

Commercials and Trademarks Recognition

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Walter Nunziati, Giuseppe Serra
{*ballan|bertini|delbimbo|nunziati|serra*}@dsi.unifi.it
MICC - Università degli Studi di Firenze

August 29, 2011

Chapter 1

Commercials and Trademarks Recognition

Abstract

In this chapter we discuss the problem of detecting and recognizing the two main categories of advertisement present in television videos: explicit advertisement in the form of commercials, i.e. short video sequences advertising a product or a service, and indirect advertisement, i.e. placement of trademarks and logos. A thorough review on the current state-of-the-art algorithms and systems for commercial and trademark recognition in a variety of different video sequences, is provided. In addition, the chapter discusses an in-depth analysis of two real-time systems, one for detecting commercials, and another for trademark recognition.

1.1 Introduction

In this chapter we discuss the problem of detecting and recognizing the two main categories of advertisement present in television videos: explicit advertisement in the form of commercials,



Figure 1.1: Examples of TV advertisement: *left*) still image of the Apple “1984” commercial; *center*) trademark placement in a sport event; *right*) logo bug in a news programme.

i.e. short video sequences advertising a product or a service, and indirect advertisement, i.e. placement of trademarks and logos, that is associated to sponsorships (particularly in sports videos) and product placement.

The first type of advertisement is still the most common one, but the latter type is now evolving in a form that is being shown more and more, mostly for advertising TV shows on the same channel, as an ad overlay at the bottom of the TV screen. In this new form the trademark is transformed in a “banner”, or “logo bug”, as they are called by media companies. Fig. 1.1 shows three frames: one taken from a commercial, one from a sport video showing placed trademarks and one grabbed from a news program showing a logo bug in the lower right corner of the image.

There are several motivations for the development of methods for the recognition of commercials and trademarks in TV videos, mostly related to economical reasons. In fact, given the high costs for commercial broadcasts and trademark placement, media agencies and sponsors are extremely keen to verify that their brand has the level of visibility they expect for such an expenditure, and want to be sure that if they paid to have a commercial aired at a certain time it was effectively done so [1]. Other motivations may include marketing analysis, e.g. to evaluate the impact of a campaign measuring sales [2] or other behaviours like website access after the airing of a commercial, or for business intelligence, e.g. to estimate the expenditures in advertisement of a competitor (also called “competitive advertising”). Motivations may also be related to sociological studies analysing, for example, the impact of certain commercials on people’s behaviour [3]. The techniques that will be described in the following sections are intended to be used in application scenarios where the only available

media are the (somehow digitized) audio/video stream as received by set-top boxes and TV, without relying, for instance, on cues or metadata that could be purposely embedded in the digital media used for transmission.

The two types of advertisement introduced above require different and specialized techniques, in order to be recognized: in the first case we have to deal with a short span of television programming (called “commercial”, “TV ad” or “spot”), whose length may vary from a few seconds to a few minutes, and our goal is to identify its appearance in a video stream. The problem of recognizing the presence of a commercial segment in a video stream can be assimilated to that of dealing with near duplicate and content-based copy detection: the commercial to be detected is the “original” video, while the broadcasted version may differ due to some editing or transmission artifacts. In the second type of advertisement we have to deal with a small portion of a video frame, and the problem is to match some views of the trademark or logo bug with parts of the frame, identifying its appearance both in space and time. Detecting the presence of a specific logo is highly correlated with the problem of (specific) object recognition: issues related to occlusion, to scale, pose and lighting variations, fully apply to our case.

Furthermore, several other practical problems have to be dealt with: the video quality is generally low, especially when considering interlaced video, and the number of high-definition television channels (HDTV) is still relatively small; these facts lead directly to the need of developing techniques that are invariant to the largest possible set of photometric and geometric disturbances. On the other hand, due to the large number of broadcast material to be analyzed in commercial/trademark automatic analysis applications, it is required to have computationally efficient methods and compact representations of the models to be retrieved.

The chapter is structured as follows: the problem of commercials recognition is addressed in Sect. 1.2, with a discussion of state-of-the-art approaches on detection (Sect. 1.2.1), recognition and modelling (Sect. 1.2.2), semantic analysis (Sect. 1.2.3); then a system for fast commercial detection based on a combination of compact descriptors based on the MPEG-7 standard is thoroughly discussed in Sect. 1.2.4. Trademark detection is analysed in Sect. 1.3.

Early works on the problem are presented in Sect. 1.3.1, while state-of-the-art approaches that can be applied in real-world images and videos are presented in Sect. 1.3.2. A thorough presentation of a complete system that can be executed in quasi-real time is given in Sect. 1.3.3. Conclusions for the whole chapter are drawn in Sect. 1.4.

1.2 Commercials recognition

The problem of recognizing commercials is related to many general video processing and analysis tasks, like scene detection, video segmentation, feature extraction and indexing [4,5]. Scene segmentation and shot detection are required to identify the segments of the video stream that are candidate to contain a spot [6,7]. Typically spots are grouped together in commercial segments; each commercial is often separated from the others and from the normal TV program by a number of black frames [8,9]. In other cases broadcasters hide their superimposed logo during the commercials [10,11]. Also audio features can be used to detect the presence of a commercials segment: the black frames that separate them have no audio [12] while during the commercials the audio level is typically higher than during the normal TV programme [1]. Audio and visual features can be fused together to improve commercials segmentation and identification [13].

Once a commercial segment candidate has been selected it is required to compute a similarity measure with the video segments taken from (usually large) database, and the system must return relevant matches. For instance, a company may need to search for its video commercials (or for other companies' commercials) within an entire day of broadcast transmission, from several TV channels, and without browsing the entire recorded video material. As such, a *video signature* (also called *fingerprint*) is typically extracted from each segment, and is stored into the database. At run time, a query video signature is compared with all the signatures stored in the database, and most similar results are returned. The problems that have to be solved are thus that of extracting a suitable signature from a video clip, and how to use it to perform matching [14].

There are some general requirements that the signature should meet. First, the signature

must be representative of the clip it refers, while being at the same time as compact as possible [15,16]. Second, the signature must be robust to a wide range of disturb, generated from different encoding schemes, different image and video formats, etc. Finally, it would be desirable to take advantages of well established, standard image and video descriptors, in order to easily compare videos coming from heterogeneous sources.

Other important considerations are related to the metric used to compare signatures extracted from two different video sequences [17–19]. As for all modern multimedia retrieval systems, the metric must be flexible enough to enable similarity matching. Another desirable property would be to allow the user to query the system with a short, “summarized” version of the video clip of interest (e.g., the trailer of a movie, or a short version of a commercial), and let the system retrieve the video the clip belongs to. Considering “spots” it is quite common that for a spot there exist several versions with different durations, e.g. a 30 seconds spot has reduced versions of 20 or 15 seconds, created by further editing of the original video.

Commercials/trademark applications need also to face scalability-related issues, i.e. how to deal with very large (end ever growing) video archives; this problem is more relevant when addressing near duplicate and copy video detection [20–23] in web-scale archives like YouTube, but it must be considered (for example to reach real-time performance) in the case of commercial detection [24,25].

1.2.1 Commercials detection

The first problem to address in commercials recognition is the identification of the candidate sequences that may contain spots in the broadcast video stream; this problem is akin to scene recognition, although it can be eased by the presence of certain characteristics that favour the identification, either imposed by local regulation that, for example, require the presence of a commercials block introduction sequence as in Germany [8], or by common practices of broadcasters, that add visual cues to distinguish the normal programmes from the commercials.

The most common practices are the introduction of a certain number of black frames

(typically between 5-12) between TV programmes and commercials blocks, and within the blocks themselves; another common practice is the disappearance of the broadcaster's logo during the commercials. These characteristics have been exploited by several authors, either alone or combined, and have been observed in TV streams of Germany, Italy, Spain, USA, China, Japan and Australia, among the others. [4,8,26–28] used the detection of monochrome black frames. The detection of presence and disappearance of TV logos has been used in [5, 10, 13].

Another common practice, perhaps the most irritating for viewers, is the increase of the audio volume during commercials, combined with silence during the sequences of black frames between the spots. This characteristic has been recently used in [29,30], where audio energy windows are used for advertisement boundary detection. A combination of visual and audio characteristics for commercials segmentation has been used in [9,31].

Some approaches skip the problem of commercials detection and rely solely on shot segmentation, either checking each sequence of the video stream [7,24,28,32,33] or by learning, using a fusion of audio-visual cues, which shots are more likely to contain commercials, as in [12].

1.2.2 Commercials representation and modelling

Descriptors

The problem of defining a video clip's signature has been widely investigated, and is still currently researched because of its applicability to content-based copy and near duplicate detection, that have recently become urgent problems since the inception of systems that allow web-based video distribution and collection, as YouTube.

In early approaches, several researchers have proposed various kinds of simple global keyframe descriptors, based usually on color and edge cues. In [8], the color coherence vector was used to characterize keyframes of commercials clip. A similar approach was followed in [34], with the addition of the use of the principal components of the color histograms of

keyframes for commercial recognition. Color histograms, compressed with PCA, have been used in [6]. More recent works along these lines introduced more sophisticated descriptors, or combinations of descriptors. In [35], the combination of color and edge-based features was considered. This work also proposed the use of inverted indices to detect copies of a video clip. A hashing mechanism based on color moment vectors to efficiently retrieve repeated video clips was proposed in [36] and also in [24]. In [37], a lookup table was used to store fingerprints based on mean luminance of image blocks. In [38] and [28] color was proposed in combination with the *ordinal measure* (originally proposed for the problem of stereo matching), to define a binary-valued signature of the video. Dominant color, Gabor filters and edge densities have been used in [5]. HSV color and edge histograms have been recently used in [33]. A combination of global visual features, like MPEG-7 Color Layout Descriptor, Edge Histogram Descriptor and others, has been recently proposed for duplicate video detection in large databases in [39]. Typical limitations of approaches based on global image descriptors are related to the difficulties of dealing with image distortions occurring either at global or local level. In addition, the discriminatory of some of these techniques rapidly decrease with the growing of the collection to be searched.

To overcome limitations of image descriptors, people have proposed to include motion-based feature in the repertoire of descriptors, either at frame level, or at local, object-like level. For instance, in [40], the concept of *video strands* was used to create spatiotemporal descriptions of video data. These video strands encapsulate the movement of objects within half-second segments of a video sequence. In [4], has been proposed the use of video editing features like cuts, considering the fact that, at least when compared to news videos, commercials have a faster paced editing and a larger number of cuts. A drawback is related to the fact that not every video segment might be suitable for this type of techniques.

Some approaches rely on audio only. In particular, [32] uses 64 dimensions Zero-Crossing Rate features; [29] uses units of energy envelope, to cope with the fact that energy is less robust to noise than frequency domain features; Fourier coefficients have been used in [30] because of their direct applicability to raw signal, and low computational cost. Audio-based approaches obviously fail when it is required to search a video collection which includes contents that have been dubbed in different languages.

Audio and visual features have also been combined to increase the robustness of the system. In [12], global visual features like edge change ratio and frame differences are combined with audio features like MFCCs and short-time energy. In [31], audio features were also used as part of the signature. The authors also performed a comprehensive experimental validation on the TRECVID dataset.

The use of local features, like interest points, is more common in the approaches for near-duplicate sequences detection, like in [41, 42], where trajectories of interest points are used as features to create the video fingerprint or like in [43] where interest points selected using Hessian-Affine detector and represented with PCA-SIFT descriptor, have been used.

Modeling and matching

In [38], commercials are compared only to sequences that have the same length, using a specific metric to handle the compact ordinal descriptor of each frame. The same approach, that does not consider possible errors in video segmentation or the fact that commercials may be edited, has been followed in [29]. Comparison of sequences of keyframes has been used in [7]; this approach allows to cope only with the shortening of the shots of the commercial.

A way to handle videos that may differ for some subsequences, e.g. like those due to re-editing of the original video stored in the database, is to use a metric belonging to the class of the edit distances. The use of the edit distance in the context of video matching was proposed in [17] and, specifically for commercials recognition, in [8]. Since then, variations have been proposed over the basic formulation [19]. Typically features' descriptors are quantized to obtain a sequence of discrete symbols (sometime referred as *visual words*), prior to perform matching. Approximate string matching has been used also in [16, 27], and has been adapted to video copy detection in [21]. Dynamic Time Warping has been used to find the optimal alignment of audio signatures of commercials in [30].

Hashing has been used in [24] because of its constant computational complexity, and the authors experimented with different hashing functions. Locality-Sensitive Hashing (LSH) has been used in [32, 33]. More recently [44, 45], similarity measures for near-duplicate detection

have been derived from metrics used in information retrieval, such as *tf/idf* weighting: in text mining application, this weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. This definition has been translated in the image context using the concept of visual words introduced above.

SVM classifiers have been proposed in [12] to classify shots containing commercials and those containing normal programmes.

Scalability

A critical problem of near-duplicate and content-based copy detection is how to deal with web-scale databases. This problem is less important when dealing with commercials recognition since a real-time system just needs to check the presence of a relatively limited number of commercials in a video stream. To speed the similarity evaluation, based on edit distance, in [21] has been introduced a pruning mechanism inspired by the FASTA algorithm employed in bioinformatics to find similar DNA sequences. In [28], differences of sequence durations are used to reduce comparisons, although this risks to miss the detection of edited spots. A dual-stage temporal recurrence hashing algorithm for commercial recognition in very large databases has been recently proposed in [25]. The issue of descriptor size has been addressed in [23], where compact frame descriptors through aggregation of local descriptors. With the growing interest of major search engines in duplicate or near-duplicate image and video detection, parallel algorithms have been introduced to perform analysis of massive quantities of data *in the cloud*, in particular by adapting *map/reduce* strategies to decompose the problem [46]: in this framework, the problem is initially chopped into a number of smaller sub-problems (*map* step), which are distributed to *worker* computational nodes. The output of these nodes is then fed back to the master node, which combines the results in a way to get the output of the original problem (*reduce* step).

1.2.3 Semantic analysis

Some authors have tried to directly encode semantic aspects in the retrieval process [13,47–49]: in these approaches, the visual content of the image or video is mapped to semantic concepts that are derived by various sources of information. In [47], the principles of *semiotics* are used to characterize the video at semantic level, by means of a weighted average of individual semiotic categories. In [48], categorization is obtained by querying public textual search engines using keyword automatically extracted from transcripts or audio features. A similar type of multi-modal approach has been proposed in [13], where a set of mid-level features are used to categorize the commercial in one of five predefined classes, following a supervised-learning strategy. More recently [49], semantics have been modeled as a latent random variable using the *Probabilistic Latent Semantic Analysis* framework, or PLSA [50]. PLSA was originally proposed as a generative model to discover topics from document corpus. As a statistical model, a PLSA model attempts to associate a latent variable (or aspect) with each observation (occurrence of a word in a document). Again, in the image context words are substituted by the occurrences of visual symbol derived from quantization of local features.

1.2.4 A sample approach

As case study let us review thoroughly a real-time approach for commercials recognition [27]. One of the main issues that has to be solved is the decision on the type of features to be used: they need to have a low computational cost to achieve real-time performance and do not have to be ad-hoc to be usable in a domain that sports a large variability like commercials. To this end we have selected an effective and efficient combination of features and features' descriptors, taken from the well established MPEG-7 standard. Using these descriptors allows video content provider to easily index their content with freely available tools, without ambiguity; moreover it also enable to match clips taken from various sources and from different organizations, provided that a suitable general purpose metric is employed, using a common representation based on a ISO standard.

The proposed method is based on the combination of three MPEG-7 descriptors, namely

the *Scalable Color Descriptor* and the *Color Layout Descriptor* for global color and color layout, and by the *Edge Histogram Descriptor* for textures. These descriptors are collected for each frame, and their sequences constitute the clip signature. To match two signatures, a variation of the edit distance [51] is employed, in order to cope with clips that may differ for small subsequences. The edit distance, or Levenshtein distance, is a metric widely used in information theory to compute the distance between two strings. It is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character; these operations have a counterpart in video editing techniques and thus edit distance is very well suited to cope with them, more than other techniques that are more suitable to handle time or speed variations like Dynamic Time Warping. However the use of edit distance has some potential drawbacks, since it is not completely clear how one should choose the number of symbols for videos of generic type. In contrast to the existing works based on edit distance, and as a second contribution of the proposed approach, we avoid this discretization step, relying directly on the distance used to compare two “symbols” (in our case, two frame descriptors, which are compared using an appropriate metric) for deciding on the cost of transforming one string into another.

The video clip matching system is performed in two phases. In the first one an indexing process generates the signatures, composed by MPEG-7 descriptors, for the clips that are to be retrieved. In the second phase the target video that has to be analyzed is processed to extract the same features, and these are used to measure similarity with the signatures. There are no special requirements on the video format used to create the index or the target video, since the descriptors are extracted from decompressed frames. Moreover experiments showed that even frame sizes as low as PAL QCIF (192×144) can be used in both phases, thus easing the achievability of real-time performance. The descriptors used capture different aspects of the video content [52], namely global color, color layout and texture and are computationally inexpensive. Motion descriptors have not been used in order to be able to perform recognition in real-time. However, the temporal aspects of video are implicitly considered using the edit distance to match the video sequences.

The MPEG-7 features that have been selected are suitable for the creation of fingerprint since they meet the important requirements of fast calculation, compact representation, good

discriminative power and tolerance to small differences due to signal degradation. To reduce the space occupation of the stored MPEG-7 descriptors, due to the verbosity of the XML format, it is possible to use the BiM (Binary Format for MPEG-7) framework; in fact BiM enables compression of any generic XML document, reaching an average 85% compression ratio of MPEG-7 data, and allows the parsing of BiM encoded files, without requiring their decompression. In the following we provide a short discussion on these descriptors and the metrics used to match them.

Scalable Color Descriptor (SCD) The SCD is a color histogram in the HSV color space, uniformly quantized into bins according to tables provided in the MPEG-7 standard normative part, encoded using the Haar transform. Its binary representation is scalable since it can be represented with different bits/bins, thus reducing the complexity for feature extraction and matching operations. Increasing the number of bits used improves retrieval accuracy. This descriptor can be extended into the GoF/GoP (Group of Frames/Group of Pictures) color descriptor, thus allowing it to be applied to a video segment. In this case two additional bits allow to define how the histogram is calculated before applying the Haar transform. The standard allows to use average, median or intersection. In the first case, adopted in this work, averaging of the counters of each bin is performed; the result is equivalent to computing the aggregate histogram of the group of pictures and performing normalization. The median histogram is equivalent to compute the median of each counter value of the bins, and may be used to achieve more robustness w.r.t. outliers in intensity values. The intersection histogram requires the calculation of the minimum counter value of each bin, and thus the result is representative of the “least common” color traits of the group of pictures.

SCD descriptors can be matched both in the histogram domain and in the Haar domain using the L1 norm, although it has to be noted that results of the L1 norm-based matching in the Haar domain are not the same of the histogram. Generation of the Haar coefficients is computationally marginal w.r.t. histogram creation, and their matching is equal in complexity to histogram matching, thus to avoid the reconstruction of the histogram from the descriptor we have used matching in the Haar domain, using 128 bits/histogram.

Color Layout Descriptor (CLD) This descriptor represents the spatial distribution of color in an extremely compact form (as low as 8 bytes per image can be used), and thus is particularly interesting for our scope, because of computational cost of matching and space occupation. The input picture is divided in an 8×8 grid and a representative color in the YCrCb color space for each block is determined, using a simple color averaging. The derived colors are then transformed into a series of coefficients using a 8×8 DCT. A few low-frequency coefficients are selected using zigzag scanning and then quantized. Since the calculation of the descriptor is based on a grid it is independent from the frame size.

To match two CLDs ($\{DY, DCr, DCb\}$ and $\{DY', DCr', DCb'\}$) the following distance measure is used [53]:

$$D = \sqrt{\sum_i w_{yi}(DY_i - DY'_i)^2} + \sqrt{\sum_i w_{bi}(DCb_i - DCb'_i)^2} + \sqrt{\sum_i w_{ri}(DCr_i - DCr'_i)^2}$$

where DY_i , DCb_i and DCr_i are the i -th coefficients of the Y, Cr and Cb color components, and w_{yi} , w_{bi} and w_{ri} are the weighting values, that decrease according to the zigzag scan order.

Edge Histogram Descriptor (EHD) This descriptor represents the spatial distribution of five types of edges (four directional and one non-directional). This distribution of edges is a good texture signature even in the case of not homogeneous texture, and its computation is straightforward. Experiments conducted within the MPEG-7 committee have shown that this descriptor is quite effective for representing natural images. To extract it, the video frame is divided into 4×4 blocks, and for each block an edge histogram is computed, evaluating the strength of the five types of edges and considering those that exceed a certain preset threshold. Values of the bins are normalized to $[0, 1]$, and a non linear quantization of the bin values results in a 3 bits/bin representation. Overall the descriptor consists of 80 bins (16 blocks and 5 bins per block), and is thus quite compact.

The simplest method to assess similarity between two EHDs is to consider the 3-bit numbers as integer values and compute the L1 distance between the EHDs.

The combination of the three descriptors discussed above creates a robust signature, that comprises global and local features that describe syntactic aspects of video content, and yet is still compact and computationally inexpensive. Moreover is standard and can be easily reproduced using the MPEG-7 XM Experimentation Model software, freely available.

Video clip matching

Our goal is to be able to perform approximate clip matching, evaluating similarity of video sequences even in case that the original video has been re-edited. This case may occur since often several variations of the same commercial are produced, usually creating shorter versions from a longer one. This may happen also with *video rushes* (the first unedited sequence that is filmed), from which a smaller section is usually selected to be used. Another case may be that of identifying sequences that have a variable length such as those containing anchormen in a news video, those that compose a dialog scene in a movie, or slow motion versions of a sequence like the replay of a sport highlight.

In our approach we extract the features used to create the commercial's fingerprint from the clip A that has to be analyzed, and consider both its feature vector and the fingerprint of the commercial B to be recognized as vectors composed by three strings, one for each feature. All the strings of A will have length m and those of B will have length n .

To evaluate the similarity of the video clips we consider each corresponding couple of corresponding strings and calculate an approximate distance. The three distances are used to calculate the Manhattan distance between the clips, and if the distance is bigger than a minimum percentage of the length of the clips then they are matched.

The approximate distance between the strings is evaluated using the Sellers algorithm [54]. This distance is similar to the Levenshtein edit distance, and adds a variable cost adjustment to the cost of gaps, i.e. to insertions and deletions. Using this distance, and tailoring the

costs of the edit operation appropriately it is possible to adjust the system to the specific clip matching task. For example if there is need to detect appearances of a long rush sequence, that is likely to be shortened in the video editing process, deletions could be considered less expensive than insertions. A simple dynamic programming implementation of the algorithm, as that shown in [51], is $O(mn)$ in time and $O(\min(m, n))$ in space, but other algorithms can reduce time and space complexity. Given the short length of commercials the complexity of the algorithm does not pose any problem. From the edit operations cost formula of [51], and considering the cost matrix C that tracks the costs of the edit operations needed to match two strings, we can then write the cost formula for the alignment of the a_i and b_j characters of two strings as:

$$C_{i,j} = \min(C_{i-1,j-1} + \delta(a_i, b_j), C_{i-1,j} + \delta_I, C_{i,j-1} + \delta_D)$$

where $\delta(a_i, b_j)$ is 0 if the distance between a_i and b_j is close enough to evaluate $a_i \approx b_j$ or the cost of substitution otherwise, δ_I and δ_D are the costs of insertion and deletion, respectively. Fig.1.2 shows a simplified example of edit distance calculation between a part of a commercial and a shortened version of the same commercial, using the CLDs.

The alphabet of the strings has size equal to the dimensionality of features, but it has to be noted that it has no effect in terms of computational time or size on the string similarity algorithm, since only the cost of the edit operations are kept in memory, and the only operation performed on the alphabet characters is the check of their equality, using the appropriate distance described for each feature in the above paragraphs. This approach allows us to overcome a limitation of string matching; in fact usually a difference between symbols is evaluated using the same score, that is the distance between symbol 1 and 2 is the same between symbol 1 and 10. In our case instead, when comparing two CLDs, for example, a close match could be evaluated as an equality, without penalizing the distance.

The calculation of similarity can be stopped earlier, when the required similarity threshold has been reached, to speed up processing.





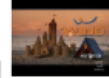





						
	0	1	2	3	4	5
	1	0	1	2	3	4
	2	1	1	1	2	3
	3	2	2	1	1	2

Figure 1.2: Simplified example of edit distance calculation performed using CLDs between two edited versions of the same commercial. The top row shows part of a commercial, the left column shows a reduced version. The table contains the $C_{i,j}$ costs, where $\delta(a_i, b_i)$, δ_I and δ_D have value 1. The circled number is the distance between the two sequences.

Experimental results

The clip matching approach described in the previous section is independent w.r.t. the domain of the videos. According to each domain the most appropriate clip selection algorithm should be used, ranging from simple shot detection to other methods that may select sequences composed by more shots. The clip selection algorithm used to extract the commercials to be matched is based on detection of black frames that appear before each commercial, as noted in [8].

About 10 hours of videos were acquired from digital TV signal and from different European and international broadcasters, at PAL frame rate and size, and frame resolution was scaled down to PAL QCIF (192×144 pixels). 40 different commercials were selected and added to the test database. Videos were converted to MPEG-1 and 2 format using FFMpeg and Mainconcept MPEG Pro encoders, at different quality, to test the robustness with respect to compression artifacts and noise. In our prototype application (implemented in C++ under Linux, without any use of multithreading that would greatly improve speed through parallel computation of features), the system took about three seconds (on average) to extract the

signature of a PAL QCIF clip of length 30 seconds. At query time, the system took less than half second to perform a single comparison between two of such signatures.

Fig. 1.3 shows the average similarity between pairs of corresponding commercials, where one of the spots was corrupted by a range of photometric or geometric disturbance: high video compression, contrast, crop and blur. These corruptions affect differently the descriptors used in the system: *i)* blur affects mainly EHD; *ii)* contrast affects both SCD and CLD, but at a certain level starts to affect also EHD; *iii)* compression affects more EHD, due to the typical blocking effect of MPEG video; *iv)* crop affects more EHD and CLD, since they are based on local representations while SCD is less affected since it's a global feature. Since we don't have a natural way to express the entity of all the type of disturbance, the x -axis represents this measure relatively to the maximum level of the disturb that was applied. Such maximum level is shown in the right column for some sample keyframes of a test sequence. All of the corrupted versions of the commercials were obtained using the program Avidemux, on a dataset of about 100 video clips taken from various sources. Clips were originally encoded at a frame rate of 1000 kbps, and the maximum compression was obtained setting a fixed bitrate of 50 kbps. The graph shows how similarity gracefully degrades for all kind of disturb, thanks to the fact that descriptors are typically not affected all together by these disturbs. As can be expected, most critical disturbs are those that heavily influence the photometric properties, such as large changes in contrast.

Discussion

In this section we have presented a TV commercial recognition algorithm that uses a robust video fingerprint based on standard MPEG-7 descriptors. Apart from detection of commercials in TV broadcasts the proposed approach can be used also to solve the generic problem of near duplicate clip matching, such as identification of structural elements like dialogs in movies or appearance of anchorman or interviews in news videos. The descriptors that compose the fingerprint capture several syntactic aspects of videos, are easily computed and compact enough to be used in large databases. Experiments have shown that the approach is suitable for real time recognition.

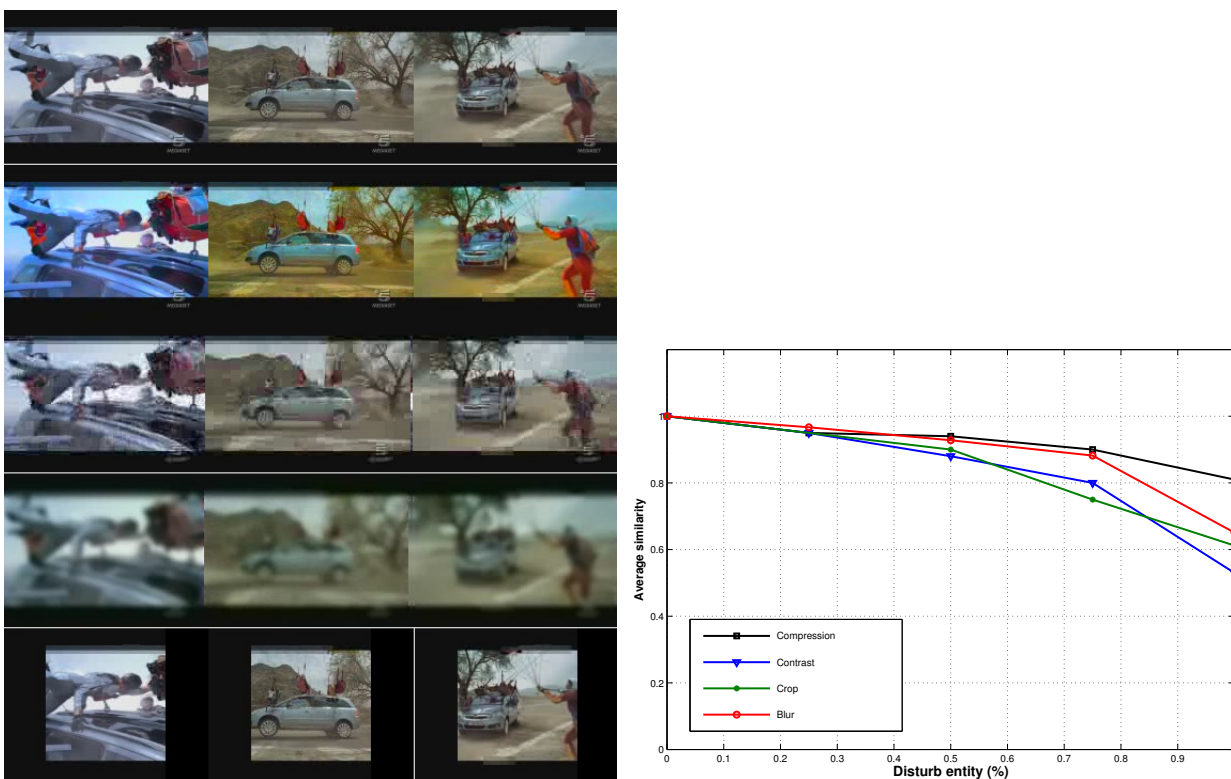


Figure 1.3: *Left)* Keyframes from a sample sequence. The top row shows the reference commercial. Subsequent rows show keyframes taken from the commercial with the maximum level of disturb applied. From top to bottom: original, contrasted, compressed, blurred, and cropped versions - *Right)* Average similarity between 100 corresponding clips versus the entity of various disturbs on the original signal.

Use of edit distance as a base for the approximate match allows to cope with re-edited clips (e.g. the common shorter versions of commercials), or clips that have different frame rates. While the proposed method can be directly adopted to solve the matching problem efficiently in small databases, additional effort must be made to complete the representation with a suitable indexing scheme, that would make possible matching clips in very large video databases, like online video sharing services, without performing exhaustive search.

1.3 Trademark detection

Currently, verification of brand visibility in a video is done manually by human annotators that view a broadcast event and annotate every appearance of a sponsor's trademark. The

annotation performed on these videos is extremely labor-intensive, usually requiring the video to be viewed in its entirety several times, and subjective. Moreover, manual annotations of this type are usually limited to the annotation of the *appearances* of a given trademark (i.e., a particular trademark *appears* at a particular timecode). Because of their popularity and the amount of related sponsorships investments, sports videos are for sure the ones that are more investigated for this kind of analysis [55–59]. Just as an example, the television coverage of the 2006 FIFA World Cup was aired in a total of 43,600 broadcasts across 214 countries and territories, generating a total coverage of 73,072 hours; this is an increase of 76% on the 2002 event and a 148% increase on 1998. The analysis of a similar huge amount of material is nowadays unfeasible for a manual procedure. This fact becomes evident as we observe that the best human annotators can annotate approximately a sports video by considering only four different trademarks in real time (i.e., one hour of video requires one hour of annotation for four trademarks). But annotations are typically required for between twenty and thirty trademarks for each sport video, requiring the annotator to view it multiple times or requires that multiple human annotators check the same video in parallel. In any case, a one hour video typically requires around six man-hours to be fully annotate.

Automatic annotation of trademark promises to significantly reduce the human labor involved in annotating. Furthermore, automatic annotation can provide a richer information than those currently performed by humans. Some methods are able, for example, to compute metrics on the duration of each trademark appearance as well as an estimation of the size it occupies in the image or frame.

1.3.1 Early approaches

The early work on automatic trademark detection and recognition addressed the problem of assisting the registration process. Since a trademark has to be formally registered, the idea of these approaches is to compare a newly designed trademark with archives of already registered ones, in order to ensure that it is sufficiently distinctive and avoid confusion [60]. Historically, the earliest approach was Kato’s Trademark system [61]. Its idea is to map normalized trademark images to an 8×8 pixel grid, and calculate an histogram (called

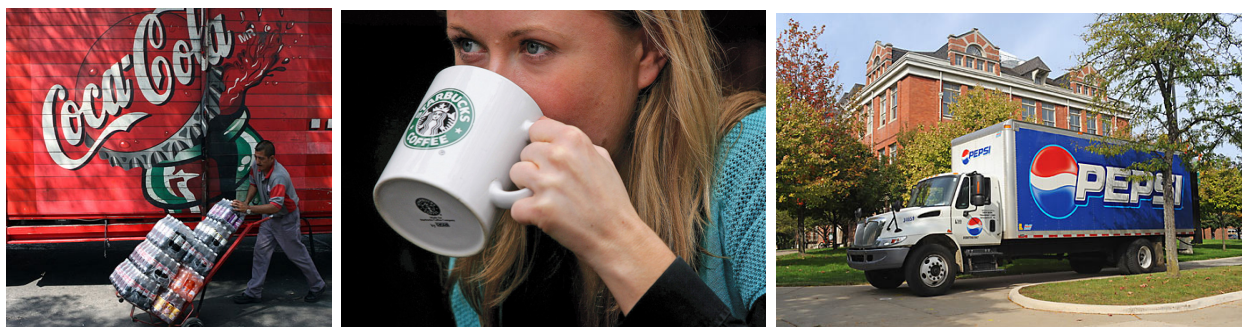


Figure 1.4: Realistic examples of trademark images characterized respectively by a bad light condition (Coca-Cola), a deformation (Starbucks), and an occlusion (Pepsi).

GF -vector) for each image from frequency distributions of black and edge pixels appearing in each cell of the grid. Matching between logos was performed by comparing the respective GF -vectors. An other notable system was Artisan [62] that achieves trademark retrieval using shape similarity. In this approach Gestalt principles were used in order to derive rules allowing individual image components to be grouped into perceptually significant parts. In [63], the authors proposed a method to retrieve trademarks using query by rough sketches. They characterize regions as either solid or line-like, extracting boundary contours for the former and skeletons for the latter. These descriptions are used to compare the overall image similarity between the query and stored images. More recently, in [64] the authors proposed a system that combines global Zernike moments, local curvatures and distance to centroid features in order to describe logos. In [65], a query-by-example retrieval system was proposed; logos are described by a variant of the shape context descriptor [66] and are then organized by a locality-sensitive hashing indexing structure, aiming to perform approximate k -NN search in high dimensional spaces in sub-linear time.

In this scenario it is usually assumed that the image acquisition and processing chain is controlled so that the images are of acceptable quality and are not distorted. For this reason, all these methods use synthetic images and rely on global logo descriptions, usually related to their contours or to particular shape descriptors.

1.3.2 Trademark detection and recognition in real-world images/videos

In the last years novel works on logo detection and recognition in real world images/videos have emerged; they are mainly targeted to automatically identify products (such as groceries in stores for assisting the blind or product on the web) [67, 68] or, as previously introduced, to verify the visibility of advertising trademarks (e.g. billboards or banners) in sports events [55–58]. Nevertheless, there are few publicly available dataset to evaluate and compare the most recent retrieval approaches, except the recent *BelgaLogos* dataset [59].

However, the trademark detection and recognition problem in natural image collections is extremely hard, due to the relatively low resolution and quality of images (e.g. due to compression artifacts, color sub-sampling, motion blur, etc.) and also to the fact that trademarks are often small and may contain very few information. Moreover their appearance is often characterized by occlusions, perspective transformations and deformations (see the examples in Fig. 1.4). The problem of detecting and tracking billboards in soccer videos has been initially studied in [55], with the goal of superimposing different advertisements according to the different audiences. Billboards are detected using color histogram back-projection and represented using a Probability Density Function in an invariant color space estimated from manually annotated video frames. The focus of this work is on detection and tracking rather than recognition. In [69], logo appearance is detected by analyzing sets of significant edges and applying heuristic techniques to discard small or sparsely populated edge regions of the image. Subsequently, the same authors extended this work ([70]) by dealing with logos appearing on rigid planar surfaces with an homogeneously colored background. Video frame are binarized and logo regions are combined using some heuristics. The Hough transform space of the segmented logo is then searched for large values, in order to find the image intensity profiles along lines, and logo recognition is performed by matching these lines with the line profiles of the models. In [71], candidate logo regions are detected using color histogram back-projection and then are tracked. Multidimensional receptive field histograms are finally used to perform logo recognition; for every candidate region the most likely logo is computed and thus, if a region does not contain a logo, the precision of identification is reduced. In [72, 73], logos are represented defining an extension of the Color-Edge Co-occurrence Histogram (CECH) [74], which captures both the color and spatial relations

between color pixels, by introducing a color quantization method based on the HSV color space.

Recently, interest points and local descriptors have been successfully used in order to describe logos and obtain flexible matching techniques that are robust to partial occlusions as well as linear and non linear transformations. The first approach of this type, proposed in [57], achieves trademark detection and localization in sports videos; each trademark is described as a bag of local features (SIFT points [75]) which are classified and matched with the bags of SIFT features in video frames. Localization is performed through robust clustering of matched features. Following a similar approach, a SIFT point representation is exploited in [59] for detecting logos in natural images. In order to refine their detection results, authors also included geometric consistency constraints by estimating affine transformations between queries and retrieved images. Furthermore, they use a contrario adaptive thresholding in order to improve the accuracy of visual query expansion. This method puts a model assumption about the possible transformation (i.e. homography) between reference logos and test images. Though it might not capture the actual inter-logo transformation, for example when logos are deformed (e.g. a logo on a t-shirt), perspective deformations are captured by a homography model. This is an important point, since in practice logos are usually affected by this kind of transformations. In [76], the authors proposed a logo detection method, following the idea previously introduced by [77] for object categorization and retrieval, by data mining association rules that capture frequent spatial configuration of quantized SIFT features at multiple resolutions. These collections of rules are indexed in order to retrieve representative training templates for matching, nevertheless image resolution is a major limitation. In [78], is presented a two-stage logo detection algorithm which also achieves localization by adapting a spatial-spectral saliency to improve the matching precision. A spatial context descriptor is introduced in order to estimate the spatial distribution of the set of matching points. The system is able to find the minimum boundary round of matched points and to partition it into nine areas. This information is used to describe the distribution of these feature points using a simple nine-dimensional histogram.

1.3.3 A sample approach

As case study we review in detail an automatic system for detecting and retrieving trademark appearances in sports videos [57]. This system has been designed to attain a quasi-real time processing of videos. Broadcast sports video is recorded directly to DVD. This video, and a collection of static trademark images, are then processed to extract a compact representation based on SIFT feature points. The results of this processing are stored in a database for later retrieval. All of the trademarks are then matched against the content extracted from every frame of the video to compute a “match score” indicating the likelihood that the trademark occurs at any given point in the video. Localization is performed through robust clustering of matched feature points in the video frames. These time series are used to retrieve intervals of the video likely to contain the trademark image. Retrieved segments are shown in a user interface used by a human annotator who can then validate this automatic annotation (see Fig. 1.5 for a screenshot of the application).

Trademark appearances

One of the distinctive aspects of trademarks is that they are usually planar objects and contain both text and other high-contrast features such as graphic logos. In sports videos they are often occluded by players or other obstacles between the camera and the trademark. Often the appearance of trademarks are also characterized by: *i) perspective deformation* due to placement of the camera and the vantage from which it images advertisements in the field; *ii) motion blur* due to camera motion, or motion of the trademark in the case of trademarks placed, for example, on Formula One cars or jerseys of soccer players. Since blur is indistinguishable from a change in scale, a scale-invariant representation is essential.

To obtain a matching technique that is robust to partial occlusions, we use local neighborhood descriptors of salient points. By combining the results of local, point-based matching we are able to match entire trademarks. Local texture and important aspects in trademarks are compactly represented using SIFT features [75], because they are robust to changes in scale, rotation and (partially) affine distortion. Trademarks are so represented as a bag of

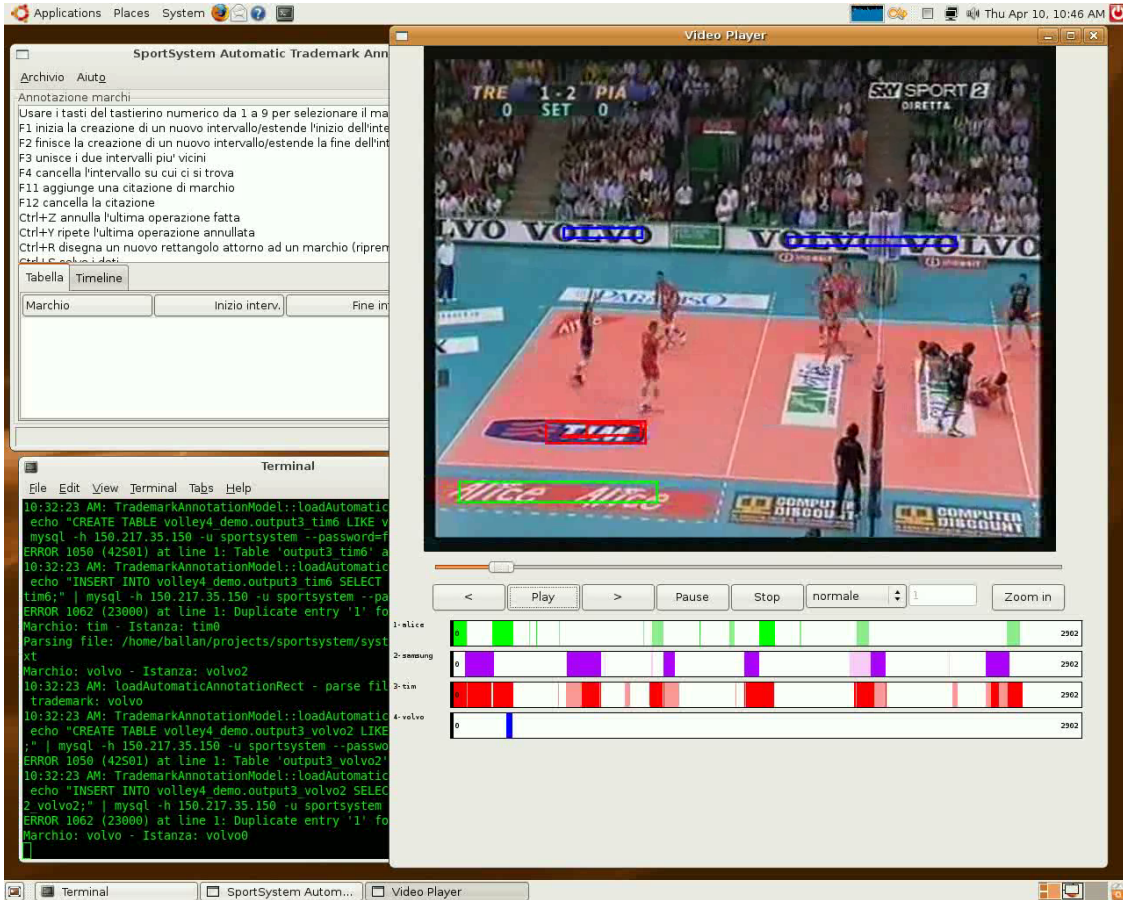


Figure 1.5: A screenshot of the visualization application. The user can configure how detected trademarks are visualized on the video frame. The rows at the bottom indicate, in different colors, the timeline of detected trademarks in the video.

SIFT feature points. Each trademark is represented by one or more image instances. More formally, trademark T_i is represented by the N_i SIFT feature points detected in the image:

$$T_i = \{(x_k^t, y_k^t, s_k^t, d_k^t, \mathbf{O}_k^t)\}, \text{ for } k \in \{1, \dots, N_i\},$$

where x_k^t , y_k^t , s_k^t , and d_k^t are, respectively, the x- and y-position, the scale, and the dominant direction of the k th detected feature point. The element \mathbf{O}_k^t is a 128-dimensional local edge orientation histogram of the SIFT point. The superscript t is used only to distinguish points from trademarks and video frames. An individual point k from Trademark i is denoted by T_i^k .

Each frame, V_i , of a video is represented similarly as a bag of M_i SIFT-feature points

detected in frame i :

$$T_i = \{(x_k^v, y_k^v, s_k^v, d_k^v, \mathbf{O}_k^v)\}, \text{ for } k \in \{1, \dots, M_i\},$$

and where each element is defined as above for trademarks. Again, the superscript is used to distinguish video frame points from points detected in trademark images.

The local orientation histogram portions (\mathbf{O}_k^v and \mathbf{O}_k^t) of the feature points are then used for the matching procedure, as described in the next section. This allows to define a feature descriptor that may results robust to geometric distortions and scale changes. Note also that the coordinates x and y of the feature points are not used during the matching phase, while they are then utilized only for trademark localization.

Detection and localization of trademarks

Trademark detection is performed by matching the bag of local features representing the trademark with the local features detected in the video frames. In order to minimize false positive detections we use a very conservative threshold. In particular, for every point detected in trademark T^j we compute its two nearest neighbors in the points detected in video frame V_i :

$$\begin{aligned} N_1(T_j^k, V_i) &= \min_q \|\mathbf{O}_q^v - \mathbf{O}_k^t\| \\ N_2(T_j^k, V_i) &= \min_{q \neq N_1(T_j^k, V_i)} \|\mathbf{O}_q^v - \mathbf{O}_k^t\|. \end{aligned} \quad (1.1)$$

Next, for every point in the video frame we compute its *match score*:

$$M(T_j^k, V_i) = \frac{N_1(T_j^k, V_i)}{N_2(T_j^k, V_i)}, \quad (1.2)$$

that is the ratio of the distances to the first and second nearest neighbors.

SIFT points are selected as being good candidate matches on the basis of their match



Figure 1.6: Two examples of the “traditional” SIFT matching technique.

scores. The *match set* for trademark T_j in frame V_i is defined as:

$$M_i^j = \{k \mid M(T_j^k, V_i) < \tau_1\}, \quad (1.3)$$

where τ_1 is a suitable chosen threshold (we empirically fix it to 0.8 in all of our experiments).

This methodology gives very good results in terms of robustness. In fact, a correct match needs to have the closest matching descriptor significantly closer than the closest incorrect match, while false matches, due to the high dimensionality of the feature space, have a certain number of other close false matches. Fig. 1.6 shows two matching examples in different sport domains.

In order to make the final decision that the trademark T_j is present in the frame V_i , a certain percentage of the feature points detected in the trademark has to be matched according to equation 1.3. This task is done by thresholding the *normalized match score*:

$$\frac{|M_i^j|}{|T_j|} > \tau_2 \iff \text{trademark } T_j \text{ present in frame } V_i.$$

Analysis of the precision–recall curves obtained using different values of τ_2 , and different trademarks, allows to determine the best choice for this threshold (see section 1.3.3). Experiments on several different trademarks and sports have shown that a value of 0.2 – 0.25

is a reasonable choice.

Trademark localization is performed by a robust estimate of the features point cloud. Let $F = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be the matched point locations in the frame. The robust centroid estimate is computed by iteratively solving for (μ_x, μ_y) in

$$\sum_{i=1}^n \psi(x_i; \mu_x) = 0, \quad \sum_{i=1}^n \psi(y_i; \mu_y) = 0$$

where the influence function ψ used is the Tukey biweight:

$$\psi(x; m) = \begin{cases} (x - m)(1 - \frac{(x-m)^2}{c^2})^2 & \text{if } |(x - m)| < c \\ 0 & \text{otherwise} \end{cases} \quad (1.4)$$

The scale parameter c is estimated using the *median absolute deviation from the median*:

$$\text{MAD}_x = \text{median}_i(|x_i - \text{median}_j(x_j)|).$$

Once the estimation of the robust centroid is done, the distance of each matched point to the robust centroid is computed according to the influence function (1.4). Points with a low influence are excluded from the final match set. Fig. 1.7 reports a schematization of the robust trademark localization procedure.

Some example matches found in different sports videos using this technique are shown in Fig. 1.8. Notice that the technique is quite robust to occlusions, scale variation, and perspective distortion. Note also that the model trademark used in the second row is a synthetic trademark image. The distinctive structure in the text of the trademark is enough to discriminate it from other trademarks and background noise.

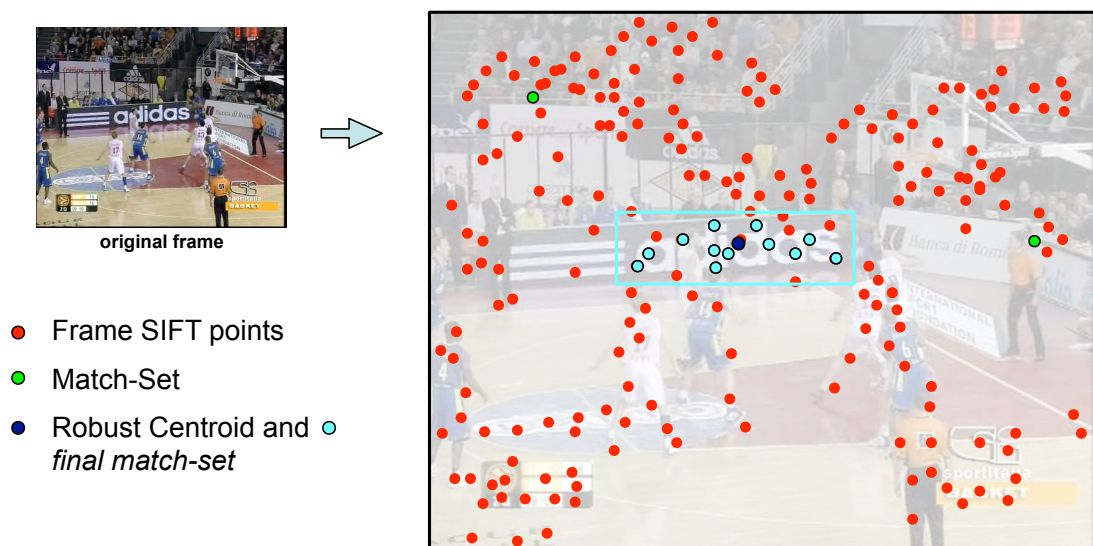


Figure 1.7: An example of robust trademark localization. Points in cyan are those selected as finale trademark points set; points in green are SIFT matched points with a low influence and so excluded from the final match set.

Experimental results

To calibrate and evaluate our proposed technique we perform several experiments. Three full videos of three different sports (each one is approximately 90 minutes long) were used for an evaluation of the performance of the approach. The first video is related to a MotoGP motorcycle race. The second one is related to a volleyball match and contains significantly different trademarks and characteristics than the MotoGP video. In fact, sports like volleyball and basketball presents a lot of situations with occlusions or partial appearance of the trademarks. The last one is of a soccer match (taken from italian *Serie A*); in this case there are often trademarks at low resolution with few SIFT feature points. We refer to this collection of videos as SPORTS-VIDEOS dataset. The examples in the top row of Fig. 1.8 are from the MotoGP video, those in the middle row are from the volleyball and those in the bottom row from the soccer video.

The MotoGP video was manually annotated for the presence of a number of trademarks, to analyse the effects of all the parameters of the proposed algorithms. These annotations were performed at the frame level, and each trademark appearance is associated with an interval in the ground-truth. The performance of the approach is evaluated in terms of two



Figure 1.8: Some example matches. The leftmost column contains the trademark model annotated with its detected SIFT feature points. The other three columns contain a portion of a video frame where a match was found. Points indicated in cyan are those selected as “good” matches according to equations (1.3, 1.4).

standard metrics i.e. *precision* and *recall*, defined as:

$$\begin{aligned} \text{precision} &= \frac{\# \text{ correct trademark detections}}{\# \text{ trademark detections}}, \\ \text{recall} &= \frac{\# \text{ correct trademark detections}}{\# \text{ trademark appearances}}. \end{aligned}$$

Fig. 1.9 gives an overview of the performance of the algorithm on the MotoGP video for six trademarks over a range of normalized match score thresholds. Also shown in the plots of Fig. 1.9 are the precision and recall performances as a function of the frame sampling rates. Results are shown for 2.5fps , 5fps , and 10fps . Note that in these plots, the recall plots are the ones that start at or around 1.0 and *decrease* as the normalized match threshold is increased. In most cases, a recall of about 85% can be obtained at a precision of around 80% with values of τ_2 varying between 0.2 – 0.25. The experiments performed on soccer and

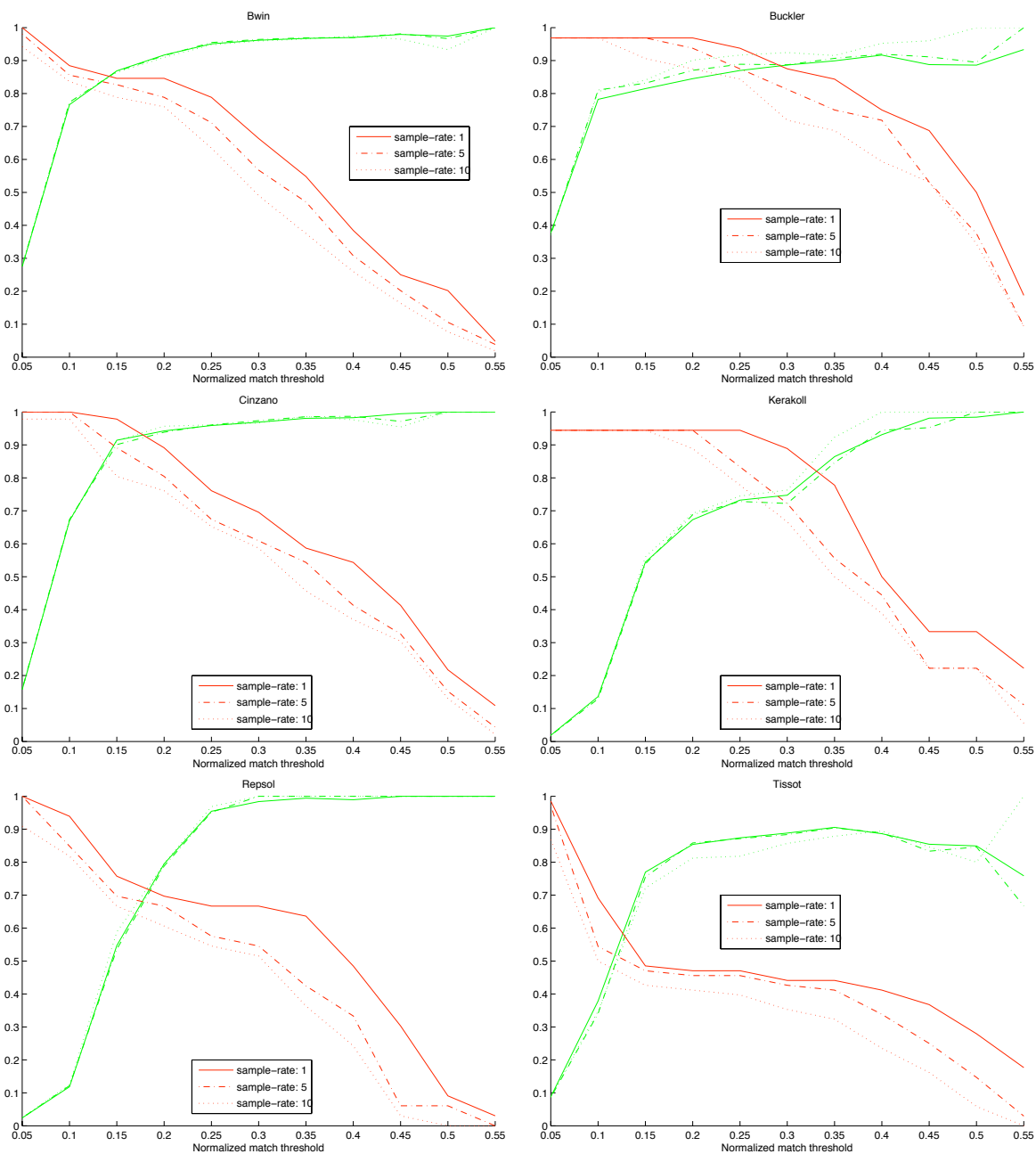


Figure 1.9: Precision and recall as a function of the normalized match threshold. Note that as the threshold increases, more matches are *excluded*. Because of this, recall usually begins at or around 1.0 and is inversely proportional to the normalized match threshold.

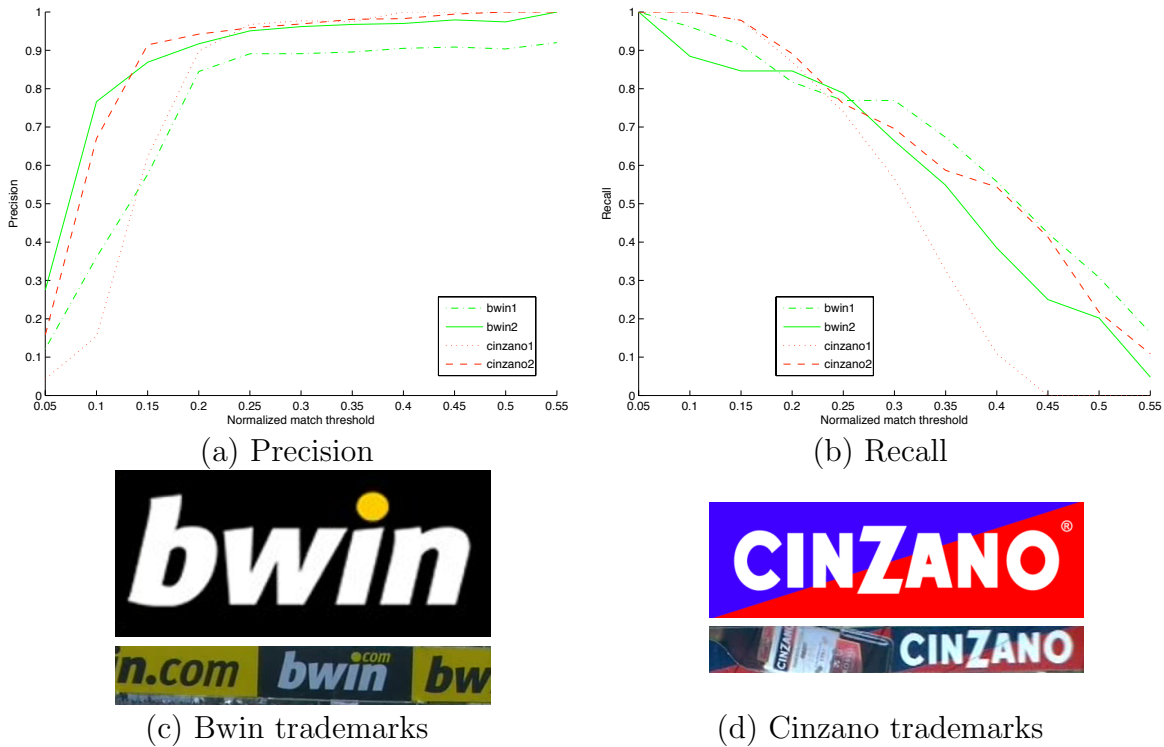


Figure 1.10: Comparison of precision between synthetic trademarks and trademarks cropped from actual video frames.

volleyball videos have shown that this value of the threshold can be used also for these other sports. Increasing the frame sampling rate predictably impacts the recall of the results. It is interesting, however, that the precision of the retrieved results is not adversely affected. This indicates that the matching technique has a very low false-positive rate. In cases such as Tissot and Kerakoll, the poor performance is related to the fact that the model trademarks have relatively few feature points detected in them, causing the normalized match score to become unreliable.

We have also experimented with different types of trademark prototypes used for matching. In some cases, the textual and graphical structure of a trademark is enough to distinguish it. In Fig. 1.10 are shown results comparing matching performance on synthetic trademarks and on trademark instances cropped directly from the video. In these plots, “bwin1” and “cinzano1” refer to the synthetic images (shown to the right of Fig. 1.10). In the case of precision, we can see that, for low values of the normalized match threshold, the synthetic images perform worse than those selected from the video itself. This is due to the fact that

many other trademarks consisting of mostly text and graphics are confused for the synthetic trademark models. Recall is affected as well, though not as significantly as precision.

Finally, we have conducted some preliminary experiments using SURF features instead of SIFT since they can speed-up our method. To this end, we have replicated the same experimental setup on the SPORTS-VIDEOS dataset. Our results shows that at the same recall values there is a breakdown in precision performance of around 11%.

Discussion

In this study case we analysed an approach to automatically detect and retrieve trademark occurrences in sports videos. A compact representation of video and trademark content, based on SIFT features, ensures that the technique is robust to occlusions, scale and perspective variations. In our SPORTS-VIDEOS dataset, on average, a recall rate of better than 85% can be achieved with a precision of approximately 60%. Our robust clustering procedure enables accurate localization of trademark instances and makes the technique robust to spuriously matched points in the video frame, by requiring spatial coherence in the cluster of matched points. Preliminary experiments on sports videos in different domains confirm that the technique can effectively retrieve trademarks in a variety of situations. Results on Formula One races, for example, are comparable to the results presented here for MotoGP. For sports such as soccer and volleyball, however, the approach suffers from the fact that trademarks are usually viewed from a wide-angle vantage and appear at a much lower resolution than in MotoGP and Formula One. This fundamentally limits the ability to detect enough feature points on the trademarks in the video. A possible solution to this problem is to double the resolution of each video frame before processing. However, this has the adverse affect of greatly increasing the amount of time required to detect feature points and perform matching on the (greatly inflated) sets of features.

1.4 Conclusions

In this chapter we have reviewed the state-of-the-art in commercials and trademark recognition in videos, providing an in-depth analysis of two real-time and quasi real-time systems. The main issues that have to be solved for commercials recognition regard the problem of scalability, i.e. how to recognize the presence of a commercial in a web-scale archive; this problem will become more and more important as video is going to be distributed also through internet channels and not only using TV channels. Regarding trademark recognition the main issues are the improvement of recall and scalability. In this case the problem of scalability is related not only with the number of videos but also with the high number of variations of the same trademark, and with the limits of the robustness of local descriptors to the strong scene variations that occur in sports videos.

References

- [1] B. Satterwhite and O. Marques, "Automatic detection of TV commercials," *IEEE Potentials*, vol. 23, no. 2, pp. 9–12, April-May 2004.
- [2] J. Y. Ho, T. Dhar, and C. B. Weinberg, "Playoff payoff: Super bowl advertising for movies," *International Journal of Research in Marketing*, vol. 26, no. 3, pp. 168 – 179, 2009.
- [3] H. G. Dixon, M. L. Scully, M. A. Wakefield, V. M. White, and D. A. Crawford, "The effects of television advertisements for junk food versus nutritious food on children's food attitudes and preferences," *Social Science & Medicine*, vol. 65, no. 7, pp. 1311 – 1323, 2007.
- [4] J.-H. Yeh, J.-C. Chen, J.-H. Kuo, and J.-L. Wu, "TV commercial detection in news program videos," in *Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2005.
- [5] L.-Y. Duan, Y.-T. Zheng, J. Wang, H. Lu, and J. Jin, "Digesting commercial clips from TV streams," *IEEE MultiMedia*, vol. 15, no. 1, pp. 28–41, Jan-Mar 2008.
- [6] J. Sánchez, X. Binefa, J. Vitrià, and P. Radeva, "Local color analysis for scene break detection applied to tv commercials recognition," in *Proc. of International Conference on Visual Information Systems (VISUAL)*, 1999.
- [7] J. M. Sánchez, X. Binefa, and J. Vitrià, "Shot partitioning based recognition of tv commercials," *Multimedia Tools and Applications*, vol. 18, pp. 233–247, 2002.
- [8] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," in *Proc. of IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, 1997.

- [9] D. A. Sadlier, S. Marlow, N. O'Connor, and N. Murphy, "Automatic TV advertisement detection from MPEG bitstream," *Pattern Recognition*, vol. 35, no. 12, pp. 2719 – 2726, 2002.
- [10] A. Albiol, M. J. Ch, F. A. Albiol, and L. Torres, "Detection of TV commercials," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
- [11] J. Wang, L. Duan, Z. Li, J. Liu, H. Lu, and J. Jin, "A robust method for TV logo tracking in video streams," in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2006.
- [12] X.-S. Hua, L. Lu, and H.-J. Zhang, "Robust learning-based TV commercial detection," in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2005.
- [13] L.-Y. Duan, J. Wang, Y. Zheng, J. S. Jin, H. Lu, and C. Xu, "Segmentation, categorization, and identification of commercial clips from tv streams using multimodal analysis," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2006.
- [14] A. Hampapur and R. Bolle, "Comparison of distance measures for video copy detection," in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2001.
- [15] T. C. Hoad and J. Zobel, "Fast video matching with signature alignment," in *Proc. of ACM Int.'l Workshop on Multimedia Information Retrieval (MIR)*, 2003.
- [16] J. Zobel and T. C. Hoad, "Detection of video sequences using compact signatures," *ACM Transactions on Information Systems*, vol. 24, pp. 1–50, January 2006.
- [17] D. A. Adjeroh, I. King, and M. C. Lee, "A distance measure for video sequences," *Computer Vision and Image Understanding (CVIU)*, vol. 75, no. 1, 1999.
- [18] S. H. Kim and R.-H. Park, "An efficient algorithm for video sequence matching using the modified hausdorff distance and the directed divergence," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 12, no. 7, 2002.
- [19] Y.-T. Kim and T.-S. Chua, "Retrieval of news video using video sequence matching," in *Proc. of Multimedia Modelling Conference*, 2005.
- [20] J. L.-t. A. Joly and N. Boujemaa, "INRIA-IMEDIA TRECVID 2008: Video copy detection," in *TREC Video Retrieval Evaluation Notebook*, 2008.
- [21] M.-C. Yeh and K.-T. Cheng, "Video copy detection by fast sequence matching," in *Proc. of ACM International Conference on Image and Video Retrieval (CIVR)*, 2009.
- [22] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua, "Scalable detection of partial near-duplicate videos by visual-temporal consistency," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2009.
- [23] M. Douze, H. Jégou, C. Schmid, and P. Pérez, "Compact video description for copy detection with precise temporal alignment," in *Proceedings of the 11th European conference on Computer vision: Part I*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 522–535. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1886063.1886103>

- [24] A. Shivadas and J. Gauch, “Real-time commercial recognition using color moments and hashing,” in *Proc. of Canadian Conference on Computer and Robot Vision (CRV)*, 2007.
- [25] X. Wu, N. Putpuek, and S. Satoh, “Commercial film detection and identification based on a dual-stage temporal recurrence hashing algorithm,” in *Proc. of International Workshop on Very-Large-Scale Multimedia Corpus, Mining and Retrieval (VLS-MCMR’10)*, 2010.
- [26] A. G. Hauptmann and M. J. Witbrock, “Story segmentation and detection of commercials in broadcast news video,” in *Proc. of Advances in Digital Libraries Conference*, 1998.
- [27] M. Bertini, A. Del Bimbo, and W. Nunziati, “Video clip matching using MPEG-7 descriptors and edit distance,” in *Proc. of ACM International Conference on Image and Video Retrieval (CIVR)*, 2006.
- [28] Y. Li, D. Zhang, X. Zhou, and J. S. Jin, “A confidence based recognition system for TV commercial extraction,” in *Proc. of ADC*, 2007.
- [29] D. Zhao, X. Wang, Y. Qian, Q. Liu, and S. Lin, “Fast commercial detection based on audio retrieval,” in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2008.
- [30] H. Duxans, D. Conejero, and X. Anguera, “Audio-based automatic management of TV commercials,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.
- [31] P. Duygulu, M.-Y. Chen, and A. G. Hauptmann, “Comparison and combination of two novel commercial detection methods,” in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2004.
- [32] N. Liu, Y. Zhao, and Z. Zhu, “Coarse-to-fine based matching for audio commercial recognition,” in *Proc. of International Conference on Neural Networks and Signal Processing (ICNNSP)*, 2008.
- [33] —, “Commercial recognition in TV streams using coarse-to-fine matching strategy,” in *Proc. of Advances in Multimedia Information Processing (PCM)*, 2010.
- [34] R. Mohan, “Video sequence matching,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.
- [35] A. Hampapur and R. Bolle, “Feature based indexing for media tracking,” in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2000.
- [36] K. M. Pua, J. M. Gauch, S. E. Gauch, , and J. Z. Miadowicz, “Real time repeated video sequence identification,” *Computer Vision and Image Understanding (CVIU)*, vol. 93, no. 3, 2004.
- [37] J. Oostveen, T. Kalker, and J. Haitsma, “Feature extraction and a database strategy for video fingerprinting,” in *Proc. of International Conference on Visual Information Systems (VISUAL)*, 2002.
- [38] Y. Li, J. Jin, and X. Zhou, “Matching commercial clips from TV streams using a unique, robust and compact signature,” in *Proc. of Digital Image Computing: Techniques and Applications (DICTA)*, 2005.
- [39] A. Sarkar, V. Singh, P. Ghosh, B. Manjunath, and A. Singh, “Efficient and robust detection of duplicate videos in a large database,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 20, no. 6, pp. 870 –885, June 2010.

- [40] D. DeMenthon and D. Doermann, "Video retrieval using spatio-temporal descriptors," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2003.
- [41] S. Satoh, M. Takimoto, and J. Adachi, "Scene duplicate detection from videos based on trajectories of feature points," in *Proc. of ACM Int.'l Workshop on Multimedia Information Retrieval (MIR)*, 2007.
- [42] X. Wu, M. Takimoto, S. Satoh, and J. Adachi, "Scene duplicate detection based on the pattern of discontinuities in feature point trajectories," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2008.
- [43] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2007.
- [44] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *Proc. of ACM International Conference on Image and Video Retrieval (CIVR)*, 2007.
- [45] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-hash and tf-idf weighting," in *Proc. of British Machine Vision Conference (BMVC)*, 2008.
- [46] E. Y. Chang, H. Bai, and K. Zhu, "Parallel algorithms for mining large-scale rich-media data," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2009.
- [47] C. Colombo, A. Del Bimbo, and P. Pala, "Retrieval of commercials by semantic content: The semiotics perspective," *Multimedia Tools and Applications*, vol. 13, no. 1, pp. 93–118, 2001.
- [48] Y. Zheng, L. Duan, Q. Tian, and J. Jin, "Tv commercial classification by using multi-modal textual information," in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2006.
- [49] J. Wang, L. Duan, L. Xu, H. Lu, and J. S. Jin, "TV ad video categorization with probabilistic latent concept learning," in *Proc. of ACM Int.'l Workshop on Multimedia Information Retrieval (MIR)*, 2007.
- [50] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of ACM Int.'l Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.
- [51] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, 2001.
- [52] B. Manjunath, J.-R. Ohm, and V. Vasudevan, "Color and texture descriptors," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 11, no. 6, June 2001.
- [53] E. Kasutani and A. Yamada, "The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, 2001.
- [54] P. H. Sellers, "The theory and computation of evolutionary distances: Pattern recognition," *Journal of Algorithms*, vol. 1, no. 4, pp. 359 – 373, 1980. [Online]. Available: <http://www.sciencedirect.com/science/article/B6WH3-4D7JN56-4T/2/2b859d72a68774d0a51959fda148c6b3>
- [55] F. Aldershoff and T. Gevers, "Visual tracking and localisation of billboards in streamed soccer matches," in *Proc. of SPIE*, 2004.
- [56] N. Ichimura, "Recognizing multiple billboard advertisements in videos," in *Proc. of IEEE Pacific-Rim Symposium*

on Image and Video Technology, December 2006.

- [57] A. D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, “Trademark matching and retrieval in sports video databases,” in *Proc. of ACM Int.’l Workshop on Multimedia Information Retrieval (MIR)*, 2007.
- [58] A. Watve and S. Sural, “Soccer video processing for the detection of advertisement billboards,” *Pattern Recognition Letters*, vol. 29, no. 7, 2008.
- [59] A. Joly and O. Buisson, “Logo retrieval with a contrario visual query expansion,” in *Proc. of ACM Multimedia (ACM MM)*, 2009.
- [60] J. Schietse, J. P. Eakins, and R. C. Veltkamp, “Practice and challenges in trademark image retrieval,” in *Proc. of ACM International Conference on Image and Video Retrieval (CIVR)*, 2007.
- [61] T. Kato, “Database architecture for content-based image retrieval,” *Proc. of SPIE Image Storage and Retrieval Systems*, 1992.
- [62] J. P. Eakins, J. M. Boardman, and M. E. Graham, “Similarity retrieval of trademark images,” *IEEE Multimedia*, vol. 5, no. 2, pp. 53–63, 1998.
- [63] W. H. Leung and T. Chen, “Trademark retrieval using contour-skeleton stroke classification,” in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2002.
- [64] C.-H. Wei, Y. Li, W.-Y. Chau, and C.-T. Li, “Trademark image retrieval using synthetic features for describing global shape and interior structure,” *Pattern Recognition*, vol. 42, no. 3, pp. 386–394, 2009.
- [65] M. Rusiñol and J. Lladós, “Efficient logo retrieval through hashing shape context descriptors,” in *Proc. of International Workshop on Document Analysis Systems*, 2010.
- [66] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 4, pp. 509–522, 2002.
- [67] M. Merler, C. Galleguillos, and S. Belongie, “Recognizing groceries in situ using in vitro training data,” in *Proc. of IEEE CVPR SLAM-Workshop*, June 2007.
- [68] Y. Jing and S. Baluja, “Pagerank for product image search,” in *Proc. of WWW*, 2008.
- [69] B. Kovar and A. Hanjalic, “Logo appearance statistics in a sport video: Video indexing for sponsorship revenue control,” in *Proc. of SPIE*, 2002.
- [70] R. J. M. Den Hollander and A. Hanjalic, “Logo recognition in video by line profile classification,” in *Proc. of SPIE*, 2004.
- [71] F. Pelisson, D. Hall, O. Riff, and J. L. Crowley, “Brand identification using gaussian derivative histograms,” *Machine Vision and Applications*, vol. 16, pp. 41–46, 2003.
- [72] R. Phan, J. Chia, and D. Androutsos, “Unconstrained logo and trademark retrieval in general color image database using color edge gradient co-occurrence histograms,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.
- [73] R. Phan and D. Androutsos, “Content-based retrieval of logo and trademarks in unconstrained color image

- databases using color edge gradient co-occurrence histograms,” *Computer Vision and Image Understanding (CVIU)*, vol. 114, no. 66–84, 2010.
- [74] J. Luo and D. Crandall, “Color object detection using spatial-color joint probability functions,” *IEEE Transactions on Image Processing (TIP)*, vol. 15, no. 6, pp. 1443–1453, 2006.
- [75] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [76] J. Kleban, X. Xie, and W.-Y. Ma, “Spatial pyramid mining for logo detection in natural scenes,” in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2008.
- [77] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool, “Efficient mining of frequent and distinctive feature configurations,” in *Proc. of ICCV*, 2007.
- [78] K. Gao, S. Lin, Y. Zhang, S. Tang, and D. Zhang, “Logo detection based on spatial-spectral saliency and partial spatial context,” in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, 2009.