [163]. Interestingly, LSCOM-Full is one of the few annotation efforts that closely involved end-user communities in the selection of concepts to annotate. As highlighted in Section 2.5, LSCOM connected these concepts to the Cyc ontology. In all concept annotation efforts, the number of positive examples vary heavily between concepts. Surprisingly, only few collaborative efforts have mentioned statistics on the quality of the concept annotations. A notable exception is [251], who report an average inter-user disagreement of 3.1% for the annotation of 61.904 sub-shots with 39 concepts.

Although fixed lexicons are preferred, alternatives to scale-up the number of concepts do exist. The LSCOM-Full set, for example, is expanded further by connecting it to the Cyc ontology. Alternatively, the annotation initiative by Lin et al. [133] allowed users to provide annotations using free-text for concepts not previously defined. Similar to social tagging, however, free-text annotations are known to be ambiguous and susceptible to mistakes [115, 251]. To assess the reliability of concept annotation efforts, it was suggested in [115] to check whether concept frequency conforms to Zipf's law. Unfortunately, efforts to correct unreliable annotations are rare. Notable exceptions are [7, 8, 252], who exploit inter-rater agreement between annotators [252] and active learning strategies to determine complex key frames requiring agreement among more users [7, 8]. We summarize several concept annotation efforts on various editions of TRECVID video data sets in Table 4.6.

Table 4.6 Overview of TRECVID-inspired concept annotation efforts. Pointers to obtain the annotation data are available in the references and at http://trecvid.nist.gov/trecvid.data.html.

	Concept annotation efforts				
	Annotation effort		***Concepts***		
Edition	*Initiative*	*Reference*	*Granularity*	*Scale*	*Number*
2003	VCA forum	[133]	Sub-shot	Region	133
2004	MediaMill	[224]	Shot	Global	32
2005	LSCOM-Lite	[159, 251]	Sub-shot	Global	39
2005	LSCOM-Full	[163]	Sub-shot	Global	449
2005	LSCOM-Event	[116]	Sub-shot	Global	24
2005	MediaMill	[230]	Shot	Global	101
2005	Tsinghua/Intel	[291]	Sub-shot	Region	27
2007	ICT-CAS	[235]	Sub-shot	Global	36
2007	LIG	[7, 8]	Sub-shot	Global	36
2007	Tsinghua/Intel	[293]	Sub-shot	Region	27
2008	ICT-CAS	[235]	Sub-shot	Global	20
2008	LIG	[7, 8]	Sub-shot	Global	20

4.3.2 Search Topic Annotations

Because search topics often describe a complex visually oriented information need, requiring careful inspection, initiatives to label video search topics are rare. Notable exceptions include the efforts by Naphade et al., Snoek et al., Yan and Hauptmann [163, 222, 283]. Yan and Hauptmann [283] supplemented the ground truth of official TRECVID search topics on test data, by also annotating the corresponding training sets. A collaborative effort was initiated in [222], providing ground truth annotations on TRECVID test data for 15 search topics not covered by TRECVID. The most extensive labeling effort for search topics to date is again undertaken by the LSCOM initiative [163]. They provide ground truth labels for 83 queries on the TRECVID 2005 training set. The topics are inspired by common broadcast news stories such as those related to elections, warfare, and natural disasters. We summarize the topic annotation efforts on various editions of TRECVID video data sets in Table 4.7.

4.4 Baselines

In addition to the concept and topic annotations, many TRECVID participants have donated shot segmentations, key frames, features,

Table 4.7 Overview of TRECVID-inspired topic annotation efforts. Pointers to obtain the annotation data are available in the references and at http://trecvid.nist.gov/trecvid.data.html.

	Search topic annotation efforts			
	Annotation effort		***Topics***	
Edition	*Initiative*	*Reference*	*Granularity*	*Number*
2003	Carnegie Mellon University	[283]	Shot	20
2004	Carnegie Mellon University	[283]	Shot	24
2005	Carnegie Mellon University	[283]	Shot	24
2005	LSCOM Use Case Queries	[163]	Sub-shot	83
2006	VideOlympics 2007	[222]	Shot	3
2007	VideOlympics 2008	[222]	Shot	12
2008	VideOlympics 2009	[222]	Shot	9

speech recognition results, and machine translations in the course of the benchmark. Moreover, all participants share their submitted concept detection results on TRECVID test data. It has, however, been identified that many concept detectors are required for concept-based video retrieval to be effective [85, 227]. To prevent expensive duplicate efforts, while simultaneously offering the community a means to compare their concept detectors in a large-scale setting, several concept detection baselines have been made publicly available.

In addition to a video data set and labeled examples, the common elements of a concept detection baseline are pre-computed low-level features, trained classifier models, and baseline experiment performance for several of the components defined in Section 2. This stimulates further investigations by offering fellow researchers the opportunity to replace one or more components of the provided baseline implementation for one or more of their own algorithms, and to compare performance directly. Furthermore, baselines lower the threshold for novice researchers from other disciplines to enter the field of concept-based video retrieval. Finally, the baseline concept detectors can be a valuable resource for (interactive) video retrieval experiments.

Although all concept detection baselines share many commonalities, each has a different emphasis. The MediaMill Challenge by Snoek et al. [230] emphasizes repeatability. By decomposing the generic concept detection problem into a visual-only, textual-only, multimodal feature fusion, supervised classifier fusion, and unsupervised classifier fusion

step, they identify five repeatable experiments. For each concept detection experiment, they provide a baseline implementation together with established performance figures for a total of 101 semantic concepts. The baseline provided by Columbia University [286] emphasizes utility for video search. They provide 374 detectors based on color, texture, and edge features, and their unsupervised classifier fusion. In a similar vein, VIREO [109] released the same set of detectors, with a special emphasis on keypoint-based features. These detectors provide better performance than the ones by Columbia, but require more computation time. Naturally, the results of various concept detection baselines may be combined, this is suggested by Jiang et al. [110] who simply averages the detectors of both Columbia and VIREO into a more robust set named CU-VIREO374. Indeed, the fusion yields improved performance over the individual baselines for the 20 concepts evaluated in TRECVID 2006 and 2007. We summarize several publicly available concept detection baselines in Table 4.8.

It can be argued that sharing concept detector results is only useful for somewhat limited experiments. In general, it would be more useful to provide concept detection software that can also classify other video frames, allowing for performance comparisons on other video data sets, for example in earlier or subsequent TRECVID editions and other benchmarks. Recently, researchers have started to share software components [28, 33, 111, 216, 245, 301] that allow to construct concept detectors easily. In addition, some coordinated software sharing efforts have just started [80]. We anticipate more open-source software initiatives on this topic to emerge in the near-future.

Table 4.8 Overview of TRECVID-inspired concept detection baselines using the annotations from Table 4.6. Pointers to obtain the annotation data are available in the references and at http://trecvid.nist.gov/trecvid.data.html.

	Concept detection baselines				
	Baseline		***Concept detectors***		
Edition	*Initiative*	*Reference*	*Annotations*	*Provided*	*Evaluated*
2005	MediaMill Challenge	[230]	MediaMill	101	101
2006	Columbia374	[286]	LSCOM-Full	374	20
2006	VIREO374	[109]	LSCOM-Full	374	20
2008	CU-VIREO374	[110]	LSCOM-Full, LIG	374	20

In principle, the topic annotations listed in Table 4.7 form a valuable resource for concept-based video retrieval research, especially when used in concert with the various concept detection baselines. Surprisingly, however, baselines for video retrieval hardly exist currently. The one exception is the baseline by Huurnink et al. [100], who provide two repeatable experiments against which concept detector selection algorithms can be compared. Given the widespread availability of video data, concept annotations, concept detectors, and annotated search topics more baselines for video retrieval are to be expected in the near-future.

4.5 Results

As we observed in Table 4.3 many people have participated in various tasks of the TRECVID benchmark over the past years. We summarize all submitted and evaluated concept-based video retrieval results obtained within the TRECVID benchmark in Figure 4.2. As the results for all tasks depend on many variables, one should take care when interpreting these performance figures. Because the video data, tasks, and performance measures change over the years, comparing results across multiple editions of TRECVID is especially tricky. Being aware of this limitation, we nevertheless discuss the results, highlight the techniques that perform the best, and identify trends per task.

Concept detection task: The results on the concept detection task show a clear top-performer each year, mostly without positive or negative outliers. Therefore, we interpret the concept detection task as a healthy and competitive benchmark. All the best performing systems are generic and based on the complex architectures as discussed in Section 2.6. We further observe that in the first cycle of TRECVID, the best performing systems fused text and visual information together with supervised machine learning, which we attribute to the good quality of the speech recognition results for the US broadcast news data. In later cycles, the influence of textual features diminished in favor of visual-only approaches, as the speech recognition and machine translations of non-English video footage is susceptible to errors. In the third

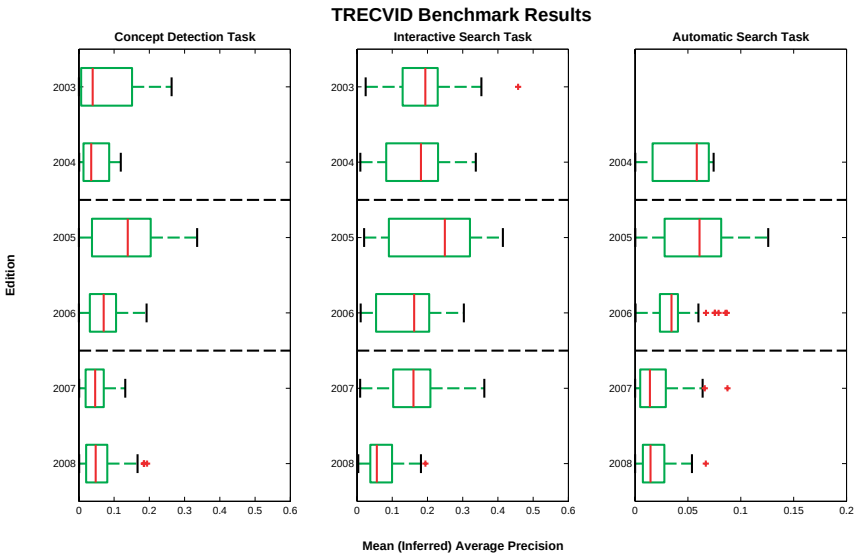


Fig. 4.2 Box plot overview of all submitted and evaluated TRECVID benchmark runs from Table 4.3 for the concept detection task, the interactive search task, and the automatic search task, grouped by video data set cycle (dashed lines). The vertical line on each box marks the median performer, the whiskers indicate the non-outlier minimum and maximum performers, respectively, and the positive outliers are marked individually (+). Note the change in performance scale for the automatic search task results.

cycle of TRECVID, the best performing concept detection systems rely increasingly on keypoint-based visual features and their fusion. We observe differences in terms of mean average precision between the best and median performer, which seems especially apparent in the first year of each cycle. We believe this indicates all participants profit from a positive learning effect in the second year of each cycle. Interestingly, the box plot for the concept detection task is skewed towards the left for the first cycle and to the right for the other cycles. We attribute this to the fact that in the first cycle of TRECVID only few participants detected semantic concepts in a generic fashion, while most participants in later cycles of the task adapted such an approach. Fueled by several annotation efforts, as listed in Table 4.6, generic concept detection really took off in 2005. Indeed, in terms of performance, 2005 appears to be the most interesting year of the task to date. After 2005, results seem to converge towards the median, indicating that systems are becoming

more and more similar. The positive outliers in 2008, employed their concept detection analysis on multiple frames per shot, indicating the potential of temporal visual analysis. We anticipate that fresh perspectives and innovative ideas will make the difference to stand out in the concept detection task in the future.

Interactive search task: The interactive search task results show a relatively wide variation in performance over all editions in all cycles. Apart from the difference in the number of available concept detectors, the text retrieval algorithms, the detector selection strategies, and the interfaces, this is probably also caused by the fact that some systems rely on experts for the retrieval task where others use novice users and less interaction time. The Informedia classic system obtained an interesting result in this task. Their many years of video retrieval expertise made them a positive outlier in 2003 [79]. Hence, their approach inspired many in later editions. A common denominator in the best performing interactive video search engines is the usage of multiple query methods, an increasing lexicon of concept detectors, advanced efficient video browsers, and retrieval by an expert user (see Sections 3.5 and 3.6). Similar to the concept detection task, 2005 was the most interesting year in terms of performance. We attribute this performance gain largely to the usage of many concept detectors at retrieval time. The trend in the results after 2005 indicates some convergence to the median, but sufficient differentiation between interactive systems still exists. We explain the positive outlier in 2008 by the fact that this system relied upon a very efficient active learning mechanism in addition to the "standard" components. We observe further that recent innovations in this area do not always comply with the task guidelines.

Automatic search task: Compared to both concept detection and interactive search, performance on the automatic search task is humble, e.g., compared to interactive search the performance is approximately a factor three lower for all editions. We attribute this to the general complexity of the problem and the increasingly difficult visually oriented search topics used over the years. Despite the visual orientation, text retrieval has played an important part in many TRECVID video

search systems. The best performing systems combine text retrieval with query classes, selection of detectors, and query combinations, as explained in Sections 3.3 and 3.4. Similar to both the concept detection and the interactive search task, overall performance was best in 2005. Indeed, the top performers included many concept detectors in their search strategy. While the absolute performance difference between the best and the median approach is only small, some positive outliers exist since 2006. Due to effective search strategy combinations the best performers in the most recent editions profited from very high results on specific topics, which had a positive effect on their mean average precision. In both 2007 and 2008, the median performance is skewed to the left. We attribute this to the lack of reliable speech transcripts and machine translations for the Dutch language, the restricted number of concept detectors that the community could learn on Sound and Vision data (due to lack of training examples), and the limited overlap between the few concept detectors and the search topics.

From a critical point of view, it is hard to get excited about concept-based video retrieval results when the median systems stand at a mean inferred average precision of 0.05, or lower, for all TRECVID benchmark tasks in 2008. Also, it seems that since 2005 there is hardly any visible progress in performance. However, as indicated earlier, care should be taken when interpreting these results. As the video data, tasks, concepts, search topics, system variables, and performance measures change over the years, comparing results across multiple editions of TRECVID is extremely hard, perhaps impossible even. It is interesting to note, however, that several participants have redone their old TRECVID submissions using new versions of their concept-based video search engines, and have reported substantial increases in performance, (see [171]). Moreover, in 2008, the best performing systems for all tasks were positive outliers in terms of their performance, indicating that after many years the benchmark still spurs innovation in concept-based video retrieval. There is, however, no reason to be overly positive about the results. We have made progress, but there is a long way to go before concept-based video retrieval is as accepted as text retrieval on the Internet.

4.6 Discussion

By defining a collaborative research agenda, TRECVID has nurtured many positive developments in concept-based video retrieval, especially with respect to the increasing availability of large-scale research resources. Indeed, many researchers have coined the benchmark priceless. It must be said, however, that there exists also a note of criticism, which mainly addresses the result-driven nature of TRECVID. It is presumably causing a convergence of approaches and, as such, killing innovation. In this respect, it is interesting to note that as long as new participants team up each year for a particular task, fresh and novel ideas are guaranteed to pop up. The uttered criticism should, nevertheless, be taken to heart, and exploited to adjust the benchmark tasks were needed.

An often heard criticism on the concept detection task is that the developed and evaluated methodologies experience difficulties when applied to other domains [288]. One of the possible reasons contributing to this behavior is the fact that the current video data sets used in TRECVID contain many repetitive pieces of video, like commercials, signature tunes, and similar studio settings. When a concept appears in such a repetitive video segment, detection boils down to (near) copy detection. Indeed, performance can be relatively good in terms of TRECVID evaluation criteria, but the detector will not generalize outside the time period and the particular video data set. To a large extent performance of a concept detector is related to the number of learning examples used at training time, see Figure 4.3. The good performing detectors have many examples and are expected to generalize to some extent. This is confirmed by experiments in [218], where detectors trained with many examples on international broadcast news generalized well to BBC stock video material, without any optimization. Note that the (few) good performing detectors, trained using a small set of examples, in Figure 4.3, are (near) copy detectors of concepts appearing in commercials. These will not generalize at all. For the large set of modest performing detectors, alternative solutions are needed. Apart from improving the (visual) analysis, as suggested in Section 2.7, possible solutions could be to adapt the

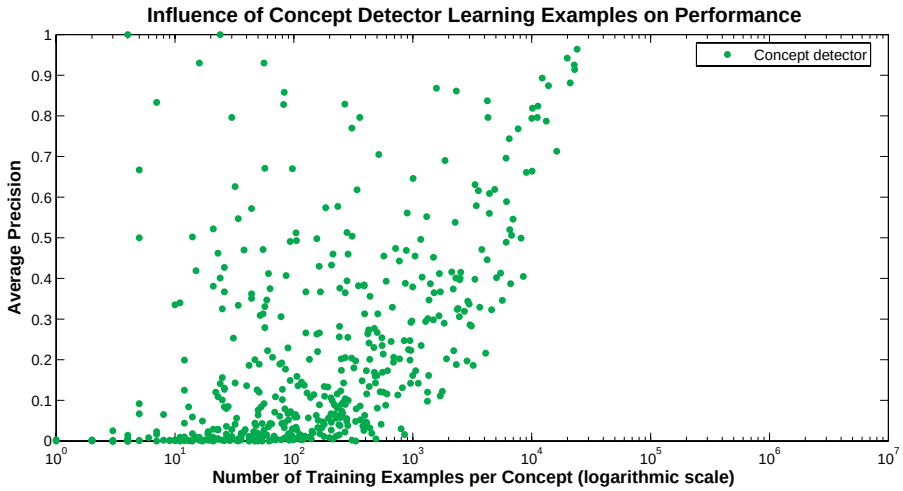


Fig. 4.3 Detection results [219] for a lexicon [163, 230] of approximately 500 concepts on TRECVID 2006 video data. Good performance is obtained for detectors of concepts with limited variability in their visual appearance, such as broadcast news specific concepts, and concept detectors which are trained on many examples like *face* and *outdoor*. Most concept detectors perform moderately because they have been trained on only a few examples.

concept detector to the domain of interest by adjusting the classifier [108, 289], or by exploiting massive sets of image examples obtained from the (social-tagged) web [131, 238, 260] in combination with semi-supervised learning [37, 303]. When concepts are domain-dependent, it is better to provide concept detectors with domain-specific examples, like was done for pop concerts [229] and visual lifelogs [29]. Broad domain applicability will remain an important topic for the years to come.

TRECVID's emphasis on retrieval performance in the interactive search task is not without criticism either [44]. Since there is a human in the loop, interactive video retrieval results also depend on the user interface and the expertise of the searcher. These important factors are, however, not evaluated directly in the benchmark. To address these factors to some extent, Snoek et al. [222] have initiated the VideOlympics: a real-time evaluation showcase of present day video search engines. Apart from having systems compete simultaneously on a common task, and having all systems solve the tasks on the same data set, the authors

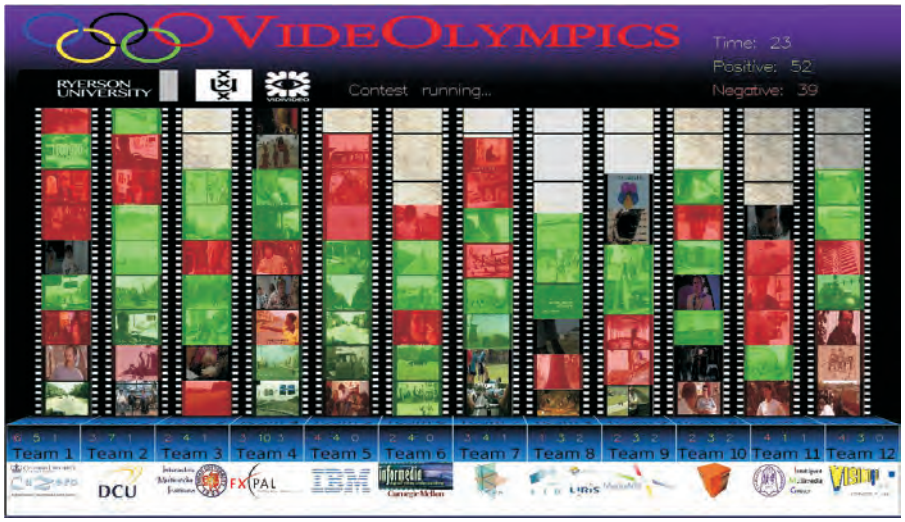


Fig. 4.4 Scoreboard used at the VideOlympics [222]. Retrieved results are displayed for individual teams as soon as they are submitted to the VideOlympics evaluation server. Based on available ground truth, immediate visual feedback is given to the audience whether retrieved shots are correct (green), wrong (red), or unknown (gray).

argue that audience involvement is achieved by communicating overall results in real-time using a scoreboard, see Figure 4.4. This allows for on the spot comparison of the performance of different prototype systems; and it also allows evaluating their interfaces and ease of use. Hence, it offers the audience a good perspective on the possibilities and limitations of current video search engines. The first two editions of the VideOlympics served as a demonstration of video search by expert users, starting in 2009 the event will also include search rounds with novice users to study the influence of user bias in video search engine comparisons. While the VideOlympics is a first effort, quantifying the human factor in interactive video search is an open scientific problem and unlikely to be resolved soon.

Criticism on the automatic search task has mainly considered the search topics and the evaluation measure used [83, 265]. As related to the search topics, it has been argued that they are overly complex, limited in number, and drifting away from real-world video retrieval scenarios. TRECVID's response by adding more, and less complex,

topics to the automatic search task is certainly a step in the right direction. Regarding the evaluation measure, the criticism mainly targets at the inclusion of recall, which causes the low overall performance numbers. In this respect, precision at n is perhaps a better evaluation measure for the automatic search task, resembling more closely real world usage. Apart from criticism related to the topics and the evaluation measures, we would like to add the difficulty to replicate automatic search experiments. Increasingly, the best performing automatic search systems rely on many information sources, analysis components, and their associated parameters. It is therefore, hard to quantify what factors affect performance most, and as such limiting progress in this area. The availability of baseline systems, similar in spirit as the ones for concept detection, could further foster research in automatic search.

For all tasks, the benefits of TRECVID clearly outweigh the criticism. We even believe it is fair to state that research in concept-based video retrieval has been boosted by the evaluation efforts carried out in TRECVID.

5

Conclusions

We have presented a review on concept-based video retrieval covering more than 300 references. In striving to bridge the semantic gap, we structured our review by laying down the anatomy of a concept-based video search engine. Obviously, the research in concept-based video retrieval has not reached its end yet. On the contrary, the community is only beginning to understand the full potential of this novel paradigm. At the end of this review, we would like to stress that concept-based video retrieval is an exciting interdisciplinary research area with many open research questions longing for an adequate answer.

Thanks to influences from computer vision and machine learning, we have witnessed the transition from specific to generic concept detectors in just 15 years. It took even less time to embed large sets of uncertain concept detectors into well-developed document retrieval frameworks. When we add the latest insights from human–computer interaction to the mix, bridging the semantic gap completely comes within reach, albeit in an interactive manner. The opportunities offered by the heritage of these founding fields have been taken with eagerness. We expect further tangential progress along the lines of these disciplines in the near-future. In particular, we anticipate improvements from temporal

visual analysis methods and their embedding in appropriate machine learning frameworks, learning meaningful concept detector combinations at query time, and the seamless integration of active learning with visualization. We do note, however, that given the large volume of data, all these factors will be unsuccessful if they ignore efficiency. The video data deluge makes efficient execution of analysis algorithms paramount, while simultaneously the impatient interacting user is unwilling to tolerate slow response times. High-performance computing is only one part of the envisaged solution; the other part must stem from clever algorithms that maintain a robust level of performance while being efficient in their execution.

All the sketched improvements will have a positive impact on concept-based video retrieval, but the most dominant element in this video retrieval paradigm remains the availability of a large lexicon of robust concept detectors. Based on some arguable assumptions this lexicon was recently estimated to need about 5000 concepts [85], but 17,000 is likely to be a better estimate as this resembles the vocabulary-size of an average educated native English speaker [70]. For the non-visual semantic concepts in this vocabulary, reasoning with ontologies remains an appealing scenario that needs to prove itself in the future. For the visual concepts, scaling up the number of robust detectors is the biggest challenge ahead. This will only be possible if the fundamental problem in concept detection based on supervised machine learning is resolved: the lack of a large and diverse set of manually labeled visual examples to model the diversity in appearance adequately. A new direction in tackling this fundamental problem is employing user tagged visual data provided by online services. These annotations are less accurate than the current practice in concept-based video retrieval, but the amount of training samples is several orders of magnitude larger. We believe, the decisive leap forward in concept-based video retrieval will stem from an increase in the amount of training data of at least two orders of magnitude beyond current practice, see also Figure 4.3.

Progress in concept-based video retrieval would not have been possible without the NIST TRECVID benchmark. Due to their definition of several challenging tasks, their independent evaluation protocol, and their world wide adaptation by the research community, the evaluation

campaign continues to be important in shaping the field. So far, the benchmark has emphasized analysis of professional video content from broadcast video archives. Video, however, is no longer the sole domain of broadcast professionals. Other domains, such as consumer, medical, surveillance, and intelligence are equally important. To realize powerful concept-based video search, domain-specific user needs and application requirements need to be consolidated. From an information retrieval perspective, we consider particularly the online domain of importance. A treasure of professional and consumer video data alike is waiting to be found on the web. The research community cannot afford to neglect this wealth of online information, with all its broad domain and real use benchmark challenges and opportunities.

In conclusion, among the many open research challenges and opportunities, one of the most interesting roads ahead for the video retrieval community drives to the Internet. To reach this destination, the community needs to realize the transition from retrieval solutions for narrow domain broadcast video to broad domain narrowcast video. For this transition to be effective, the latest insights from information retrieval, computer vision, machine learning, and human–computer interaction need to be considered jointly in an efficient fashion. In particular, we need to harness freely available images and videos labeled by the masses, using them to create a more complete and accurate lexicon of concept detectors. When we can realize the transition towards this online ecosystem, for concept-based video retrieval the best is yet to come.

Acknowledgments

The authors are grateful to Arnold Smeulders from the University of Amsterdam for insightful discussions and to Paul Over from NIST for TRECVID benchmark information. This research is supported by the Dutch Technology Foundation STW, the Dutch BSIK MultimediaN project, the Dutch/Flemish IM-PACT program, the European IST-CHORUS project, and the European VIDI-Video project.

References

[1] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. P. Natsev, M. R. Naphade, C. Neti, H. J. Nock, H. H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, T. V. Ashwin, and D. Zhang, "IBM research TREC-2002 video retrieval system," in *Proceedings of the 11th Text Retrieval Conference*, Gaithersburg, USA, 2002.

[2] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 170–185, 2003.

[3] J. Adcock, M. Cooper, and F. Chen, "FXPAL MediaMagic video search system," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 644–644, Amsterdam, The Netherlands, 2007.

[4] J. Adcock, M. Cooper, and J. Pickens, "Experiments in interactive video search by addition and subtraction," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 465–474, Niagara Falls, Canada, 2008.

[5] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, pp. 832–843, 1983.

[6] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. R. Naphade, A. P. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang, "IBM research TRECVID-2003 video retrieval system," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[7] S. Ayache and G. Quénot, "Evaluation of active learning strategies for video indexing," *Image Communication*, vol. 22, nos. 7–8, pp. 692–704, 2007.

[8] S. Ayache and G. Quénot, "Video corpus annotation using active learning," in *European Conference on Information Retrieval*, pp. 187–198, Glasgow, UK, 2008.

[9] S. Ayache, G. Quénot, and J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," in *European Conference on Information Retrieval*, pp. 494–504, Rome, Italy, 2007.

[10] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Transactions on Multimedia*, vol. 4, pp. 68–75, 2002.

[11] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted American football video by highlight selection," *IEEE Transactions on Multimedia*, vol. 6, pp. 575–586, 2004.

[12] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *International Joint Conference on Artificial Intelligence*, pp. 805–810, Acapulco, Mexico, 2003.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008.

[14] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation," in *Proceedings of SPIE Conference on Internet Multimedia Management Systems*, Boston, USA, 2000.

[15] M. Bertini, A. D. Bimbo, and C. Torniai, "Automatic video annotation using ontologies extended with visual information," in *Proceedings of the ACM International Conference on Multimedia*, pp. 395–398, Singapore, 2005.

[16] A. D. Bimbo, *Visual Information Retrieval*. Morgan Kaufmann, 1999.

[17] A. D. Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 121–132, 1997.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, 2006.

[19] H. M. Blanken, A. P. de Vries, H. E. Blok, and L. Feng, eds., *Multimedia Retrieval*. Springer, 2007.

[20] T. Bompada, C.-C. Chang, J. Chen, R. Kumar, and R. Shenoy, "On the robustness of relevance measures with incomplete judgments," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 359–366, Amsterdam, The Netherlands, 2007.

[21] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 55–73, 1990.

[22] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck-Jones, and S. J. Young, "Automatic content-based retrieval of broadcast news," in *Proceedings of the ACM International Conference on Multimedia*, San Francisco, USA, 1995.

[23] R. Brunelli, O. Mich, and C. M. Modena, "A survey on the automatic indexing of video data," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 78–112, 1999.

[24] E. Bruno, N. Moenne-Loccoz, and S. Marchand-Maillet, "Design of multimodal dissimilarity spaces for retrieval of multimedia documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1520–1533, 2008.

[25] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 25–32, Sheffield, UK, 2004.

[26] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, pp. 13–47, 2006.

[27] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[28] G. J. Burghouts and J.-M. Geusebroek, "Performance evaluation of local color ivariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

[29] D. Byrne, A. R. Doherty, C. G. M. Snoek, G. J. F. Jones, and A. F. Smeaton, "Validating the detection of everyday concepts in visual lifelogs," in *Proceedings International Conference on Semantics and Digital Media Technologies*, pp. 15–30, Berlin, Germany: Springer-Verlag, 2008.

[30] M. Campbell, A. Haubold, S. Ebadollahi, D. Joshi, M. R. Naphade, A. P. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tešić, and L. Xie, "IBM research TRECVID-2006 video retrieval system," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[31] J. Cao, Y. Lan, J. Li, Q. Li, X. Li, F. Lin, X. Liu, L. Luo, W. Peng, D. Wang, H. Wang, Z. Wang, Z. Xiang, J. Yuan, W. Zheng, B. Zhang, J. Zhang, L. Zhang, and X. Zhang, "Intelligent multimedia group of Tsinghua University at TRECVID 2006," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[32] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1026–1038, 2002.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm/, 2001.

[34] S.-F. Chang, W. Chen, H. J. Men, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatio-temporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 602–615, 1998.

[35] S.-F. Chang, J. He, Y.-G. Jiang, E. E. Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky, "Columbia University/VIREO-CityU/IRIT TRECVID-2008 high-level feature extraction and interactive video search," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[36] S.-F. Chang, W. Hsu, W. Jiang, L. S. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, "Columbia University TRECVID-2006 video search and high-level feature extraction," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[37] O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. Cambridge, USA: The MIT Press, 2006.

[38] M.-Y. Chen, M. G. Christel, A. G. Hauptmann, and H. Wactlar, "Putting active learning into multimedia applications: Dynamic definition and refinement of concept classifiers," in *Proceedings of the ACM International Conference on Multimedia*, pp. 902–911, Singapore, 2005.

[39] M.-Y. Chen and A. G. Hauptmann, "Multi-modal classification in digital news libraries," in *Proceedings of the Joint Conference on Digital Libraries*, pp. 212–213, Tucson, USA, 2004.

[40] M. G. Christel and R. M. Conescu, "Mining novice user activity with TRECVID interactive retrieval tasks," in *CIVR*, pp. 21–30, Springer-Verlag, 2006.

[41] M. G. Christel and A. G. Hauptmann, "The use and utility of high-level semantic features," in *CIVR*, pp. 134–144, Springer-Verlag, 2005.

[42] M. G. Christel, A. G. Hauptmann, H. D. Wactlar, and T. D. Ng, "Collages as dynamic summaries for news video," in *Proceedings of the ACM International Conference on Multimedia*, pp. 561–569, Juan-les-Pins, France, 2002.

[43] M. G. Christel, C. Huang, N. Moraveji, and N. Papernick, "Exploiting multiple modalities for interactive video retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1032–1035, Montreal, Canada, 2004.

[44] T.-S. Chua, "Towards the next plateau — Innovative multimedia research beyond TRECVID," in *Proceedings of the ACM International Conference on Multimedia*, Augsburg, Germany, 2007.

[45] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives — Techniques, experience and trends," in *Proceedings of the ACM International Conference on Multimedia*, pp. 656–659, New York, USA, 2004.

[46] T.-S. Chua et al., "TRECVID-2004 search and feature extraction task by NUS PRIS," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2004.

[47] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1210–1224, 2005.

[48] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences and trends of the new age," *ACM Computing Surveys*, vol. 40, pp. 1–60, 2008.

[49] G. Davenport, T. G. A. Smith, and N. Pincever, "Cinematic principles for multimedia," *IEEE Computer Graphics & Applications*, vol. 11, pp. 67–74, 1991.

[50] M. Davis, "Editing out video editing," *IEEE MultiMedia*, vol. 10, pp. 54–64, 2003.

[51] F. M. G. de Jong, J. L. Gauvain, J. den Hartog, and K. Netter, "OLIVE: Speech-based video retrieval," in *European Workshop on Content-Based Multimedia Indexing*, Toulouse, France, 1999.

[52] O. de Rooij, C. G. M. Snoek, and M. Worring, "Query on demand video browsing," in *Proceedings of the ACM International Conference on Multimedia*, pp. 811–814, Augsburg, Germany, 2007.

[53] O. de Rooij, C. G. M. Snoek, and M. Worring, "Balancing thread based navigation for targeted video search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 485–494, Niagara Falls, Canada, 2008.

[54] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 800–810, 2001.

[55] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 881–884, Toronto, Canada, 2006.

[56] P. Enser, "Visual image retrieval: Seeking the alliance of concept-based and content-based paradigms," *Journal of Information Science*, vol. 26, pp. 199–210, 2000.

[57] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref, and L. Wu, "ClassView: Hierarchical video shot classification, indexing and accessing," *IEEE Transactions on Multimedia*, vol. 6, pp. 70–86, 2004.

[58] J. Fan, H. Luo, Y. Gao, and R. Jain, "Incorporating concept ontology for hierarchical video classification, annotation and visualization," *IEEE Transactions on Multimedia*, vol. 9, pp. 939–957, 2007.

[59] C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*. Cambridge, USA: The MIT Press, 1998.

[60] J. M. Ferryman, ed., *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. Rio de Janeiro, Brazil: IEEE Press, 2007.

[61] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, vol. 28, pp. 23–32, 1995.

[62] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, nos. 1–2, pp. 89–108, 2002.

[63] J.-M. Geusebroek, R. Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1338–1350, 2001.

[64] J.-M. Geusebroek and A. W. M. Smeulders, "A six-stimulus theory for stochastic texture," *International Journal of Computer Vision*, vol. 62, nos. 1–2, pp. 7–16, 2005.

[65] T. Gevers, "Adaptive image segmentation by combining photometric invariant region and edge information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 848–852, 2002.

[66] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, pp. 453–464, 1999.

[67] T. Gevers and A. W. M. Smeulders, "PicToSeek: Combining color and shape invariant features for image retrieval," *IEEE Transactions on Image Processing*, vol. 9, pp. 102–119, 2000.

[68] K.-S. Goh, E. Y. Chang, and W.-C. Lai, "Multimodal concept-dependent active learning for image retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 564–571, New York, USA, 2004.

[69] S. A. Golder and B. A. Huberman, "The structure of collaborative tagging systems," *Journal of Information Science*, vol. 32, pp. 198–208, 2006.

[70] R. Goulden, P. Nation, and J. Read, "How large can a receptive vocabulary be?," *Applied Linguistics*, vol. 11, pp. 341–363, 1990.

[71] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu, "Multi-layer multi-instance learning for video concept detection," *IEEE Transactions on Multimedia*, vol. 10, pp. 1605–1616, 2008.

[72] A. Gupta and R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, pp. 70–79, 1997.

[73] M. Guy and E. Tonkin, "Folksonomies: Tidying up tags?," *D-Lib Magazine*, vol. 12, Available at: http://www.dlib.org/dlib/january06/ guy/01guy.html, 2006.

[74] A. Hanjalic, *Content-Based Analysis of Digital Video*. Boston, USA: Kluwer Academic Publishers, 2004.

[75] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?," *Proceedings of the IEEE*, vol. 96, pp. 541–547, 2008.

[76] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, pp. 143–154, 2005.

[77] A. Haubold and J. R. Kender, "VAST MM: Multimedia browser for presentation video," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 41–48, Amsterdam, The Netherlands, 2007.

[78] A. Haubold and A. P. Natsev, "Web-based information content and its application to concept-based video retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 437–446, Niagara Falls, Canada, 2008.

[79] A. G. Hauptmann, R. V. Baron, M.-Y. Chen, M. G. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C. G. M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. D. Wactlar, "Informedia at TRECVID-2003: Analyzing and searching broadcast news video," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[80] A. G. Hauptmann and S.-F. Chang, "LIBSCOM: Large analytics library and scalable concept ontology for multimedia research," http://www.libscom.org/, 2009.

[81] A. G. Hauptmann and M. G. Christel, "Successful approaches in the TREC video retrieval evaluations," in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, 2004.

[82] A. G. Hauptmann, M. G. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proceedings of the IEEE*, vol. 96, pp. 602–622, 2008.

[83] A. G. Hauptmann and W.-H. Lin, "Assessing effectiveness in video retrieval," in *CIVR*, pp. 215–225, Springer-Verlag, 2005.

[84] A. G. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen, "Extreme video retrieval: Joint maximization of human and computer performance," in

Proceedings of the ACM International Conference on Multimedia, pp. 385–394, Santa Barbara, USA, 2006.

[85] A. G. Hauptmann, R. Yan, W.-H. Lin, M. G. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Transactions on Multimedia*, vol. 9, pp. 958–966, 2007.

[86] D. Heesch, "A survey of browsing models for content based image retrieval," *Multimedia Tools and Applications*, vol. 40, pp. 261–284, 2008.

[87] D. Heesch and S. Rüger, "Image browsing: A semantic analysis of NN^k networks," in *CIVR*, pp. 609–618, Springer-Verlag, 2005.

[88] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, 1994.

[89] M. A. Hoang, J.-M. Geusebroek, and A. W. M. Smeulders, "Color texture measurement and segmentation," *Signal Processing*, vol. 85, pp. 265–275, 2005.

[90] L. Hollink, M. Worring, and G. Schreiber, "Building a visual ontology for video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 479–482, Singapore, 2005.

[91] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, "Video content annotation using visual analysis and a large semantic knowledgebase," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 327–334, Madison, USA, 2003.

[92] R. Houghton, "Named faces: Putting names to faces," *IEEE Intelligent Systems*, vol. 14, pp. 45–50, 1999.

[93] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Reranking methods for visual search," *IEEE MultiMedia*, vol. 14, pp. 14–22, 2007.

[94] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, pp. 179–187, 1962.

[95] J. Huang, S. R. Kumar, W.-J. Z. M. Mitra, and R. Zabih, "Color-spatial indexing and applications," *International Journal of Computer Vision*, vol. 35, pp. 245–268, 1999.

[96] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Active learning for interactive multimedia retrieval," *Proceedings of the IEEE*, vol. 96, pp. 648–667, 2008.

[97] M. Huijbregts, R. Ordelman, and F. M. G. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proceedings International Conference on Semantics and Digital Media Technologies*, pp. 78–90, Berlin: Springer-Verlag, 2007.

[98] W. Hürst, "Video browsing on handheld devices — Interface designs for the next generation of mobile video players," *IEEE MultiMedia*, vol. 15, pp. 76–83, 2008.

[99] B. Huurnink and M. de Rijke, "Exploiting redundancy in cross-channel video retrieval," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 177–186, Augsburg, Germany, 2007.

[100] B. Huurnink, K. Hofmann, and M. de Rijke, "Assessing concept selection for video retrieval," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 459–466, Vancouver, Canada, 2008.

[101] E. Hyvönen, S. Saarela, A. Styrman, and K. Viljanen, "Ontology-based image retrieval," in *Proceedings of the International World Wide Web Conference*, Budapest, Hungary, 2003.

[102] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. P. Khudanpur, D. Klakow, M. R. Krause, R. Manmatha, H. J. Nock, D. Petkova, B. Pytlik, and P. Virga, "Joint visual-text modeling for automatic retrieval of multimedia documents," in *Proceedings of the ACM International Conference on Multimedia*, pp. 21–30, Singapore, 2005.

[103] 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.

[104] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern Recognition*, vol. 24, pp. 1167–1186, 1991.

[105] A. K. Jain and A. Vailaya, "Shape-based retrieval: A case study with trademark image databases," *Pattern Recognition*, vol. 31, pp. 1369–1390, 1998.

[106] R. Jain and A. Hampapur, "Metadata in video databases," *ACM SIGMOD Record*, vol. 23, pp. 27–33, 1994.

[107] W. Jiang, S.-F. Chang, and A. C. Loui, "Context-based concept fusion with boosted conditional random fields," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 949–952, Honolulu, USA, 2007.

[108] W. Jiang, E. Zavesky, S.-F. Chang, and A. C. Loui, "Cross-domain learning methods for high-level visual concept classification," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 161–164, San Diego, USA, 2008.

[109] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 494–501, Amsterdam, The Netherlands, 2007.

[110] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, "CU-VIREO374: Fusing Columbia374 and VIREO374 for large scale semantic concept detection," Technical Report 223-2008-1, Columbia University ADVENT Technical Report, 2008.

[111] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning*, (B. Schölkopf, C. Burges, and A. J. Smola, eds.), pp. 169–184, Cambridge, USA: The MIT Press, 1999.

[112] M. Kankanhalli and Y. Rui, "Application potential of multimedia information retrieval," *Proceedings of the IEEE*, vol. 96, pp. 712–720, 2008.

[113] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text and vehicle detection and tracking in video: Data, metrics and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 319–336, 2009.

[114] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database — Query by visual example," in *Proceedings of the International Conference on Pattern Recognition*, pp. 530–533, The Hague, The Netherlands, 1992.

[115] J. R. Kender and M. R. Naphade, "Visual concepts for news story tracking: analyzing and exploiting the NIST TRECVID video annotation experiment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1174–1181, Washington, DC, USA, 2005.

[116] L. S. Kennedy, "Revision of LSCOM event/activity annotations," Technical Report 221-2006-7, Columbia University ADVENT Technical Report, 2006.

[117] L. S. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 333–340, Amsterdam, The Netherlands, 2007.

[118] L. S. Kennedy, S.-F. Chang, and A. P. Natsev, "Query-adaptive fusion for multimodal search," *Proceedings of the IEEE*, vol. 96, pp. 567–588, 2008.

[119] L. S. Kennedy, A. P. Natsev, and S.-F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proceedings of the ACM International Conference on Multimedia*, pp. 882–891, Singapore, 2005.

[120] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 489–497, 1990.

[121] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[122] M. Larson, E. Newman, and G. Jones, "Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content," in *Working Notes for the Cross-Language Evaluation Forum Workshop*, Aarhus, Denmark, 2008.

[123] L. J. Latecki, R. Lakaemper, and U. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 424–429, Hilton Head Island, USA, 2000.

[124] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, New York, USA, 2006.

[125] H. Lee and A. F. Smeaton, "Designing the user-interface for the Físchlár digital video library," *Journal of Digital Information*, vol. 2, 2002.

[126] D. Lenat and R. Guha, eds., *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project*. Reading, USA: Addison-Wesley, 1990.

[127] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the International Conference on Systems Documentation*, pp. 24–26, Toronto, Canada, 1986.

[128] M. S. Lew, ed., *Principles of Visual Information Retrieval*. Springer, 2001.

[129] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on*

Multimedia Computing, Communications and Applications, vol. 2, pp. 1–19, 2006.

[130] J. Li, W. Wu, T. Wang, and Y. Zhang, "One step beyond histograms: Image representation using markov stationary features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[131] X. Li, C. G. M. Snoek, and M. Worring, "Annotating images by harnessing worldwide user-tagged photos," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009.

[132] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 603–610, Amsterdam, The Netherlands, 2007.

[133] C.-Y. Lin, B. L. Tseng, and J. R. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[134] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, pp. 267–276, 2007.

[135] W.-H. Lin and A. G. Hauptmann, "News video classification using SVM-based multimodal classifiers and combination strategies," in *Proceedings of the ACM International Conference on Multimedia*, Juan-les-Pins, France, 2002.

[136] H. Liu and P. Singh, "ConceptNet: A practical commonsense reasoning toolkit," *BT Technology Journal*, vol. 22, pp. 211–226, 2004.

[137] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li, "Video search re-ranking via multi-graph propagation," in *Proceedings of the ACM International Conference on Multimedia*, pp. 208–217, Augsburg, Germany, 2007.

[138] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Transactions on Multimedia*, vol. 10, pp. 240–251, 2008.

[139] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[140] G. Lu, "Indexing and retrieval of audio: A survey," *Multimedia Tools and Applications*, vol. 15, pp. 269–290, 2001.

[141] H.-B. Luan, S.-Y. Neo, H.-K. Goh, Y.-D. Zhang, S.-X. Lin, and T.-S. Chua, "Segregated feedback with performance-based adaptive sampling for interactive news video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 293–296, Augsburg, Germany, 2007.

[142] H. Luo, J. Fan, J. Yang, W. Ribarsky, and S. Satoh, "Analyzing large-scale news video databases to support knowledge visualization and intuitive retrieval," in *IEEE Symposium on Visual Analytics Science and Technology*, pp. 107–114, Sacramento, USA, 2007.

[143] W.-Y. Ma and B. S. Manjunath, "NeTra: A toolbox for navigating large image databases," *Multimedia Systems*, vol. 7, pp. 184–198, 1999.

[144] J. Magalhães and S. Rüger, "Information-theoretic semantic multimedia indexing," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 619–626, Amsterdam, The Netherlands, 2007.

[145] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 836–842, 1996.

[146] B. S. Manjunath, P. Salembier, and T. Sikora, eds., *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[147] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[148] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proceedings ACM International Conference on Hypertext and Hypermedia*, pp. 31–40, Odense, Denmark, 2006.

[149] K. K. Matusiak, "Towards user-centered indexing in digital image collections," *OCLC Systems & Services*, vol. 22, pp. 263–296, 2006.

[150] K. McDonald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *CIVR*, (W.-K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, eds.), pp. 61–70, Heidelberg, Germany: Springer, 2005.

[151] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z. Zha, Y. Liu, Z. Gu, G. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu, "MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2007.

[152] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Computer Vision and Image Understanding*, vol. 94, nos. 1–3, pp. 3–27, 2004.

[153] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, pp. 121–143, 2008.

[154] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1632–1646, 2008.

[155] F. Nack and W. Putz, "Saying what it means: Semi-automated (news) media annotation," *Multimedia Tools and Applications*, vol. 22, pp. 263–302, 2004.

[156] M. R. Naphade, "On supervision and statistical learning for semantic multimedia analysis," *Journal of Visual Communication and Image Representation*, vol. 15, pp. 348–369, 2004.

[157] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering and retrieval," *IEEE Transactions on Multimedia*, vol. 3, pp. 141–151, 2001.

[158] M. R. Naphade and T. S. Huang, "Extracting semantics from audiovisual content: The final frontier in multimedia retrieval," *IEEE Transactions on Neural Networks*, vol. 13, pp. 793–810, 2002.

[159] M. R. Naphade, L. S. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. G. Hauptmann, "A light scale concept ontology for multimedia understanding for TRECVID 2005," Technical Report RC23612, IBM T. J. Watson Research Center, 2005.

[160] M. R. Naphade, I. V. Kozintsev, and T. S. Huang, "A factor graph framework for semantic video indexing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 40–52, 2002.

[161] M. R. Naphade, A. P. Natsev, C.-Y. Lin, and J. R. Smith, "Multi-granular detection of regional semantic concepts," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 109–112, Taipei, Taiwan, 2004.

[162] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, 2004.

[163] M. R. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, pp. 86–91, 2006.

[164] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 991–1000, Augsburg, Germany, 2007.

[165] A. P. Natsev, M. R. Naphade, and J. Tešić, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proceedings of the ACM International Conference on Multimedia*, pp. 598–607, Singapore, 2005.

[166] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *CIVR*, (H. Sundaram et al., eds.), pp. 143–152, Heidelberg, Germany: Springer-Verlag, 2006.

[167] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin, "ETISEO, performance evaluation for video surveillance systems," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 476–481, London, UK, 2007.

[168] G. P. Nguyen and M. Worring, "Interactive access to large image collections using similarity-based visualization," *Journal of Visual Languages and Computing*, vol. 19, pp. 203–224, 2008.

[169] G. P. Nguyen, M. Worring, and A. W. M. Smeulders, "Interactive search by direct manipulation of dissimilarity space," *IEEE Transactions on Multimedia*, vol. 9, pp. 1404–1415, 2007.

[170] H. T. Nguyen, M. Worring, and A. Dev, "Detection of moving objects in video using a robust motion similarity measure," *IEEE Transactions on Image Processing*, vol. 9, pp. 137–141, 2000.

[171] NIST, "TRECVID video retrieval evaluation — Online proceedings," http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, 2001–2008.

[172] T. O'Reily, "What is web 2.0," http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web%-20.html, 2005.

[173] P. Over, G. Awad, T. Rose, J. Fiscus, W. Kraaij, and A. F. Smeaton, "TRECVID 2008 — Goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[174] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton, "TRECVID 2005 An Overview," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2005.

[175] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton, "TRECVID 2006 an overview," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[176] P. Over, A. F. Smeaton, and P. Kelly, "The TRECVID 2007 BBC rushes summarization evaluation pilot," in *Proceedings of the International Workshop on TRECVID Video Summarization*, pp. 1–15, 2007.

[177] S. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, USA: The MIT Press, 1999.

[178] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the ACM International Conference on Multimedia*, pp. 65–74, Boston, USA, 1996.

[179] Z. Pecenovic, M. Do, M. Vetterli, and P. Pu, "Integrated browsing and searching of large image collections," in *Proceedings of the International Conference on Advances in Visual Information Systems*, Lyon, France, 2000.

[180] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, pp. 233–254, 1996.

[181] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2004.

[182] E. G. M. Petrakis, A. Diplaros, and E. Milios, "Matching and retrieval of distorted and occluded shapes using dynamic programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1501–1516, 2002.

[183] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, Y. Kompatsiaris, and S. Staab, "Knowledge representation and semantic annotation of multimedia content," *IEE Proceedings of Vision, Image and Signal Processing*, vol. 153, pp. 255–262, 2006.

[184] J. C. Platt, "Probabilities for SV machines," in *Advances in Large Margin Classifiers*, (A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, eds.), pp. 61–74, Cambridge, USA: The MIT Press, 2000.

[185] E. Pogalin, A. W. M. Smeulders, and A. H. C. Thean, "Visual Quasi-Periodicity," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[186] G. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang, "Correlative multilabel video annotation with temporal kernels," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 5, 2009.

[187] G. M. Quénot, D. Moraru, L. Besacier, and P. Mulhem, "CLIPS at TREC-11: Experiments in video retrieval," in *Proceedings of the 11th Text Retrieval Conference*, (E. M. Voorhees and L. P. Buckland, eds.), Gaithersburg, USA, 2002.

[188] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[189] T. Randen and J. H. Husøy, "Filtering for texture classification: A comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 291–310, 1999.

[190] N. Rasiwasia, P. L. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, pp. 923–938, 2007.

[191] M. Rautiainen, T. Ojala, and T. Seppanen, "Cluster-temporal browsing of large news video databases," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[192] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, pp. 115–118, Trento, Italy, 2008.

[193] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *International Joint Conference on Artificial Intelligence*, pp. 448–453, Montréal, Canada, 1995.

[194] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," in *Proceedings of the Text Retrieval Conference*, pp. 73–96, Gaithersburg, USA, 1996.

[195] A. Rosenfeld, "Picture processing by computer," *ACM Computing Surveys*, vol. 1, pp. 147–176, 1969.

[196] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[197] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 644–655, 1998.

[198] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[199] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, USA: McGraw-Hill, 1983.

[200] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE MultiMedia*, vol. 6, pp. 22–35, 1999.

[201] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga, "Ontology-based photo annotation," *IEEE Intelligent Systems*, vol. 16, pp. 66–74, 2001.

[202] N. Sebe and M. S. Lew, "Texture features for content-based retrieval," in *Principles of Visual Information Retrieval*, (M. S. Lew, ed.), pp. 51–86, Springer, 2001.

[203] F. J. Seinstra, J.-M. Geusebroek, D. Koelma, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "High-performance distributed image and video content analysis with parallel-horus," *IEEE MultiMedia*, vol. 14, pp. 64–75, 2007.

[204] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu, "Watch what I watch: Using community activity to understand content," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 275–284, Augsburg, Germany, 2007.

[205] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of the IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition, pp. 1–8, Anchorage, USA, 2008.

[206] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *International Journal of Computer Vision*, vol. 67, pp. 189–210, 2006.

[207] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," *Proceedings of the IEEE*, vol. 96, pp. 548–566, 2008.

[208] A. F. Smeaton, "Large scale evaluations of multimedia information retrieval: The TRECVid experience," in *CIVR*, pp. 19–27, Springer-Verlag, 2005.

[209] A. F. Smeaton, "Techniques used and open challenges to the analysis, indexing and retrieval of digital video," *Information Systems*, vol. 32, pp. 545–559, 2007.

[210] A. F. Smeaton, C. Foley, D. Byrne, and G. J. F. Jones, "iBingo mobile collaborative search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 547–548, Niagara Falls, Canada, 2008.

[211] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.

[212] A. F. Smeaton, P. Over, and W. Kraaij, "High level feature detection from video in TRECVid: A 5-year retrospective of achievements," in *Multimedia Content Analysis, Theory and Applications*, (A. Divakaran, ed.), Springer, 2008.

[213] 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.

[214] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE MultiMedia*, vol. 4, pp. 12–20, 1997.

[215] J. R. Smith, M. R. Naphade, and A. P. Natsev, "Multimedia semantic indexing using model vectors," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 445–448, Baltimore, USA, 2003.

[216] J. R. Smith, A. P. Natsev, J. Tešić, L. Xie, and R. Yan, "IBM multimedia analysis and retrieval system," http://www.alphaworks.ibm.com/ tech/imars/, 2008.

[217] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Transactions on Multimedia*, vol. 9, pp. 975–986, 2007.

[218] C. G. M. Snoek, J. C. van Gemert, J.-M. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. de Rooij, F. J. Seinstra, A. W. M. Smeulders, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2005 semantic video search engine," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2005.

[219] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2006 semantic video search engine," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[220] C. G. M. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Transactions on Multimedia*, vol. 7, pp. 638–647, 2005.

[221] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, pp. 5–35, 2005.

[222] C. G. M. Snoek, M. Worring, O. de Rooij, K. E. A. van de Sande, R. Yan, and A. G. Hauptmann, "VideOlympics: Real-time evaluation of multimedia retrieval systems," *IEEE MultiMedia*, vol. 15, pp. 86–91, 2008.

[223] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, and F. J. Seinstra, "On the surplus value of semantic video analysis beyond the key frame," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Amsterdam, The Netherlands, 2005.

[224] 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.

[225] C. G. M. Snoek, M. Worring, and A. G. Hauptmann, "Detection of TV news monologues by style analysis," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[226] C. G. M. Snoek, M. Worring, and A. G. Hauptmann, "Learning rich semantics from news video archives by style analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, pp. 91–108, 2006.

[227] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Transactions on Multimedia*, vol. 9, pp. 280–292, 2007.

[228] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the ACM International Conference on Multimedia*, pp. 399–402, Singapore, 2005.

[229] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, and B. Freiburg, "The role of visual content and style for concert video indexing," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 252–255, Beijing, China, 2007.

[230] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proceedings of the ACM International Conference on Multimedia*, pp. 421–430, Santa Barbara, USA, 2006.

[231] C. G. M. Snoek et al., "The MediaMill TRECVID 2008 semantic video search engine," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[232] M. J. Swain and D. H. Ballard, "Color Indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[233] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *IEEE International Workshop on Content-based Access of Image and Video Databases, in Conjunction with ICCV'98*, Bombay, India, 1998.

[234] J. Tague-Sutcliffe, "The pragmatics of information retrieval experimentation, revisited," *Information Processing & Management*, vol. 28, pp. 467–490, 1992.

[235] S. Tang et al., "TRECVID 2008 high-level feature extraction By MCG-ICT-CAS," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[236] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. A. Bouman, and E. J. Delp, "ViBE: A compressed video database structured for active browsing

and search," *IEEE Transactions on Multimedia*, vol. 6, pp. 103–118, 2004.

[237] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki, "Structured video computing," *IEEE MultiMedia*, vol. 1, pp. 34–43, 1994.

[238] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1958–1970, 2008.

[239] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, 2007.

[240] B. L. Tseng, C.-Y. Lin, M. R. Naphade, A. P. Natsev, and J. R. Smith, "Normalized classifier fusion for semantic visual concept detection," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 535–538, Barcelona, Spain, 2003.

[241] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, pp. 177–280, 2008.

[242] J. Urban, X. Hilaire, F. Hopfgartner, R. Villa, J. M. Jose, S. Chantamunee, and Y. Gotoh, "Glasgow University at TRECVID 2006," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[243] 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.

[244] A. Vailaya, A. K. Jain, and H.-J. Zhang, "On image classification: City images vs. landscapes," *Pattern Recognition*, vol. 31, pp. 1921–1936, 1998.

[245] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for object and scene recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[246] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *European Conference on Computer Vision*, Marseille, France, 2008.

[247] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," in *International Workshop on Semantic Learning Applications in Multimedia, in Conjunction with CVPR'06*, New York, USA, 2006.

[248] J. C. van Gemert, C. G. M. Snoek, C. Veenman, and A. W. M. Smeulders, "The influence of cross-validation on video classification performance," in *Proceedings of the ACM International Conference on Multimedia*, pp. 695–698, Santa Barbara, USA, 2006.

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

[250] R. C. Veltkamp and M. Hagedoorn, "State-of-the-art in shape matching," in *Principles of Visual Information Retrieval*, (M. S. Lew, ed.), pp. 87–119, Springer, 2001.

[251] T. Volkmer, J. R. Smith, A. P. Natsev, M. Campbell, and M. R. Naphade, "A web-based system for collaborative annotation of large image and video

collections," in *Proceedings of the ACM International Conference on Multimedia*, pp. 892–901, Singapore, 2005.

[252] T. Volkmer, J. A. Thom, and S. M. M. Tahaghoghi, "Modelling human judgement of digital imagery for multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 9, pp. 967–974, 2007.

[253] L. von Ahn, "Games with a purpose," *IEEE Computer*, vol. 39, pp. 92–94, 2006.

[254] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, USA: The MIT Press, 2005.

[255] H. D. Wactlar, M. G. Christel, Y. Gong, and A. G. Hauptmann, "Lessons learned from building a terabyte digital video library," *IEEE Computer*, vol. 32, pp. 66–73, 1999.

[256] D. Wang, X. Li, J. Li, and B. Zhang, "The importance of query-concept-mapping for automatic video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 285–288, Augsburg, Germany, 2007.

[257] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang, "Video diver: Generic video indexing with diverse features," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 61–70, Augsburg, Germany, 2007.

[258] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 947–963, 2001.

[259] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proceedings of the ACM International Conference on Multimedia*, pp. 862–871, Augsburg, Germany, 2007.

[260] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: Image auto-annotation by search," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1483–1490, New York, USA, 2006.

[261] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, pp. 12–36, 2000.

[262] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Transactions on Multimedia*, vol. 10, pp. 1085–1096, 2008.

[263] D. Weinberger, *Everything is Miscellaneous*. New York, USA: Times Books, 2007.

[264] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," in *Proceedings of the ACM International Conference on Multimedia*, pp. 71–80, Vancouver, Canada, 2008.

[265] T. Westerveld, "Using generative probabilistic models for multimedia retrieval," PhD thesis, University of Twente, 2004.

[266] T. Westerveld and A. P. de Vries, "Multimedia retrieval using multiple images," in *CIVR*, pp. 344–352, Springer-Verlag, 2004.

[267] P. Wilkins, T. Adamek, N. E. O'Connor, and A. F. Smeaton, "Inexpensive fusion methods for enhancing feature detection," *Image Communication*, vol. 7–8, pp. 635–650, 2007.

[268] P. Wilkins, A. F. Smeaton, N. E. O'Connor, and D. Byrne, "K-Space interactive search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 555–556, Niagara Falls, Canada, 2008.

[269] P. Wilkins et al., "K-Space at TRECVid 2008," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[270] M. Worring and G. Schreiber, "Semantic image and video indexing in broad domains," *IEEE Transactions on Multimedia*, vol. 9, pp. 909–911, 2007.

[271] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Novelty and redundancy detection with multimodalities in cross-lingual broadcast domain," *Computer Vision and Image Understanding*, vol. 110, pp. 418–431, 2008.

[272] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proceedings of the ACM International Conference on Multimedia*, pp. 572–579, New York, USA, 2004.

[273] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[274] L. Xie and S.-F. Chang, "Pattern mining in visual concept streams," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 297–300, Toronto, Canada, 2006.

[275] L. Xie, H. Sundaram, and M. Campbell, "Event mining in multimedia streams," *Proceedings of the IEEE*, vol. 96, pp. 623–647, 2008.

[276] X. Xie, L. Lu, M. Jia, H. Li, F. Seide, and W.-Y. Ma, "Mobile search with multimodal queries," *Proceedings of the IEEE*, vol. 96, pp. 589–601, 2008.

[277] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Transactions on Multimedia*, vol. 10, pp. 421–436, 2008.

[278] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Transactions on Multimedia*, vol. 10, pp. 1342–1355, 2008.

[279] H. Xu and T.-S. Chua, "Fusion of AV features and external information sources for event detection in team sports video," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, pp. 44–67, 2006.

[280] R. Yan, M.-Y. Chen, and A. G. Hauptmann, "Mining relationship between video concepts using probabilistic graphical models," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 301–304, Toronto, Canada, 2006.

[281] R. Yan and A. G. Hauptmann, "The combination limit in multimedia retrieval," in *Proceedings of the ACM International Conference on Multimedia*, Berkeley, USA, 2003.

[282] R. Yan and A. G. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 324–331, Seattle, USA, 2006.

[283] R. Yan and A. G. Hauptmann, "A review of text and image retrieval approaches for broadcast news video," *Information Retrieval*, vol. 10, nos. 4–5, pp. 445–484, 2007.

[284] R. Yan, A. G. Hauptmann, and R. Jin, "Negative pseudo-relevance feedback in content-based video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, Berkeley, USA, 2003.

[285] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights for automatic video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, 2004.

[286] A. Yanagawa, S.-F. Chang, L. S. Kennedy, and W. Hsu, "Columbia University's baseline detectors for 374 LSCOM semantic visual concepts," Technical Report 222-2006-8, Columbia University ADVENT Technical Report, 2007.

[287] J. Yang, M.-Y. Chen, and A. G. Hauptmann, "Finding person X: Correlating names with visual appearances," in *CIVR*, pp. 270–278, Springer-Verlag, 2004.

[288] J. Yang and A. G. Hauptmann, "(Un)Reliability of video concept detection," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 85–94, Niagara Falls, Canada, 2008.

[289] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proceedings of the ACM International Conference on Multimedia*, pp. 188–197, Augsburg, Germany, 2007.

[290] E. Yilmaz and J. A. Aslam, "Estimating average precision when judgments are incomplete," *Knowledge and Information Systems*, vol. 16, pp. 173–211, 2008.

[291] J. Yuan, H. Wang, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang, "Tsinghua University at TRECVID 2005," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2005.

[292] 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.

[293] J. Yuan et al., "THU and ICRC at TRECVID 2007," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2007.

[294] E. Zavesky and S.-F. Chang, "CuZero: Embracing the frontier of interactive visual search for informed users," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 237–244, Vancouver, Canada, 2008.

[295] E. Zavesky, S.-F. Chang, and C.-C. Yang, "Visual islands: Intuitive browsing of visual search results," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 617–626, Niagara Falls, Canada, 2008.

[296] Z. J. Zha, T. Mei, Z. Wang, and X.-S. Hua, "Building a comprehensive ontology to refine video concept detection," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 227–236, Augsburg, Germany, 2007.

[297] H.-J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, pp. 10–28, 1993.

[298] H.-J. Zhang, S. Y. Tan, S. W. Smoliar, and Y. Gong, "Automatic parsing and indexing of news video," *Multimedia Systems*, vol. 2, pp. 256–266, 1995.

[299] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, pp. 213–238, 2007.

[300] R. Zhang, R. Sarukkai, J.-H. Chow, W. Dai, and Z. Zhang, "Joint categorization of queries and clips for web-based video search," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 193–202, Santa Barbara, USA, 2006.

[301] W.-L. Zhao and C.-W. Ngo, "LIP-VIREO: Local interest point extraction toolkit," Software available at http://www.cs.cityu.edu.hk/∼wzhao2/lip-vireo.htm, 2008.

[302] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, pp. 536–544, 2003.

[303] X. Zhu, "Semi-supervised learning literature survey," Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.