Università degli Studi di Firenze
Corso di Dottorato in Ingegneria Informatica,
Multimedialità e Telecomunicazioni
Media Integration and Communication Center (MICC)
ING-INF/05

Object and event recognition in multimedia archives using local visual features

Candidate
Lamberto Ballan

Supervisors
Prof. Alberto Del Bimbo

Dr. Marco Bertini

PhD Coordinator
Prof. Giacomo Bucci

---

ciclo XXIII, 2008-2010

Università degli Studi di Firenze, Media Integration and Communication Center (MICC).

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Engineering, Multimedia and Telecommunication.

A mio padre

Acknowledgments

I would like to acknowledge the efforts and input of my supervisor, Prof. Alberto Del Bimbo, and all my colleagues of the Media Integration and Communication Center (MICC) who were of great help during my research. In particular my thanks go to Marco Bertini, Giuseppe Serra and Lorenzo Seidenari who collaborated on the main parts of my research work. I would like to thank also Dr. Hichem Sahbi for his kind support and hospitality during my stay at Télécom ParisTech.

Contents

Contents **v**

1 Introduction **1**
1.1 The objective 1
1.2 Contributions 2

2 Literature review **7**
2.1 Recognition of object instances 7
2.1.1 Local visual features 8
2.2 Recognition of object categories 11
2.2.1 Codebooks 12
2.3 Semantic video annotation 14
2.3.1 Actions and events 15
2.3.2 Spatio-temporal features 17
2.3.3 Classification of composite events 23
2.4 Ontologies 28

3 Trademark retrieval in sports video archives **33**
3.1 Introduction 33
3.2 Image and video features 36
3.3 Detection and retrieval of trademarks 37
3.4 Experimental results 40
3.4.1 Implementation 40
3.4.2 Test data and experiment design 43
3.4.3 Results 43
3.5 Conclusion 46

4 Context-dependent trademark matching and retrieval 49
4.1 Introduction . . . 49
4.2 Context-dependent similarity . . . 53
4.2.1 Context . . . 53
4.2.2 Similarity design . . . 55
4.2.3 Solution . . . 56
4.3 Logo detection and consistency . . . 57
4.3.1 Matching . . . 57
4.3.2 Logo detection . . . 58
4.3.3 Similarity invariance . . . 60
4.4 Benchmarking . . . 60
4.4.1 Test data and settings . . . 60
4.4.2 Performance, comparison and discussion . . . 61
4.5 Conclusion . . . 64

5 A SIFT-based forensic method for copy-move detection 65
5.1 Introduction . . . 66
5.2 SIFT Features for Image Forensics . . . 68
5.2.1 Our contribution . . . 70
5.3 The proposed method . . . 71
5.3.1 SIFT features extraction and multiple keypoint matching 72
5.3.2 Clustering and forgeries detection . . . 73
5.3.3 Geometric transformation estimation . . . 75
5.4 Experimental results . . . 77
5.4.1 Settings for forgery detection . . . 78
5.4.2 Test on multiple copied regions . . . 84
5.4.3 Test on a large dataset . . . 84
5.4.4 Image splicing . . . 90
5.5 Conclusion . . . 93

6 Video event classification using string kernels 95
6.1 Introduction . . . 96
6.2 Related works . . . 98
6.3 Event representation and classification . . . 100
6.3.1 Frame representation . . . 101
6.3.2 Video representation . . . 102
6.4 Classification using string kernels . . . 105

6.5 Experimental results 108
6.5.1 Experiment 1: characters distance and codebook size 111
6.5.2 Experiment 2: comparison with kNN classifier 112
6.5.3 Experiment 3: comparison with a traditional keyframe-based BoW approach 113
6.6 Conclusion 114

7 Effective codebooks for human action categorization 117
7.1 Introduction and previous work 118
7.2 Detector and descriptors 121
7.2.1 Detector 121
7.2.2 Descriptors 122
7.3 Action representation and categorization 124
7.3.1 Codebook formation 124
7.3.2 Codeword assignment 126
7.4 Experimental results 127
7.4.1 Evaluation of our descriptor 129
7.4.2 Performances obtained by effective codebooks 130
7.4.3 Comparison to state-of-the-art 132
7.5 Conclusion 133

8 Video annotation using ontologies and rule learning 135
8.1 Introduction 135
8.2 Related work 138
8.3 Automatic rule learning using first order logic 139
8.3.1 Improving performance 142
8.3.2 Rule learning example 143
8.4 Experimental results 145
8.5 The Sirio web-based search engine 149
8.6 Conclusion 151

9 Conclusion 153
9.1 Summary of contribution 153
9.2 Directions for future work 155

A Appendix 157
A.1 Proof of proposition 2 in Section 4.3.1 157

B Publications **161**

Bibliography **165**

Chapter 1

Introduction

The *digital revolution* has converted old, analog technologies into a digital format. The sweeping changes brought about by digital computing and communication technology during the latter half of the 20th century, have opened exciting new challenges [157]. In this context, due to the widespread availability of personal and professional imaging devices, the low cost of multimedia storage and ease of content transmission and sharing, the need to automatically analyze and organize large amounts of visual data becomes more and more prominent.

1.1 The objective

While data processing capabilities of machines are truly impressive if compared to a human, data interpretation skills are very poor. Just as an example, considering an archive of sports videos, retrieving the winning goal from the 2010 Word Cup is a hard task requiring human intervention when the data has not been manually annotated.

Our goal is to enable *visual search* of videos and image collections. This task may consist of determining whether the visual data contains some specific property, object or activity (Figures 1.1, 1.2, 1.3 and 1.4). Nowadays, currently available text-based image and video search engines (such as Google and Bing) are based on queries that are mainly given by few textual words, and images and videos are searched based on their metadata and surrounding text. Although these systems are (relatively) effective and powerful, their potential is limited by the information given in the text. Moreover

Figure 1.1: Example search for the query "panorama" with the particular property of having red color.

this textual information is often imprecise, ambiguous and overly personalized (think for example at social websites for media sharing like YouTube and Flickr). Moreover, systems that enable fast, automatic visual search of images and videos are very appealing because of their relevance to many real applications, from semantic image and video indexing to intelligent video-surveillance and advanced human-computer interaction. For these reasons, content-based image/video retrieval and understanding receives a lot of attention from both the scientific community and industry. Although a lot of work has been done [203, 55], this task remains very challenging. It is mainly due to the fact that machines can only compute low level properties of data that have no clear relation with high level conceptual semantics. This is a well-known problem in the literature, and it has been formalized as the *semantic gap*:

> *"The semantic gap is the lack of coincidence between the information that machines can extract from the visual data and the interpretations the user may give to the data."*

Therefore, the problem of *visual search and recognition* could be summarized in the question: "how can we bridge the semantic gap for image and video understanding?".

1.2 Contributions

We present in this thesis a step-by-step methodology to reduce the semantic gap and to achieve automatic annotation and retrieval of visual content. The contribution of this research work is divided into two main themes. The first

Figure 1.2: Example search for a particular object instance (in this case the logo "Starbucks").

Figure 1.3: Example search for a particular object category (in this case "airplane").

Figure 1.4: Example search for a particular activity (in this case the action "shake hand").

one is related to the recognition of objects in images and videos, while the second one to the recognition of dynamic concepts such as actions and activities. Although these two main themes can be individuated, the common idea of our work is the usage of *local visual features*. Local representations

are a robust and general solution since they are able to describe the visual observations as a collection of independent local patches. Objects and events are fully represented by the appearance of hundreds of local visual features. This approach is well suited in generic real-world scenarios, since it exhibits a strong robustness against several geometrical transformations, such as rotation and scaling, partial occlusions and clutter. This basic representation is then extended by introducing spatial and temporal constraints between local features in order to obtain more complete information and to manage ambiguity. A formal representation of knowledge is introduced by a multimedia ontology that connects visual observations (the *perceptual level*) to linguistic and abstract concepts and relations (the *semantic level*). The structure of the ontology itself, together with reasoning, can be used to perform higher-level annotation of the visual data, to generate complex queries that comprise concepts, their relations and temporal evolutions, and to create extended text commentaries.

The rest of the thesis is organized as follows[1]. We start in Chapter 2 with a survey of related work on object and event recognition in multimedia archives. Special attention is given to visual recognition approaches using local representations, which forms the basis for our work.

Chapters 3 and 4 deal with the particular problem of detection and retrieval of trademarks in images and videos. First, we propose a real system for logo recognition in large sports video archives (Chapter 3). Here classification of trademarks is performed by matching a set of SIFT feature descriptors [137] for each trademark instance against the set of features detected in each frame of the sport video. Localization is performed through robust clustering of matched feature points in the video frame. Experimental results are provided, along with an analysis of the precision and recall in a real application scenario. In Chapter 4, we extend this retrieval approach by introducing a robust *context-dependent* similarity measure between local descriptors. This measure takes into account not only the intrinsic visual features but also their context and spatial configuration. The main contribution of this work includes: *i*) a variational framework which makes it possible to design our similarity as the fixed point of an energy function mixing a visual "data term", a "context criterion" and a "regularization term"; *ii*) a theoretical study of the consistency of logo matching/detection and its invariance

[1]Note that each chapter is written in a self-contained fashion and can be read on its own.

to different transformations including similarity and occlusion. The validity of this method is shown through extensive experiments on a challenging logo image dataset.

Chapter 5 builds on ideas from Chapter 3 to address the problem of image forensics; to be more precise, it consists in determining if a particular image is authentic or not, which could be a crucial task when images are presented as basic evidence to influence judgment (e.g. in a court of law). Special attention has been paid to the case in which an area of an image is copied and then pasted onto another zone to make a duplication or to cancel something that was awkward (*copy-move* attack). The proposed method can determine if such tampering has occurred and which image patches are involved, and to recover which geometric transformation was used to perform cloning.

In Chapter 6 we move to video event recognition problem. Events are modeled as a sequence of histograms of "static" visual features (e.g. SIFT), computed from each video frame. The sequences are treated as strings where each histogram is considered as a character. Event classification of these strings of variable length, depending on the duration of the video clips, are performed using SVM classifiers with a novel string kernel (based on the Needlemann-Wunsch edit distance [156]). In other words, the basic idea of the approach is to represent the dynamics of the event by collecting the visual appearance in each video frames through the time. Experiments have been performed on two different domains: soccer and TRECVID 2005 news videos.

Chapter 7 focuses on categorization of human action classes from video collections. Automatic human activity recognition methods are useful in many real applications, in particular in videosurveillance scenarios. To this end, we first define a novel 3D spatio-temporal gradient descriptor that, combined with optic flow, outperforms the state-of-the-art without requiring fine parameter tuning. Second, we introduce a more effective codebook model by applying a radius-based clustering method and a soft assignment that considers the information of two or more relevant codeword candidates. We extensively test our approach on standard KTH and Weizmann datasets showing its validity and outperforming several recent approaches.

In Chapter 8 we present a novel rule-based approach to describe and recognize composite concepts and events. Our method automatically learns rules expressed in Semantic Web Rules Language (SWRL), exploiting the

knowledge embedded in a multimedia ontology. The relationship between concepts, their co-occurrence and the temporal consistency of video data are used to improve the performance of individual concept detectors.

Chapter 9 summarizes the contribution of the thesis and discusses avenues for future research. Note also that the full-list of published papers from this thesis is given in Appendix B.

Chapter 2

Literature review

This chapter gives a brief survey of related work on object and event recognition using local visual features. The first part of the chapter roughly introduces the problem of object recognition in image archives, while the second part deals with the problem of semantic video annotation. Finally, multimedia ontologies have been presented as a formal tool to enrich the semantic image/video annotation or to derive new knowledge. [1]

2.1 Recognition of object instances

The goal of object recognition in visual archives is to detect the presence of a particular object in an image, and possibly localize the object in the image and estimate its pose. This usually involves designing an object representation that can model the imaged appearance of an object under a broad class of imaging conditions, such as varying object and camera pose, scene lighting, partial occlusion and deformation. Such representation should also be robust enough to deal with large amounts of background clutter.

[1] The part of this chapter related to semantic video annotation has been published as "Event detection and recognition for semantic annotation of video" in *Multimedia Tools and Applications (Special Issue: Survey Papers in Multimedia by World Experts), vol. 51, iss. 1, pp. 279-302, 2011* [17].

2.1.1 Local visual features

While early works made strong simplification about the real world by using mostly 3D geometric object models and geometric invariants (see for example *block world* by Roberts [181] and [136]) or, more generally, global appearance models, recent works in object recognition often use local appearance models. The main reason of this trend is that global representations suffer from too simplistic assumption about object appearances and they also present problems with partial occlusion and background clutter (due to the global appearance representation used). For these reasons, this kind of models are often unsatisfactory in real-world generic scenarios.

Local representations describe the visual observations as a collection of independent local patches. These methods are at the heart of some of the most successful object instance recognition systems to date. The main idea of this kind of approaches is that objects are represented by the appearance of hundreds of local visual features. First, a database of objects is built by storing local appearances in the form of a feature vector (descriptor). In recognition, local regions are extracted from a test image and matched to the object database using their appearance descriptors. This initial set of local region matches is then disambiguated using semi-local or global geometric constraints (e.g. if the object is planar, we can use the constraint that all local features must be mapped by a planar homography).

Local visual features (e.g. SIFT, SURF, GLOH, etc.) have been widely used for the particular tasks of image retrieval and object recognition, due to their robustness to several geometrical transformations (such as rotation

(a) (b) (c)

Figure 2.1: Object recognition example from the original work of Lowe [137]. (a) Model images for two objects; (b) Test image; (c) Recognition result.

and scaling), occlusions and clutter. Most of the algorithms proposed in the literature for detecting and describing local visual features usually requires two steps. The first is the detection step, in which interest points are localized, while in the second step robust local descriptors are built so as to be invariant with respect to orientation, scale and affine transformations. A comprehensive analysis of several local descriptors is provided in [149] while local affine region detectors are surveyed in [150]. These works confirm that SIFT feaures [137] are a good solution because of their high performance and relatively low computational costs. In the following we report, as an example of these methods, a brief summary of the SIFT algorithm; for more details (obviously) refer to the original paper.

Scale-Invariant Feature Transform (SIFT)

This method can be roughly summerized as the following four steps: *i)* scale-space extrema detection; *ii)* keypoint localization; *iii)* assignment of one (or more) canonical orientation; *iv)* generation of keypoint descriptors.

In other words, given an input image I, SIFT features are detected at different scales by using a scale-space representation implemented as an image pyramid. The pyramid levels are obtained by Gaussian smoothing and subsampling of the image resolution while interest points are selected as local extrema (min/max) in the scale-space. These keypoints, also referred as $\mathbf{x}_i$ in the following, are extracted by applying a computable approximation of the Laplacian of Gaussian (LoG) called Difference of Gaussians (DoG). Specifically, a DoG image D is given by: $D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma)$, where $L(x, y, k\sigma)$ is the convolution of the original image $I(x, y)$ with the Gaussian blur $G(x, y, k\sigma)$ at scale $k\sigma$.

In order to guarantee invariance to rotations, the algorithm assigns to each keypoint a canonical orientation o. To determine this orientation, a gradient orientation histogram is computed in the neighborhood of the keypoint. Specifically, for an image sample $L(x, y, \sigma)$ at scale σ (the scale in which that keypoint was detected), the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are precomputed using pixel differences:

$$m(x,y) = \Big(\big((L(x+1,y) - L(x-1,y)\big)^2 + \\ + \big((L(x,y+1) - L(x,y-1)\big)^2 \Big)^{1/2}, \tag{2.1}$$

$$\theta(x,y) = tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right). \tag{2.2}$$

An orientation histogram with 36 bins is formed, with each bin covering approximately 10 degrees. Each sample in the neighboring window added to a histogram bin is weighted by its gradient magnitude and by a Gaussian-weighted circular window with σ equal to 1.5 times respect to the scale of the keypoint. The peaks in this histogram correspond to dominant orientations. Once these keypoints are detected, and canonical orientations are assigned, SIFT descriptors are computed at their locations in both image plane and scale-space. Each feature descriptor consists in a histogram $\mathbf{f}$ of 128 elements, obtained from a 16×16 pixels area around the corresponding keypoint. This area is selected using the coordinates (x, y) of the keypoint as the center and its canonical orientation as the origin axis. The contribution of each pixel is obtained by accumulating image gradient magnitude, $m(x, y)$, and orientation, $\theta(x, y)$, in scale-space and the histogram is computed as the local statistics of gradient orientations (considering 8 bins) in 4×4 sub-patches (see Fig. 2.2).

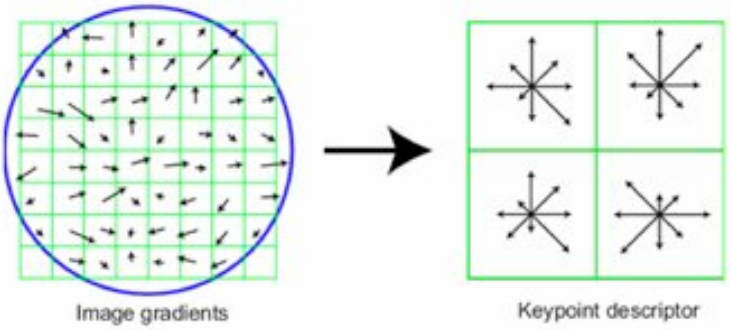


Figure 2.2: Image gradients within a patch (left) are accumulated into a coarse 4×4 spatial grid (right, only a 2×2 grid is shown). A histogram of gradient orientations is formed in each grid cell. 8 orientation bins are used in each grid cell giving a descriptor of dimension $128 = (4 \times 4 \times 8)$.

Summarizing the above, given an image I, this procedure ends with a list of N keypoints each of which is completely described by the following informations:

$$\mathbf{x}_i = \{x, y, \sigma, o, \mathbf{f}\}, \tag{2.3}$$

where (x, y) are the coordinates in the image plane, σ is the scale of the keypoint (related to the level of the image-pyramid used to compute the de-

Figure 2.3: Examples of successful face detections by the Schneiderman and Kanade detector [218].

scriptor), o is the canonical orientation (used to achieve rotation invariance) and $\mathbf{f}$ is the final SIFT descriptor.

2.2 Recognition of object categories

The recognition of object categories in images is a challenging problem in computer vision, especially when the number of categories is large. The challenge is that appearance variations among instances of an object class have to be modeled, in addition to standard problems of viewpoint and lighting changes and partial occlusion. Early works in this field usually addressed the problem of detecting very particular categories such as *faces* or *cars* (Fig. 2.3 shows an example of the results obtained using the Schneiderman and Kanade detector [193]). Often these methods involve training a sliding window classifier, which for a small image patch (e.g. 24×24 pixels) decides whether the desired object (e.g. a face) is present or not [218].

Given the big success of local representations for recognition of object instances, approaches based on local visual features have been extended in the object and scene categorization scenario by exploiting the Bag-of-Words (BoW) model [201]. The BoW in document retrieval (in particular in natural language processing) is a popular method for representing documents, which ignores the word orders. For example, "a good work" and "work good a" are the same under this model. The BoW model allows a dictionary-based modeling, and each document looks like a *bag* which contains some words from the dictionary. In the visual case, an image can be treated as a document, and visual features extracted from the image are considered as the *visual words*. Through the use of this intermediate description (the code-

book), images can be represented very compactly. The codebook is usually obtained with a vector quantization procedure exploiting some clustering algorithm (such as k-means). This intermediate description allows both fast data access, by building an inverted index [201, 164], and generalization over category of objects by representing each instance as a composition of common parts [72]. As in the textual counterpart, the bag of visual words does not retain any structural information and so, by using this representation, we actually do not care where regions occur in an image. As this comes with some advantages, like robustness to occlusions and generalization over different object and scenes layouts, there is also a big disadvantage in discarding completely image structure, since this actually removes all spatial information. A local visual words spatial layout description [189] can recover some image structure without loss of generalization power. A global approach has been proposed by Lazebnik *et al.* [124]; in their work structure is added in a multi-resolution fashion by matching spatial pyramids obtained by subsequently partitioning the image and computing bag-of-words representations for each of the sub-image partition.

2.2.1 Codebooks

Visual features are usually quantized by applying some clustering algorithm, to obtain a visual dictionary (usually called *codebook*) that is used to represent and classify categories of visual concepts [163, 72, 201, 97]. Most of the methods use the k-means algorithm for codebook creation, because of its simplicity and convergence speed. However, local representations with code-

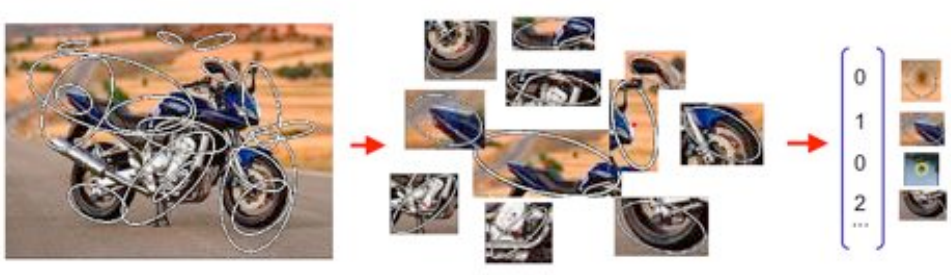


Figure 2.4: Illustration of the bag-of-visual-words model. An image is represented by a histogram of quantized local features (visual words). Note that all spatial relations between regions are lost.

books also have shown limitations. On the one hand, the k-means clustering method, despite of its popularity, is not very robust w.r.t. outliers and the number of codewords has to be known in advance, requiring an empirical evaluation of this number. Jurie and Triggs [100] have also shown that with k-means, due to the simplistic assumption of uniform distribution of the features in the descriptor space, cluster centers are selected almost exclusively around the denser regions in the descriptor space and more sparsely elsewhere, thus failing to code other informative regions. They have proposed a radius-based clustering to improve the quality of the codebook. Another critical point of bag-of-features approaches is related to the high dimensionality of the codebooks and so to their computational requirements. Reducing the codebook size allows to better fit a real-time scenario reducing action recognition time. A simple means of codebook size reduction is to use a lower number of clusters (or a bigger radius in radius-based clustering). However this approach compromises descriptor statistics by increasing quantization errors, visual words uncertainty and word discriminative power. A more principled approach is the use of Principal Component Analysis (PCA) [104,227]. In the field of text retrieval, a common technique used to project high dimensional features in a "semantic space" is Latent Semantic Indexing (LSI) [56] which, by applying Singular Value Decomposition (SVD) to the document/term matrix, attempts at finding the best linear subspace where to project all document vectors.

Moreover, the traditional codebook approach represents an image by a histogram of codeword frequencies obtained through a *hard-assignment* between features and codewords. In other words, for each codeword w in the vocabulary V, the frequency distribution in an image is computed by:

$$H(w) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 1 & \text{if } w = \underset{v \in V}{\operatorname{argmin}}(D(v, p_i)); \\ 0 & \text{otherwise;} \end{cases} \tag{2.4}$$

where n is the number of local visual features, p_i is the i^{th} features, and $D(v, p_i)$ is the distance (usually Euclidean) between the codeword v and p_i. This hard assignment, that takes account only of the closest codeword, lacks to consider two issues: codeword *uncertainty* (selection of the correct codeword when two or more candidates are relevant) and codeword *plausibility* (selection of a codeword when all codewords are too far and not representative) [215]. For this reason the distribution of the codewords in a sequence

has to contain the information of two or more relevant candidates. This can be done by applying a *soft-assignment* such as, for example, the one proposed by van Gemert *et al.* [215]; it consists in smoothing the assignment of local features to the codebook using Gaussian kernel density estimation. A similar idea has been proposed also in [172].

Finally, given the extremely high success of codebook-based solutions for visual categorization and retrieval, several researchers are trying to scale this kind of approaches to extremely large visual archives (e.g. using GPUs) [214, 211].

2.3 Semantic video annotation

Semantic annotation of video content is a fundamental process that allows the creation of applications for semantic video database indexing, intelligent surveillance systems and advanced human-computer interaction systems.

Several surveys on semantic video annotation have been recently presented in the literature. A review of multi-modal video indexing was presented in [206], considering entertainment and informative video domains. Multi-modal approaches for video classification have been surveyed in [41]. A survey on event detection has been presented in [123], focusing on modeling techniques; our work [17] extends this, providing also a review of low-level features suitable for event representation, like detectors and descriptors of interest points, as well as a review of knowledge representation tools like ontologies. A survey on behavior recognition in surveillance applications has been provided in [113], while in [176] are reported the most recent works on human action recognition.

Typically videos are automatically segmented in shots and a representative keyframe of each shot is analyzed to recognize the scene and the objects shown, thus treating videos like a collection of static images and losing the temporal aspect of the media. This approach is not feasible for the recognition of events and activities, especially if we consider videos that have not been edited and do not contain shots. Recognising the presence of concepts that have a temporal component in a video sequence, if the analysis is done using simply a keyframe, is difficult [221] even for a human annotator, as shown in Fig. 2.5. A revision of the TRECVid 2005 ground truth annotation of 24 concepts related to events and activities has shown that 22% of

the original manual annotations, performed inspecting only one keyframe per shot, were wrong [106]. An event filmed in a video is related to the temporal aspect of the video itself and to some changes in the properties of the entities and scenes represented; therefore there is need of representing and modeling time and properties' variations, using appropriate detectors, feature descriptors and models.

(a) (b) (c)

Figure 2.5: Keyframe-based video event recognition. (a) Is it *shot-on-goal* or *placed-kick*? (b) Is the person *entering* or *exiting* in/from the car? (c) Is the aircraft *landing* or *taking-off* ?

Recently, the problem of the detection and recognition of events and activities is getting a larger attention also within the TRECVid evaluation. The high-level concept detection task of TRECVid 2009 [167] considered the problem of event detection, with 7 out of 20 high-level concepts to be detected that were related to events and actions [48]. The most recent approaches proposed in this task have started to cope with the problem of representing videos considering the temporal aspects of it, analyzing more than one keyframe per shot and introducing some representation of the context [167, 234]. Since 2008 a new dataset of airport surveillance videos, to be used in a event detection task, has been added to the TRECVid evaluation campaign; the dataset focuses mostly on crowd/group actions (e.g. people meeting), human gestures (e.g. person running) and human activities (e.g. putting an object somewhere).

2.3.1 Actions and events

We refer to events as concepts with a dynamic component; an *event* is "something happening at a given time and in a given location". In the video analysis community the event recognition task has never been tackled by

proposing a generic automatic annotation tool and the proposed approaches are usually domain dependent. Video domains considered in this survey are broadcast news, sports, movies, video-surveillance and user generated content. Videos in the broadcast news, sports and movies are usually professionally edited while video-surveillance footage and user generated content are usually unedited. This editing process adds a structure [206] which can be exploited in the event modeling (as explained in the following Sections 2.3.3 and 2.4). Automatic annotation systems are built so as to detect events of interest. Therefore we can firstly split events in *interesting* and *non-interesting*; in the case of video-surveillance interesting events can be specific events such as "people entering a prohibited area", "person fighting" or "person damaging public property", and so on; sometimes defining a-priori these dangerous situations can be cumbersome and, of course, there is the risk of the non exhaustivity of the system; therefore it can be useful to detect *anomalous* vs. *non-anomalous* (i.e. normal) events [196, 143]. In this case an event is considered interesting without looking at its specific content but considering how likely is given a known (learnt) statistics of the regular events. Also in the sport domain the detection of rare events is of interest, but systems need to detect events with a specific content (typically called *highlights*, [32]) such as "scoring goal", "slam dunk", "ace serve", etc. Most of the domains in which video-analysis is performed involve the analysis of human motion (sports, video-surveillance, movies). Events originated by human motion can be of different complexity, involving one or more subjects and either lasting few seconds or happening in longer timeframes. *Actions* are short task oriented body movements such as "waving a hand", or "drinking from a bottle". Some actions are atomic but often actions of interest have a cyclic nature such as "walking" or "running"; in this case detectors are built to recognise a single phase of it. Actions can be further decomposed in *action primitives*, for example the action of running involves the movement of several body limbs [74]. This kind of human events are usually recognized using low-level features, which are able to concisely describe such primitives, and using per-action detectors trained on exemplar sequences. A main difficulty in the recognition of human actions is the high intra-class variance; this is mainly due to variation in the appearance, posture and behaviour (i.e. "the way in which one acts or conducts oneself") of the "actor"; *behaviour* can thus be exploited as a biometric cue [102].

Events involving multiple people or happening in longer timeframes can

be referred as *activities* [176]. Activity analysis requires higher level representations usually built with action detectors and reasoning engines. Events can be defined activities as long as there is not excessive inter-person occlusion and thus a system is able to analyse each individual motion (typically in sequences with two to ten people). In case of presence of a large amount of people, the task is defined as *crowd analysis* [240]: persons are no more considered as individuals but the global motion of a crowd is modelled [147]. In this case the detection of anomalous events is prominent because of its applicability to surveillance scenarios and because of the intrinsic difficulty of precisely defining crowd behaviours. *Human actions* are extremely useful in defining the video semantics in the domains of movies and user generated content. In both domains the analysis techniques are similar and challenges arise mainly from the high intra-class variance. Contextual information such a static features or scene classifiers may improve event recognition performance [146, 132, 91].

2.3.2 Spatio-temporal features

Recognition of events in video streams depends on the ability of a system to build a discriminative model which has to generalise with respect to unseen data. Such generalization is usually obtained by feeding state-of-the art statistical classifiers with an adequate amount of data. We believe that the key to solve this issue is the use of sufficiently invariant and robust image descriptors. While tackling a problem such as single-object recognition (i.e. find instances of "this object" in a given collection of images or videos) image descriptors are required to yield geometric and photometric invariance in order to match object instances across different images, possibly acquired with diverse sensors in different lighting environment and in presence of clutter and occlusions. An elegant way of dealing with clutter, occlusion and viewpoint change is the use of region descriptors [137,149]; image regions can be normalized [150] to obtain invariance to deformations due to viewpoint change, other normalization can be applied to obtain rotation and partial photometric invariance [137].

This kind of description has been extended in the object and scene categorization scenario exploiting the bag-of-words framework (as previously introduced). Given the success of bag of keypoints representations in static concept classification, efforts have been also made to introduce this frame-

work in event categorization. The first attempt in video annotation has been made by Zhou *et al.* [243], describing a video as a bag of SIFT keypoints. Since keypoints are considered without any spatial or temporal location (neither at the frame level) it is possible to obtain meaningful correspondences between varying length shots and shots in which similar scenes occur in possibly different order. Again, the structure is lost but this allows a robust matching procedure. Anyway temporal structure of videos carries rich information which has to be considered in order to attain satisfactory video event retrieval results. This information can be recovered using sequence kernels, as reviewed in Sect. 2.3.3. A different temporal information lies at a finer grained level and can be captured directly using local features. This is the case of gestures, human actions and, to some extent, human activities. Since gestures and actions are usually composed of *action primitives*, which occur in a short span of time and involve limb movements, their nature is optimally described by a local representation.

As in static keypoint extraction frameworks, the approach consists of two stages, detection and description. The detection stage aims at producing a set of "informative regions" for a sequence of frames, while the goals of the description stage are to gain invariance with respect to several region transformations caused by the image formation process, and to obtain a feature representation that enables matching through some efficiently computable metric.

Detectors

Space-time interest points located by detectors should contain information on the objects and their motion in the world. Detectors are thus functions computed over the image plane and over time that present higher values in presence of local structures undergoing non-constant motion. These structures in the image should correspond to an object part that is moving in the world. Since they deal with dynamic content they need to be robust to motion generated by camera movements; these noisy detections have to be filtered without damaging detector ability to extract interesting image structures.

Local dynamic representations have been mostly derived directly from their static counterparts [118, 228, 226, 166] while the approaches presented in [59, 48] are explicitly designed for space-time features. Laptev extended

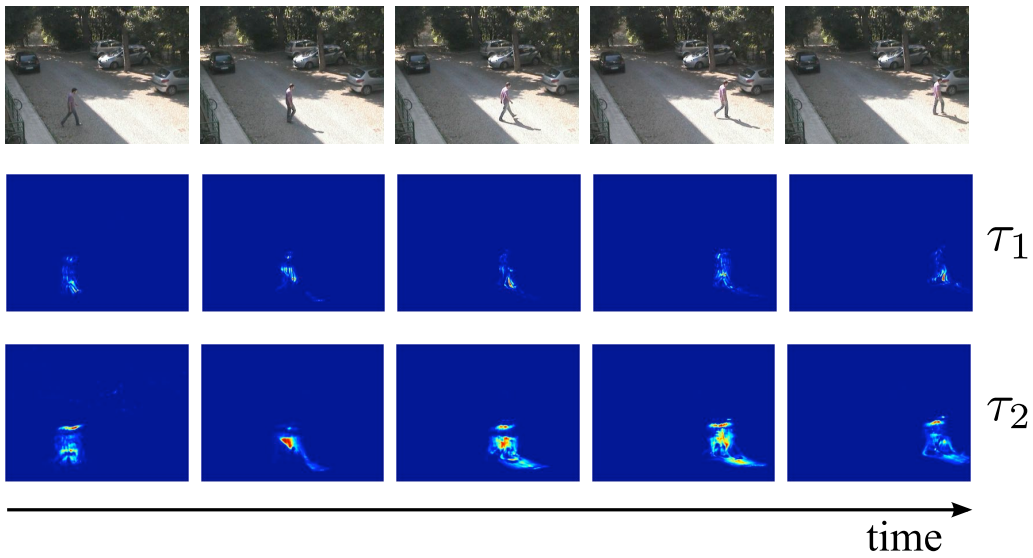


Figure 2.6: Spatio-temporal interest point detector [16] running at different temporal scales (blue low response, red high response); first row: original video frames, second row detector response at temporal scale τ_1 (mostly due to the limbs), third row: detector response temporal scale τ_2 (mostly due to the torso), with $\tau_1 < \tau_2$. Frames taken from the ViSOR video repository [217].

Harris corners keypoints to the space-time domain [118]; space-time corners are corner-like structures undergoing an inversion of motion. Wong *et al.* employed a difference-of-Gaussian operator on space-time volumes, after a preprocessing with non-negative matrix factorisation, in order to exploit the global video structure. Willems extended the SURF [22] detector using box filters and integral videos in order to obtain almost real time feature extraction; finally, the saliency measure originally proposed by Kadir and Brady [101] have been extended by Oikonomopoulos *et al.* [166]. The detector proposed by Dollár *et al.* [59] separates the operator which process the volume in space and time; the spatial dimension is filtered with a Gaussian kernel while the temporal dimension is processed by Gabor filters in order to detect periodic motion. A similar approach, specifically designed for the spatio-temporal domain, has been proposed by Chen *et al.* [48], which exploits a combination of optical flow based detectors with the difference of Gaussian detector used by SIFT.

Region scale can be selected by the algorithm [118,226,228] both in space

and time or may simply be a parameter of it [120, 59]; moreover scale for space and time can be fixed as in [59] or a dense sampling can be performed to enrich the representation [120,16]. Fig. 2.6 shows an example of the response of the detectors presented in [16], applied to the video surveillance domain. All the above approaches model the detector as an analytic function of the frames and scales, some other approaches instead rely on learning how to perform the detection using neural networks [109] or extending boosting and Haar features used for object detection [218]. Kienzle *et al.* trained a feed-forward neural network using, as a dataset, human eye fixations recorded with an headmounted tracker during the vision of a movie.

Recent detectors and approaches lean toward a denser feature sampling, since in the categorisation task a denser feature sampling yields a better performance [165]. State-of-the art image classifiers are, by now, performing feature sampling over regular multi-scale overlapped grids. This kind of approach is probably still too computational expensive to be performed on a sequence composed of hundred of frames. Finally, to the end of extracting as much information as possible, multiple feature detectors, either static or dynamic, have been used in conjunction [146, 132, 151].

Descriptors

The regions extracted by detectors need to be represented compactly. Descriptors are usually computed using a common pipeline as outlined in [227] for static features and, partially, in [119] for dynamic ones: preprocessing, non-linear transformation, pooling and normalisation. The preprocessing stage is usually a smoothing operation performed using a 3-dimensional Gaussian kernel [118, 111]. In order to obtain more robust descriptors a region normalisation can be applied [118]; the normalisation procedure proposed by Laptev attempt to obtain camera-motion invariant regions in order to increase the matching procedure reliability. Regions are transformed by computing an image measurement; typical choices are: normalised brightness [59], image gradients [118], spatio-temporal gradients [111, 59, 16, 195] and optical flow [16, 59, 118]. Gradients are used to provide photometric invariance, 3-dimensional gradients are capable of representing appearance and motion concisely. Optical flow descriptors can offer very informative low dimensional representations in case of smooth motion patterns, but in presence of noise the performance may degrade. Even if both carry mo-

tion information these two descriptions have been found to be complementary [16] and the fusion is beneficial for recognition. After computing this region transformation, the descriptor size is still very high dimensional and there is no invariance to small deformations (due for example to viewpoint change). Typically either global [119, 59] or local [16, 195, 111] histograms of gradient/optical flow orientation are computed. The use of local statistics contribute to obtain invariance to little viewpoint changes. A simpler approach is to apply PCA to the concatenated brightness, gradient or optical flow values [119, 59]. A different technique is to compute higher order derivatives of image intensity values [118]. Finally, following the approach of SIFT a descriptor normalisation and clipping can be applied to obtain robustness w.r.t. contrast change [111]. As shown in [227], for static feature descriptors, parameters can be learnt instead of "handcrafted"; Marszalek *et al.* performed such an optimisation by training on datasets [146]. This technique shows an improvement over the handcrafted values but it is also shows sensitivity to data: descriptors trained over Hollywood movies[2] dataset does not perform as well on videos of the KTH dataset [3] and vice-versa. Fig. 2.7 shows sample frames of these two datasets.

(a)

(b)

Figure 2.7: Sample frames from actions in KTH (a) and Hollywood (b) datasets.

[2]http://www.irisa.fr/vista/actions/
[3]http://www.nada.kth.se/cvap/actions/

Action representation

Actions can be represented as a collection of space-time pixel neighbourhoods descriptors. Statistical classification frameworks require an instance-to-instance or an instance-to-class matching procedure. Local feature matching can be done using simple metrics such as the Euclidean distance and exploiting [137] nearest neighbour distances to remove outliers. This technique is highly effective in the single-object recognition task but can deliver poor performance when generalisation power is needed as in a category recognition problem. As in object category recognition the intermediate codebook representation can offer together generalisation power and dimensionality reduction; in fact features which are often high dimensional (200+) are replaced with a code corresponding to a visual word in the dictionary. As stated previously bag-of-words representations completely lack any notion of the global features layout or their correlations. In action representation the visual words are often associated with an action primitive such as "raising an arm" or "extending a leg forward" and their spatio-temporal dependence is a strong cue. These relations can be modelled in the codebook formation [195, 133] or encoded in the final action representation [188, 229, 162, 151]. Scovanner *et al.* [195] have grouped co-occurring visual words to capture spatio-temporal feature correlations. Liu *et al.* have acted similarly on the dictionary by iteratively grouping visual words that maximise the mutual information. Niebles *et al.* [162] and Wong *et al.* [229] exploited graphical models to introduce a structural representation of the human action by modelling relations among body parts and their motion. Savarese *et al.* [188] augmented the action descriptor by computing visual words spatio-temporal correlograms instead of a flat word-count. Finally Mikolajczyk and Uemura [151] exploited vocabulary forest together with a star-shape model of the human body to allow localisation together with recognition. All these structural representations deal with relations between the feature themselves and are suitable in the analysis of isolated actions or behaviours. In the case of unconstrained scenarios, global layout representation can be a better choice [121, 120, 70]. The main advantage is their reduced computational cost. Moreover their coarse description can deal better with a higher intra-class variation. These approaches split the video volume with a coarse spatio-temporal grid, which can have a uniform [121, 70] or non-uniform layout [120], and by binning features in space and time, position dependent feature statistics is computed.

2.3.3 Classification of composite events

Events that are characterised by complex or composite evolution are often modelled by using a mid-level representation of the particular domain which eases the event recognition. Therefore many works try to build classifiers that are able to characterise the evolution and the interaction of particular visual features. These kinds of representations are often used in specific domains (for example in sports videos), where it is easier to define "in advance" the relations among visual features. Several different techniques have been proposed in the literature for this purpose: simple heuristic rules, finite state machines, statistical models (such as HMM or Bayesian networks) and kernel methods.

Heuristic rules and Finite State Machines

Several works in the sports video domain apply heuristics or rule-based approaches to automatically recognise simple events. An example is given by Xu *et al.* [233] in which recognition of play/break events of soccer videos is performed using classification of simple and mutually exclusive events (obtained by using a simple rule-based approach). Their method is composed by two steps; in the first step they classify each sample frame into global, zoom-in and close-up views using an unique domain-specific feature, grass-area-ratio. Then heuristic rules are used in processing the sequence of views, and obtain play/break status of the game.

More complex events can be recognised using Finite State Machines (FSMs). The knowledge of the domain is encoded into a set of FSMs and each of them is able to represent a particular video event. This approach was initially proposed by Assfalg *et al.* in [8] to detect the principal soccer highlights, such as shot on goal, placed kick, forward launch and turnover, from a few visual cues, such as playground position, speed and camera direction, etc. The idea of applying FSMs to model highlights and events has been recently followed also in [14]; scored goal, foul and generic play scenes in soccer videos have been modeled using four types of views (e.g. in-field, slow motion, etc.) for the states of the FSMs and transitions are determined by some audio-visual events such as the appearance of a caption or the whistle of the referee. Experiments have been performed using a set of manually annotated views and audio-visual events.

Markovian models

Visual events that evolve in a predictable manner are suitable for a Markovian modelling, and thus they can be detected by HMMs. Sports videos, in particular those that have a specific structure due to the rules like baseball and tennis, have been analysed using HMMs for event classification. If the events always move forward then a left-to-right model may be more suitable; in other cases, if the meaning of the states is not tangible it is better to choose a model with a sufficient number of states. A fully connected (ergodic) model is more suited for unstructured events. The feature set needs to capture the essence of the event, and features have to be chosen depending on the events being modelled. In general the steps that have to be followed when using HMMs for event classification/recognition [88] is to check if a "grammar" of the events is identifiable: this helps to identify if HMMs can model events directly or if the states within the HMM model the events. An appropriate choice of model topology, e.g. left-to-right or fully connected, has to be done. Then features have to be chosen according to the events to be modelled. Enough training data, representative of the range of manifestations of the events, has to be selected, increasing its size in case of ergodic models. In general a significant effort is required to train a HMM system, and ergodic models require more training data than left-to-right models. In [39] is noted that the conventional HMM training approaches, based on maximum likelihood such as the Baum-Welch algorithm, often produce models that are both under-fit (failing to capture the hidden structure of the signal) and over-fit (with many parameters that model noise and signal bias), thus leading to both poor predictive power and small generalisation.

A number of approaches that use HMM have been proposed to analyse sports videos, since the events that are typical for this domain are very well suited for this approach. It has to be noted that reliable event classification can be achieved if events have been accurately segmented and delimited. Classification of three placed kicks events (free, corner and penalty kick) using HMMs has been proposed by Assfalg *et al.* in [9], using a 3-state left-to-right model for each highlight, based on the consideration that the states correspond well to the evolution of the highlights in term of characteristic content. The features used are the framing term (e.g. *close-up*), camera pan and tilt (quantised in five and two levels). Similar approaches for event detection in news videos have been applied also at a higher semantic level,

using the scores provided by concept detectors as synthetic frame representations or exploiting some pre-defined relationships between concepts. For example, Ebadollahi *et al.* [65] proposed to treat each frame in a video as an observation, applying then HMM to model the temporal evolution of an event. In [232] multi-layer HMMs (called SG-HMM) have been proposed by Xu *et al.* for basket and volleyball. Each layer represents a different semantic layer, and low-level features (horizontal, vertical and radial motion and acceleration cues) are fed to the bottom layer to generate hypothesis of basic events, the upper layer gets the results of the below HMMs and each state corresponds to an HMM; this requires to treat differently these HMM: the observation probability distribution is taken from the likelihood of the basic HMMs. Fully connected HMMs, with six states, are used to model all the basic events in both sports. The Basket SG-HMM has two layers: one for sub-shot classification and the upper layer for shot classification in 16 events. The Volley SG-HMM has three layers: shots are classified in the two bottom layers, and the intermediate layer accounts for shots relationships; this allows to classify 14 events that cannot be recognised within a shot.

Bayesian networks

Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose arcs encode conditional independencies between the variables. Nodes can represent any kind of variable, be it a measured parameter, a latent variable or a hypothesis. Bayesian networks can represent and solve decision problems under uncertainty. They are not restricted to representing random variables, which represents another "Bayesian" aspect of a Bayesian network. Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (such as for example speech signals or protein sequences) are called Dynamic Bayesian Networks (DBNs). Dynamic Bayesian Networks are directed graphical models of stochastic processes. They generalise hidden Markov models (HMMs). In fact a HMM has one discrete hidden node and one discrete or continuous observed node per slice. In particular a Hidden Markov Model consists of a set of discrete states, state-to-state transition probabilities, prior probabilities for the first state and output probabilities for each state.

In [139] Bayesian Networks are used to recognise frame and clip classes (close-up, playfield centre and goal areas, medium views). In order to iden-

tify shot-on-goals the proposed system groups the clips that are preceding and following the clips classified as showing the goal areas. If a certain pattern of clips is found, and the values of a feature that corresponds to the position of the field end line follow a certain pattern, then a shot-on-goal is determined to be present. In [44] DBNs are used by Chao *et al.* to model the contextual information provided by the timeline. It is argued that HMMs are not expressive enough when using a signal that has both temporal and spatial information; moreover, DBNs allow a set of random variables instead of only one hidden state node at each time instance: this stems from the fact that HMMs are a special case of DBNs. In [44] five events are defined and are modeled considering five types of primitive scenes such as close-ups, medium views, etc. Medium level visual features such as playfield lines are used as observable features. Since all the states of the DBN are observable in the training stage it is required to learn the initial and transition probabilities among the scenes in each event separately. In the inference stage the DBN finds the most plausible interpretation for an observation sequence of features.

Kernel methods

Kernel methods are a class of algorithms for pattern analysis, whose best known element is the Support Vector Machine (SVM), a group of supervised learning methods that can be applied to classification problems. These methods map the input data into a high dimensional feature space, by doing a non-linear transformation using suitably chosen basis functions (kernel). This is known as the "kernel trick". The linear model in the feature space corresponds to a non-linear model in the input space. The kernel contains all of the information about the relative positions of the inputs in the feature space; the actual learning algorithm is based only on the kernel function and can thus be carried out without explicit use of the feature space. Since there is no need to evaluate the feature map in the high dimensional feature space, the kernel function represents a computational shortcut.

An approach that uses SVM with RBF kernel to classify sequences that contain interesting and non-interesting events was proposed in [183], showing an application to field sports such as soccer, hockey and rugby. Each shot is represented using five values, one for each feature used (e.g. speech-band audio activity, motion activity, etc.), and the maximum value of each feature

is selected as representative value for the whole shot. In this way the temporal extent and the dynamics of the event are not considered or exploited. Authors note that a classification scheme such as HMM may be more appropriate if continuous knowledge of past and present states is desired. In [91] was proposed the use of SVM models for a set of motion features, computed from MPEG motion vectors, and static features, followed by a late fusion strategy to aggregate results at the decision level.

As previously discussed, many recent methods extend the traditional BoW approach. In fact, the application of this part-based approach to event classification has shown some drawbacks with respect to the traditional image categorisation task. The main problem is that it does not take into account temporal relations between consecutive frames, and thus event classification suffers from the incomplete dynamic representation. Recently methods have been proposed to consider temporal information of static part-based representations of video frames. Xu and Chang [231] proposed to apply Earth Mover's Distance (EMD) and Temporally Aligned Pyramid Matching (TAPM) for measuring video similarity; EMD distance is incorporated in a SVM framework for event detection in news videos. In [221], BoW is extended constructing relative motion histograms between visual words (ERMH-BoW) in order to employ motion relativity and visual relatedness. Zhou *et al.* [243] presented a SIFT-Bag based generative-to-discriminative framework for video event detection, providing improvements on the best results of [231] on the same TRECVid 2005 corpus. They proposed to describe video clips as a bag of SIFT descriptors by modeling their distribution with a Gaussian Mixture Model (GMM); in the discriminative stage, specialised GMMs are built for each clip and video event classification is performed. Ballan *et al.* [20] modelled events as a sequence composed of histograms of visual features, computed from each frame using the traditional bag-of-words. The sequences are treated as strings where each histogram is considered as a character. Event classification of these sequences of variable length, depending on the duration of the video clips, are performed using SVM classifiers with a string kernel that uses the Needlemann-Wunsch edit distance. Hidden Markov Model Support Vector Machine (SVMHMM), which is an extension of the SVM classifier for sequence classification, has been used in [96] to classify the behaviour of caged mice.

2.4 Ontologies

In many image/video content-based applications there is need of methodologies for knowledge representation and reasoning, to analyse the context of an action in order to infer an activity. This has led to an increasing convergence of research in the fields of video analysis and knowledge management. This knowledge can include heterogeneous information such as video data, features, results of video analysis algorithms or user comments. Logical-based methods for activity recognition have been proposed, to represent domain knowledge and model each event. In these approaches an event is generally specified as a set of logical rules that allow to recognise them by using logical inference techniques, such as resolution or abduction [198, 60, 169, 7]. In particular, Shet *et al.* [198] proposed a framework that combines computer vision algorithms with logic programming to represent and recognise activities in a parking lot in the domain of video surveillance. Lavee *et al.* [122] have proposed the use of Petri-Nets, and provided a methodology on how to transform ontology definitions in a Petri-Net formalism. Artikis *et al.* [7] and Paschke *et al.* [169] presented two different activity recognition systems based both on a logic programming implementation of an Event Calculus dialect [116]. The Event Calculus is a set of first-order predicate calculus, including temporal formalism, for representing and reasoning about events and their effects. These approaches do not consider the problems of noise or missing observations, that always exist in real world applications. To cope with these issues, some extensions to logic approaches have been presented. Tran *et al.* [212] described a domain knowledge as first-order logic production rules with associated weights to indicate their confidence. Probabilistic inference is performed using Markov-logic networks. While logic-based methods are an interesting way of incorporating domain knowledge, they are limited in their utility to specific settings for which they have been designed. Hence, there is need of a standardised and shareable representation of activity definitions.

Recently, ontologies have been regarded as the appropriate tool for domain knowledge representation because of several advantages. Their most important property is that they provide a formal framework for supporting explicit, shareable, machine-processable semantics definition of domain knowledge, and they enable the derivation of implicit knowledge through automated inference. In particular, an ontology is a formal specification of a

shared conceptualisation of a domain of interest [85] and form an important part of the emerging semantic web, in which ontologies allow to organise contents through formal semantics. Ontology Web Language (OWL) and Semantic Web Rule Language (SWRL) have been proposed by the World Wide Web Consortium (W3C) as language standards for representing ontologies and rules, respectively. SPARQL Protocol and RDF Query Language (SPARQL) has been approved as W3C recommendation as query language for the Semantic Web technologies. An overview of such languages is presented in [152]. These languages enable autonomic agents to reason about Web content and to carry out more intelligent tasks on behalf of the user. Thus, ontologies are suitable for expressing video content semantics.

For these reasons, many researches have exploited ontologies to perform semantic annotation and retrieval from video digital libraries [114]. Ontologies that can be used for semantic annotation of videos are those defined by the Dublin Core Metadata Initiative [1], TV Anytime [2] - they have defined standardised metadata vocabularies - and the LSCOM initiative [154] - that has created a specialised vocabulary for news video. Other ontologies provide structural and content-based description of multimedia data, similarly to the MPEG-7 standard [79, 213, 6]. Other approaches have directly included in the ontology an explicit representation of the visual knowledge [34, 145]. Dasiopoulou *et al.* [53] have included in the ontology instances of visual objects. They have used as descriptors qualitative attributes of perceptual properties like colour homogeneity, low-level perceptual features like components distribution, and spatial relations. Semantic concepts have been derived from colour clustering and reasoning. In the attempt of having richer annotations, other authors have explored the usage of reasoning over multimedia ontologies. In this case spatial relationships between concept occurrences are analysed so as to distinguish between scenes and provide more precise and comprehensive descriptions. Hollink *et al.* [94] defined a set of rules in SWRL to perform semi-automatic annotation of images. Jain *et al.* [126] have employed a two-level ontology of artistic concepts that includes visual concepts such as colour and brushwork in the first level, and artist name, painting style and art period for the high-level concepts of the second level. A transductive inference framework has been used to annotate and disambiguate high-level concepts. In Staab *et al.* [54] automatically segmented image regions are modeled through low-level visual descriptors and associated to semantic concepts using manually labelled regions as training set.

Context information is exploited to reduce annotation ambiguities. The labelled images are transformed into a constraint satisfaction problem (CSP), that can be solved using constraint reasoning techniques.

Several authors have exploited ontologies for event recognition. These methods have to deal with two issues: how to represent the entities and events of the considered domain in the ontology, and how to use the ontology for improving the video event analysis results. For solving the first issue, researchers have proposed ontologies to describe several domains, e.g. for visual surveillance analysis. In particular, Hakeen and Shah [87] have defined a meeting ontology that is determined by the knowledge base of various meeting sequences. Chen *et al.* [45] proposed an ontology for analysing social interaction of the patients with one another and their caregivers in a nursing home, and Georis *et al.* [81] for describing bank attack scenarios. Akdemir *et al.* [3] drew on general ontology design principles and adapted them to the specific domains of human activity, bank and airport tarmac surveillance. Moreover, a special formal language to define ontologies of events, that uses Allen's logic to model the relations between the temporal intervals of elementary concepts so as to be able to assess complex events in video surveillance has been proposed by Francois et al. [160, 76]. More recently, Scherp *et al.* [190] defined a formal model of events that allows interchange of event information between different event-based systems, causal relationships between events, and interpretations of the same event by different humans. A more generic approach has been followed in [170], where a verb ontology has been proposed to better describe the relations between events, following Fellbaum's verb entailments [71]. This ontology is used to classify events that may help the comprehension of other events (e.g. when an event is a precondition of another one). The outcomes of event classification are then used to create hyperlinks between video events using MPEG-7 video annotations, to create a hypervideo.

Solutions for the second issue have also been explored. Neumann and Möller [159] have proposed a framework for scene and event interpretation using Description Logic reasoning techniques over "aggregates"; these are composed of multiple parts and constrained by temporal and spatial relations to represent high-level concepts, such as objects configurations, events and episodes. Another solution was presented by Bertini *et al.* in [30], using generic and domain specific descriptors, identifying visual prototypes as representative elements of visual concepts and introducing mechanisms for

their updating, as new instances of visual concepts are added to the ontology; the prototypes are used to classify events and objects observed in video sequences. Bai *et al.* [13] defined a soccer ontology and applied temporal reasoning with temporal description logic to perform event annotation in soccer videos. Snidaro *et al.* [204] addressed the problem of representing complex events in the context of security applications. They described a complex event as a composition of simple events, thus fusing together different information, through the use of the SWRL language. SWRL rules have been also employed to derive complex events in soccer domain [18]. In [187] the authors proposed an ontology that integrates two kinds of knowledge information: the scene and the system. Scene knowledge is described in terms of objects and relations between them. System knowledge is used to determine the best configuration of the processing schemas for detecting the objects and events of the scene.

Finally, in this research work we have presented an ontology-based framework for semantic video annotation by learning spatio-temporal rules (see Chapter 8 and [19, 33]). In our approach, an adaptation of the First Order Inductive Learner to the Semantic Web technologies (FOILS) is used to learn SWRL rule patternsthat have been then validated on a few TRECVid 2005 and CAVIAR video events.

Chapter 3

Trademark retrieval in sports video archives

In this chapter we describe a system for detection and retrieval of trademarks appearing in sports videos. We propose a compact representation of trademarks and video frame content based on SIFT feature points. This representation can be used to robustly detect, localize, and retrieve trademarks as they appear in a variety of different sports video types. Classification of trademarks is performed by matching a set of SIFT feature descriptors for each trademark instance against the set of SIFT features detected in each frame of the video. Localization is performed through robust clustering of matched feature points in the video frame. Experimental results are provided, along with an analysis of the precision and recall. Results show that the our proposed technique is efficient and effectively detects and classifies trademarks. [1] [2]

3.1 Introduction

Every year sponsors spend millions of euros on sports marketing, a large portion of which is spent on placement of billboards, banners, and other physical

[1] This chapter has been published as "Trademark Matching and Retrieval in Sports Video Databases" in *Proc. of ACM Multimedia Information Retrieval (MIR), 2007* [10].

[2] *Acknowledgments:* this work was partially supported by Sport System Europe srl, Bologna, Italy.

Figure 3.1: Example trademarks. *i)* top row: graphic version; *ii)* bottom row: trademarks extracted from actual videos.

advertising media positioned in and around soccer and football fields, formula one race circuits, tennis courts, etc. These physical advertising artifacts are usually emblazoned with the sponsor's name, logo, and their trademark brand in general. Given these astronomical numbers, sponsors are extremely keen to verify that their brand has the level of visibility they expect for such an expenditure. Such verifications are essential to major sponsors in order to justify advertising budgets and ensure their brands achieve the desired level of visibility.

Currently, verification of brand visibility is done manually by human annotators that view a broadcast sporting event and annotate every appearance of a sponsor's trademark in the broadcast. The annotation performed on these videos is extremely labor-intensive, usually requiring the video to be viewed in its entirety several times. Manual annotation of this type is often limited to annotations of the *appearance* of trademarks (i.e., that a trademark *appears* at a given timecode).

The problem of automatic trademark and logo detection and recognition belongs to the broader problem of object recognition that has been studied following many different approaches in recent decades. The two primary types of features used are geometric and photometric object features: the former rely on properties of objects such as lines, vertexes, curves and shapes [117, 26], while the latter are computed from pixel values (luminance or color) of the imaged object [192, 82, 137]. Object detection and recognition using photometric features has been the subject of much recent research due to the fact that if these features are computed locally they can cope with the problem of occlusion and are able to distinguish similar objects much better [192].

Most of the work related to trademark recognition deals with the problem of content-based indexing and retrieval in logo databases, with the goal of

assisting in the detection of trademark infringement by comparing a newly designed trademark with archives of already registered logos [110, 238, 63, 216]. In this case it can be assumed that the image acquisition and processing chain is controlled so that the images are of acceptable quality and are not distorted. The problem of trademark recognition in videos is inherently harder, since the whole process is not controlled and several limitations of the imaging equipment introduce considerable distortion and loss of quality of the original logos (e.g. video interlacing, color sub-sampling, motion blur, etc.)

In [4] the problem of detecting and tracking billboards in soccer videos has been studied, with the goal of superimposing different advertisements according to the different audiences. Billboards are detected using colour histogram back-projection and represented using a *PD* 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 [115] 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. The logo recognition method proposed in [58] extends the work presented in [57] and deals with logos appearing on rigid planar surfaces that have an homogeneously colored background; the video frame is binarized and logo regions are combined using heuristics. The Hough transform space of the segmented logo is then searched for large values to find the image intensity profiles along lines. Logo recognition is performed by matching these lines with the line profiles of the models. In [171] candidate logo regions are detected using color histogram back-projection and then they are tracked. Multidimensional receptive field histograms are then 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 [108] the architecture for a system for media monitoring is presented. The system provides logo detection and recognition functionalities, and the authors briefly discuss a variation of the SIFT algorithm to select and track keypoints in videos. The points are used for trademark recognition, but the logo matching algorithm is not described, and very few results of the proposed variation are provided.

In this chapter we propose a system for automatically detecting and retrieving trademark appearances in sports videos. In brief, broadcast sports video is recorded directly to DVD. This video, and a collection of static

trademark images, are then processed to extract a compact, salient-point representation. 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. These time series are used to retrieve intervals of the video likely to contain the trademark image. Retrieved segments are used to drive a user interface used by a human annotator who can then validate this automatic annotation.

In the next section we describe the compact representation used to model trademarks and video frame content. Section 3.3 details the matching procedure that is used to detect and localize trademarks in video streams. In section 3.4 we present the experimental results we obtained with the approach. Finally, in section 3.5 we conclude with an analysis of the technique and indications for future work.

3.2 Image and video features

Figure 3.1 contains several representative examples of the types of trademarks we wish to detect. The top row of figure 3.1 contains clean, synthetic versions of each trademark, while the bottom row contains example trademarks extracted from frames of actual sports videos.

The appearance of trademarks in sports videos are often characterized by:

- **Perspective deformation** due to placement of the camera and the vantage from which it images advertisements in the field.
- **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.
- **Occlusion** caused by players or other obstacles between the camera and the trademark. In many sports, soccer for instance, trademarks are occluded more often than not.

Since blur is indistinguishable from a change in scale, a scale-invariant representation is essential. To render our matching technique 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. One the most distinguishing aspects of trademarks is that they usually contain both text and other high-contrast features such as graphic logos. They are also usually planar objects. Because of this, and because of the observations made above, we use SIFT points and SIFT feature descriptors as a compact representation of the important aspects and local texture in trademarks [137]. These feature points are robust to changes in scale, perspective, and rotation.

Trademarks are represented as a bag of SIFT feature points. Each trademark is represented by one or more graphical instances. Trademark T_i is represented by the N_i SIFT feature points detected in the image:

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

and 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 kth detected feature point. The element 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:

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

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.

For the matching procedure described in the next section will be using only the local orientation histogram portions ($\mathbf{O}_k^v$ and $\mathbf{O}_k^t$) of the feature points described above. This renders the feature descriptors robust to geometric distortions and scale changes. The geometric elements of the feature point descriptors are used only for visualization of match results.

3.3 Detection and retrieval of trademarks

Trademark matching is done by comparing the bag of local features representing the trademark with the local features detected in the frames of the

video. This is done using a very conservative threshold in order to minimize false positive detections. For every point detected in trademark T^j we compute its two nearest neighbors in the points detected in video frame V_i:

$$\begin{aligned} N_1(T_j^k, V_i) &= \min_q ||\mathbf{O}_q^v - \mathbf{O}_k^t|| \\ N_2(T_j^k, V_i) &= \min_{q \neq N_1(T_j^k, V_i)} ||\mathbf{O}_q^v - \mathbf{O}_k^t||. \end{aligned} \tag{3.1}$$

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

$$M(T_j^k, V_i) = \frac{N_1(T_j^k, V_i)}{N_2(T_j^k, V_i)}, \tag{3.2}$$

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

Points are selected as being good candidate matches on the basis of their match scores. The *match set* for trademark T_j in frame V_i is:

$$M_i^j = \{k \,|\, M(T_j^k, V_i) < \tau_1\}, \tag{3.3}$$

where τ_1 is a suitable chosen threshold (0.8 in all of our experiments).

This approach gives very good results in terms of robustness. It performs well because a correct match needs to have the closest matching descriptor significantly closer than the closest incorrect match, while false matches have a certain number of other close false matches, due to the high dimensionality of the feature space (Figure 3.2 for some examples).

The final determination of whether a frame V_i contains trademark T_j is made by thresholding the *normalized match score*:

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

This threshold requires that a certain percentage of the feature points detected in the trademark be matched according to equation 3.3 in order to make that final determination that the trademark T_j is in fact 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 3.4). Experiments have shown that a value of $0.2 - 0.25$ is a reasonable choice for several different sports.

In order to localize the trademark in the original frame V_i and to approximate its area, we compute a robust estimate of the feature point cloud.

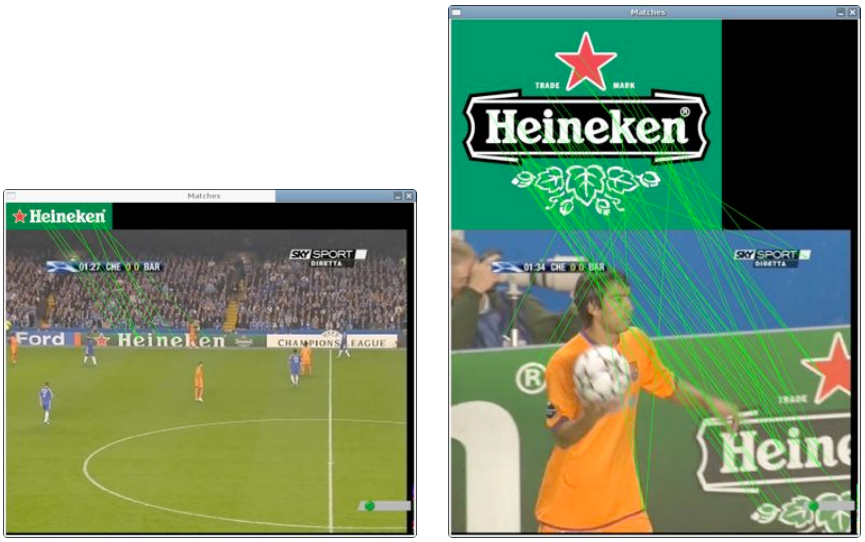


Figure 3.2: Two examples of the "traditional" SIFT matching technique.

The current feature point locations are so denoted as:

$$F = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

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:

$$\psi(x; m) = \begin{cases} (x-m)(1 - \frac{(x-m)^2}{c^2})^2 & \text{if } |(x-m)| < c \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

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

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

After the robust centroid is estimated, the distance of each matched point to the robust centroid is computed according to the influence function (3.4). Points with a low influence are excluded from the final match set (see figure 3.3).

Figure 3.3: 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.

Some example matches found in videos using this technique are shown in figure 3.4. 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.

3.4 Experimental results

We have implemented the approach described above and here we describe a number of experiments we have performed to calibrate and evaluate the performance of the matching technique.

3.4.1 Implementation

The steps involved in our matching procedure are as follows:

1. **Acquisition**
   Videos are processed directly from DVD in MPEG2 format. Prepro-

Figure 3.4: 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 (3.3, 3.4).

cessing consists of de-interlacing each video frame.

2. **Feature point detection**
   The SIFT feature point detection algorithm is run on each frame of the video. Because of the large quantity of data generated by this process, all feature points are stored in a database for retrieval later. The SIFT detector that we use is implemented in C++ and is able to process video at about 2.5*fps*.

3. **Matching of trademarks**
   Each trademark in the database is matched against the feature points detected in the previous step. Again, match information corresponding to equations (3.2) and (3.1) are stored in another table in the database.

4. **Retrieval of matches**
   Trademarks are retrieved from the database by supplying the thresh-

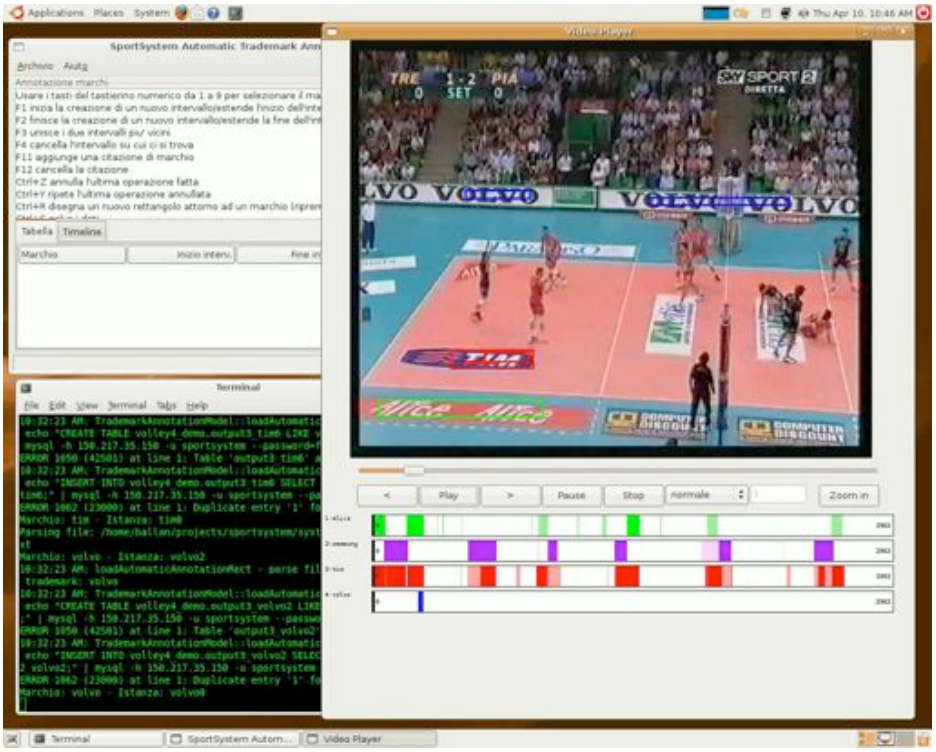


Figure 3.5: A screenshot of the visualization application. The user can configure how detected trademarks are visualized on the video frame. The rows at the bottom indicate, in different colors, the timeline of detected trademarks in the video.

olds τ_1 and τ_2 on the match score and normalized match score, respectively. A list of candidate frames is returned by this process. These frames are grouped temporally to define intervals where the trademark is (believed to be) present.

5. **Visualization**
   Match results are displayed in an applications that also doubles as a manual annotation tool for trademarks in sports videos. The tool allows a user to inspect and correct the automatic match results, adjust the thresholds τ_1 and τ_2, and save the resulting annotation in MPEG7 format. A screenshot of the visualization/annotation application is shown in figure 3.5.

The most time consuming steps in this procedure are the detection of the SIFT feature points and the matching of trademarks against every frame

in the video. One hour of video at 25*fps* contains around 90,000 frames, in each of which SIFT features must be detected. Each frame contains around 1,000 feature points, on average, and each trademark around one hundred. Consequently, the time required for detection and matching is directly proportional to the number of video frames processed.

3.4.2 Test data and experiment design

Three videos of three different sports were used for a preliminary evaluation of the performance of the proposed technique. The first video, of a MotoGP motorcycle race, is approximately one hour long. The second video is of 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; in this case there are often trademarks at low resolution with few SIFT feature points. The examples in the top row of figure 3.4 are from the MotoGP video, those in the middle row are from the volleyball and those in the bottom row from the soccer video.

To evaluate the effects of all the parameters of the proposed algorithms, the MotoGP video was manually annotated for the presence of a number of trademarks. 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 technique is evaluated in terms of precision and recall:

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

3.4.3 Results

Figure 3.6 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 figure 3.6 are the precision and recall performances as a function of the frame sampling rates. Results are shown for 2.5*fps*, 5*fps*, and 10*fps*. Note that in these plots, the recall plots are the ones that start at or around 1.0 and *decrease* as the normalized match

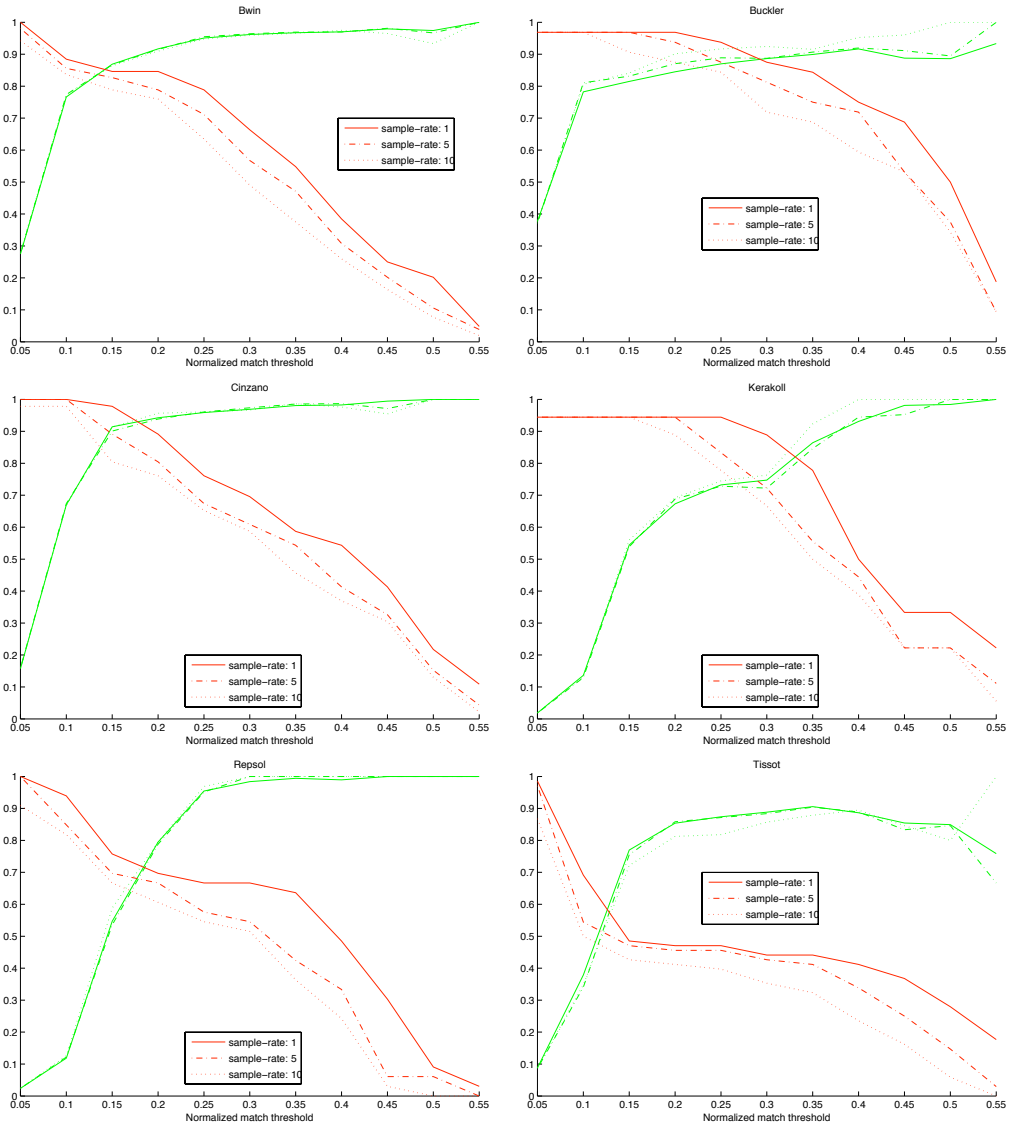
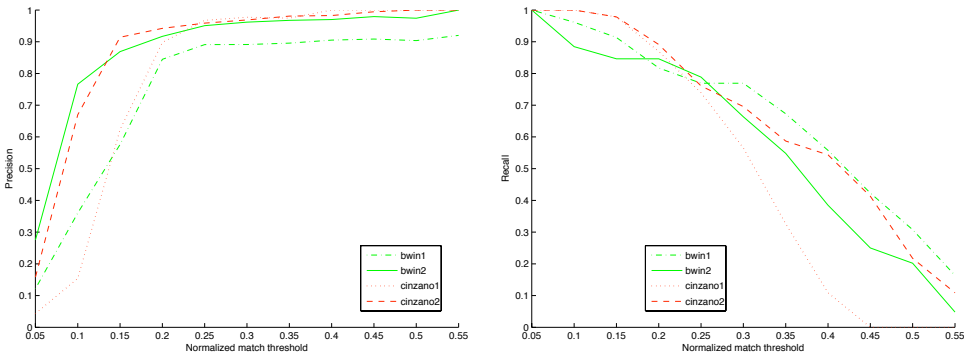
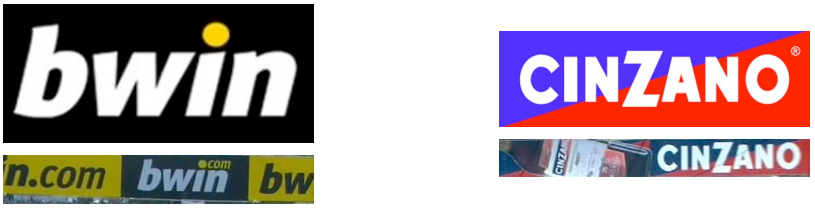


Figure 3.6: Precision and recall as a function of the normalized match threshold. Note that as the threshold increases, more matches are *excluded*. Because of this, recall usually begins at or around 1.0 and is inversely proportional to the normalized match threshold.



(a) Precision (b) Recall



(c) Bwin trademarks (d) Cinzano trademarks

Figure 3.7: Comparison of precision between synthetic trademarks and trademarks cropped from actual video frames.

threshold is increased. In most cases, a recall of about 85% can be obtained at an precision of around 80% with values of τ_2 varying between $0.2 - 0.25$. The preliminary experiments performed on soccer and volleyball videos have shown that this value of the threshold can be used also for these other sports.

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. 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.

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 figure 3.7 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 figure 3.7). 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.

3.5 Conclusion

In this work we introduced an approach for automatically detecting and retrieving trademark occurrences in sports video. The use of SIFT features as a compact representation of video and trademark content ensures that technique is robust to occlusions, scale and perspective variations. In most cases, a recall rate of better then 85% can be achieved with a precision of approximately 80%. Our experiments indicate that SIFT-features a good representation for many types of trademarks consisting of text and logos. The high-contrast nature of this type of trademark design makes these features very distinguishable from background noise. Experiments on clean, synthetic trademarks also indicate that in many cases these can be used as trademark prototypes for matching. A process of robust clustering 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.

While in many cases a satisfactory level of precision and recall can be achieved by selecting the appropriate trademark directly from the video, additional trademark descriptors will certainly be required to guarantee a high level of recall and simultaneously maintaining a low false-positive rate. Color trademark descriptors are the most promising approach to maintaining precision when using less conservative normalized match thresholds. Preliminary results on other types of sports videos confirm that the technique is capable of effectively retrieving 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. One solution to this is to double the resolution of each video frame before processing. 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.

The best human annotators can annotate a sports video for four different trademarks in real time (i.e., one hour of video requires one hour of annotation for four trademarks). Annotations are typically required for between twenty and thirty trademarks for each video, requiring the annotator to view it multiple times or for annotation to be performed in parallel by multiple human annotators. In any case, each video typically requires around six man-hours to fully annotate. Automatic annotation of sports videos, as we propose in this article, promises to significantly reduce the labor involved in annotating for trademark visibility. Furthermore, automatic annotation can provide richer annotations than those currently performed by humans. For example, our technique is able to compute metrics on the duration of each trademark appearance as well as an estimation of the size it occupies in the frame.

Future work will include an investigation of how the approach performs on other types of sports videos. We are also investigating different metrics that can be computed on automatically annotated videos. To this end, we are particularly interested in refining the localization of trademarks so that a visibility metric can be computed in terms of how much visible space is occupied by a sponsor's brand during the course of a broadcast sporting event.

Chapter 4

Context-dependent trademark matching and retrieval

In this chapter we introduce a novel multiple-logo matching and detection algorithm based on a new class of similarity functions referred to as context dependent. Our approach is based on designing a similarity measure, involving interest points, which takes into account not only their intrinsic visual features but also their context and spatial configuration. The main contribution of this work includes (i) a variational framework which makes it possible to design our similarity as the fixed point of an energy function mixing a visual "data term", a "context criterion" and a "regularization term" and, (ii) a theoretical study of the consistency of logo matching/detection and its invariance to different transformations including similarity and occlusion. Finally, we will show the validity of the method through extensive experiments on challenging logo images. [1]

4.1 Introduction

Automatic image and video annotation has received an increasing attention from the research and industrial community in the recent years [55]. This is

[1]Part of this work was conducted while the author was a visiting Ph.D. student at Télécom ParisTech, Paris (France), from April to June 2010 (working with. Dr. Hichem Sahbi). This chapter previously appeared as research report n. 2010D009, Télécom ParisTech [186].

Figure 4.1: Realistic examples of trademark images characterized respectively by a bad light condition (Coca-Cola), occlusions (McDonald's), a deformation (Starbucks) and a small size (Ferrari).

mainly due to the growing request for content based search and retrieval of interesting visual elements, resulting from the exponential growth of multimedia sharing systems such as Flickr and YouTube. In particular, a really challenging task is the detection and recognition of advertising trademarks/logos, which are of great interest for several real world applications. In fact, logos are key elements for companies and play essential role in industry and commerce; they also recall the expectations associated with a particular product or service.

The early work on 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 [191]. Historically, the earliest approach was Kato's Trademark system [103,110]. Its idea is to map normalized trademark images to an 8×8 pixel grid, and calculate a *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 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. More recently, Wei *et al.* [223] proposed a system that combines global Zernike moments and local curvature and distance to centroid features in order to describe logos. All these methods use synthetic images and rely on global logo descriptions, usually related to their contours or to particular shape descriptors (such as *shape context* [26]), so they require logos to be fully visible.

In the last years, other work on logo detection and recognition, in real world images/videos, has emerged and is targeted to automatically identify products (such as groceries in stores for assisting the blind) [148, 98] or to verify the visibility of advertising trademarks (e.g. billboards or banners) in sports events [10, 222]. This problem is much harder, due to the relatively low resolution and quality of images (e.g. compression artifacts, color subsampling, motion blur, etc.) and also to the fact that trademarks are often small and may contain few information. Moreover their appearance is often characterized by occlusions, perspective transformations and deformations (see the examples in Fig. 4.1).

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 liner and non linear transformations. The first approach, proposed by Bagdanov *et al.* [10], achieves trademark detection and localization in sport video; each trademark is described as a bag of local features (SIFT points [137]) which are classified and matched with the bags of SIFT features in video frames. Localization is performed through robust clustering of matched SIFT features. Following the same approach, Joly and Buisson [99] exploit SIFT point representation in order to detect logos in natural images. In order to refine their detection results, they also include 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.

More recently, interesting work includes spatial informations into logo representations in order to improve the detection performances. Kleban *et al.* . [112] introduced a method for logo detection based on association rules that capture frequent spatial configurations of local features at multiple resolutions. These configurations are indexed in order to retrieve representative training templates for matching, nevertheless image resolution is a major limitation. Gao *et al.* [78] presented a two-stage logo detection algorithm which also achieves localization by adapting a spatial-spectral saliency in order to improve the matching precision. They proposed a spatial context descriptor in order to estimate the spatial distribution of the set of matching points. In particular, they find minimum boundary round of matched points and partition it into nine areas. Finally, they describe the distribution of these points using a nine-dimensional histogram. However, this global logo

representation is sensitive to occlusion.

Following the above discussed work, logo detection algorithms, based on interest points, are known to be very effective and also flexible in order to handle invariance (including occlusion and affine transformations). Nevertheless, their success strongly depends on the quality of matching (also referred to as alignment) mainly when images contain repeatable and redundant structures. On the one hand, a *naive* matching strategy, which given two images (a reference logo and a test image), looks for all pairs of interest point matches, using a context[2]-free similarity, such as the laplacian or the Gaussian, might result into many false matches. Figure (4.3, left) illustrates the deficiency of such naive approach when used between two groups of interest points; any slight perturbation of the values of the underlying features will result into unstable matching results *if no context* is taken into account. On the other hand, putting strong model assumptions about possible transformations (homography, affine, etc.) between reference logos and test images, might not capture the actual inter-logo transformations; for instance when logos deform.

In this chapter, we introduce an alternative matching framework, for logo detection, based on a new class of similarity functions, called "context-dependent" (CD) and defined as the fixed-point of an energy function which balances a "fidelity" term, a "context" criterion and an "entropy" term. The fidelity term is inversely proportional to the expectation of the Euclidean distance between the most likely aligned interest points while the context criterion measures the spatial coherence of the alignments, i.e., how good two interest points, with close context, match. Given a pair of interest points (f_p, f_q) with a high alignment score (defined by our "CD" values), the context criterion is proportional to the alignment scores of all the pairs close to (f_p, f_q) *but with a given spatial configuration*. The "entropy" term considers that without any a priori knowledge about the alignment scores between pairs of interest points, the joint probability distribution related to these scores should be as flat as possible so this term acts as a *regularizer*. In a second major part of this work, we introduce a matching procedure based on our "CD" similarity and we show, under the hypothesis of the existence of reference logos into test images, the probability of success of this procedure

[2]Given a set of interest points $\mathcal{X}$, the context of $x \in \mathcal{X}$ is defined as the set of points spatially close to x and with some particular geometrical constraints (see section 4.2.1 for a detailed and a formal definition of the context.)

is high, which is also corroborated through experiments. Notice also that the proposed alignment and logo detection method is model-free, i.e., it is not based on any a priori alignment model such as homography which might not capture the actual inter-logo transformations.

We consider the following organization of the chapter; we first discuss in Section 2, our energy function which makes it possible to design our context-dependent similarity, then we show in Section 3, the application of this similarity in order to align interest points and perform logo detection. We will also show some theoretical properties about our alignment procedure mainly its probability of success even in challenging conditions such as presence of partial occlusion and its rotation, scale and translation invariance. In Section 4, we show logo detection results and comparison on challenging logo images, and we conclude in Section 5 while providing possible extensions for a future work.

4.2 Context-dependent similarity

Let $\mathcal{S}_p = \{x_1^p, \ldots, x_n^p\}$, $\mathcal{S}_q = \{x_1^q, \ldots, x_m^q\}$ be the list of interest points taken respectively from a reference logo and a test image ($n \ll m$ and the value of n, m may vary with the objects p, q). The set $\mathcal{X}$ of all possible interest points is the union over all possible objects p, q of $\mathcal{S}_p$, $\mathcal{S}_q$:

$$\mathcal{X} = \left\{ \cup_p \mathcal{S}_p \right\} \cup \left\{ \cup_q \mathcal{S}_q \right\}.$$

We consider $k : \mathcal{S}_p \times \mathcal{S}_q \rightarrow \mathbb{R}$ as a function which, given two interest points (x_i^p, x_j^q), provides a similarity measure between them. This will be designed as shown in Section (4.2.2).

4.2.1 Context

Formally, an interest point x is defined as $x = (\psi_g(x), \psi_f(x), \psi_o(x), \omega(x))$ where the symbol $\psi_g(x) \in \mathbb{R}^2$ stands for the $2D$ coordinates of x while $\psi_f(x) \in \mathbb{R}^s$ corresponds to the feature of x (in practice the 128 coefficients of the SIFT; [137]). We have an extra information about the orientation of x (denoted $\psi_o(x) \in [-\pi, +\pi]$) which is provided by the SIFT gradient. Finally, we use $\omega(x)$ to denote the object from which the interest point comes from, so that two interest points with the same location, feature and orientation

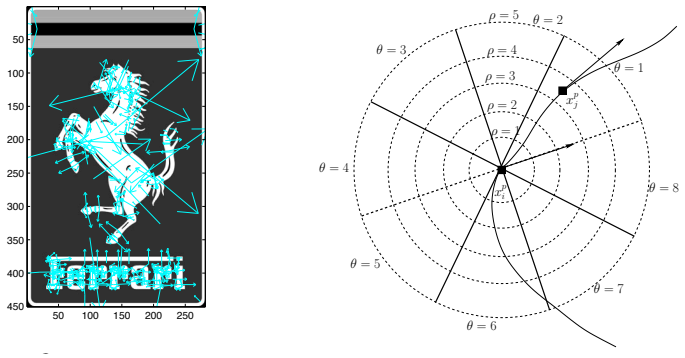


Figure 4.2: This figure shows a collection of SIFT interest points (with their locations, orientations and scales) (left) and the partitioning of the context (also referred to as neighborhood) of an interest point into different sectors for orientations and bands for locations (right).

are considered different when they are not in the same image (this is not surprising since we want to take into account the context of the interest point in the image it belongs to).

Let $d(x, x') = \|\psi_f(x) - \psi_f(x')\|_2$ measure the dissimilarity between two interest point features, $\|\cdot\|_2$ is the "entrywise" L_2-norm (i.e., the sum of the square values of vector coefficients). Introduce the context of x

$$\mathcal{N}^{\theta,\rho}(x) = \{x' : \omega(x') = \omega(x), x' \neq x \text{ s.t. (i) and (ii) hold}\},$$

with

$$\frac{\rho - 1}{N_r}\,\epsilon_p \leq \|\psi_g(x) - \psi_g(x')\|_2 \leq \frac{\rho}{N_r}\,\epsilon_p, \qquad \text{(i)}$$

and

$$\frac{\theta - 1}{N_a}\pi \leq \angle\big(\psi_o(x), \psi_g(x') - \psi_g(x)\big) \leq \frac{\theta}{N_a}\pi. \qquad \text{(ii)}$$

Here ϵ_p is the radius of a neighborhood disk surrounding x and $\theta = 1, ..., N_a$, $\rho = 1, ..., N_r$ correspond to indices of different parts of that disk (see Fig. 4.2). In practice, N_a and N_r correspond to 8 sectors and 8 bands. The definition of neighborhoods $\{\mathcal{N}^{\theta,\rho}(x)\}_{\theta,\rho}$ reflects the co-occurrence of different interest points with particular spatial geometric constraints (see again Fig. 4.2).

4.2.2 Similarity design

For a finite collection of interest points, the sets $\mathcal{S}_p$, $\mathcal{S}_q$ are finite. Provided that we put some (arbitrary) order on $\mathcal{S}_p$, $\mathcal{S}_q$, we can view a function k on $\mathcal{S}_p \times \mathcal{S}_q$ as a matrix $\mathbf{K}$ in which the "$(x, x')-$element" is the similarity between x and x': $\mathbf{K}_{x,x'} = k(x, x')$. Let $\mathbf{P}_{\theta,\rho}$ be the intrinsic adjacency matrices respectively defined as $\mathbf{P}_{\theta,\rho,x,x'} = g_{\theta,\rho}(x, x')$, where g is a decreasing function of any (pseudo) distance involving (x, x'), *not necessarily symmetric*. In practice, we consider $g_{\theta,\rho}(x, x') = \mathbb{1}_{\{\omega(x)=\omega(x')\}} \times \mathbb{1}_{\{x' \in \mathcal{N}^{\theta,\rho}(x)\}}$. Let $\mathbf{D}_{x,x'} = d(x, x')$. We propose to use the function on $\mathcal{S}_p \times \mathcal{S}_q$ defined by solving

$$\begin{aligned} \min_{\mathbf{K}} \quad & \mathrm{Tr}\big(\mathbf{K}\ \mathbf{D}^{'}\big) \ + \ \beta\ \mathrm{Tr}\big(\mathbf{K}\ \log\ \mathbf{K}^{'}\big) \\ & - \alpha \sum_{\theta,\rho} \mathrm{Tr}\big(\mathbf{K}\ \mathbf{P}_{\theta,\rho}\ \mathbf{K}'\ \mathbf{P}'_{\theta,\rho}\big) \qquad (4.1) \\ \text{s.t.} \quad & \begin{cases} \mathbf{K} \geq 0 \\ \|\mathbf{K}\|_1 = 1 \end{cases} \end{aligned}$$

Here $\alpha, \beta \geq 0$ and the operations $\sqrt{\ }$, log and $\geq$ are applied individually to every entry of the matrix (for instance, $\log \mathbf{K}$ is the matrix with $(\log \mathbf{K})_{x,x'} = \log k(x, x')$), $\| \cdot \|_1$ is the "entrywise" L_1-norm (i.e., the sum of the absolute values of the matrix coefficients) and Tr denotes matrix trace. The first term, in the above constrained minimization problem, measures the quality of matching two features $\psi_f(x)$, $\psi_f(x')$. In the case of SIFT, this is considered as the distance, $d(x, x')$, between the 128 SIFT coefficients of x and x'. A high value of $d(x, x')$ should result into a small value of $k(x, x')$ and vice-versa.

The second term is a regularization criterion which considers that without any a priori knowledge about the aligned interest points, the probability distribution $\{k(x, x')\}$ should be flat so the negative of the entropy is minimized. This term also helps defining a direct analytic solution of the constrained minimization problem (4.1). The third term is a neighborhood criterion which considers that a high value of $k(x, x')$ should imply high values in the neighborhoods $\mathcal{N}^{\theta,\rho}(x)$ and $\mathcal{N}^{\theta,\rho}(x')$. This criterion also makes it possible to consider the spatial configuration of the neighborhood of each interest point in the matching process.

We formulate the minimization problem by adding an equality constraint and bounds which ensure a normalization of the similarity values and allow to see $\{k(x, x')\}$ as a probability distribution on $\mathcal{S}_p \times \mathcal{S}_q$.

4.2.3 Solution

Proposition 1. *Let* $\mathbf{u}$ *denote the matrix of ones and introduce*

$$\zeta = \frac{\alpha}{\beta} \sum_{\theta,\rho} \|\mathbf{P}_{\theta,\rho}\mathbf{u}\mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{u}\mathbf{P}_{\theta,\rho}\|_\infty,$$

where $\|\cdot\|_\infty$ *is the "entrywise" L_∞-norm. Provided that the following two inequalities hold*

$$\zeta \exp(\zeta) < 1 \tag{4.2}$$

$$\|\exp(-\mathbf{D}/\beta)\|_1 \geq 2 \tag{4.3}$$

the optimization problem (4.1) *admits a unique solution* $\tilde{\mathbf{K}}$*, which is the limit of*

$$\mathbf{K}^{(t)} = \frac{G(\mathbf{K}^{(t-1)})}{\|G(\mathbf{K}^{(t-1)})\|_1}, \tag{4.4}$$

with

$$G(\mathbf{K}) = \exp\left\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\theta,\rho} \left(\mathbf{P}_{\theta,\rho}\mathbf{K}\mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{K}\mathbf{P}_{\theta,\rho}\right) \right\}, \tag{4.5}$$

and

$$\mathbf{K}^{(0)} = \frac{\exp(-\mathbf{D}/\beta)}{\|\exp(-\mathbf{D}/\beta)\|_1}$$

Besides $\mathbf{K}^{(t)}$ *satisfy the convergence property:*

$$\|\mathbf{K}^{(t)} - \tilde{\mathbf{K}}\|_1 \leq L^t \|\mathbf{K}^{(0)} - \tilde{\mathbf{K}}\|_1. \tag{4.6}$$

with $L = \zeta \exp(\zeta)$.

Proof. see an extended version [184] of our previous work [185]. □

By taking not too large β, one can ensure that (4.3) holds. Then by taking small enough α, Inequality (4.2) can also be satisfied. Note that $\alpha = 0$ corresponds to a similarity which is not context-dependent: the similarities between neighbors are not taken into account to assess the similarity between two interest points. Besides our choice of $\mathbf{K}^{(0)}$ is exactly the optimum (and fixed point) for $\alpha = 0$.

To have partitioned the neighborhood into several cells corresponding to different degrees of proximity (as shown in Fig. 4.2) has lead to significant

improvements of our experimental results. On the one hand, the constraint (4.2) becomes easier to satisfy, for larger α with partitioned neighborhood, compared to [185]. On the other hand, when the context is split into different parts, we end up with a context term, in the right-hand side of the exponential (4.5), which grows slowly compared to the one presented in our previous work [185] and grows only *if similar spatial configurations* of interest points have high similarity values. Therefore, numerically, the evaluation of that term is still tractable for large values of α which apparently produces a more positively influencing (and precise) context-dependent term, i.e., last term in (4.1) (see also equation (4.9) and discussion in Section 4.3.1).

4.3 Logo detection and consistency

Let X, Y be two random variables standing respectively for interest points in $\mathcal{S}_p$, $\mathcal{S}_q$, and $\{X_1, \ldots, X_n\}$ (resp. $\{Y_1, \ldots, Y_m\}$) as n (resp. m) realizations with the same distribution as X (resp. Y). Define also H_1 (resp. H_0) as the set of all possible matching points (resp. non matching points) taken from $\{\mathcal{S}_p\} \times \{\mathcal{S}_q\}$ according to a well defined ground truth.

4.3.1 Matching

Given X, a good matching strategy as will be shown in the remainder of this section, consists in declaring Y_J as a match iff the conditional probability on (X, Y_J) is larger than the sum of the conditional probabilities on $\{(X, Y_j),\ j \neq J\}$; leading to

$$\mathbf{K}_{Y_J|X} \ > \ \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X}, \tag{4.7}$$

here $\mathbf{K}_{Y|X} = \mathbf{K}_{X,Y} / \big(\sum_{j=1}^{m} \mathbf{K}_{X,Y_j} \big)$; the intuition behind choosing the above criterion comes from the fact that when $\mathbf{K}_{Y_J|X} \ \gg \ \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X}$, the entropy of the conditional probability distribution $\mathbf{K}_{.|X}$ will be close to 0, so given X, the uncertainty about its possible matches will be reduced.
Considering (4.7), we define its probability of success

$$p_s = P\bigg(\mathbf{K}_{Y_J|X} \ > \ \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X} \bigg), \tag{4.8}$$

this probability is with respect to $\{X, Y_1, \ldots, Y_m\}$. In the remainder of this section, we will discuss the consistency of the matching criterion (4.7) under H_1 and H_0.

Proposition 2. *given X, under the hypothesis of existence of a reference logo into a test image (i.e. $\exists Y_J$ s.t. $(X, Y_J) \in H_1$), the probability of success p_s (in 4.8) is*

$$\frac{\exp\big(n(1-1/Q)\big)}{\exp\big(n(1-1/Q)\big)+(m-1)}, \tag{4.9}$$

and if $Q > 1$, ($Q = N_r N_a$, see (i),(ii)), then p_s is exponentially-convergent to 1 with respect to n and decreasing with respect to m.

And under the hypothesis of non existence of a reference logo into a test image (i.e. $\nexists Y_J$, s.t. $(X, Y_J) \in H_1$), the probability of success (4.8) is $1/m$ which is convergent to 0 as m increases.

Proof. see appendix A. □

It results from the above proposition, that under H_1 (in contrast to H_0), p_s is an increasing function of n, Q and a decreasing function of m. For instance, if $m = 10.000$, $Q = 64$, then p_s reaches 1 with only $n \geq 20$ sample points in the reference logo. Clearly, this shows that the procedure is able to correctly match very few interest points (in $\mathcal{S}_p$) into a very large collection (in $\mathcal{S}_q$) as also corroborated through experiments.

4.3.2 Logo detection

Given a test image $\mathcal{S}_q$ and a reference logo $\mathcal{S}_p$, the latter is declared as present into $\mathcal{S}_q$ *if the number of times the inequality (4.7) is satisfied is larger than* τn ($\tau \in]0, 1]$); here $(1-\tau)$ is the amount of occlusion that $\mathcal{S}_p$ might have in $\mathcal{S}_q$ while still can be detected[3]. Let $\mathcal{X}_s$ ($\mathcal{X}_s \rightarrow \mathcal{B}(n, p_s)$) be a binomial random variable standing for the number of times good matches are found in $\mathcal{S}_q$, for the n points in $\mathcal{S}_p$, using (4.7). In this section, we are interested in lower bounding

$$P\big(\mathcal{X}_s \geq \tau n\big), \quad \tau \in]0, 1], \tag{4.10}$$

[3] *It reasonable to set $\tau = 0.5$, which means that a reference logo is still detectable event-though half-occluded in a test image.*

here P is the probability distribution of $\mathcal{X}_s$. Now, we provide our main result which allows us under some conditions to lower bound (4.10):

Proposition 3. *fix τ and consider $\mathcal{X}_s$ as a binomial random variable with parameter p_s. If p_s ($\in [0,1]$) is at least $\sqrt{-\frac{\ln(\delta/2)}{2n}} + \tau$, then*

$$P\big(\mathcal{X}_s \geq \tau n\big) \geq 1 - \delta \tag{4.11}$$

here $\delta \ll 1$ is a fixed error rate.

Proof. the left-hand side of the above inequality is equal to

$$\begin{aligned} & P\left(\sum_i^n Z_i \geq \tau n\right), \text{ here } \mathcal{X}_s = \sum_i^n Z_i,\ Z_i \to \mathcal{B}(1, p_s) \\ = \ & 1 - P\left(p_s - \frac{1}{n}\sum_i^n Z_i \geq p_s - \tau\right) \\ \geq \ & 1 - 2\exp\Big(-2n(p_s - \tau)^2\Big), \quad \text{(by Hoeffding's inequality)} \end{aligned} \tag{4.12}$$

the sufficient condition is

$$2\exp\Big(-2n(p_s - \tau)^2\Big) \leq \delta \Rightarrow p_s \geq \sqrt{-\frac{\ln(\delta/2)}{2n}} + \tau, \tag{4.13}$$

when $n \to +\infty$, and if p_s is at least equal to τ, then

$$P\big(\mathcal{X}_s \geq \tau n\big) \underset{n\to+\infty}{\longrightarrow} 1 \tag{4.14}$$

□

Now combining (4.9) and (4.13), the sufficient condition which guarantees (4.11) becomes under H_1

$$\frac{\exp\big(n(1-1/Q)\big)}{\exp\big(n(1-1/Q)\big) + (m-1)} \ \geq \ \sqrt{-\frac{\ln(\delta/2)}{2n}} + \tau, \tag{4.15}$$

which holds true mainly for larger n, Q, but $\tau < 1$, and even large m. For instance if $n = 20$, $m = 10.000$, $Q = 64$, the left hand side is very close to 1 and hence the inequality (4.15) will be satisfied even when $\tau \to 1$ (low occlusion factor) and $\delta \to 0$ (high lower bound).

4.3.3 Similarity invariance

The adjacency matrices $\mathbf{P}_{\theta,\rho}$, in $\mathbf{K}$, provide the intrinsic properties and also characterize the geometry of logos $\{\mathcal{S}_p\}$ in $\mathcal{X}$. It is easy to see that $\mathbf{P}_{\theta,\rho}$ is translation and rotation invariant and can also be made scale invariant when ϵ_p (see (i)) is adapted to the scales of $\psi_g(\mathcal{S}_p)$. It follows that the right-hand side of our similarity $\mathbf{K}$ is invariant to any $2D$ similarity transformation. Notice, also, that the left-hand side of $\mathbf{K}^{(t)}$ may involves similarity invariant features $\psi_f(.)$ (actually SIFT features), so $\mathbf{K}^{(t)}$ (and also the matching process) is similarity invariant.

4.4 Benchmarking

4.4.1 Test data and settings

In order to show the extra-value of our context dependent matching strategy both with respect to context free one and other approaches, we evaluate the performances of multiple-logo detection on the TradeMark-720 database containing 13 trademark classes each one represented with $14-87$ real world pictures, resulting into a collection of 720 images. 13 reference logos are used and correspond to trademarks: 1: "agip", 2: "apple", 3: "barilla", 4: "birra_moretti", 5: "cinzano", 6: "cocacola", 7: "esso", 8: "ferrari", 9: "heineken", 10: "marlboro", 11: "mcdonald", 12: "pepsi", 13: "starbucks". Note that each reference logo is synthetically transformed in order to generate 4 affine transformations. Interest points are extracted from test images as well as reference logos and encoded using the usual SIFT features.

Each test image $\mathcal{S}_q$ is processed in order to evaluate the similarity function $\mathbf{K}$ (shown in 4.4) with respect to each reference logo $\mathcal{S}_p$, using Gaussian power assist setting, i.e., $\mathbf{K}^{(0)}_{x,x'} = \exp(-d(x,x')/\beta)$. Our goal is to show the improvement brought when using $\mathbf{K}^{(t)}$, $t \in \mathbb{N}^+$, so we tested it against the standard context-free similarity (i.e., $\mathbf{K}^{(t)}$, $t = 0$). First, the setting of β is performed by maximizing the performance of the Gaussian similarity as the latter corresponds to the left-hand side (and the baseline form) of $\mathbf{K}^{(t)}$, i.e., when $\alpha = 0$.[4] For our database, we found that the best performances are achieved for $\beta = 0.1$ and this also guarantees condition (4.3) in practice.

[4]Notice that selecting β independently from α is obviously *"not sub-optimal"* for the context dependent similarity but *"sub-optimal"* for the Gaussian similarity.

The influence (and the performance) of the right-hand side of $\mathbf{K}^{(t)}, \alpha \neq 0$ increases as α increases nevertheless and as shown earlier, the convergence of $\mathbf{K}^{(t)}$ to a fixed point is guaranteed only if (4.2) is satisfied. Intuitively, the weight parameter α should then be relatively high while also satisfying condition (4.2). In practice, we found that the best α is 0.1; for a matter of space, more details about the setting of α, β can be found in a research report [184].

4.4.2 Performance, comparison and discussion

We used criteria (4.7), (4.10) in order to decide whether a given reference logo $\mathcal{S}_p$ exists into a test image $\mathcal{S}_q$. Different values were experimented for the tolerance factor τ and performances are measured using False Acceptance (FAR) and False Rejection Rates (FRR) defined as

$$\begin{aligned} \text{FAR} &= \mathbb{E}\left(\frac{\text{false positive}}{\text{false positive+ true negative}}\right) \\ \text{FRR} &= \mathbb{E}\left(\frac{\text{false negative}}{\text{false negative + true positive}}\right), \end{aligned} \tag{4.16}$$

here the expectation is with respect to all possible test images. Diagrams in (4.4), show the FAR, FRR errors for different classes (trademarks) of our test set; we clearly see the out-performance and the improvement of the our context dependent similarity function (i.e., $\mathbf{K}^{(t)}$, $t \in \mathbb{N}^+$), in logo detection, with respect to the baseline, i.e., context-free similarity ($\mathbf{K}^{(0)}$). For almost all the classes of the test set, the improvement brought by the "CD" similarity is clear and consistent; except the classes "apple" and "mcdonald" as their reference logos contain very few interest points ($n < 12$), and this makes (*consistently with our theoretical analysis*) inequality (4.15) difficult to satisfy mainly for high expectations about the lower bound in (4.11) (i.e., low δ) and when τ is relatively high.

Table. 4.1 shows a comparison of our context dependent similarity for logo matching and detection with respect to other techniques including SIFT matching and also with respect to (iterative) Ransac matching using the inliers/outliers of SIFT matching. Even though, the FAR and FRR results are variable depending on the setting of τ, in all these cases, the average error rates, defined as (FAR+FRR)/2, of our method are lower than those reported for SIFT matching and Ransac.

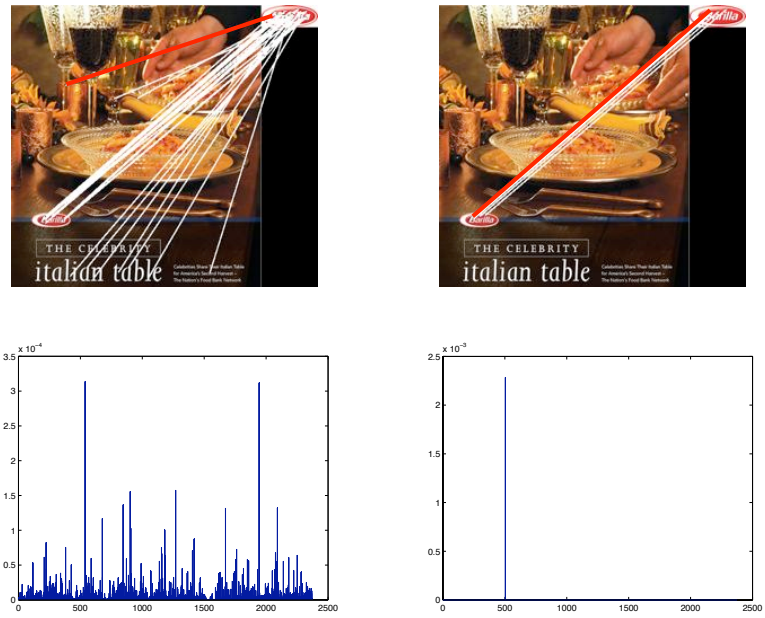


Figure 4.3: This figure shows a comparison of the matching results when using a naive matching strategy without context and our context dependent matching. (Bottom figures) show the conditional probability distribution $K_{.|X}$ for a particular interest point X in the reference logo. This distribution is peaked when using context dependent similarity so the underlying entropy is close to 0 and the uncertainty about possible matches is dramatically reduced. (Top figures) show the matching results between the reference logo and the test image which are correct using the context dependent matching framework.

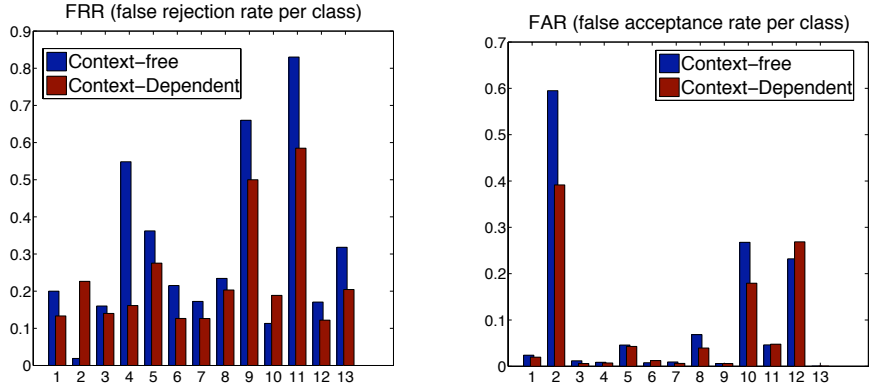


Figure 4.4: This figure shows a comparison of logo detection using our (i) context-dependent similarity and (ii) context-free one (actually Gaussian). FAR and FRR rates are shown for each class. In these experiments, $\beta = \alpha = 0.1$ and $\tau = 0.5$ while n and m vary of course with reference logos and test images. Excepting the logos "Apple" and "Mc Donald's" (which contain very few interest points $n < 12$), the FRR errors are almost always significantly reduced while FAR is globally reduced.

Figure 4.5: These pictures shows logo detection results; in all these pictures, all the 13 reference logos were checked using criterion (4.10). Match points are also displayed.

Thresholds (τ)	0.1	0.2	0.3	0.4	0.5
Errors	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR
Sift Matching	0.299/0.312	0.509/0.148	0.632/0.090	0.730/0.054	0.773/0.039
Ransac Matching	0.395/0.087	0.527/0.035	0.620/0.014	0.705/0.005	0.753/0.003
CD Matching	**0.095/0.279**	**0.109/0.220**	**0.111/0.200**	**0.120/0.188**	**0.120/0.182**
Thresholds (τ)	0.6	0.7	0.8	0.9	1
Errors	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR
Sift Matching	0.822/0.026	0.856/0.020	0.893/0.015	0.924/0.011	0.950/0.009
Ransac Matching	0.802/0.002	0.827/0.002	0.848/0.001	0.866/0.001	0.877/0.0006
CD Matching	**0.128/0.177**	**0.132/0.173**	**0.131/0.171**	**0.135/0.167**	**0.137/0.1650**

Table 4.1: This table shows a comparison of our method, with respect to Sift and Ransac matching; in all these experiments we clearly see that the global error rates (defined as $\frac{1}{2}(FAR+ERR)$) of our method are better than those reported for standard matching techniques. Notice also that FAR is an increasing function of the occlusion factor $(1-\tau)$ while FRR is a decreasing function.

4.5 Conclusion

We introduced in this work a novel logo detection and localization approach based on a new class of similarities referred to as context dependent. The strength of the proposed method resides in several aspects (i) the inclusion of the information about the spatial configuration in similarity design as well as visual features (ii) the ability to control the regularization of the solution via our energy function (iii) the invariance to many transformations including translation, scale, rotation and also partial occlusion, and (iv) the theoretical groundness of the matching framework which shows that under the hypothesis of existence of a reference logo into a test image, the probability of success of matching and detection is high while very low under background.

Further extensions of this work include the application of the method to logo retrieval in videos and also the refinement of the definition of context in order to handle other rigid and non-rigid logo transformations.

Chapter 5

A SIFT-based forensic method for copy-move detection

One of the principal problem image forensics has to deal with is determining if a particular image is authentic or not. This task is very important in all those fields where is crucial to use such digital content as evidence like, for instance, in a court of law. To carry out such forensic analysis various technological instruments have been developed in literature. Many of them try to reveal if some modifications have been performed thus assessing that something of suspect could have been made, other ones search for comprehending what has happened and possibly which relations there are with other linked photos. In this chapter the problem of detecting if a feigned image has been created is investigated; in particular, attention has been paid to the case in which an area of an image is copied and then pasted onto another zone to make a duplication or to cancel something that was awkward. To detect such modifications we propose a new methodology based on SIFT features. Our method allows both to understand if a copy-move attack has occurred and which are the image points involved, and, furthermore, to recover which has been the geometric transformation happened to perform cloning.[1]

[1] A preliminary version of the work presented in this chapter has been published as "Geometric tampering estimation by means of a SIFT-based forensic analysis" in *Proc. of IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010*, [5] and submitted to IEEE Trans. on Information Forensics and Security (TIFS).

5.1 Introduction

Digital crime, together with constantly emerging software technologies, is growing at a rate that far surpasses defensive measures. Sometimes a digital image or a video are incontrovertible evidence of a crime or the proof of a malevolent action. By looking at a digital content as a digital clue, *multimedia forensics* aims at introducing novel methodologies to support clue analysis and to provide an aid for making a decision on a crime. Multimedia forensics [142, 67, 179] deals with developing technological instruments operating in the absence of any watermark [52, 21] or signature inserted in the image. In fact, diversely from digital watermarking, forensics means are defined as "passive" because they can formulate an assessment on a digital document only by resorting to such digital asset. These techniques basically allow the user to determine if a certain content has been tampered [68, 175] or which has been the adopted acquisition device [210, 47]. In particular, by focusing on the task of acquisition device identification, two are the main aspects that have to be studied: the first one is to understand which kind of device has generated that digital image (e.g. a scanner, a digital camera or is a computer graphics product) [107, 42], while the second one is to succeed in determining which specific camera or scanner (by recognizing model and brand) has acquired that particular content [210, 47].

The other main multimedia forensics topic is about image tampering detection [68], assessing the authenticity or not of a digital image. Information integrity is fundamental in a trial, but it is clear that the advent of digital pictures and relative ease of digital image processing makes today this authenticity uncertain. Two examples of this problem, that recently appeared in newspapers and TV news [2], are given in Fig. 5.1 and Fig. 5.2. Modifying a digital image to change the meaning of what is represented in it, could be crucial when it is used in a court of law, where images are presented as basic evidences to influence the judgement. Furthermore, it would be interesting, once established that something has been manipulated, to understand exactly what is happened: if an object or a person has been covered, if a part of the image has been cloned, if something has been copied from another image or, even more, if a combination of these processes have been carried

[2] See *http://thelede.blogs.nytimes.com/2008/07/10/in-an-iranian-image-a-missile-too-many/* and *http://littlegreenfootballs.com/weblog/?entry=24492_Iranian_Fauxtography_Bust&only/*

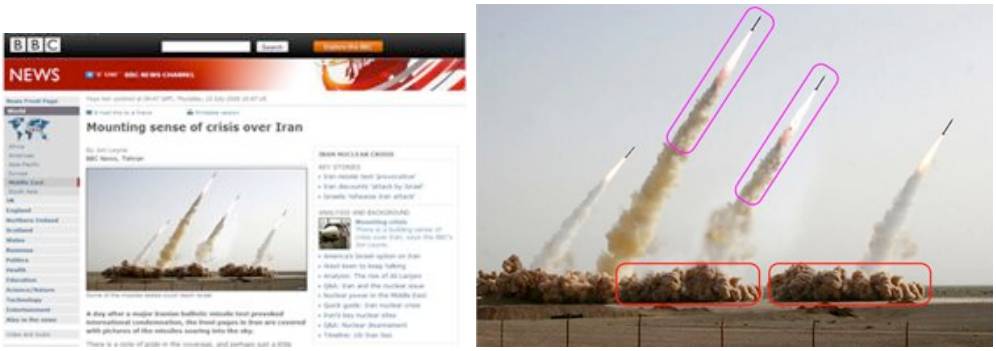


Figure 5.1: An example of image tampering appeared on press in July 2008. The feigned image (on the right) shows four Iranian missiles but only three of them are real; two different sections (encircled in red and purple respectively) replicate other image sections by applying a copy-move attack.

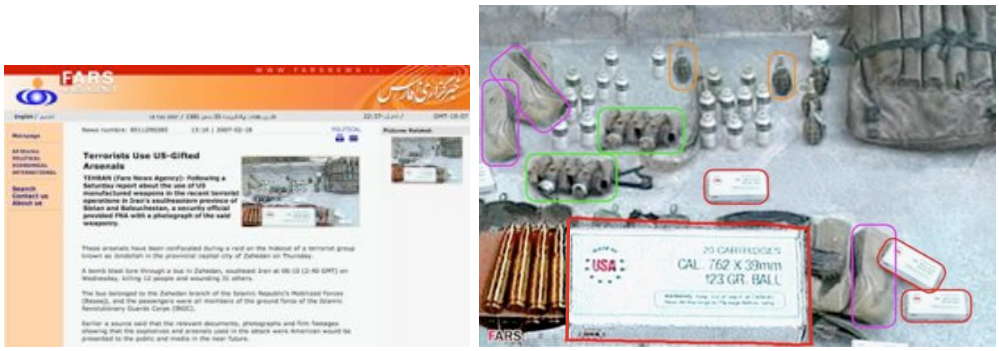


Figure 5.2: A close look at this picture, appeared on news in 2007 (Fars News Agency, Tehran), shows that many elements are cloned over and over. Also in this case the cloned sections are encircled in different colors.

out. In particular, when an attacker creates his feigned image by cloning an area of the image onto another zone (*copy-move* attack), he is often obliged to apply a geometric transformation to satisfactorily achieve his aim.

In this chapter this issue is investigated, and the proposed method is able to individuate if the copy-move tampering has taken place and also to estimate the parameters of the transformation occurred (i.e. horizontal and vertical translation, scaling factors, rotation angle). On the basis of our preliminary work [5], a new methodology which answers to this requirement is presented hereafter. Such a technique is based on Scale Invariant Features Transform (SIFT) [137], which are used to robustly detect and

describe clusters of points belonging to cloned areas. Successively, these points are exploited to reconstruct the parameters of the occurred geometric transformation. The proposed technique has also been tested against *splicing* attack (i.e. when an image block is duplicated onto another different image). In fact, in a context where the source image is available (e.g. the forensic analyst has to check a suspect dataset which contains both the source and the destination image) this methodology can be still applied. The rest of the chapter is structured as follows: Section 5.2 presents related works regarding copy-move forgery detection. Moreover, the contribution and the novelty of our approach respect to the state-of-the-art is discussed. Section 5.3 presents the proposed method in its three main stages, while experimental results on forgery detection and on applied transformation parameters estimation are presented in Section 5.4. Conclusions are finally drawn in Section 5.5.

5.2 SIFT Features for Image Forensics

One of the most common image manipulations is to clone (copy and paste) portions of the image, for instance, to conceal a person or an object in the pictured scene. When this is done with care, and retouch tools are used, it can be very difficult to detect cloning. Moreover, since the copied parts are from the same images some components (e.g noise and color) will be compatible with the rest of the image and thus will not be detectable using methods that look for incompatibilities in statistical measures in different parts of the image [23, 69]. Furthermore, since the cloned regions can be of any shape and location, it is computationally impossible to search all possible image locations and sizes with an exhaustive search as pointed out in [77].

The problem of copy-move forgery detection has been faced by proposing different approaches each of these based on the same concept: a copy-move forgery introduces a *correlation* between the original image area and the pasted one. Several methods search this dependence dividing the image into overlapping blocks and then applying a feature extraction process in order to represent the image blocks by using a low dimensional representation. In [140] the averages of red, green and blue components are chosen together with other four features computed on overlapping blocks, obtained by calculating the energy distribution of luminance along four different directions. A different approach is presented in [127] in which the features are represented

by the Singular Value Decomposition (SVD) performed on low-frequency coefficients of the block-based Discrete Wavelet Transform (DWT). The authors in [144] proposed a block representation calculated using blur invariants. Their specific aim is to find features invariant to the presence of blur artifacts that a falsifier can apply to make detection of forgery more difficult. Then they used Principal Component Analysis (PCA) to reduce the number of features and a k-tree to identify the interested regions. In [61] authors present a technique to detect cloning when the copied part has been modified using two specific tools, the *Adobe Photoshop* healing brush and the Poisson cloning. Others two algorithms [77] and [174] based on using low dimensional representation of blocks and fast sorting to improve efficiency have been developed to detect copy-move image regions. In particular, the authors in [77] apply a Discrete Cosine Transform (DCT) to the block. Duplicated regions are then detected by lexicographically sorting the DCT block coefficients and grouping similar blocks with the same spatial offset in the image. While in [174] the authors apply PCA on image blocks to yield a reduced-dimension representation. Duplicated regions are again detected by lexicographically sorting and grouping all of the image blocks. A related approach is the method in [25] where a Fourier Mellin Transform is applied on each block. A forgery decision is made when there are more then a given number of blocks that are connected to each other and the distance between block pairs is the same. To create a convincing forgery, it is often necessary to resize, rotate, or stretch portions of an image. For example, when creating a composition of two objects, one object may have to be resized to match the relative heights. This process requires re-sampling of the original image introducing specific periodic correlations between neighboring pixels. The presence of these correlations due to the re-sampling can be used to detect that something happened to the image [173] but not to detect the specific manipulation.

So a good copy-move forgery detection should be robust to some types of transformations as rotation and scaling and also to some manipulations including JPEG compression, Gaussian noise addition and gamma correction. Most of the existing methods do not deal with all these manipulations and are often computationally prohibitive. In particular the method in [174] is not able to detect scaling or rotation transformation, whereas with the methods in [77] and [25] only small variations in rotation and scaling are identifiable as reported in [24]. The authors in [182] make an attempt to

overcome this problem solving copy-move identification when only rotation of the copied area takes place by using Zernike moments. This issue is also discussed in [129] where rotation transformation and JPEG compression and Gaussian noise manipulations are analyzed to understand how they could affect the copy-move detection. Authors in [40] instead propose a method to detect duplicated and transformed regions through the use of a block description invariant to reflection and rotation such as the log-polar block representation summed along its angle axis. Finally a comparison among some of copy-move methods described above has been reported in [49] evaluating the performance of each methods with and without geometric transformation applied to the copied patch.

Nowadays local visual features (e.g SIFT, SURF, GLOH, etc.) have been widely used for the particular tasks of image retrieval and object recognition, due to their robustness to several geometrical transformations (such as rotation and scaling), occlusions and clutter. More recently few attempts have been done to apply this kind of features also in the digital forensics domain; in fact, SIFT features have been used for fingerprint detection [199], shoeprint image retrieval [209], and also for copy-move detection.

5.2.1 Our contribution

A very preliminary work on copy-move forgery detection based on SIFT features was proposed in [95], but in that paper no estimation of the parameters of the applied geometric transformation is performed and, furthermore, extended numerical results to evaluate real performances of the methodology (e.g. True/False Positive Rates) are not provided. Another very recent work has been presented in [168], but, though the technique is able to deal with region extraction by resorting to a correlation map, it can not manage affine transformation and, also in this case, quantitative results on the reliability of the estimate of geometric transformation parameters are not given; in addition to this, the approach adopts many different empirical thresholds whose setting seems to be not completely unsupervised. Moreover none of these contributions considers accurately the case of multiple copy-move forgeries. As we will show furthermore, this is a key point in a realistic forensic scenario since often a forged image contains several cloned areas (like in the case of Fig. 5.2).

In this scenario is placed the proposed method that is able to detect and

then to estimate the geometrical transformation occurred in a copy-move forgery attack. Multiple copy-move forgeries are managed by performing a robust feature matching procedure and then a clustering on the keypoints coordinates in order to separate the different cloned areas. These two tasks are fundamental since otherwise, in case of multiple cloning, it is often impossible to detect and separate each forgery and also to estimate the geometric transformation. Estimating the geometric parameters with accuracy is deemed as a fundamental task not only to understand how the cloned patch has been processed [224] and possibly to infer which was the counterfeiter's motive, but also to compare the original source block of image and the forged one on a common ground; furthermore a reliable estimate of the transformation permits to register the two patches for a possible deeper forensics analysis [138]. The method proposed hereafter is able to deal with affine geometric transformations and, as witnessed by experimental results, can grant a reliable estimate of the transformation parameters. Such a technique acts by relying on a unique empirical threshold which regulates clustering operation, and that has been determined by a training procedure on a general dataset. This is a very important issue also in comparison with other similar techniques like that in [168].

5.3 The proposed method

The proposed approach is based on the SIFT algorithm to extract robust features which can allow to discover if a part of an image was copy-moved and furthermore which geometrical transformation was applied. In fact, the copied part has basically the same appearance of the original one, thus keypoints extracted in the forged region will be quite similar to the originals. Therefore, matching among SIFT features can be adopted for the task of determining a possible tampering. A simple schematization of the whole system is shown in Fig. 5.3: the first step consists of SIFT features extraction and keypoint matching, the second step is devoted to cluster such keypoints and assess forgeries detection, while the third one is in charge to estimate the occurred geometric transformation, if a tampering has been individuated.

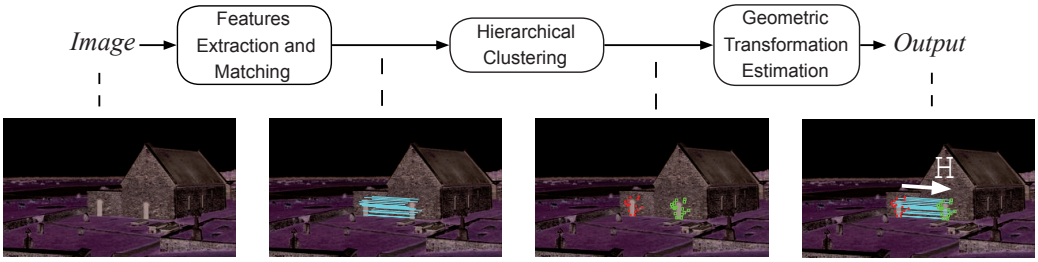


Figure 5.3: Overview of the proposed system. SIFT matched pairs and clusters.

5.3.1 SIFT features extraction and multiple keypoint matching

Given a test image, a set of keypoints $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with their corresponding SIFT descriptors $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ is extracted. A matching operation is performed in the SIFT space among the $\mathbf{f}_i$ vectors of each keypoint to identify similar local patches in the test image. The best candidate match for each keypoint $\mathbf{x}_i$ is found by identifying its nearest neighbor from all the other $(n-1)$ keypoints of the image, which is the keypoint with the minimum Euclidean distance in the SIFT space. In order to decide for a matching between two keypoints (i.e. "are these two descriptors the same or not?"), simply evaluating the distance between two descriptors with respect to a global threshold does not perform well. This is due to the high-dimensionality of the feature space (128) in which some descriptors are much more discriminative than others.

We can obtain a more effective procedure, as suggested in [137], by using the ratio between the distance of the closest neighbor to that of the second-closest one, and comparing it with a threshold T (often fixed to 0.6). For the sake of clarity, given a keypoint we define a similarity vector $\mathbf{D} = \{d_1, d_2, \ldots, d_{n-1}\}$ that represents the sorted euclidean distances with respect to the other descriptors. Following this idea, the keypoint is matched only if this constraint is satisfied:

$$\frac{d_1}{d_2} < T \quad where \quad T \in (0, 1). \tag{5.1}$$

We refer to this procedure as 2NN test. But also this matching procedure shows a main drawback: it is unable to manage multiple keypoint matching. This is a key aspect in case of copy-move forgeries since it may happen that

the same image area is cloned over and over (see for example Fig. 5.2). In other words, it only finds matches between keypoints whose SIFT descriptors are very different from those of the rest of the set (i.e. features that are *globally distinctive*). Therefore, the case of cloned patches is very critical since the keypoints detected in those regions are very similar to each other.

For this reason we propose a novel matching procedure, that is a generalization of (5.1), and is able to deal with multiple copies of the same features. Our generalized 2NN test (referred as g2NN) starts from the observation that in a high-dimensionality feature space such as that of SIFT features, keypoints that are different from the inspected one share very high and very similar values (in terms of Euclidean distances) among them. Instead, similar features show low Euclidean distances respect to the others. The idea of the 2NN test is that the ratio between the distance of the candidate match and the distance of the $2nd$ nearest neighbor is low in the case of a match (e.g. lower than 0.6) and very high in case of two "random features" (e.g. greater than 0.6). Our generalization consists in iterating the 2NN test between d_i/d_{i+1} until this ratio is greater than T (in our experiments we set this value to 0.5). If k is the value in which the procedure stops, each keypoint in correspondence to a distance in $\{d_1, \ldots, d_k\}\,(where\ 1 \leqslant k < n)$ is considered as a match for the inspected keypoint.

Finally, by iterating on each keypoint belonging to $\mathbf{X}$, we can obtain the set of matched points. All the matched keypoints are held, instead isolated ones are no more considered in the following processing steps. Already at this stage a draft idea of the authenticity of the image is provided. But it can happen that images that are legitimately containing areas with very similar texture, can yield to matched keypoints that might induce false alarms: the following two steps of the proposed methodology try to reduce this possibility.

5.3.2 Clustering and forgeries detection

To identify possible cloned areas, an *agglomerative hierarchical clustering* [90] is performed on spatial locations (i.e. x, y coordinates) of the matched points. Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure. The algorithm starts by assigning each keypoint to a cluster; then it computes all the reciprocal spatial distances among clusters, finds the closest pair of clusters, and finally merges them into a single

cluster. Such computation is iteratively repeated until a final merging situation is achieved. The way such final merging can be accomplished is basically conditioned both by the linkage method adopted and by the threshold used to stop clusters' grouping.

Several linkage methods exist in the literature and our experiments evaluate their performance and then estimate the best cut-off threshold T_h (see Subsection 5.4.1 for a detailed description of such experiments) for forgery detection. In particular, three different linkage methods have been taken into account: *Single*, *Centroid* and *Ward's* linkage. Given two clusters P and Q, respectively containing n_P and n_Q objects (where $\mathbf{x}_{Pi}$ and $\mathbf{x}_{Qj}$ indicate the i^{th} and the j^{th} object in the clusters P and Q), the diverse linkage method operates as it follows:

- *Single* linkage uses the smallest euclidean distance between objects in the two clusters:

$$\begin{aligned} dist(P,Q) &= \min(\|\mathbf{x}_{Pi}, \mathbf{x}_{Qj}\|_2) \\ with\ i &= [1, n_P],\ j = [1, n_Q]. \end{aligned} \tag{5.2}$$

- *Centroid* linkage uses the euclidean distance between the centroids of the two clusters:

$$dist(P,Q) = \|\overline{\mathbf{x}}_P - \overline{\mathbf{x}}_Q\|_2 \tag{5.3}$$

where

$$\overline{\mathbf{x}}_P = \frac{1}{n_P}\sum_{i=1}^{n_P}\mathbf{x}_{Pi} \quad \text{and} \quad \overline{\mathbf{x}}_Q = \frac{1}{n_Q}\sum_{i=1}^{n_Q}\mathbf{x}_{Qi}. \tag{5.4}$$

- *Ward's* linkage evaluates the increment/decrement (5.5) in the *Error Sum of Squares* (ESS) after merging the two clusters into a single one with respect to the case of two separated clusters:

$$\Delta_{dist}(P,Q) = ESS(PQ) - [ESS(P) + ESS(Q)] \tag{5.5}$$

where

$$ESS(P) = \sum_{i=1}^{n_P}|\mathbf{x}_{Pi} - \overline{\mathbf{x}}_P|^2, \tag{5.6}$$

$\overline{\mathbf{x}}_P$ is the centroid (again) and PQ indicates the combined cluster.

According to the adopted linkage method, a specific tree structure is obtained. In addition to this, the proper choice of the threshold T_h, to determine where to cut the tree and consequently which is the final number of clusters, is crucial. The parameter which is utilized to be compared with T_h is the *Inconsistency Coefficient* (IC) which characterizes each clustering operation; the higher the value of this coefficient, the less similar the objects connected by the link, thus when it exceeds the threshold T_h clustering stops. IC takes basically into account the average distance among clusters and does not allow to join clusters spatially too far at that level of hierarchy. It is easy to understand that an appropriate assumption of T_h directly influences tampering detection performances. At the end of clustering procedure, however clusters which do not contain a significant number (more than three) of matched keypoints are eliminated. On this basis, to optimize detection performances and consequently to the carried out experimental tests (see again Subsection 5.4.1), it has been established to consider that an image has been altered by a copy-move attack, if the method detects two (or more) clusters with at least three pairs of matched points that link a cluster to another one. This aspect has been investigated and this assumption grants a good trade-off between the need to provide a low false alarm rate.

It is worthy to point out that can occur the case where no matched keypoints are obtained, mainly because salient features are not revealed in the forged patch (e.g. when an object is hidden with a flat patch): anyway this is a very well-known open issue in SIFT-related scientific literature.

5.3.3 Geometric transformation estimation

When an image has been classified as non-authentic, the proposed method allows to determine which is the geometrical transformation occurred between the original area and its copy-moved version. Let the matched point coordinates be, for the two areas, $\tilde{\mathbf{x}}_i = (x, y, 1)^{\mathtt{T}}$ and $\tilde{\mathbf{x}}'_i = (x', y', 1)^{\mathtt{T}}$ respectively, their geometric relationships can be defined by an affine homography which is represented by a 3×3 matrix $\mathtt{H}$ as:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \mathtt{H} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{5.7}$$

This matrix can be computed by resorting at three matched points at least. In particular, we determine $\mathtt{H}$ by using Maximum Likelihood estimation of

the homography [89]. This method seeks homography $\mathtt{H}$ and pairs of perfectly matched points $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}'_i$ that minimizes the total error function as in Equation 5.8:

$$\sum_i \left[d(\mathbf{x}_i, \hat{\mathbf{x}}_i)^2 + d(\mathbf{x}'_i, \hat{\mathbf{x}}'_i)^2 \right] \text{ subject to } \hat{\mathbf{x}}'_i = \mathtt{H}\hat{\mathbf{x}}_i \ \forall i. \tag{5.8}$$

However mismatched points (*outliers*) can severely disturb the estimated homography. For this purpose we perform the previous estimation by applying the RANdom SAmple Consensus algorithm (RANSAC) [75]. Such algorithm randomly selects a set (in our case three pairs of points) from the matched points and estimates the homography H, then all the remained points are transformed according to H and compared in terms of distance with respect to their corresponding matched ones. If this distance is under or above a certain threshold β, they are catalogued as *inliers* or *outliers* respectively. After a pre-defined number N_{iter} of iterations, the estimated transformation which is associated with the higher number of inliers is chosen. In our experimental tests, N_{iter} has been set to 1000 and the threshold β to 0.05; this is due to the fact that we used a standard method of normalization of the data for homography estimation. The points are translated so that their centroid is at the origin and then they are scaled so that the average distance from the origin is equal to $\sqrt{2}$. This transformation is applied to both of the two areas $\mathbf{x}_i$ and $\mathbf{x}'_i$ independently.

Once the affine homography is found, rotation and scaling transformations can be computed by its decomposition, while translation can be determined by considering the centroids of the two matched clusters. In particular, $\mathtt{H}$ can be represented as:

$$\mathtt{H} = \begin{bmatrix} \mathtt{A} & \mathbf{t} \\ \mathbf{0}^\mathtt{T} & 1 \end{bmatrix} \quad \text{where} \quad \mathtt{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{5.9}$$

The matrix $\mathtt{A}$ is the composition of rotation and non-isotropic scaling transformations. In fact, it can always be decomposed as

$$\mathtt{A} = \mathtt{R}(\theta)(\mathtt{R}(-\Phi)\mathtt{S}\mathtt{R}(\Phi)) \tag{5.10}$$

where $\mathtt{R}(\theta)$ and $\mathtt{R}(\Phi)$ are rotations by θ and Φ respectively, and $\mathtt{S} = \mathtt{diag}(s_1, s_2)$ is a diagonal matrix for the scaling transformation. Hence, the $\mathtt{A}$ defines the concatenation of a rotation by Φ, a scaling by s_1 and s_2 respectively in the rotated x and y directions; a rotation back by $-\Phi$; and finally another

rotation by θ. This decomposition is computed directly by the SVD (Singular Value Decomposition). In fact, the matrix $\mathtt{A}$ can be also rewritten as: $\mathtt{A} = \mathtt{USV}^{\mathtt{T}} = (\mathtt{UV}^{\mathtt{T}})(\mathtt{VSV}^{\mathtt{T}}) = \mathtt{R}(\theta)(\mathtt{R}(-\Phi)\mathtt{SR}(\Phi))$ since $\mathtt{U}$ and $\mathtt{V}$ are orthogonal matrices.

5.4 Experimental results

In this section we evaluate the proposed methodology providing two main kinds of the tests: firstly, on a small dataset named MICC-F220, a benchmarking of the technique is done to properly set the operative threshold T_h and to compare it with other methods known in the literature; secondly, on a larger dataset named MICC-F2000, a complete evaluation is carried out by testing the system against different types of modifications. Both datasets are composed by images with different contents coming from the Columbia photographic images repository [161] and from a personal collection. The first dataset MICC-F220 is composed by 220 images: 110 are tampered images and 110 are originals. The images resolution varies from 722×480 to 800×600 pixels and the size of the forged patch covers, on the average, 1.2% of the whole image. The second dataset MICC-F2000 is composed by 2000 photos of 2048×1536 pixels (3M pixels) and the forgery is, on the average, 1.12% of the whole image: so it is again quite small and similar to the MICC-F220 dataset case. To reproduce as much as possible a practical situation, the number of original and altered images belonging to the MICC-F2000 dataset is not the same: 1300 original images and 700 tampered images have been taken. The forged images are obtained, in both the datasets, by randomly selecting (both as location and as dimension) an image area (squared or rectangular) and copy-pasting it over the image after having applied a number of different attacks such as translation, rotation, scale (symmetric/asymmetric) or a combination of them.

Table 5.1 and Table 5.2 summarize the geometric transformations for the attack applied in the MICC-F220 dataset (10 attacks, from A to J in Table 5.1) and in the MICC-F2000 (14 attacks, from a to o in Table 5.2) respectively. In particular, for each attack, is reported the rotation θ expressed in degrees and the scaling factors s_x, s_y applied to the x and y axis of the cloned image part (e.g. in the attack G, the x and y axes are scaled by 30%, and no rotation is performed).

Attack	θ °	s_x	s_y
A	0	1	1
B	10	1	1
C	20	1	1
D	30	1	1
E	40	1	1

Attack	θ °	s_x	s_y
F	0	1.2	1.2
G	0	1.3	1.3
H	0	1.4	1.2
I	10	1.2	1.2
J	20	1.4	1.2

Table 5.1: The 10 different combinations of geometric transformations applied to the original patch for the MICC-F220 dataset.

Attack	θ °	s_x	s_y
a	0	1	1
b	0	0.5	0.5
c	0	0.7	0.7
d	0	1.2	1.2
e	0	1.6	1.6
f	0	2	2
g	0	1.6	1.2

Attack	θ °	s_x	s_y
h	0	1.2	1.6
i	5	1	1
j	30	1	1
l	70	1	1
m	90	1	1
n	40	1.1	1.6
o	30	0.7	0.9

Table 5.2: The 14 different combinations of geometric transformations applied to the original patch for the MICC-F2000 dataset.

5.4.1 Settings for forgery detection

First of all the proposed method is analyzed to determine the best settings for the cut-off threshold T_h introduced in Section 5.3.2 according to the chosen linkage method. Such values will be set up for the successive phase of experiments and comparisons. To address this problem, the following experiment has been set-up applying a *4*-fold cross-validation process: from the database of 220 images (MICC-F220), 165, that is 3/4 of the image set, (82 tampered and 83 original) have been randomly chosen to perform a training to find the best threshold T_h for each of the three considered linkage methods (*Single*, *Centroid*, *Ward's*); the remaining 55 images (1/4 of the whole set) have been used in a successive testing phase to evaluate detection performances of the proposed technique. The experiment was repeated four times, by cyclically exchanging the four image sub-sets belonging to the training (3 sub-sets) and to the testing set (1 sub-set), and the results have been averaged. Detection performances have been measured in terms of

True Positive Rate (TPR) and False Positive Rate (FPR), where TPR is the fraction of tampered images correctly identified as such, while FPR is the fraction of original images that are not correctly identified as such:

$$\begin{aligned} \text{TPR} &= \frac{\#\text{ images detected as forged being forged}}{\#\text{ forged images}}, \\ \text{FPR} &= \frac{\#\text{ images detected as forged being original}}{\#\text{ original images}}. \end{aligned}$$

We underline that has been assumed to consider that an image has been altered by a copy-move attack, if the method detects two (or more) clusters with at least three pairs of matched points that link a cluster to another one (as debated in Subsection 5.3.2).

In Table 5.3, for each linkage method, the TPR and the FPR obtained during the training phase are reported with respect to the threshold T_h which varies in the interval [0.8, 3] with steps of 0.2.

	Single		*Centroid*		*Ward's*	
T_h	FPR(%)	TPR(%)	FPR(%)	TPR(%)	FPR(%)	TPR(%)
0.8	2.729	41.827	1.822	23.626	0.911	10.906
1	5.455	70.001	4.547	56.373	3.636	32.739
1.2	8.180	89.994	7.273	90	7.273	82.714
1.4	8.180	95.456	8.180	95.456	7.273	90.905
1.6	***8.180***	***98.185***	7.273	97.274	8.180	97.274
1.8	7.269	96.360	***8.180***	***98.182***	9.088	99.089
2	6.362	91.820	7.269	95.456	9.088	100
2.2	5.451	82.721	5.451	92.723	***8.177***	***100***
2.4	4.544	63.639	4.544	84.536	7.269	96.364
2.6	2.726	48.185	2.729	70.897	7.273	89.998
2.8	0.911	22.726	1.822	46.360	3.640	78.170
3	0.911	15.461	0.911	18.179	3.640	61.813

Table 5.3: Training phase: TPR and FPR values (in percentage) for each metric with respect to T_h.

The goal was to minimize the FPR while maintaining a very high TPR; as it can be seen FPR is almost always very low, on the contrary TPR is very variable, so the optimal threshold T_h has been chosen, as evidenced

in Table 5.3, for the maximum value of TPR that means 1.6 for the *Single* linkage method, 1.8 for the *Centroid* and 2.2 for the *Ward's* linkage. Finally, the test phase has been launched for the best metrics by using the T_h previously obtained in the training phase. The final detection results are reported in Table 5.4. These values show that the proposed method performs satisfactorily providing a low FPR though maintaining an high rate of correct tampering detection basically for all the used linkage method, though *Ward's* metric seems to be slightly better. It is possible to conclude that the choice of linkage method is not so fundamental while T_h setting is crucial.

	Single	*Centroid*	*Ward's*
FPR (%)	8.16	8.16	8
TPR (%)	98.21	98.17	100

Table 5.4: Test phase on MICC-F220 dataset: detection results in terms of FPR and TPR.

Furthermore, for the cases of correctly detected forged images, the estimation of the geometric transformation parameters which bring the original patch onto the forged one has also been computed. The Mean Absolute Error (MAE) between each of the true values of the transformation parameters and the estimated ones are reported in Table 5.5. As in the previous tables, s_x and s_y refers to the scaling factors occurred in the transformation; θ refers to the rotation (in degrees) while t_x and t_y are translation on x/y direction respectively.

MAE (t_x)	MAE (t_y)	MAE (θ)	MAE (s_x)	MAE (s_y)
4.04	2.48	0.94	0.021	0.015

Table 5.5: Transformation parameters estimation errors for the MICC-F220 (*Single* linkage method with $T_h = 1.6$, as previously underlined other metrics give similar performances). The values t_x and t_y are expressed in pixels while θ in degrees.

Results show an high degree of precision in the estimate of the various parameters of the affine transformation. In addition to this, Table 5.6 reports for one of the test image belonging to the MICC-F220, named *Cars* (see Fig. 5.4 first column), each transformation parameter (the original value applied to the patch, the estimated one and the absolute error ($|e|$)). It can

A	t_x	$\hat{t}_x$	$\|e\|$	t_y	$\hat{t}_y$	$\|e\|$	θ	$\hat{\theta}$	$\|e\|$	s_x	$\hat{s}_x$	$\|e\|$	s_y	$\hat{s}_y$	$\|e\|$
A	304	304.02	0.02	80.5	81.01	0.51	0	0.040	0.040	1	1.004	0.004	1	0.998	0.002
B	304	305.20	1.20	80.5	82.42	1.92	10	9.963	0.037	1	1.001	0.001	1	0.999	0.001
C	304	305.55	1.55	80.5	82.64	2.14	20	20.009	0.009	1	1.006	0.006	1	0.998	0.002
D	304	305.04	1.04	80.5	82.49	1.99	30	30.092	0.092	1	1.002	0.002	1	0.998	0.002
E	304	306.08	2,08	80.5	78.43	2.07	40	39.932	0.067	1	1.007	0.007	1	1.004	0.004
F	304	304.88	0.88	80.5	80.41	0.09	0	0.080	0.080	1.2	1.202	0.002	1.2	1.198	0.002
G	304	305.07	1.07	80.5	79.87	0.63	0	0.108	0.108	1.3	1.304	0.004	1.3	1.303	0.003
H	304	305.78	1.78	80.5	80.18	0.32	0	0.037	0.037	1.4	1.403	0.003	1.2	1.206	0.006
I	304	305.23	1.23	80.5	81.76	1.26	10	9.910	0.090	1.2	1.203	0.003	1.2	1.201	0.001
J	304	305.02	1.02	80.5	80.82	0.32	20	20.067	0.067	1.4	1.404	0.004	1.2	1.198	0.002

Table 5.6: Transformation parameters estimation on image *Cars*. The values t_x and t_y are expressed in pixels while θ in degrees.

be observed how reliable the estimate is, specifically for the scale parameters and also for an asymmetric scaling combined with a rotation.

Qualitative evaluation

Hereafter, some experimental results on images where a copy-move attack has been performed by taking into account the context are reported. In this case the patch is selected according to the specific goal to be achieved and, above all, transformed by paying attention to perfectly conceal the occurred modification. Alterations are not recognizable at least at a first rough watch and a forensic tool could help in investigation action. In Fig. 5.4, four of these specific cases are pictured by presenting the tampered image and the corresponding one where matched keypoints and clusters, extracted by the proposed method, are highlighted. Interesting situation concerns the individuation of a cloned patch for the image named *Dune* (second column) where, though the duplicated area is quite flat, the method is able to detect a sufficient number of matched keypoints. On the contrary, an opposite case is registered for the image named *Santorini* (last column), where a very high amount of matched keypoints is obtained; now the cloned block is very textured and though it has undergone a geometrical transformation to be properly adapted to the context, the SIFT algorithm is so robust not to be disturbed.

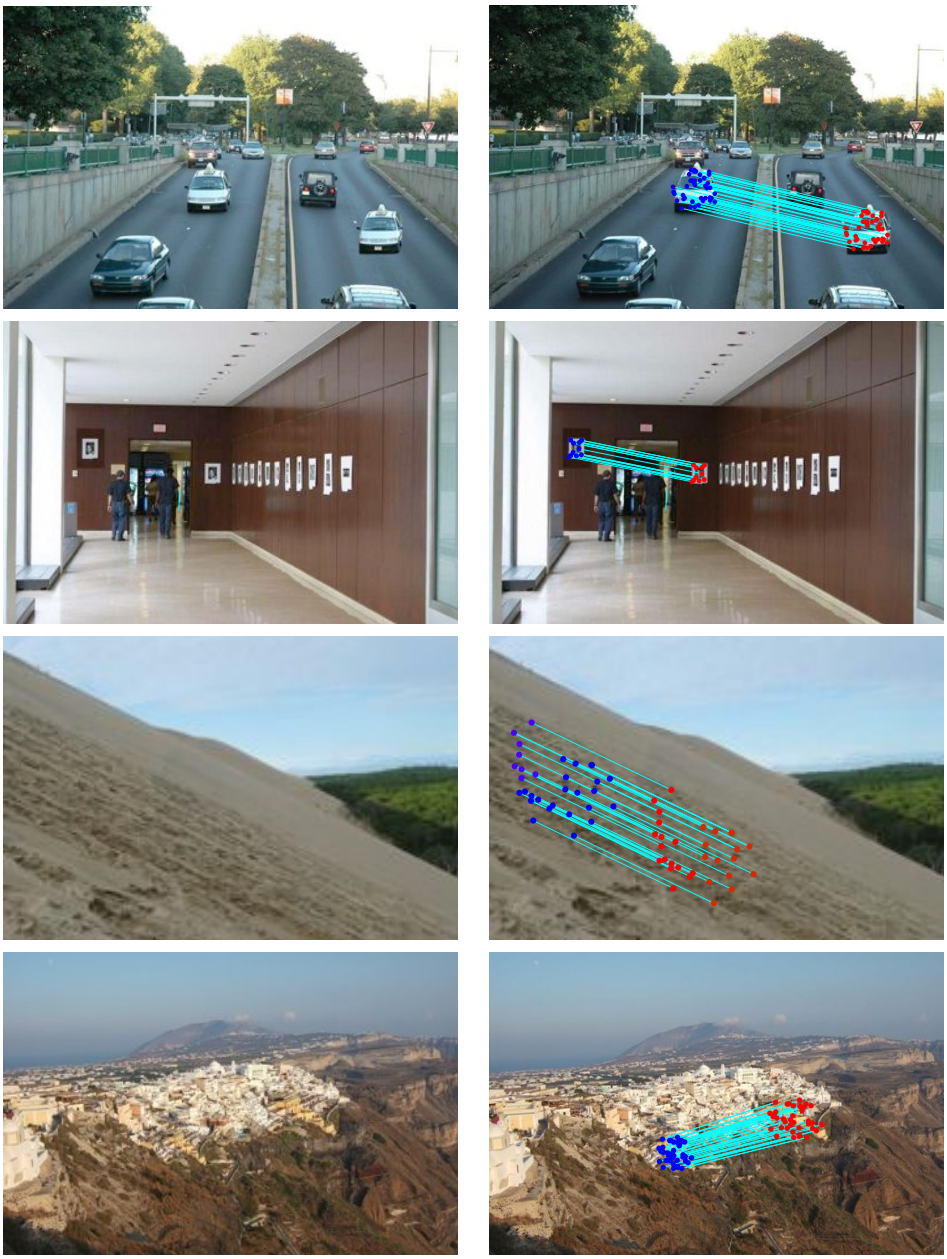


Figure 5.4: Some examples of tampered images are pictured in the first column, while the corresponding detection results are reported in the second column.

Copy-move methods comparison

The proposed approach has been compared to the results obtained with our implementations of the methods presented in [77], based on Discrete Cosine Transform (DCT), and in [174], based on Principal Component Analysis (PCA) (please note that both have been previously introduced in Section 5.2). The input parameters required by the two methods are set as it follows: $b = 16$ (number of pixels per block), $N_n = 5$ (number of neighborhood rows to search in the lexicographically sorted matrix), $N_f = 1000$ (threshold for the minimum frequency) and $N_d = 22$ (threshold to determine a duplicated block). These parameters are used in both the algorithms, while $e = 0.01$ (fraction of the ignored variance along the principle axes after PCA is computed) and $Q = 256$ (number of the quantization bins) are only used for the method presented in [174]. In our method, the *Ward's* linkage with a threshold $T_h = 2.2$ has been assumed.

The experiments have been launched on the whole MICC-F220 image database on a machine with an *Intel Q6600 quad core with 4-GB RAM (linux os)* and the FPR, TPR and the processing time have been evaluated. Table 5.7 shows the detection performance and the processing time on average (in seconds) for an image relatively to each methodology.

Method	FPR (%)	TPR (%)	Time (s)
Fridrich *et al.* [77]	84	89	294.69
Popescu and Farid [174]	86	87	70.97
Our method	8	100	4.94

Table 5.7: TPR, FPR values (%) and processing time (one image averagely) for each method.

The results point out that the proposed method performs better with respect to the others methods; in fact the processing time (per image) is on average about 5 seconds, whereas the other two take more than 1 minute and almost 5 minutes respectively. Furthermore DCT and PCA methods, though presenting an acceptable TPR, fail when a decision about original images is required (high FPR values in Table 5.7). Anyway this is basically due to the incapacity of such methods to properly deal with cases where a geometrical transformation which is not just a translation is applied to the copy-moved patch. For the specific case of simple patch translation FPR is 0% for all the three methods.

5.4.2 Test on multiple copied regions

In this experiment we analyze the performance of our method in presence of tampered images which have multiple copies of a same region. This test has been performed on ten photos of 2048×1536 pixels. In these pictures, one or more image areas are copied and pasted in several different positions over the image, taking into account the context in order to hide, at the first glance, the forgery.

In this scenario, as we have previously highlighted in Subsection 5.3.1, the standard 2NN matching procedure is a critical point for copy-move forgery detection methods based on SIFT features [95, 168]. In fact, comparing the standard SIFT matching technique with our g2NN strategy, we determine that our method increases (averagely) of 195% the number of the extracted matches. A high number of matches is fundamental in order to have sufficient information for a correct estimation of the geometric transformation, but it can introduce false alarms. To this end, we tested our matching strategy on MICC-F220 dataset in order to evaluate how these new matches influence the results. We obtained that we lose on average 3% in terms of FPR with the same results in terms of TPR.

Fig. 5.5 shows a qualitative comparison between the two techniques. It is interesting to note that the number of the matched keypoints between the gun a and c, obtained by the standard 2NN technique, are very few (2 matches) with respect to g2NN (54 matches). In this case the technique, based on standard matching, fails to detect the relationship between the two guns. Finally, Fig. 5.6 shows some examples of multiple cloning obtained with the g2NN test. Detection results are reported by highlighting matched keypoints and clusters.

5.4.3 Test on a large dataset

In this section, experimental results obtained on a larger dataset, named MICC-F2000, to verify the behavior of the proposed technique are presented; detection performances and geometric transformation parameters estimation are investigated as well. Furthermore tests to check the robustness of the method against usual operations such as JPEG compression or noise addition, an image can undergo, have been carried out; such kinds of processing have been considered as applied both to the whole forged image and only to

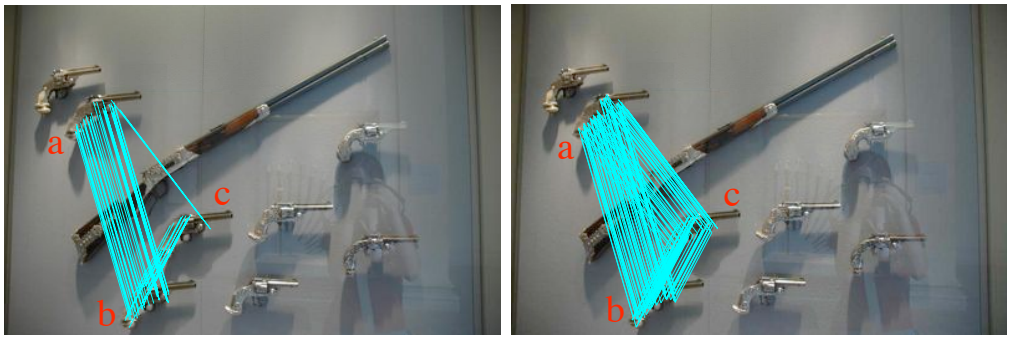


Figure 5.5: Matched keypoints computed by the 2NN standard SIFT matching technique (left) and our g2NN strategy (right).

the altered image patch.

	Single		*Centroid*		*Ward's*	
T_h	FPR(%)	TPR(%)	FPR(%)	TPR(%)	FPR(%)	TPR(%)
0,8	3.41	51.86	1.69	32.29	0.54	11.43
1	5.56	70.19	4.92	62.43	3	51.29
1,2	10.28	89.95	10.31	87.43	9.54	83.86
1,4	10.95	91.24	12.15	90.14	11.62	88.43
1.6	**10.97**	**93**	13.23	93.57	13.15	93.14
1.8	9.46	91	**12.46**	**93.43**	14.54	93.86
2	7.46	84.43	11.23	92.29	13.85	93.86
2.2	4.79	72.38	9.00	89.43	**11.62**	**93.43**
2,4	2.72	54.43	6.46	78.43	9.85	91.29
2,6	1.00	29.14	3.23	62.86	8.46	87.71
2,8	0.21	19.86	1.23	40.86	5.62	79.43
3	0.08	12.86	0.38	23.29	3.38	67.43

Table 5.8: Training phase on MICC-F2000 dataset: TPR and FPR values (in percentage) for each metric with respect to T_h.

First of all, we have tried to set up again an experiment for the determination of the best threshold T_h, according to the three linkage methods, as done in Subsection 5.4.1 for the MICC-F220; this has been made to further check if the established thresholds were correct. To do that a 4-fold cross-valitation process has been carried out. Results are listed in Table 5.8. It can be observed that a similar behavior to that obtained with MICC-F220 is registered and, above all, that the values chosen in Subsection 5.4.1 for

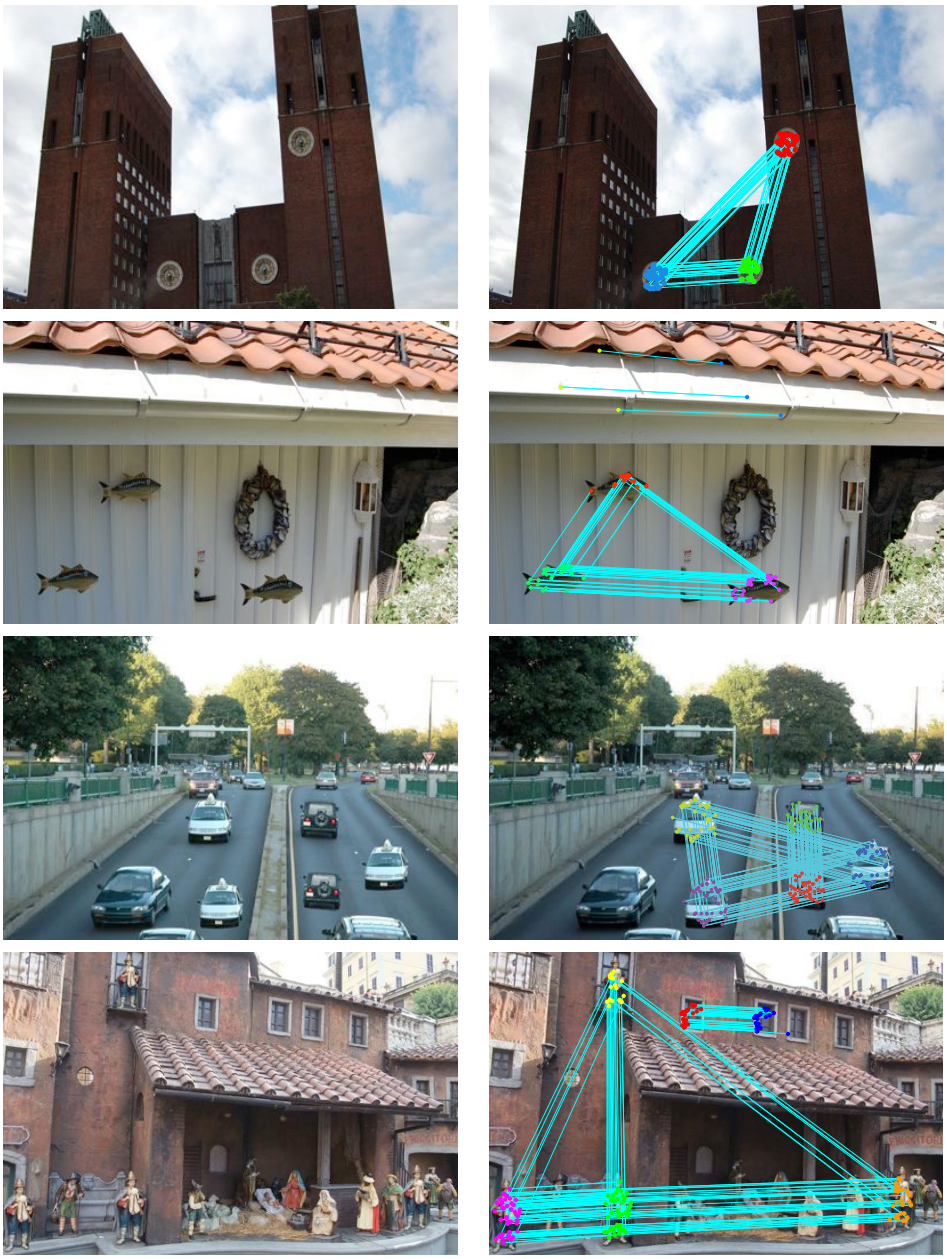


Figure 5.6: Examples of tampered images with multiple cloning are shown in the first column, while the detection results are reported in the second column.

T_h (1.6 for *Single*, 1.8 for *Centroid* and 2.2 for *Ward's*) still grant about the higher performances in terms of TPR and FPR. After this, the test phase is launched by setting such values for T_h and in Table 5.9 the detection rates are reported demonstrating both the effectiveness of the proposed method which achieves a TPR around 93% for all the three metrics and its robustness obtaining again performances very coherent to those presented in Table 5.8 for these fixed thresholds.

	Single	*Centroid*	*Ward's*
FPR (%)	10.99	12.45	11.61
TPR (%)	92.99	93.23	93.42

Table 5.9: Test phase on MICC-F2000 dataset: detection results in terms of FPR and TPR obtained with $T_h = 1.6$, $T_h = 1.8$ and $T_h = 2.2$ for the three linkage methods respectively.

Going into detail, in Fig. 5.7 the number of errors for each attack is listed with regard to tampered images not detected as such. The most critical attacks seem to be the f ($\theta = 0°$, $s_x = 2$ and $s_y = 2$) and the n ($\theta = 40°$, $s_x = 1.1$ and $s_y = 1.6$) which increase twice the patch dimension and apply a 40 degrees rotation combined with a consistent variation on scale respectively. The histogram in Fig. 5.7 shows that these two kinds of attacks generate everyone around the 30% of the total errors.

In Table 5.10 are then reported the estimate errors for the geometric transformation parameters averaged on all the 500 test images. The Mean Absolute Error (MAE) still remains small enough although the transformations applied to the images in this circumstance for MICC-F2000 dataset are more challenging with respect to the case of MICC-F220 dataset.

MAE(t_x)	MAE(t_y)	MAE(θ)	MAE(s_x)	MAE(s_y)
22.49	8.49	1.55	0.27	0.2

Table 5.10: Transformation parameters estimation errors. The values t_x and t_y are expressed in pixels while θ in degrees.

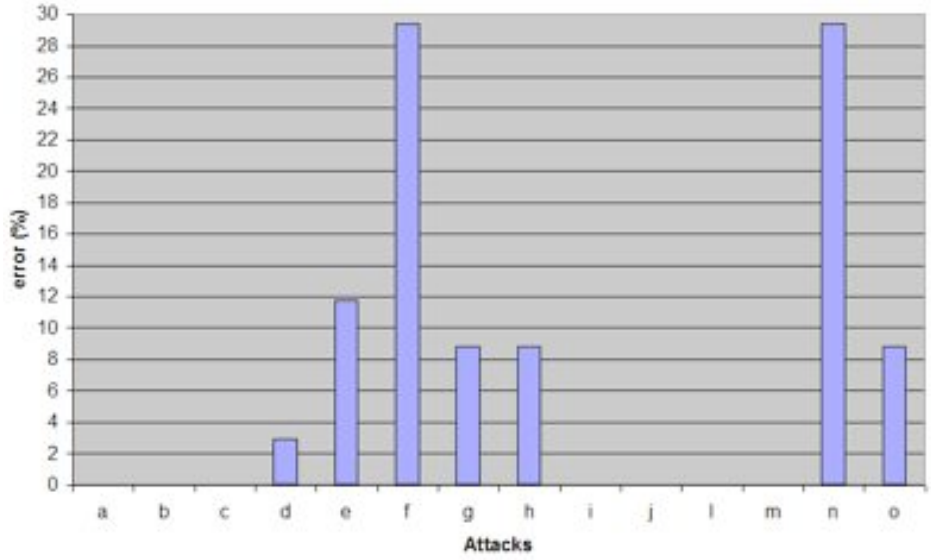


Figure 5.7: Error analysis of tampered images misdetection for each different attack (in percentage).

JPEG compression and noise addition

The proposed methodology has also been tested in terms of detection performances from a robustness point of view; in particular, the impact of JPEG compression and then of noise addition on all the 2000 images of the MICC-F2000 dataset has been investigated. In the first experiment all the images which were originally in the JPEG format (quality factor of 100), have been compressed in JPEG format with a decreasing quality factor of 75, 50, 40 and 20. Table 5.11 (left) shows the FPR and TPR (*Ward's* linkage method with $T_h = 2.2$) for all the diverse JPEG quality factors; it can be seen that FPR is practically stable while the TPR tends to slightly diminish when image quality decreases. In the second experiment, in the same way as before, the images of MICC-F2000 dataset are distorted by adding a Gaussian noise obtaining different final decreasing signal-noise-ratios (SNR) of 50, 40, 30 and 20 db. Noisy images are obtained by adding white Gaussian noise to the image with a JPEG quality factor of 100. In Table 5.11 (right), obtained results are shown and it can be noticed that the TPR is over 90% till a SNR of 30 dB while FPR is again quite stable, though it seems to even improve.

JPEG quality	FPR	TPR
100	11.61	93.42
75	12.07	93.42
50	11.15	93.16
40	11.38	92.14
20	10.46	87.15

SNR (dB)	FPR	TPR
50	11.46	93.71
40	11.69	94.14
30	11.46	92.00
20	8.15	82.42

Table 5.11: Detection performances against JPEG compression (left) and noise addition (right).

JPEG compression, noise addition, gamma correction on cloned patches

The duplicated patches are often modified by applying some further processing such as brightness/contrast adjustment, gamma correction, noise addition and so on, in order to adjust the patch with respect to the image area where it has to be located. So to explore this scenario the following experiment has been made. Starting from 10 original images, a block is randomly (as explained before) selected for each of them and 4 geometric transformations (a, d, j and o from Table 5.2) are applied to every of these patches. Furthermore, before pasting them, 4 different gamma corrections with values [2.2, 1.4, 0.7, 0.45] are applied to each single block. Finally, 160 tampered images are obtained. In the same way, the final stage of gamma correction is firstly substituted by JPEG compression with different quality factors [75, 50, 40, 20] and secondly by Gaussian noise addition with SNR (dB) equal to 50, 40, 30, 20. For every case, 160 fake images have been created. So for each of the three situations (gamma correction, JPEG compression and noise addition), a dataset composed by 160 fake images and by 350 original ones randomly taken from the MICC-F2000 database is built. In Table 5.12, performances in terms of TPR and FPR are reported.

These experiments show that the proposed method maintains its level of accuracy though some diverse kinds of post-processing are applied to the duplicated patch in addition to a geometric transformation, to adapt it to the image context where it is pasted.

Kind of processing	FPR	TPR
Gamma correction	9.23	99.37
JPEG	11.38	100.00
SNR (dB)	12.00	100.00

Table 5.12: Detection performances against gamma correction, JPEG compression and noise addition applied to the duplicated and geometrically transformed patch.

5.4.4 Image splicing

Though the proposed technique has been presented to operate in a copy-move attack scenario, it can also be utilized in a context where a splicing operation has occurred. With the term splicing attack is intended that a part of an image is grabbed and, possibly after having been adapted (geometric transformed and/or enhanced), pasted onto another one to build a new fake image. In most of the cases only the final fake photo is available to the forensic analyst for inspection, the source one is often undeterminable; because of this, the SIFT matching procedure, which is the core of the proposed method could not take place and would seem that there is no room for it in such circumstance. Anyway this is not always true in practice! In fact, often, the analyst is required to give an assessment over a dataset of images for example belonging to a specific person under judgement, or that have been found in a hard disk or a pen drive, and so on. In this operative scenario, it can happen that the source image used to create a fraudulent content belongs to the image collection at disposal. It is easy to understand that the proposed method can be adopted again to determine both if within the to-be-checked collection there is a false image containing an "external" patch and, above all, where it comes from. It is interesting to highlight that succeeding in detecting such link could help investigation activities. To prove that the proposed technique can be used in such a scenario the following experimental test has been set up.

A subset of 100 images (96 original and 4 tampered with) taken from a private collection with size of 800×600 pixels has been selected. In particular, the 4 fake images have been created by pasting a patch that was cut from another image belonging to the other original 96. The proposed technique has been launched to analyze all the possible pairs of photos $\left(\frac{n \cdot (n-1)}{2} =\right.$

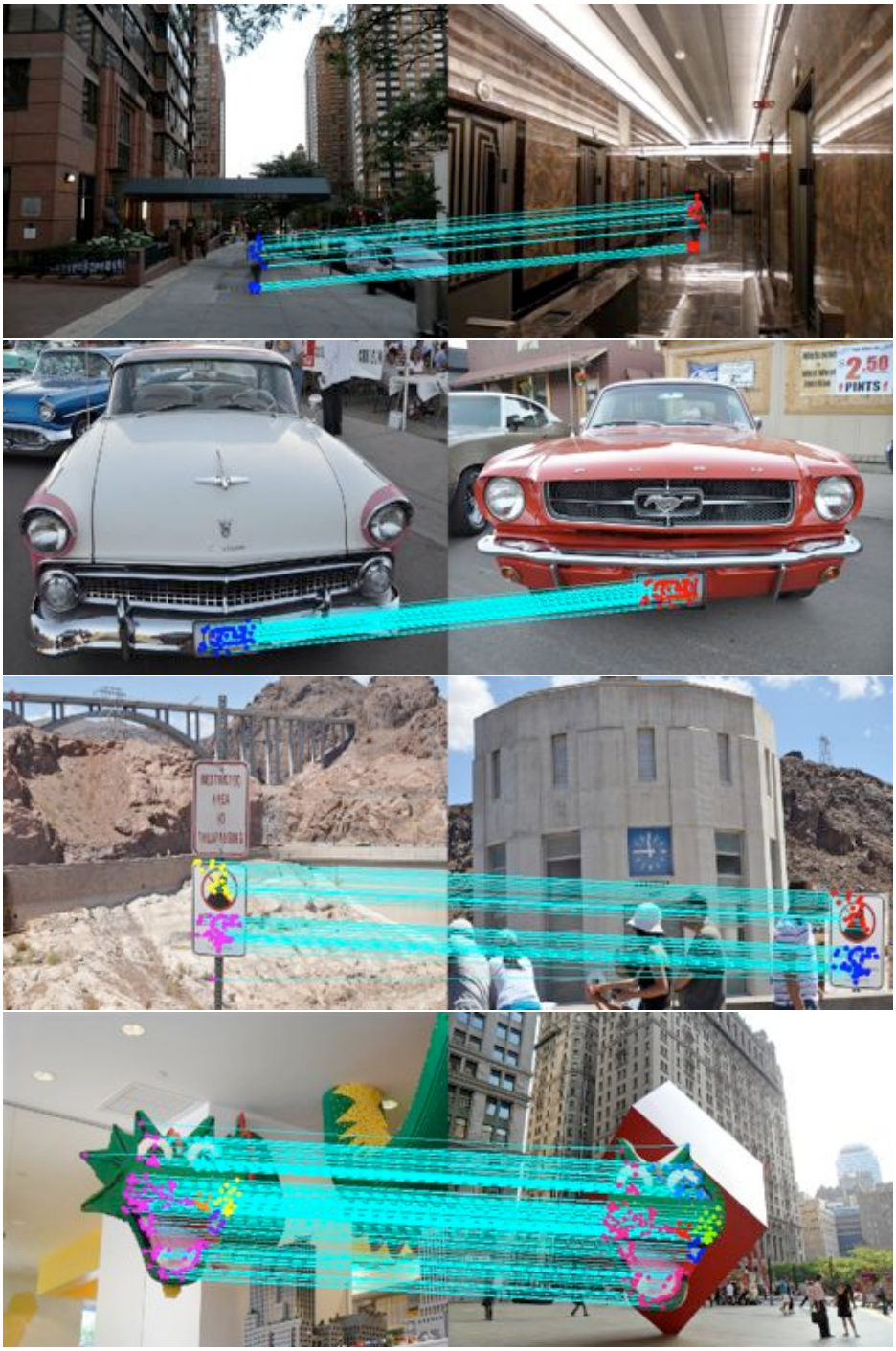


Figure 5.8: Examples of correct detection of splicing attack.

Figure 5.9: An example of wrong detection of splicing attack.

$\frac{100 \cdot 99}{2} = 4950$) within the dataset looking for duplicated areas. To allow to the presented algorithm to perform as it is the pair of images to be checked are considered as a single image with a double number of columns (size equal to $N \times 2M$); due to this fact, the detection threshold T_h has been moved up to 3.4 (it was 2.2 in the previous experimental tests of this section) for the *Ward's* linkage method which was chosen for this specific experiment. In Table 5.13 performances on FPR and TPR are reported.

Splicing attack	FPR (%)	TPR (%)
	0.04	100.00

Table 5.13: Detection performances against splicing attack (in percentage).

The method is able to correctly reveal all the four fake pairs as expected determining a link between the possible original image and the forged one, though it can not distinguish the source from the destination as well-known if other tools are not adopted. The procedure also detects as suspected two other innocent couples of images incurring in false alarms. In Fig. 5.8 the four cases of splicing attack detection are pictured, while in Fig. 5.9 one of the false alarm is illustrated. In this last circumstance, it is immediate to understand that the error is induced by the presence of the same objects (the posters over the wall of the wooden box) in both the photos taken in the same real context. However this could be the actual situation that might happen in practical scenario (e.g. establishing possible relations among photos acquired in similar environments).

5.5 Conclusion

A novel methodology to support image forensics investigation based on SIFT features has been proposed. Given a suspected photo, it allows to reliably detect if a certain region has been duplicated and, furthermore, to determine the geometric transformation applied to perform such tampering. The presented technique has shown effectiveness with respect to diverse operative scenarios such as composite processing and multiple cloning. Future works will be mainly dedicated to investigate how to improve detection phase with respect to cloned image patch with highly uniform texture where salient keypoints are not recovered by SIFT-like techniques. In particular, integration with other forensics techniques applied locally onto flat zones is envisaged. Furthermore, clustering phase will be extended by means of an image segmentation procedure.

Chapter 6

Video event classification using string kernels

Event recognition is a crucial task to provide high-level semantic description of the video content. The bag-of-words (BoW) approach has proven to be successful for the categorization of objects and scenes in images, but it is unable to model temporal information between consecutive frames. In this chapter we present a method to introduce temporal information for video event recognition within the BoW approach. Events are modeled as a sequence composed of histograms of visual features, computed from each frame using the traditional BoW. The sequences are treated as strings (phrases) where each histogram is considered as a character. Event classification of these sequences of variable length, depending on the duration of the video clips, are performed using SVM classifiers with a string kernel that uses the Needlemann-Wunsch edit distance. Experimental results, performed on two domains, soccer videos and a subset of TRECVID 2005 news videos, demonstrate the validity of the proposed approach. [1]

[1]This chapter has been published as "Video Event Classification using String Kernels" in *Multimedia Tools and Applications, vol. 48, iss. 1, pp. 69-87, 2010* [20].

6.1 Introduction

Recently it has been shown that part-based approaches are effective methods for object detection and recognition due to the fact that they can cope with partial occlusions, clutter and geometrical transformations. Many approaches have been presented, but a common idea is to model a complex object or a scene by a collection of local interest points. Each of these local features describes a small region around the interest point and therefore they are robust against occlusion and clutter. To achieve robustness to changes of viewing conditions the features should be invariant to geometrical transformations such as translation, rotation, scaling and also affine transformations. In particular, SIFT features by Lowe [137] have become the de facto standard because of their high performances and (relatively) low computational cost. In fact, SIFT features have been frequently applied to object or scene recognition and also to many other related tasks. In this field, a solution that recently has become very popular is the Bag-of-Words (BoW) approach. It has been originally proposed for natural language processing and information retrieval, where it is used for document categorization in a text corpus, where each document is represented by its word frequency. In the visual domain, an image or a frame of a video is the visual analogue of a document and it can be represented by a bag of quantized invariant local descriptors (usually SIFT), called *visual-words* or *visterms*. The main reason for its success is that it provides methods that are sufficiently generic to cope with many object types simultaneously. We are thus confronted with the problem of generic visual categorization [201, 72, 236, 242], like classification of objects or scenes, instead of recognizing a specific class of objects. The efficacy of the BoW approach is demonstrated also by the large number of systems based on this approach that participate in the PASCAL VOC and TRECVID [202] challenges.

More recently, part-based models have been successfully applied also to the classification of human actions [59, 163], typically using salient features that represent also temporal information (such as spatio-temporal gradients), and to video event recognition. These tasks are particularly important for video indexing and retrieval where dynamic concepts occur very frequently. Even if few novel spatio-temporal features have been proposed, the most common solution is to apply the traditional BoW approach using static features (e.g. SIFT) on a keyframe basis. Unfortunately, for this purpose the

(a) (b) (c)

Figure 6.1: Keyframe-based video event detection. (a) Is it *shot-on-goal* or *placed-kick*? (b) Is it *walking* or *running*? (c) Is it a *car exiting* or *entering* from somewhere?

standard BoW approach has shown some drawbacks with respect to the traditional image categorization task. Perhaps the most evident problem is that it does not take into account temporal relations between consecutive frames. In this way, event detection suffers from the incomplete dynamic representation given by the keyframe, resulting in a poor performance compared to the results obtained for the detection of static concepts. Fig. 6.1 shows a few examples of difficulties that arise when performing event detection using only keyframes. Nevertheless, only few works have been proposed to cope with this problem [221, 231].

In this chapter, we present a novel method to model actions as a sequence of histograms (one for each frame) represented by a traditional bag-of-words approach. An action is described by a "phrase" of variable length, depending on the clip's duration, thus providing a global description of the video content that is able to incorporate temporal relations. Then video phrases can be compared by computing edit distances between them and, in particular, we use the Needleman-Wunsch distance [156] because it performs a global alignment on sequences dealing with video clips of different lengths. Using this kind of representation we are able to perform classification of video events and, following the promising results obtained in text categorization [135] and in bioinformatics (e.g. protein classification) [125], we investigate the use of SVMs with a string kernel, based on edit distance, to perform classification. Experiments have been performed on soccer and news video datasets, comparing the proposed method to a baseline kNN classifier and to a traditional keyframe-based BoW approach. Experimental results obtained by SVM and string kernels outperform the other approaches and, more generally, they

demonstrate the validity of the proposed method.

The rest of the chapter is organized as follows. A review of related previous works is presented in the next section. The techniques for frame and event representation are discussed in Sect. 6.3. The classification method, including details about the SVM string kernel, is presented in Sect. 6.4. Experimental results are discussed in Sect. 6.5 and, finally, conclusions are drawn in Sect. 6.6.

6.2 Related works

Video event detection and recognition is really challenging because of complex motion, occlusions, clutter, geometric transformations and illumination changes. Nevertheless, it is an essential task for automatic video content analysis and annotation. Previous works in this field can be roughly grouped into three main categories; abnormal/unusual event detection [241, 36, 230], human action categorization [59, 163, 35, 194, 120], and video event recognition [221, 231, 105, 64, 92].

Unusual event detection and activity recognition are very active research areas in video surveillance and many different approaches have been previously proposed. Several of these works rely on HMM models or Dynamic Bayesian Networks. In [241], Zhang *et al.* used HMMs to model usual events from a large training set; unusual event models are learned in a second step through Bayesian adaptation. The problem of detecting suspicious behaviors in video sequences is addressed also by Boiman and Irani [36]. They posed the problem as an inference process in a probabilistic graphical model, used to describe large ensembles of patches at multiple spatio-temporal scales. Inferred unusual configurations are treated as suspicious behaviors.

Over the past decade, the specific problem of recognizing human actions has received considerable attention from the research community. In fact, an automatic human activity recognition method may be very useful for many applications such as video surveillance, video annotation and retrieval and human-computer interaction. The early works in this field are usually based on holistic representations. For example, Bobick *et al.* [35] proposed motion history images to encode short spans of motion; this representation is then matched using global statistics, such as moment features. Although this method is efficient, it is assumed to have a well segmented foreground and background. More recently, part-based appearance models have been

successfully applied to the human action categorization problem, because they overcome some limitations of holistic models such as the necessity of performing background subtraction and tracking. These approaches rely on salient visual features that represent also temporal information (such as spatio-temporal intensity gradients) or motion descriptors like optical flow. Laptev [118] proposed a spatio-temporal interest point detector by extending the Harris corner operator. Local features are extracted from locations of the video which exhibit strong variations of intensity both in spatial and temporal directions. Dollar *et al.* [59] applied separable linear Gabor filters, treating time differently from space and looking for locally periodic motion. These features have been frequently used by different researchers within part-based frameworks (e.g. the BoW approach) in combination to learning techniques such as support vector machines (SVM) [194] and probabilistic latent semantic analysis (pLSA) [163]. More recently, Laptev *et al.* [120] have abandoned the interest point detection approach, preferring a structural representation based on dense temporal and spatial scale sampling (inspired by spatial pyramids), providing state-of-the-art results and showing promising results also on realistic video settings. Instead, in [105] action is modeled by a space-time volume in the video sequence and volumetric features (based on optic flow) are extracted for event detection. However, the performance of these methods heavily depends on the spatio-temporal features which often privilege high-motion regions. As a result, the approach is very sensitive to motion, thus providing high performance in the recognition of motion events that are more frequent in constrained video domains such as videosurveillance.

The generalization of this approach to less constrained and more general domains, like news videos or movies, has not been demonstrated. Therefore, the most common solution is to apply the traditional BoW approach using static features (such as SIFT descriptors) on a keyframe basis; in fact, many of the methods that have been submitted to the TRECVID competition extend this idea. Unfortunately, the application of this approach to event classification has shown some drawbacks with respect to the traditional image categorization task. The main problem is that it does not take into account temporal relations between consecutive frames, and thus event classification suffers from the incomplete dynamic representation. Recently some attempts have been made to employ temporal information among static part-based representations of video frames. Xu and Chang [231] proposed

to apply Earth Mover's Distance (EMD) and Temporally Aligned Pyramid Matching (TAPM) for measuring video similarity; EMD distance is incorporated in a SVM framework for event detection in news videos. In [221], BoW is extended constructing relative motion histograms between visual words (ERMH-BoW) in order to employ motion relativity and visual relatedness. Zhou *et al.* [243] presented a SIFT-Bag based generative-to-discriminative framework for video event detection, providing improvements on the best results of [231] on the same TRECVID 2005 corpus. They proposed to describe video clips as a bag of SIFT descriptors by modeling their distribution with a Gaussian Mixture Model (GMM); in the discriminative stage, specialized GMMs are built for each clip and video event classification is performed.

Similar approaches for event detection in news videos have been applied also at a higher semantic level, using the scores provided by concept detectors as synthetic frame representations or exploiting some pre-defined relationships between concepts. For example, Ebadollahi *et al.* [64] proposed to treat each frame in a video as an observation, applying then HMM to model the temporal evolution of an event. Yang and Hauptmann [234] proposed to exploit temporal consistency between nearby shots (described by their concept score) obtaining a temporal smoothing procedure for improving video retrieval. In [76] an ontology framework (VERL) has been defined for representation and annotation of video events. Finally, Bertini *et al.* [28] have recently presented an ontology-based framework for semantic video annotation by learning spatio-temporal rules; in their approach, an adaptation of the First Order Inductive Learner (FOIL) is used to learn rule patterns that have been then validated on few TRECVID 2005 video events.

6.3 Event representation and classification

Given a set of labeled videos, our goal is to automatically learn event models to perform categorization of new videos. In this work, we investigate in particular a new way of representing an event and how to learn this representation. An overview of our approach is illustrated in Fig. 6.2.

Structurally an event is represented by a sequence of frames, that may have different lengths depending on how it has been carried out. We model an event by a sequence of visual word frequency vectors, computed from the frames of the sequence; considering each frequency vector as a *character* we call this sequence (i.e. string) *phrase*. Additionally, we define a kernel, based

on an edit-distance, used by SVMs to handle variable-length input data such as this kind of event representation.

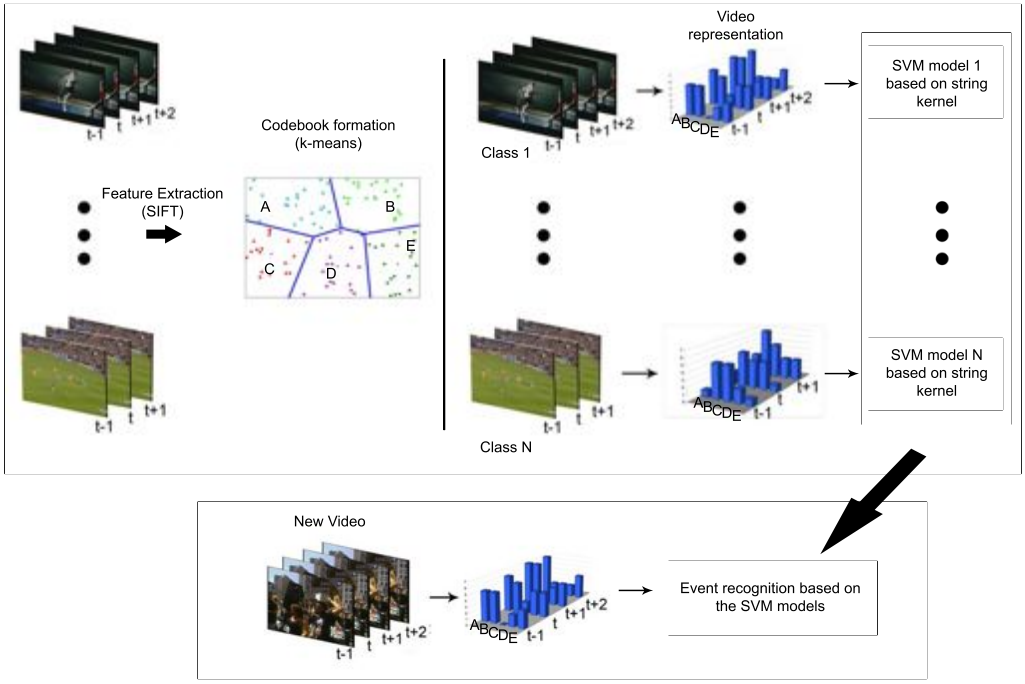


Figure 6.2: Schematization of the proposed approach. In the training stage the features (SIFT) extracted from videos are clustered into visual words (A,B,C,D,E). Each video is represented as a sequence of BoW histograms. Events are described by a *phrase* (string) of variable length, depending on the clip's duration. SVMs with string kernel are used to learn the event representation model for each class. The learned models can be used to recognize events in a new video.

6.3.1 Frame representation

Video frames are represented using bag-of-words, because this representation has demonstrated to be flexible and effective for various image analysis tasks [201, 72, 242]. First of all, a visual vocabulary is obtained through vector quantization of large sets of local feature descriptors extracted from a collection of videos. In this work, we use DoG [149] as keypoint detector and SIFT [137] as keypoint descriptor. The visual vocabulary (or codebook) is generated by clustering the detected keypoints in the feature space using the

k-means algorithm and Euclidean distance as the clustering metric. The center of each resulting cluster is defined as *visual word*. The size of the visual vocabulary is determined by the number of clusters and it is one of the main critical point of the approach. A small vocabulary may lack discriminative power since two features may be assigned to the same cluster even if they are not similar, while a large vocabulary is less generalizable. The trade-off between discrimination and generalization is highly content dependent and it is usually determined by experiments [236]. The effect of the codebook size is explored also in our experiments (see Sect. 6.5). Once a vocabulary is defined, each detected keypoint in a frame is assigned to a unique cluster membership (i.e. a particular visual word), so that a frame is represented by a visual word frequency vector. In this way, this frame representation ignores the spatial arrangement between the different words and thus between the extracted visual features. This effect brings the advantages of using a simple representation that makes learning efficient but, on the other hand, it discards useful information. Alternative approaches might include structural information by encoding information of the structure of the model, for example by modeling the geometrical arrangement of local features [73]. In most cases, the trade-off is an increased computational complexity.

6.3.2 Video representation

As previously introduced, each video shot is described as a *phrase* (string) formed by the concatenation of the bag-of-words representations of consecutive *characters* (frames). To compare these *phrases*, and consequently actions and events, we can adapt metrics defined in the information theory.

Edit distance. The edit distance between two string of characters is the number of operations required to transform one of them into the other. There are several different algorithms to define or calculate this metric, and different transformations can be used. In particular, our approach uses the Needleman-Wunsch distance [156] with the substitution, insertion and deletion transformations.The main motivation of this choice is that Needleman-Wunsch distance performs a global alignment that accounts for the structure of the strings, and the distance can be considered as a score of similarity. The basic idea of the algorithm is to build up the best alignment through optimal alignments of smaller subsequences, using dynamic programming; unlike other approaches, such as dynamic time warping, this type of edit

distance algorithms is able to cope with noise and inaccurate sequence segmentation [180]. 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. The matrix contains all possible pair combinations that can be constructed from the two sequences being compared, and every possible comparison of the sequences can be represented by a path in the matrix. Fig. 6.3 shows an example of the evaluation of the Needleman-Wunsch distance for the case of text and soccer action, respectively. The distance is the number in the lower-right corner of the cost matrix. The traceback that shows the sequence of edit operations leading to the best alignment between the sequences is highlighted in each cost matrix. The algorithm guarantees to find the best alignment of the sequences and is $O(mn)$ in time and space, where m and n are the lengths of the two strings being compared. We have used a dynamic programming implementation of the algorithm that reduces the space complexity to $O(min(m, n))$ [155].

Measuring similarity between characters. A crucial point is the evaluation of the similarity among characters ($a_i \approx b_j$). In fact, when evaluating this similarity on text it is possible to define a similarity matrix between characters, because their number is limited. Instead, in our case each frequency vectors is a different character, therefore we deal with an extremely large alphabet. This requires us to define a function that evaluates the similarity of two characters. Since in our approach each character is an histogram we have evaluated several different methods to compare the frequency vectors of two frames, p and q. In particular we have considered the following distances: *Chi-square* test, *Kolmogorov-Smirnov* test, *Bhattacharyya*, *Intersection*, *Correlation*, *Mahalanobis*. Note that in our implementation each histogram is normalized to sum to one so that it can be considered as a probability distribution.

Chi-square test is a statistical method that permits to compare an

(a) text example

		S	E	N	D
	0	1	2	3	4
A	1	1	2	3	4
N	2	2	2	2	3
D	3	3	3	3	2

(b) video example

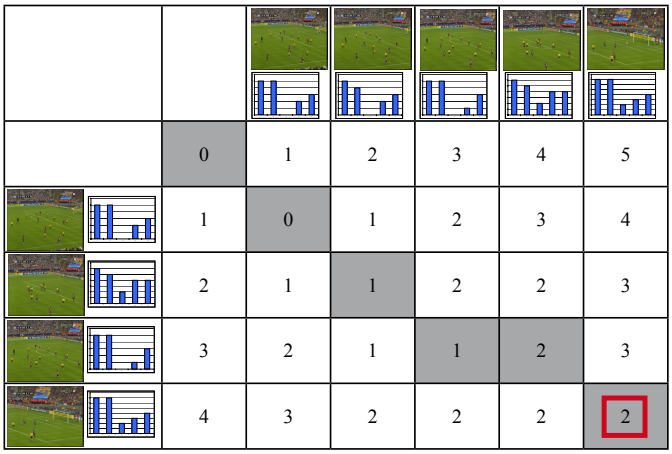


Figure 6.3: Needleman-Wunsch edit distance: (a) text and (b) video examples. Each video frame is represented using its visual word frequency vector. The highlighted path in the cost matrix shows the sequence of operations leading to the best alignment.

observed frequency with a reference frequency. It is defined as:

$$d(p,q) = \sum_{k=1}^{N} \frac{(p(k) - q(k))^2}{p(k) + q(k)}. \tag{6.1}$$

A low value means a better match than a high value.

Kolmogorov-Smirnov test is a statistical method that quantifies the distance between one cumulative distribution function and a reference cumulative distribution function. In our case it can be defined as:

$$d(p,q) = \sup_k \left| F_p(k) - F_q(k) \right|, \tag{6.2}$$

where $F_s(k) = \sum_{j=1}^{k} s(j)$.

Bhattacharyya's distance is defined equal to:

$$d(p,q) = \left(1 - \sum_{k=1}^{N} \frac{\sqrt{p(k)q(k)}}{\sqrt{\sum_{k=1}^{N} p(k) \cdot \sum_{k=1}^{N} q(k)}}\right)^{\frac{1}{2}}. \tag{6.3}$$

Using this distance a perfect match is evaluated as 0, whereas a total mismatch is 1.

Intersection distance is equal to:

$$d(p,q) = \sum_{k=1}^{N} min(p(k), q(k)). \tag{6.4}$$

The intersection of two histograms is connected to the Bayes error rate, the minimum misclassification (or error) probability which is computed as the overlap between two PDF's $P(A)$ and $P(B)$. If both histograms are normalized to 1, then a perfect match is 1 and a total mismatch is 0.

Correlation is defined as:

$$d(p,q) = \frac{\sum_{k=1}^{N} p'(k)q'(k)}{\sqrt{\sum_{k=1}^{N} p'^2(k)q'^2(k)}}, \tag{6.5}$$

where $s'(k) = s(k) - (1/N)(\sum_{j=1}^{N} s(j))$ and N equals the number of bins in the histogram. For correlation, a high score represents a better match than a low score.

Mahalanobis is a distance between an unknown sample and a set of samples which has known mean vector and covariance matrix. Formally given a sample x and a group of samples Y with mean μ and covariance matrix Σ the Mahalanobis distance is:

$$d(x,Y) = (x-\mu)'\Sigma^{-1}(x-\mu). \tag{6.6}$$

In our case this distance can be exploited to find the similarity between a frequency vector of a frame p and a set of frames $q_{-n}..,q_{-1},q,q_1,..q_n$, where q_{-n} is n^{th} frame before q. In particular n is empirically set to ten.

6.4 Classification using string kernels

In recent years, Support Vector Machines (SVMs), introduced by Vapnik *et al.* [38], have become an extremely popular tool for solving classification

problems. In their simplest version, given a set of labeled training vectors of two classes, SVMs map these vectors in a high dimensional space and learn a linear decision boundary between the two classes that maximizes the margin, which is defined to be the smallest distance between the decision boundary and any of the input samples. The result is a linear classifier that can be used to classify new input data. In the binary classification problem, suppose to have a training data set that comprises N input vectors $\mathbf{x}_1, ..., \mathbf{x}_N$, with corresponding target values $t_1, ..., t_N$ where $t_n \in \{-1, 1\}$. The SVMs approach finds the linear decision boundary $y(\mathbf{x})$ as:

$$y(\mathbf{x}) = w^T \phi(\mathbf{x}) + b \tag{6.7}$$

where ϕ denotes a fixed feature-space transformation, b is a bias parameter, so that, if the training data set is linearly separable, $y(\mathbf{x}_n) > 0$ for points having $t_n = +1$ and $y(\mathbf{x}_n) < 0$ for points having $t_n = -1$. In this case the maximum marginal solution is found by solving for the optimal weight vector $\mathbf{a} = (a_1, \ldots, a_N)$ in the dual problem in which we maximize:

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m) \rangle \tag{6.8}$$

with respect to $\mathbf{a}$, that is subject to the constraints:

$$\sum_{n=1}^{N} a_n t_n = 0, \quad a_n \geq 0 \quad \text{for} \quad n = 1, ..., N \tag{6.9}$$

where $\langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m) \rangle$ is the inner product of $\mathbf{x}_n$ and $\mathbf{x}_m$ in the feature-space. The parameters w and b are then derived from the optimal $\mathbf{a}$.

The mapping to a higher dimensionality feature-space is motivated by Cover's theorem [51]. This theorem states that a complex classification problem cast non-linearly into high dimensional space is more likely to be linearly separable than in the original low dimensional space. However, the explicit mapping of input samples in a high dimensional space, and then their inner product, generally have very high computational costs. Kernel functions have been introduced to handle this problem, since they permit to perform the inner product in the feature-space without requiring to explicitly perform the transformation $\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle$. Formally, let χ be the original input vector space and $k : \chi \times \chi \rightarrow \Re$ a function mapping pairs of input vector to real numbers. If the function satisfies the Mercer condition [27] then there

exists a feature-space and a mapping function ϕ such that k acts as a inner product in this feature-space and it is called valid kernel function [197]. In particular a necessary and sufficient condition for a function $k(\mathbf{x}_1, \mathbf{x}_2)$ to be a valid kernel is that the Gram matrix $\mathbf{K}$, whose elements are given by $k(\mathbf{x}_n, \mathbf{x}_m)$, should be positive semidefinite for all possible choices of the input samples.

Recently, many approaches in image categorization have successfully used different kernels such as linear, radial and chi-square basis functions; in particular the latter often gives the best results [242]. However, these kernels are not appropriate for event classification. In fact they deal with input vectors with fixed dimensionality, whereas vectors that represent an action usually have different lengths, depending on how the action is performed. Unlike other approaches that solve this problem simply by representing the video clips with a fixed number of samples [183], we introduce a kernel that deals with input vectors with different dimensionality, in order to account for the temporal progression of the actions. Starting from a Gaussian Kernel that takes the form:

$$k(\mathbf{x}, \mathbf{x}') = exp(-\left\|\mathbf{x} - \mathbf{x}'\right\|^2 / 2\sigma^2), \tag{6.10}$$

we replace the Euclidean with the Needleman-Wunsch (NW) [156] edit distance. Thus the proposed resulting kernel is:

$$k(\mathbf{x}, \mathbf{x}') = exp(-d(\mathbf{x}, \mathbf{x}')). \tag{6.11}$$

where $d(\mathbf{x}, \mathbf{x}')$ is the NW edit distance between $\mathbf{x}$ and $\mathbf{x}'$ input vectors.

It has been previously empirically demonstrated that this type of kernel obtains good results for classification of handwritten digits, shapes, chromosome images [12, 153, 158], despite the fact that the edit distance has not been proved to be a valid kernel. These good empirical results recently have become subject of investigation, in order to obtain a more formal theoretical understanding for the use of indefinite kernel functions [86, 46, 141]. In the cases in which the kernel does not satisfy the Mercer condition, it is possible to adjust the Gram matrix by adapting eigenvalues of this matrix to be all positive, as described in [83]. However, it should be noted that the Gram matrices we applied in our experiments did not require any adaptation.

Figure 6.4: Soccer dataset consists of four different events: shot-on-goal, placed-kick, throw-in and goal-kick.

6.5 Experimental results

We have carried out video event classification experiments on different domains, soccer videos and a subset of TRECVID 2005 video corpus, to analyze the performance of the proposed method and to evaluate its general applicability. The soccer dataset consists of 100 video clips in MPEG-2 format at full PAL resolution (720×576 pixels, 25 fps), and it contains 4 different events: *shot-on-goal*, *placed-kick*, *throw-in* and *goal-kick*. Examples of these events are shown in Fig. 6.4. The full dataset, including also ground truth, is public available on the web[2]. The sequences were taken from 5 different matches of the Italian *"Serie A"* league (season 2007/08) played by 7 different teams. For each class there are 25 clips of variable lengths, from a minimum of about 4 sec (corresponding to ~ 100 frames) to a maximum of about 10 sec (~ 2500 frames). This collection is particularly challenging because events are performed in a wide range of scenarios (i.e. different lighting conditions and different stadiums) and event classes show an high intra-class variability, because even instances of the same event may have very different progression. For our experiments videos are grouped in training and testing

[2]http://www.micc.unifi.it/ballan/research/video-events

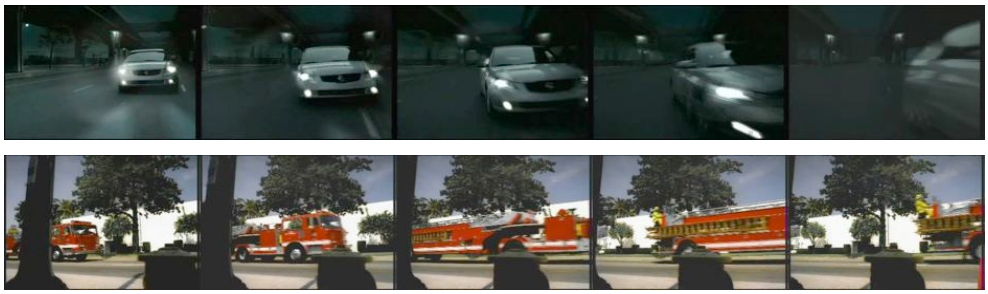
Figure 6.5: Dataset based on a subset of the TRECVID 2005 video corpus. It consists of five events: Exiting Car, Running, Walking, Demonstration or Protest and Airplane Flying.

sets, selecting for each class 15 and 10 videos respectively, and results are obtained by 3-fold cross-validation.

The second dataset is composed by a subset of the TRECVID 2005 video corpus. It is obtained selecting five classes related to a few LSCOM dynamic concepts [106]. In particular we have selected the following classes: *Exiting Car*, *Running*, *Walking*, *Demonstration or Protest* and *Airplane Flying*. Examples of these events are shown in Fig. 6.5. The resulting video collection consists of about 180 videos for each class ($\sim$ 860 in total) in MPEG-1 format with resolution 352×240 pixels and 30 fps. For each class, also in this dataset, videos have different lengths and show an high intra-class variability. Examples of high intra-class variability are shown in Fig. 6.6. Experiments are performed applying 3-fold cross-validation also for this dataset.

In the first experiment we have evaluated the effect on the classification accuracy of the metrics presented in Sect. 6.3 and of the codebook sizes. The second experiment shows the improvement obtained using SVMs, based on the proposed string kernels, with respect to the baseline kNN classifier. In particular we used the LIBSVM implementation [43] using the "one-against-all" approach for multiclass classification. Finally, in the last experiment, we show that our method outperforms the traditional keyframe-based BoW approach.

(a) Running



(b) Exiting Car

Figure 6.6: Examples of intra-class variability in two TRECVID 2005 classes: a) shows two sequences containing the *Running* action, b) shows two sequences containing an *Exiting Car*.

Metric	Th	Accuracy	Metric	Th	Accuracy
Bhattacharyya	0.5	0.47	Intersection	0.1	0.52
Chi-square	**0.13**	**0.54**	Kolmogorov-Smirnov	0.5	0.50
Correlation	0.7	0.53	Mahalanobis	7	0.37

Table 6.1: Comparison of different metrics used to compare the *characters* (frequency vectors) of the strings that represent video shots.

6.5.1 Experiment 1: characters distance and codebook size

In this experiment we evaluate what is the best *characters* distance that has to be used when computing the Needleman-Wunsch distance and the best codebook size. It has been conducted on the soccer dataset, using a kNN classifier, varying the number of visual words used to build the codebook (from 30 to 500 codewords) and the metric used to compare the *characters* of the strings that represent video shots. Classification performances shows that the best codebook size is 200, while the best distance is the *Chi-square test*, since it has a more uniform performance, for the various classes of events, that is not achieved by the others (e.g. correlation metric). Table 6.1 reports the best results obtained for each distance, with a codebook of 200 words, along with the corresponding threshold. For these reasons we select *Chi-square* as the metric used in all the following experiments, and we set to 200 the codebook size for the soccer domain.

It can be observed in table 6.2 that, unlike the case of object classification, the increase of the codebook size does not improve the performance and, instead, the effect may become negative. This can be explained by analyzing the type of views of the soccer domain: events are shown using the main camera that provides an overview of the playfield and of the ongoing event, and thus the SIFT points are mostly detected in correspondence of playfield lines, crowd and players' jerseys and shorts, as shown in Fig. 6.7, and thus the whole scene can be thoroughly represented using an histogram with a limited number of bins for the interest points. Increasing the number of bins may risk to amplify the intra-class variability and then reduce the accuracy of classification, resulting finally also in higher computational costs.

Codebook Size	Accuracy	Codebook Size	Accuracy
30	0.52	**200**	0.54
60	0.52	**300**	0.51
100	0.53	**500**	0.48

Table 6.2: Comparison of classification accuracy obtained using different codebook sizes on the soccer dataset.

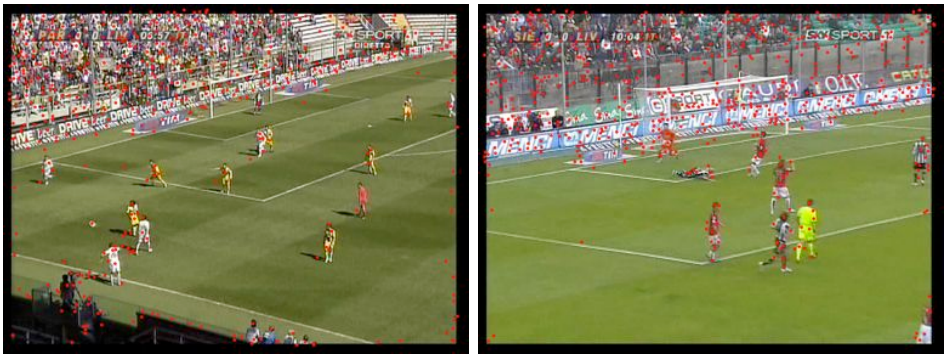
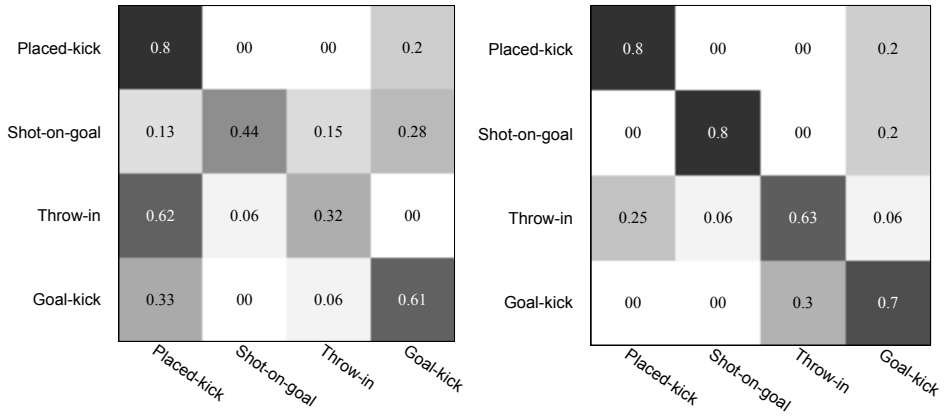


Figure 6.7: Examples of SIFT points detected in a soccer video frame.

6.5.2 Experiment 2: comparison with kNN classifier

In this experiment we have compared the results of the baseline kNN classifier with the results of the SVM classifier using the proposed kernel on the soccer dataset. The mean accuracy obtained by the SVM (0.73) largely outperforms that obtained using the kNN classifier (0.54). Fig. 6.8 reports the global accuracy and the confusion matrices for the kNN and SVM classifiers, respectively. A large part of the improvement, in terms of accuracy, is due to the fact that the SVM has a better performance on the two most critical actions: *shot-on-goal* and *throw-in*. This latter class has the worst classification results, due to the fact that it has an extremely large variability in the part of the action that follows immediately the throw of the ball (e.g. the player may choose several different directions and strengths for the throw, the defending team may steal the ball, etc.).



(a) kNN classifier (b) SVM string classifier

	kNN	**SVM**
Mean Accuracy	0.54	**0.73**

(c) Global accuracy

Figure 6.8: Confusion matrices of baseline kNN and the proposed SVM string classifiers; mean Accuracy for kNN is equal to 0.54 and 0.73 for SVM with string kernel.

6.5.3 Experiment 3: comparison with a traditional keyframe-based BoW approach

Finally, in this experiment we show the improvement of the proposed method with respect to a traditional keyframe-based BoW approach, using the TRECVID dataset. As in the first experiment, we have initially tested different vocabulary sizes, looking for the correct choice for the TRECVID 2005 videos corpus. Results show that, in this case, a vocabulary of 300 words is a good trade-off between discriminativity and generalizability.

For a direct comparison we evaluate the classification performance using the Mean Average Precision (MAP) measure, which is the standard evaluation metric employed in the TRECVID benchmark. In particular, this measure gives a single numerical figure to represent the accuracy of a ranked concept detection result. Formally, let T be the size of the test set, R the number of relevant shots and R_i the number of relevant shot in the top i shot of a query result. $C_i = 1$ if the i^{th} shot is relevant and 0 otherwise. The

average precision is defined as:

$$AP = \frac{1}{R}\sum_{i=1}^{S}\frac{R_i}{i}C_i. \tag{6.12}$$

MAP is the mean of average precision scores over a set of queries.

Table 6.3 reports the comparison results between a traditional BoW approach, as reported in [221], and the proposed method in terms of Mean Average Precision. Our method outperforms the traditional bag-of-words approach in four classes out of five, with an average improvement of 3%. We found a drop in classification performances only for the *Running* event. This is due to the fact this class shows a very high intra-class variability (see Fig. 6.6), with large differences in shot lengths. In particular, there are many different kinds of running actions in the dataset, each of which is depicted from a different camera viewpoint; for example, in several videos, often related to commercials, the running person is filmed frontally, while in many others people is filmed from the sides (e.g. in sports videos).

	Exiting Car	*Running*	*Walking*	*Demonstration or Protest*	*Airplane Flying*	**MAP**
BoW	0.25	**0.57**	0.28	0.32	0.17	0.32
Our Approach	**0.37**	0.36	**0.29**	**0.38**	**0.34**	**0.35**

Table 6.3: Mean Average Precision (MAP) for event recognition in TRECVID 2005.

6.6 Conclusion

In this work we introduced a method for event classification based on the BoW approach. The proposed system uses generic static visual features (SIFT points) that represent the visual appearance of the scene; the dynamic progression of the event is modelled as a *phrase* composed by the temporal sequence of the bag-of-words histograms (*characters*). Phrases are compared using the Needleman-Wunsch edit distance and SVMs with a string kernel have been used to deal with these feature vectors of variable length. Experiments have been performed on soccer videos and TRECVID 2005 news videos; the results show that SVM with string kernels outperform both the

performance of the baseline kNN classifiers and of the standard BoW approach and, more generally, they exhibit the validity of the proposed method. Our future work will deal with the application of this method to a broader set of events and actions that are part of the TRECVID LSCOM events/activities list, and the use of other string kernels. Moreover, we will investigate the possibility to integrate the proposed approach in an ontology-based framework [18], that exploits concept and event dependencies to improve the quality of classification.

Chapter 7

Effective codebooks for human action categorization

In this chapter we propose a new method for human action categorization, combining novel gradient and optic flow descriptors, and creating an effective bag-of-words model. Recent approaches have represented videos using bag of spatio-temporal visual words, following the successful results achieved in object and scene classification. In such cases codebooks are usually obtained by k-means clustering and hard assignment of visual features to the more representative codewords. Our main contribution is two-fold. First, we define a novel 3D gradient descriptor that combined with optic flow outperforms the state-of-the-art, without requiring fine parameter tuning. Second, we show that for spatio-temporal features the popular k-means algorithm is insufficient, because cluster centers are attracted by the denser regions of the sample distribution, providing a non-uniform description of the feature space and thus failing to code other informative regions. We obtain a more effective codebook by applying a radius-based clustering method and a soft assignment that considers the information of two or more relevant codeword candidates. [1]

[1] A preliminary version of the work presented in this chapter has been published as "Effective Codebooks for Human Action Categorization" in *Proc. of ICCV International Workshop on Video-oriented Object and Event Classification (VOEC), 2009* [15].

7.1 Introduction and previous work

Automatic human activity recognition methods are useful for many applications such as video surveillance, video annotation and retrieval and human-computer interaction. For example, in video surveillance, an automatic action classification system that alerts an operator of a possible dangerous situation can reduce human effort and mistakes. However, building a general human activity recognition and classification system is a challenging problem, because of the variations in environment, people and actions. In fact environment variation can be caused by cluttered or moving background, camera motion, illumination changes. People may have different size, shape and posture appearance. Semantically equivalent actions can manifest differently or partially; for example, imagine the different ways of running or actions that can be only partially observed due to occlusions.

Over the past decade, this problem has received considerable attention. Existing action recognition approaches can be classified as using *holistic information* or *part-based information*. An early work based on holistic representation was proposed by Bobick *et al.* [35]. They proposed the motion history images, to encode short spans of motion. For each frame of the input video the motion history image is a gray scale image that records the location of motion; recent motion results into high intensity values whereas older motion produces lower intensities. This representation can be matched using global statistics, such as moment features. Although this method is efficient, it is assumed to have a well segmented foreground and background. Efros *et al.* [66] created stabilized spatio-temporal volumes for each object whose action is to be classified. For each volume a smoothed dense optic flow field is extracted and used as descriptor. This method is particularly suited for distant objects where detailed information of the appearance is unavailable. Yilmaz and Shah [237] used a spatio-temporal volume, built stacking object regions obtained by a contour tracking method, in consecutive frames. Descriptors encoding direction, speed and local shape of the resulting surface are generated by measuring local differential geometrical properties. Gorelick *et al.* [84] analysed three-dimensional shapes induced by the silhouettes and exploited the solution to the Poisson equation to extract features, such as shape structure and orientation. These methods require robust tracking to generate the 3D volumes. Moreover most of the holistic-based approaches are computationally expensive due to the requirement of pre-processing the

input data (e.g. to perform background subtraction, shape extraction, optic flow calculation, object tracking) and they perform better in a controlled environment.

Part-based representations, that exploit interest point detectors combined with robust descriptor methods, have been used very successfully for object and scene classification tasks [72, 201, 236, 242]. Recently, part-based models have been successfully applied to the human action classification problem, because they overcome some limitations of holistic models such as the necessity of performing background subtraction and tracking. Laptev [118] proposed an extension to the Harris-Förstner corner detector for the spatio-temporal case; interesting parts are extracted from voxels surrounding local maxima of spatio-temporal corners, i.e. locations of videos which exhibit strong variations of intensity both in spatial and temporal directions. The extension of the scale-space in the temporal dimension yields a method for automatic scale-selection. Schüldt *et al.* [194] successfully used these features for human action classification by discretizing them into codewords and producing an histogram of the occurring words for each shot. Dollár *et al.* [59] have followed in principle the same approach of Laptev, but suggested to treat time differently from space and to look for locally periodic motion using a quadrature pair of Gabor filters. Their approach produces a denser sampling of the spatio-temporal volume but does not provide a scale-selection criterion. Comparison of the experimental results w.r.t. the approach of Schüldt *et al.* shows an improvement on the same dataset. Niebles *et al.* [163] have then trained an unsupervised probabilistic topic model on the same features as Dollár *et al.* , obtaining comparable classification performance. More recently, Laptev *et al.* [120] have addressed the human action recognition problem in more realistic video settings. They also abandon the scale selection approach, preferring a structural representation based on dense temporal and spatial scale sampling inspired by spatial pyramids [124], showing an improvement of the state-of-the-art results. Finally, Willems *et al.* [226] proposed a new efficient and scale-invariant spatio-temporal detector and descriptor, extending the static SURF features.

All of these part-based approaches use the codebook paradigm that allows classification by describing a video as a bag of words, where video features are represented by discrete visual codewords. These are defined beforehand in a given vocabulary. A vocabulary, in the object and scene classification domain, is commonly obtained by following one of two approaches: an

annotation approach [220] or a *data-driven approach* [37, 201, 242]. The annotation approach obtains a vocabulary by assigning meaningful labels to image patches (e.g. sky, water, vegetation, etc.) while, in contrast, a data-driven approach applies vector quantization on the features using typically k-means clustering. However, despite of its popularity, this is not the optimal solution. Jurie and Triggs [100] have shown that in k-means clustering the centres are almost exclusively around the denser regions in descriptor space and thus fail to code other informative regions. They show that k-means works well for texture analysis in homogeneous images, but the images that arise in natural scenes have far less uniform statistics. For this reason they proposed a scalable acceptance radius-based clustering that generates better codebooks. Nevertheless, all the previous part-based methods for human action recognition use the k-means algorithm for codebook creation. To the best of our knowledge, few papers address approaches to obtain an efficient codebook in human action recognition area. Liu *et al.* [133] proposed a method to automatically find the optimal number of word clusters by utilizing maximization of mutual information (MMI) between words and actions. Initially they apply k-means and then MMI clustering is used to discover a compact representation from the initial codebook of words. They show an improvement of the performance with the learned optimal number of words. A different approach has been proposed by Mikolajczyk and Uemura [151] that recently explored the idea of using a large number of features represented in many vocabulary trees instead of a single flat vocabulary.

Independently of the clustering algorithm, one of the main drawback of the codebook approach, recently pointed out in object and scene classification, is the hard assignment of image feature vectors to codewords in the vocabulary [172,215]. This hard assignment is particularly critical because of two main issues. The first one (*uncertainty*) refers to the problem of selecting the correct codeword out of two or more relevant candidates; the second one (*plausibility*) denotes the problem of selecting a codeword without a suitable candidate in the vocabulary.

In this chapter we describe a new method for classification of human actions that relies on an appropriate quantization method, dealing with the ambiguity of the traditional codebook model. Our main contribution is two-fold: *i)* the definition of gradient and optic flow descriptors that, combined together, outperform the state-of-the-art without requiring fine parameter tuning; *ii)* a radius-based clustering method and a soft assignment proce-

dure that, considering the information of two or more relevant candidates, are able to generate effective codebooks showing a further improvement of classification performances. The rest of the chapter is organized as follows. The interest point detector and descriptors are presented in the next section. The techniques for action representation and categorization, including the codebook creation, are discussed in Sect. 7.3. Experimental results, with an extensive comparison with state-of-the-art approaches, are discussed in Sect. 7.4. Finally, conclusions are drawn in Sect. 7.5.

7.2 Detector and descriptors

Following the approach commonly used for local interest points in images, the detection and description of spatio-temporal interest points are separated in two different steps. Among the different spatio-temporal interest point detectors available, the spatio-temporal corner detector proposed by Laptev *et al.* [118] provides a too sparse representation of the actions. For this reason the spatio-temporal interest points detector proposed by Dollár *et al.* [59], that is able to detect a greater number of points, has received a large attention from the scientific community and has been adopted in several recent works [133, 163].

7.2.1 Detector

In our approach we have adopted the detector proposed by Dollár *et al.* [59]. This detector applies two separate linear filters to spatial and temporal dimensions, respectively. The response function is computed as follows:

$$\begin{aligned} R = &(I(x,y,t) * g_\sigma(x,y) * h_{ev}(t))^2 \\ &+ (I(x,y,t) * g_\sigma(x,y) * h_{od}(t))^2 \end{aligned} \tag{7.1}$$

where $I(x,y,t)$ is a sequence of images over time, $g_\sigma(x,y)$ is the spatial Gaussian filter with kernel σ, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied along the time dimension. They are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$, where $\omega = 4/\tau$, and they give a strong response to the temporal intensity changes, in particular for periodic motion patterns. The interest points are detected at locations where the response is locally maximum.

The main problem of this detector is the fact that it does not cope with scale selection. However, both spatial and temporal scales have to be considered when analyzing motion activity. The spatial scale is related to the ability to detect more or less detailed visual features, while the temporal scale is related to the ability to detect actions that are performed at different speed. In order to cope with the lack of scale selection we run the detector over a set of spatial and temporal scales, to permit the recognition of the same action at different distance and velocity. In particular the spatial scales used are $\sigma = \{2, 4\}$ and the temporal scales are $\tau = \{2, 4\}$. This approach has also some other desirable properties such as a reduced computational complexity w.r.t. scale selection and the production of a richer description of the scene, using a larger number of interest points.

7.2.2 Descriptors

For each detected point a patch that contains the volume that contributed to the response function is considered. The volume is proportional to the scale at which the interest point is detected. Each volume is divided in equally sized sub-regions, three for the spatial dimensions and two for the temporal dimension. To obtain a representation for each spatio-temporal volume, we evaluate a descriptor based on gradients on x, y and t direction and an optic flow descriptor, considering also their combinations. This is motivated by the fact that these two descriptors encode different information. In fact the descriptor based on gradient encodes mostly the visual appearance of each volume, while the optic flow descriptor encodes the motion information. The two descriptors are presented in the following.

The gradient magnitude and orientations in 3D are:

$$M_{3D} = \sqrt{G_x^2 + G_y^2 + G_t^2}, \tag{7.2}$$

$$\phi = \tan^{-1}(G_t / \sqrt{G_x^2 + G_y^2}), \tag{7.3}$$

$$\theta = \tan^{-1}(G_y / G_x). \tag{7.4}$$

where G_x, G_y and G_z are respectively computed using finite difference approximations: $I(x+1, y, t) - I(x-1, y, t)$, $I(x, y+1, t) - I(x, y-1, t)$ and $I(x, y, t+1) - I(x, y, t-1)$. We compute two separated orientation histograms quantizing ϕ and θ and weighting them by the magnitude M_{3D}. The ϕ (with range, $-\frac{\pi}{2}, \frac{\pi}{2}$) and θ ($-\pi$,π) are quantized in four and eight bins

respectively. The spatio-temporal gradient is computed after smoothing the values extracted with those of two adjacent scales, to increase the robustness of the feature description. The overall dimension of the descriptor is thus $3 \times 3 \times 2 \times (8 + 4) \times 2 = 432$. This construction of the three-dimensional histogram is inspired by the approach proposed by Scovanner *et al.* [195], in which they construct a weighted three-dimensional histogram normalized by the solid angle value (instead of quantizing separately the two orientations) to avoid distortions due to the polar coordinate representation. Moreover we do not re-orient the 3D neighbourhood, since rotational invariance, which is invaluable in object detection and recognition, is not desired in an action categorization context. We have found that our method is computationally less expensive, equally effective in describing motion information given by appearance variation, and showing a better performance (see comparison results in Tab. 7.2).

The optic flow is estimated using the Lucas&Kanade algorithm. Considering the optic flow computed for each couple of consecutive frames, the relative apparent velocity of each pixel is (V_x, V_y). These values are expressed in polar coordinates as in the following:

$$M_{2D} = \sqrt{V_x^2 + V_y^2}, \tag{7.5}$$

$$\theta = \tan^{-1}(V_y/V_x). \tag{7.6}$$

We compute position dependent histograms as in the gradient based descriptor but, being the optic flow two dimensional, only a single orientation histogram is stored for each of the 18 sub-regions within the voxel. Every sample is weighted with the magnitude M_{2D}, as is done for the gradient-based descriptor. Then we have also added an extra "no-motion" bin that, in our initial experiments, has shown to greatly improve the performance. Thus the final descriptor size is $3 \times 3 \times 2 \times (8 + 1) = 162$.

We have finally analysed two possible combinations of these descriptors: *i)* a weighted concatenation of the two descriptors and *ii)* a concatenation of the histograms of the bag-of-words that have been computed from the 3D gradient descriptor and from the histogram of optic flow. In the first case the visual words, created according to the bag-of-words paradigm, are computed from a vector that has higher dimensionality, while in the second case the visual words are computed differently for each descriptor and the SVM classifiers are able to pick the best combinations of features, practically resulting in an implicit feature selection.

7.3 Action representation and categorization

The spatio-temporal bag-of-words (BoW) model is built through the creation of a discrete visual vocabulary (or codebook) and then by assigning each feature to the corresponding codeword. First of all, it is required to perform a vector quantization for large sets of feature vectors in a high dimensional space. Typically this is performed through clustering methods and the most common approach is the use of k-means clustering, because of its simplicity and convergence speed. The BoW approach then assigns each feature to the closest vocabulary word and a histogram of visual word frequencies is computed. The histogram is fed to a classifier to predict the action category. The performance of this model depends on the quantization method and on the number of words that are selected.

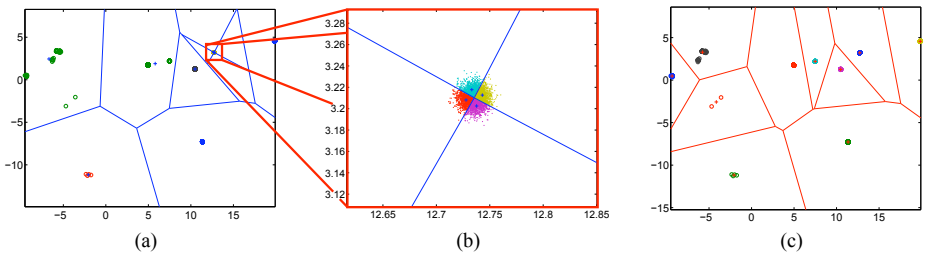


Figure 7.1: Comparison of k-means and radius-based clustering on a synthetic dataset. (a) k-means clustering; (b) k-means clustering: detail of a dense region that has been split in four clusters; (c) Radius-based clustering.

7.3.1 Codebook formation

The use of k-means clustering has some disadvantages: *i)* the cluster centers are attracted by the denser regions of the sample distribution, resulting more clustered near these regions and more sparse otherwise, thus providing a more imprecise quantization for the vectors laying in these latter regions [100]. This effect, due to the assumption of uniform distribution of the features in the descriptor space, is even more pronounced in high dimensional spaces such as those spanned by the spatio-temporal descriptors; a representation of this effect can be obtained visualizing a Voronoi tessellation of the feature space, where Voronoi cells do not uniformly cover the feature space as shown in Fig. 7.1. Other disadvantages are: *ii)* the clustering is not very robust

w.r.t. outliers, *iii)* the number of visual words has to be known in advance, requiring an empirical evaluation of this number.

Radius-based clustering. In order to overcome the limitations of k-means clustering, we explore the idea of using an on-line radius-based clustering technique following a mean-shift approach [50, 80]. In fact, as shown by Jurie and Triggs [100], in the case of dense sampling image representations, it is better to apply a radius-based clustering method. This observation is interesting also for the human action domain because, as previously introduced (Sect. 7.2), the spatio-temporal features extracted by the Dollár detector [59] can be considered as a dense representation; this fact is even more pronounced using our multi-scale approach. In this case the non-uniformity in the descriptor space, caused by densely sampled patches, is better coded using a radius-based method that is able to allocate centers more uniformly. An example of this effect is shown in Fig. 7.1 c.

The algorithm starts with an uniform random sub-sampling D_n of the original dataset points D. Given a radius R, mean shift clustering on D_n is used to find the modes of the samples distribution. A new cluster center is then allocated on the mode corresponding to the maximal density region. Data points of the original dataset D, within a distance less than R from the center, are considered members of this cluster and eliminated for the following iterations. This elimination prevents the algorithm from repeatedly assigning centers to the same high density regions. Finally, the algorithm can be stopped when a "sufficiently" large number of clusters (words) has been identified.

Visual words statistics. One of the assumption in text categorization methods is that, given a natural language textual corpus, the words frequency distribution follows the well-known *Zipf's law*. This is a critical point because, considering this empirical evidence, we can consider words at intermediate frequencies as the most informative for classification. Therefore it is interesting to see how the visual words are distributed in a visual corpus, as also noted in [100, 177, 236]. In particular, we want to know whether their distribution satisfies Zipf's law. Fig. 7.2 shows the statistics of visual words frequency using k-means and radius-based quantization on our experimental dataset (see Sect. 7.4 for details). An "ideal" Zipf's distribution must be a straight line in log-log scale. The figure shows that the

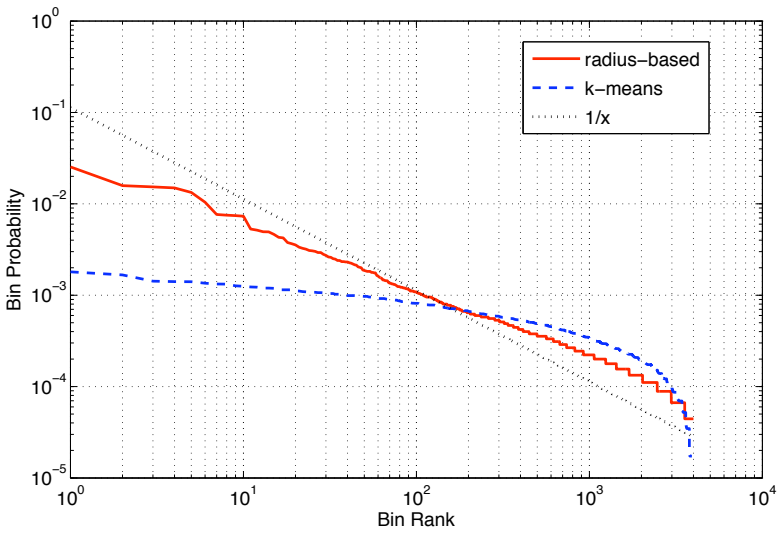


Figure 7.2: Log-log plots of visual words frequency using k-means and radius-based quantization.

distribution of visual words obtained by k-means quantization satisfies the Zipf's law only roughly. In fact, most of the bins has similar frequencies and they are distributed more evenly with respect to the expected power law. In contrast, the proposed radius-based quantization shows a statistics that fit better the expected distribution. This confirms the assumptions discussed in the previous paragraph and confirms that this approach models better medium density frequencies.

7.3.2 Codeword assignment

Given a vocabulary, the traditional codebook approach represents a video sequence containing an action by a histogram of codeword frequencies. In particular, for each word w in the vocabulary V the frequency distribution in a sequence is computed by:

$$FD(w) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1 & \text{if } w = \underset{v\in V}{\operatorname{argmin}}(D(v,p_i));\\ 0 & \text{otherwise;}\end{cases} \tag{7.7}$$

where n is the number of spatio-temporal patches in a sequence, p_i is the i^{th} spatio-temporal patch, and $D(v,p_i)$ is the distance (usually Euclidean) between the codeword v and the patch p_i. This hard assignment, that takes

account only of the closest codeword, lacks to consider two issues: codeword *uncertainty* (selection of the correct codeword when two or more candidates are relevant) and codeword *plausibility* (selection of a codeword when all codewords are too far and not representative). We observe that, in our case, the plausibility is less problematic, because the radius-based clustering method that we employ is able to allocate the centers more uniformly. On the other hand, as noted by van Gemert *et al.* [215], in a high-dimensional feature space the codeword uncertainty issue becomes very urgent. In fact, if we consider a codeword as a high-dimensional sphere in feature space, most feature points in this sphere lay near the surface and are close to the boundary between different codewords. For this reason the distribution of the codewords in a sequence has to contain the information of two or more relevant candidates. This can be done by smoothing the hard assignment of a spatio-temporal patch to the codeword vocabulary using Gaussian kernel density estimation, computing the uncertainty frequency distribution with:

$$UFD(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_\sigma(D(w, p_i))}{\sum_{j=1}^{|V|} K_\sigma(D(v_j, p_i))} \tag{7.8}$$

where D is the Euclidean distance and K_σ is the Gaussian kernel:

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \tag{7.9}$$

where σ is the scale parameter of the Gaussian kernel; this parameter has to be tuned on the training set, because dependent on the dataset, the features length and their range values.

7.4 Experimental results

We tested our approach on two datasets commonly used for human action recognition: the KTH and Weizmann datasets. The KTH dataset contains 2391 video sequences showing six actions: walking, running, jogging, hand-clapping, hand-waving, boxing. They are performed by 25 actors under four different scenarios of illumination, appearance and scale change. The video resolution is 160×120 pixel. The Weizmann dataset contains 93 video sequences showing nine different people, each performing ten actions such as run, walk, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop-sideways, wave-two-hands, wave-one-hand and bend. The

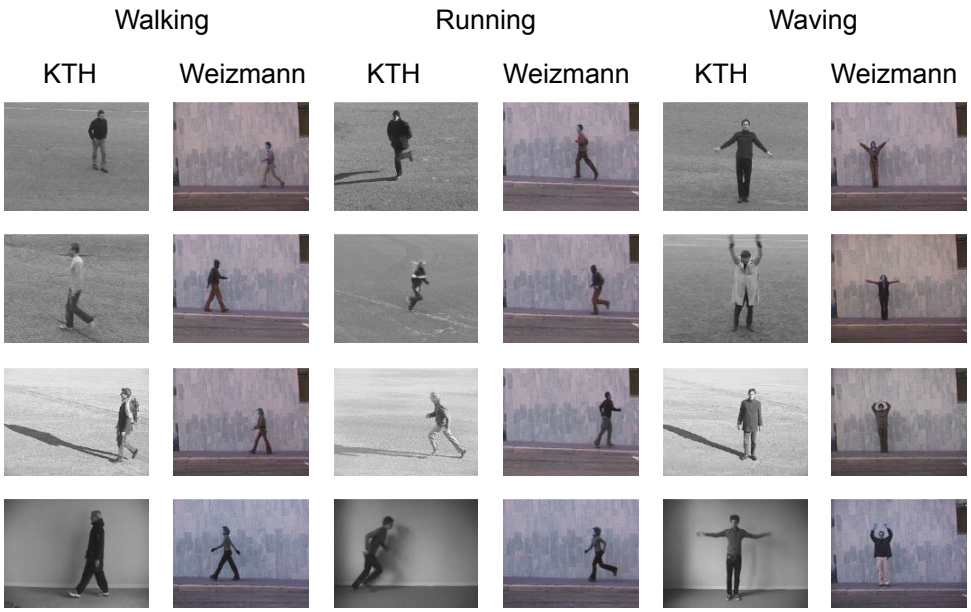


Figure 7.3: Sample frames from the KTH and Weizmann datasets (Walking, Running and Waving actions).

video resolution is 180×144 pixel. An example of the differences between the two dataset is shown in Fig. 7.3, where sample frames selected from videos containing the same action in the two sets are compared each other.

Two approaches were followed during the training phase, due to the different sizes in the datasets. The SVM classifiers used for the KTH dataset were trained on videos of 16 actors and the performance was evaluated using the videos of the remaining 9 actors. Measures have been taken according to a five-fold cross-validation. Due to the small size of the Weizmann dataset the classifiers were trained on actions from eight actors and tested on the remaining one. Measures have been taken using the leave-one-out cross-validation. This setup is identical to the most recent works in action recognition domain and thus is suitable for a direct comparison [111,120,163,228]. Classification is performed using non-linear SVMs with the χ^2 kernel [242]:

$$K_{\chi^2}(p,q) = \exp\left(-\frac{1}{2\gamma}\sum_{k=1}^{N}\frac{(p_k - q_k)^2}{(p_k + q_k)}\right) \tag{7.10}$$

where N is the vocabulary size, p and q are histograms of word occurrences. The value of the kernel parameter γ is obtained by cross-validation on the

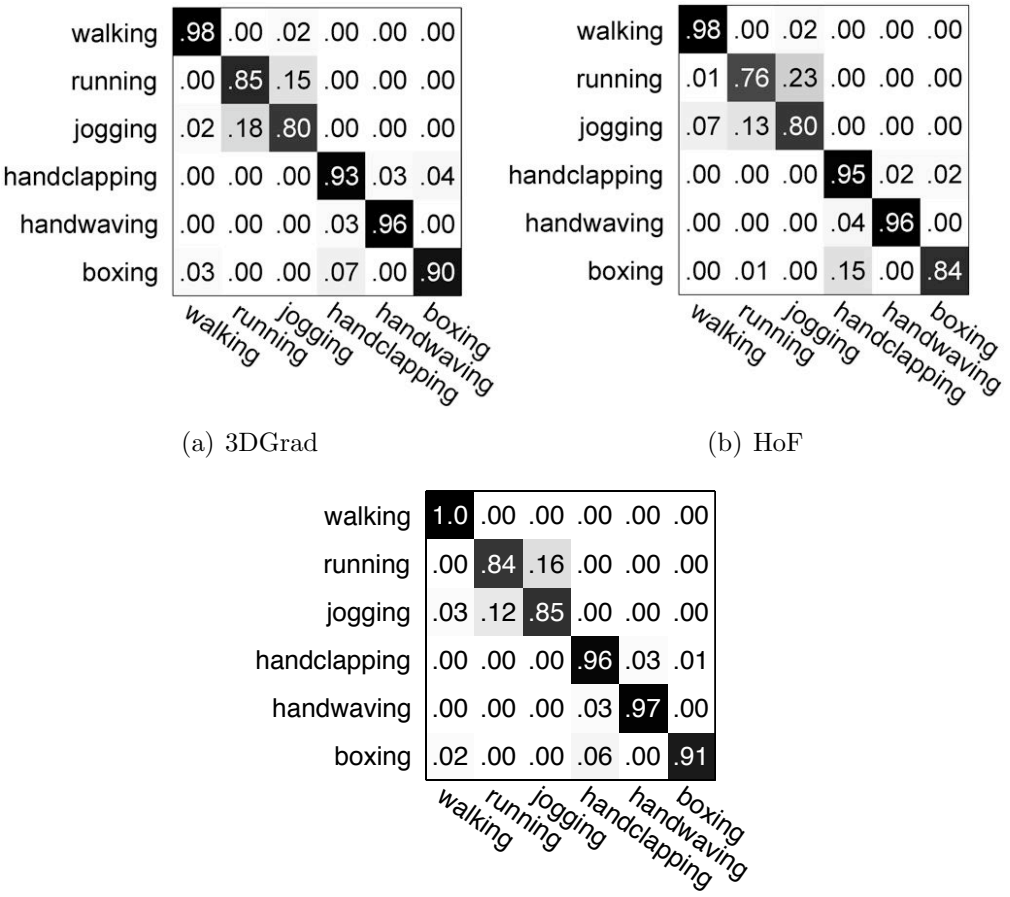
Descriptor	**KTH**	**Weizmann**
3DGrad	90.38 ± 0.8	92.30±1.6
HoF	88.04 ± 0.7	89.74±1.8
3DGrad_HoF combination	91.09 ± 0.4	92.38±1.9
3DGrad+HoF combination	**92.10 ± 0.4**	**92.41 ±1.9**

Table 7.1: Comparison of our descriptors, alone and combined, on the KTH and Weizmann datasets.

training set. For multi-class classification, we use the *one-vs-one* approach.

7.4.1 Evaluation of our descriptor

Table 7.1 evaluates the performance of our proposed descriptors, comparing the performance of each descriptor alone and the two possible combinations discussed in Sect. 7.2. The experiments have been carried on using the setup described above, and the quantization approach used is k-means clustering (using 4000 words for KTH and 1000 for Weizmann), in order to be directly comparable with other approaches. In the first two rows we report results obtained using only one of the two descriptors, 3D gradient in the first row and histogram of optic flow in the second. In the third row are reported the results for the descriptor that is obtained through a weighted concatenation of the two descriptors, while in row four the descriptor is composed by the concatenation of the histograms of the bag-of-words that have been computed from the 3D gradient descriptor and from the histogram of optic flow. The best result, on both datasets, is achieved by the concatenation of the histograms of the BoWs computed from both descriptors. This is due to the fact that the performance of 3D gradient and HoF are quite complementary (see Fig. 7.4). For example, the action recognition performance for the boxing class on the KTH dataset is better when using the 3D gradient instead of the HoF description, while for handclapping is the opposite case. It can be observed (Fig. 7.4 c) that the concatenation of the histograms improves the performance for all the classes except one, running class. In the Weizmann dataset we obtain a smaller improvement, with the concatenation of histogram, probably caused by the smaller training set that is available and the increased size of the representation.



(a) 3DGrad (b) HoF

(c) 3DGrad+HoF combination

Figure 7.4: Confusion matrices on the test set KTH actions.

7.4.2 Performances obtained by effective codebooks

In this set of experiments we evaluate the different codebook creation approaches presented in Sect. 6.3. The datasets used are the KTH and Weizmann with the same experimental setup described above, and the descriptor is the concatenation of the histograms of bag-of-words computed from 3D gradient and optic flow descriptors (3DGrad+HoF). Fig. 7.5 compares the classification performances obtained by the standard k-means and hard assignment approach - commonly used by previous works - with the proposed radius-based clustering and soft assignment. The graph reports the variation in accuracy w.r.t. the number of visual words, up to the number of words (4000 for KTH, 1000 for Weizmann) that were used in the previous

experiments.

With a very low number of words the soft radius-based clustering method has a lower performance than k-means, since in this approach the words that are used are those that are more common (i.e. those that provide less discriminative information). However, this effect disappears rapidly (above 1500 words for KTH and 400 words for Weizmann) due to the more effective choice of the words, as discussed in Sect. 7.3.2. The radius-based clustering extended so as to account for codeword uncertainty outperforms k-means clustering and classification results are improved in both datasets; in particular, it has a better performance even with a relatively low number of visual words (e.g. $\sim$ 2000 for KTH and $\sim$ 500 for Weizmann). Indeed the radius-based clustering better encodes sparser regions while the soft assignment is able to moderates uncertainty in the denser ones, leading thus to more effective codebooks.

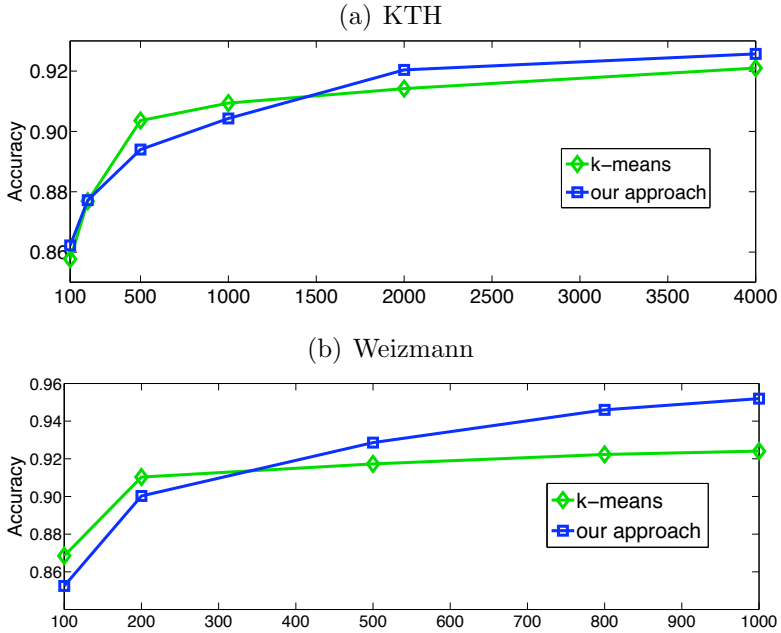
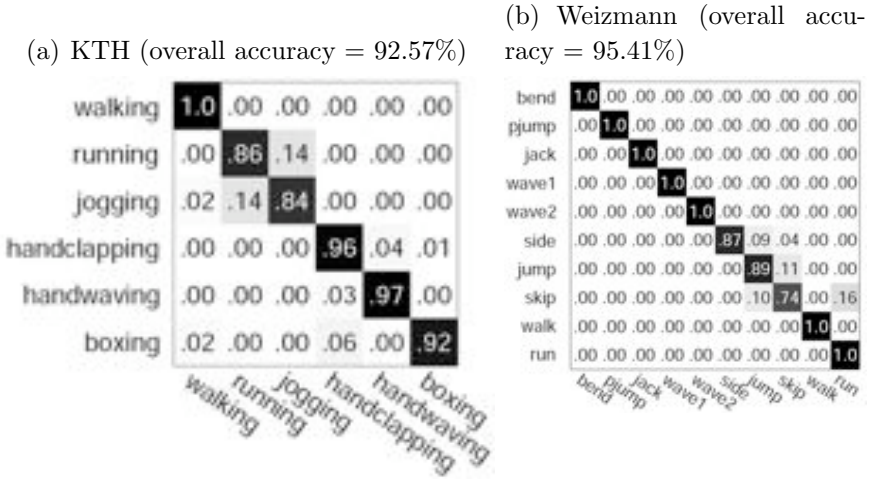


Figure 7.5: Comparison of classification accuracies on KTH and Weizmann datasets using the combined descriptor (3DGrad+HoF) and *i)* k-means based codebooks and *ii)* our effective codebooks approach (i.e. radius-based clustering + soft assignment).

We report on Fig. 7.6 the final classification performance on KTH and Weizmann datasets, obtained using the proposed soft radius-based quan-



(a) KTH (overall accuracy = 92.57%)

(b) Weizmann (overall accuracy = 95.41%)

Figure 7.6: Final confusion matrices on KTH and Weizmann.

tization, as confusion matrices. Interestingly, the major confusion occurs between similar classes (running-jogging on KTH and jump-skip on Weizmann). The overall accuracy on KTH is 92.57% while on Weizmann is 95.41%.

7.4.3 Comparison to state-of-the-art

In Table 7.2 we report a comparison of the average class accuracy of our approach with state-of-the-art results, reported by other researchers.

Results obtained on KTH using our combined descriptor (3DGrad+HoF) united with the proposed effective codebook formation outperform previous works based on standard BoW models [194, 59, 111, 120, 163, 226, 228], even those that employ fine tuning of parameters or additional structural descriptors. Note that the previous state-of-the-art result (91.8%), achieved by Laptev *et al.* [120] using their best combination of features, is obtained performing a greedy search on different combination of descriptors (HoG and HoF) and grids, which add structural information. Our results outperform also those of Kläser *et al.* [226] (91.4%) that use a single 3D gradient descriptor but with a heavy optimization of eight descriptor parameters, resulting in a high dependence on the dataset used.

Also when considering the Weizmann dataset we outperform previous BoW-based works [111, 163, 195], and also the results reported by Liu *et al.* [131] (90.4%) obtained combining and weighting multiple features. How-

ever, we cannot compare to results by Gorelick *et al.* [84] or Fathi and Mori [70] because they use an holistic representation and more data given by segmentation masks.

Method	**KTH**	**Weizmann**
Our method	**92.57**	**95.41**
Laptev *et al.* [120]	91.8	-
Dollár *et al.* [59]	81.2	-
Wong and Cipolla [228]	86.62	-
Scovanner *et al.* [195]	-	82.6
Niebles *et al.* [163]	83.33	90
Liu *et al.* [131]	-	90.4
Kläser *et al.* [111]	91.4	84.3
Willems *et al.* [226]	84.26	-
Schüldt *et al.* [194]	71.7	-

Table 7.2: Comparison of our method to state-of-the-art.

7.5 Conclusion

In this chapter we have presented a novel method for human action categorization based on a new descriptor for spatio-temporal interest points, that combines appearance (3D gradient descriptor) and motion (optic flow descriptor), and on an effective codebook formation. We replaced the traditional codebook quantization method using a radius-based clustering algorithm and a soft assignment of features descriptors to codewords. The approach was validated on two popular datasets (KTH and Weizmann), showing results that outperform state-of-the-art BoW approaches, without requiring parameter tuning employed by the previous best results. The proposed approach is modular and each contribution of this chapter can be adapted to any framework based on interest points and BoW. Our future work will deal with evaluation on real world videos.

Chapter 8

Video annotation using ontologies and rule learning

In this chapter we present an approach for automatic annotation and retrieval of video content, based on ontologies and semantic concept classifiers. A novel rule-based method is used to describe and recognize composite concepts and events. Our algorithm learns automatically rules expressed in Semantic Web Rules Language (SWRL), exploiting the knowledge embedded into the ontology. The relationship between concepts, their co-occurrence and the temporal consistency of video data are used to improve the performance of individual concept detectors. Finally, we present a web video search engine, based on ontologies, that permits queries using a composition of boolean and temporal relations between concepts. [1]

8.1 Introduction

Whereas understanding of the semantic meaning of video content is immediate for humans, for a computer this is far from true. This discrepancy is commonly referred to as the "semantic gap". A recent trend for bridging this gap is to define a large set of semantic concept detectors, each of which automatically detects the presence of a semantic concept such as "*indoor*",

[1] This chapter has been published as "Video Annotation and Retrieval Using Ontologies and Rule Learning" in *IEEE MultiMedia, vol. 17, iss. 4, pp. 80-88, 2010* [19].

"*face*", "*person*" or "*airplane flying*". Typically these detectors learn the mapping between a set of low-level visual features, such as local descriptors, color and texture, and a concept from examples. Many approaches have been proposed to design these detectors, but the common idea is to apply machine-learning techniques (typically SVM) to automatically learn this mapping from the data, thus obtaining a large set of binary classifiers. Among the others, the most popular solution is to use the Bag-of-Words (BoW) approach [201] in which an image, or a video frame, is treated as the visual analogue of a document and it is represented by a bag of quantized descriptors (e.g. SIFT), referred to as visual-words. This representation of the visual content is used to compute histograms of visual-word frequencies, that are used to train appropriate classifiers. Another approach that has been proved to be extremely successful for detection of specific object classes such as "*face*" or "*person*", is the Viola and Jones detector [219]. These approaches have been implemented by several systems that participated to visual concept recognition challenges, such as PASCAL-VOC and TRECVid. However, semantic concepts are in general still difficult to be accurately detected, so that their detection in video remains a challenging problem to be solved. Observing the results provided by state-of-the-art detectors, the accuracy of detection (measured by Average Precision) can range from less than 0.1, for semantic concepts such as "*people marching*" or "*fire weapon*", to above 0.6 for a concept such as "*face*". Despite the fact that continuous performance improvement has been reported in the last years, and a large effort has been devoted to extend the number of different concept classifiers, important questions are still open. How many concept detectors are really useful [93] and how much reliable [235] should they be? Moreover, concept classifiers are usually drawn from a particular domain, but how well are they able to generalize across different domains?

On the other hand, exploitation of the semantic relationships between concepts is recently receiving a large attention from the scientific community, since it can improve the detection accuracy of concepts and obtain a richer semantic annotation of a video. To this end, ontologies are expected to improve the capability of computer systems to automatically detect even complex concepts and events from visual data with higher reliability. Ontologies consist of concepts, concept properties, and relationships between concepts. They organize semantic heterogeneity of information, using a formal representation, and provide a common vocabulary that encodes semantics and

supports reasoning. Few attempts have been done to integrate high-level semantic concepts provided by an ontology with their visual representation. In the most common approach, the ontology provides the conceptual view of the domain at the schema level, and appropriate concept detectors play the role of observers of the real world sources, classifying an observed entity or event in the nearest concept of the ontology. In this way concept detectors have the responsibility of implementing invariance with respect to several conditions while, once the observations are classified, the ontology is exploited to have a more complete semantic annotation, establishing links to other concepts and disambiguating the results of classification. Among the recent works that follow this approach, Snoek *et al.* [205] defined "semantically enriched detectors" by linking a general-purpose ontology (obtained from WordNet) to a set of detectors (with several hundreds of concepts), obtaining an improvement with respect to TRECVid 2005 classification results. Zha *et al.* [239] also followed this approach, using hierarchical relationships and pairwise correlations between concepts, to refine confidence scores of concept detectors. In a different approach, proposed by Bertini *et al.* [29], the ontology includes also visual data instances related to high-level concepts, identifying their spatio-temporal patterns; visual prototypes, representative of these patterns, are then defined and used for automatic annotation.

However, in real applications there is need to detect and recognize complex concepts and situations where multiple elementary concepts are in mutual relation in time and space. Therefore, ontologies have to be extended to define these higher level concepts, adding sets of rules that encode spatio-temporal relationships among individual concepts. As the number of these concepts grows, the number of rules for their detection becomes high. Thus, the definition of the rules by human experts is not practical; the appropriate solution is to learn automatically a set of rules for each composite concept to be detected.

In this chapter we present an approach for automatic annotation and retrieval of video content, based on ontologies and semantic concept classifiers. Several elements are included to support effective annotation and retrieval. First of all, automatic determination of semantic linguistic relations between concepts (*is a*, *has part*, *is part of*) is performed, using WordNet, to define the ontology schema; the concept detectors are then linked to the corresponding concepts in the ontology. We propose a novel rule-based method for automatic semantic annotation of composite concepts and events in videos;

our algorithm learns automatically rules expressed in Semantic Web Rules Language (SWRL), exploiting the knowledge embedded into the ontology. Moreover, the concepts' relationship of co-occurrence and the temporal consistency of video data are used to improve the performance of individual concept detectors. Finally, we present a web video search engine, based on ontologies, that permits queries using a composition of boolean and temporal relations between concepts; this system exploits the ontology structure permitting also, for example, to expand queries to synonyms and concept specializations.

8.2 Related work

The usefulness of the construction of large sets of automatic video concept classifiers and the evaluation of the number of detectors needed for effective video retrieval has been studied in [93, 207, 235]. In [93] Hauptmann *et al.* report that concept-based video retrieval (with fewer than 5000 concepts detected) with a minimal 0.1 Mean Average Precision is likely to provide high accuracy results in news video retrieval. Snoek and Worring [207] have confirmed the positive correlation between the number of concept detectors and video retrieval performance, as well as the improvement of the pair-wise combination of detectors, using a set of 363 concept detectors. Presently, the performance of video search engines is still far to be acceptable. In fact, their performance in terms of Mean Average Precision, as resulting from the TRECVid 2008 on 20 concepts of the Large Scale Concept Ontology for Multimedia lexicon (LSCOM), varies in a range from 0.19 to 0.13. Ontologies and concept relations have been recently proposed to improve the performance of the concept detectors. Zha *et al.* [239] defined an ontology to provide a simple structure to LSCOM [154], using pairwise correlations between concepts and hierarchical relationships to refine concept detection of support vector machine classifiers. Wei *et al.* [225] have proposed two semantic spaces (Ontology-enriched Semantic Space (OSS) and Ontology-enriched Orthogonal Semantic Space (OS^2)) to facilitate the selection and fusion of concept detectors for video search.

To obtain richer annotations, other authors have explored the usage of rule-based reasoning over objects and events in different domains. Hollink *et al.* [94] defined a set of SWRL rules to perform semi-automatic annotation of images of pancreatic cells. Bai *et al.* [13] defined a soccer ontology

and applied temporal reasoning, with temporal description logic, to perform event annotation in soccer videos. All these approaches expect that rules are created by human experts; thus, they are not practical for the definition of a large set of rules. Automatic learning of rules has been proposed by Shyu *et al.* [200]. The authors proposed a method to annotate rare events and concepts based on set of rules that use low-level and middle-level features. A decision tree algorithm is applied to the rule learning process. Moreover they addressed the imbalance problem of positive and negative examples in the case of rare event/concept using data mining techniques. Liu *et al.* [134] proposed a method to enhance accuracy of semantic concepts detection, using association mining techniques to imply the presence of a concept from the co-occurrence of other high-level concepts. However, these methods have shown to be insufficiently expressive to describe composite concepts and events, since they do not take into account spatio-temporal relations between individual concepts.

8.3 Automatic rule learning using first order logic

In our approach, first order logic rules defined in SWRL are automatically learned from the knowledge that is embedded in the ontology. To this end our ontology contains abstract concepts, the ontology schema (based on concepts that are detected by semantic classifiers, their linguistic relations, namely: *is a, has part, is part of*, as encoded in WordNet), and, for each concept a set of the concept instances that have been observed. Rules are learned using FOILS, a new algorithm obtained as an adaptation of the First Order Inductive Learner technique (FOIL [178]) to ontologies and Semantic Web technologies.

To describe in detail the algorithm, let us introduce some basic terminology from formal logic. All the expressions are composed of constants, variables, predicate symbols and function symbols. The difference between predicates and functions is that predicates (in the following written with upper-case first letter) can assume only boolean values, whereas functions (in the following written in lower-case) may have any constant as their value. A term is any constant, any variable, or any function. A literal is any predicate, or its negation, applied to any term. If a literal contains a negation

symbol ($\neg$), it is called *negative literal*, otherwise *positive literal*. A *clause* is any disjunction of literals, where all variables are assumed to be universally quantified. A *Horn clause* is a clause containing at most one positive literal, as in:

$$H \vee \neg L_1 \vee \neg L_2 \ldots \vee \neg L_n$$

where H is the positive literal, and $\neg L_1 \vee \neg L_2 \ldots \vee \neg L_n$ are negative literals. It is equivalent to:

$$(L_1 \wedge L_2 \ldots \wedge L_n) \rightarrow H$$

which is equivalent to "`IF` $(L_1 \wedge L_2 \ldots \wedge L_n)$ `THEN` H". The Horn clause precondition $L_1 \wedge L_2 \ldots \wedge L_n$ is called *body*, while the literal H, that forms the post-condition, is called *head*. As an example of *Horn clause* consider the sentence that describes the composite concept: "*a person is in a secured area*" ("`IF` a person and a secured area instances occur in a shot and the bounding box of that person is in the bounding box of that secured area `THEN` that person is in secured area"). This sentence can be translated in the following fragment in first-order logic:

$$\begin{aligned} &Person(p) \ \wedge \ SecuredArea(s) \ \wedge \\ &HasBoundingBox(p, pBox) \ \wedge \ HasBoundingBox(s, sBox) \ \wedge \\ &BoxIsInBox(pBox, sBox) \ \rightarrow PersonIsInSecuredArea(p) \end{aligned}$$

where p and s are variables that can be bound to any person and any secured area respectively, while $sBox$ and $pBox$ are their bounding boxes.

The hypotheses learned by FOILS are sets of rules that are Horn clauses. The algorithm starts with an initial rule, written in SWRL, composed by the *head* (i.e. the target composite concept) and an empty or initial *body*, and an ontology with a set of instances that are positive and negative examples of the target concept. As an example, the initial rule for the composite concept "*a person enters in a secured area*" could be:

$$Person(p) \ \wedge SecuredArea(s) \ \rightarrow \ PersonEntersSecuredArea(p).$$

The algorithm iterates searching new literals that have to be added to the *body*. This is a general-to-specific search through the space of hypotheses, beginning with the most general preconditions possible (the empty or initial precondition), and adding literals one at a time to specialize the rule until it avoids all negative examples, or when no more negative examples are

excluded for a certain number of iterations l. A schema of the algorithm is shown in Alg. 1. Two issues have to be addressed: the generation of hypothesis candidates and the choice of the most promising candidate.

Pos ← Positive examples
Neg ← Negative examples
Rule ← Initial rule
repeat
  Candidate_literals ← Generating hypothesis candidates
  Best_literal ← $\arg\max_{L}$ Rule_Gain(L,Rule)
  Add Best_literal to Rule preconditions
  Pos ← subset of Positive examples that satisfy Rule
  Neg ← subset of Negative examples that does not satisfy Rule
until Neg is empty or no more Neg examples are excluded for l iterations

Algorithm 1: FOILS algorithm

Suppose that at the iteration i^{th}, the current rule R_i being considered is $(L_1 \wedge L_2 \ldots \wedge L_i) \rightarrow H(x_1, x_2, \ldots, x_k)$, where $(L_1 \wedge L_2 \ldots \wedge L_i)$ are literals forming the current rule preconditions and $H(x_1, x_2, \ldots, x_k)$ is the *head*. FOILS generates candidate specializations of this rule by considering as new literals L_{i+1} any predicate occurring in the ontology (i.e. all concepts and concepts relations), where at least a variable already exists in the rule. A special literal, $Equal(x_j, x_k)$ where x_j and x_k are variables already present in the rule, can be considered, because it can happen that variables created at different iterations could have the same meaning.

To select the most promising literal from the candidates generated at each step, the algorithm considers the performance of the rule over the instances stored in the ontology. The evaluation function used to estimate the utility of adding a new literal is based on the number of positive and negative bindings covered before and after adding this new literal. Let us consider a rule R_i and a candidate literal L_{i+1} that might be added to the *body* of the rule. The evaluation function is defined as:

$$Rule\_Gain(L_{i+1}, R_i) \equiv t\left(log_2 \frac{p_1}{p_1 + n_1} - log_2 \frac{p_0}{p_0 + n_0}\right) \tag{8.1}$$

where p_0 and n_0 are the number of positive and negative bindings of R_i, while p_1 and n_1 are the number of positive and negative bindings of the new rule R_{i+1} (resulting from the addition of L_{i+1}). Finally, t is the number of positive binding of rule R that are still covered after adding literal L_{i+1} to R_i.

8.3.1 Improving performance

The performance of composite concept annotation is tightly related to the reliability of the semantic classifiers. This can be improved considering the probability of contemporary presence of pairs of individual concepts, as well as their temporal consistency. To this end we included in the ontology the relation of concepts co-occurrence, expressed using mutual information, that measures the mutual dependence of a pair of concepts. This quantity is computed from the analysis of the concept instances as:

$$MI(C_i, C_j) = \sum_{k,l \in \{0,1\}} P(C_i = k, C_j = l) \frac{P(C_i = k, C_j = l)}{P(C_i = k)P(C_j = l)} \tag{8.2}$$

where $MI(C_i, C_j)$ is the mutual information between concept C_i and C_j. The value of $P(C_i = k)$ for $k \in \{0, 1\}$ is the probability of the presence or absence of C_i in the videos. The probability values $P(C_i = k)$, $P(C_j = l)$ and $P(C_i = k, C_j = l)$ for $k, l \in \{0, 1\}$ are computed from ground truth. Following the approach introduced in [239] it is possible to exploit the mutual information to refine the confidence values of the detected concept instances. Given $P_i = P(C_i = 1|S)$ the confidence score of a detector for the concept C_i in shot S and $P = [P_1, \ldots, P_n]^T$ the confidence score vector for all concepts in S, it is possible to refine the confidence scores P^+ with:

$$P^+ = (1 - \alpha)P + \alpha MP \tag{8.3}$$

where $\alpha \in [0, 1]$ weights the contribution of the mutual information and M is a matrix, whose entries have been computed using the mutual information, with the diagonal elements set to 0 to avoid self-reinforcement.

The confidence score of a detector for concept C can be improved considering the fact that the presence of a semantic concept generally spans multiple consecutive shots [134]. In particular for each concept we re-evaluate

its confidence values at each shot using:

$$P^T(C^t = 1|S^t) = \sum_{i=-d}^{+d} w_i P(C^t = 1|C^{t-i} = 1) P(C^{t-i} = 1|S^{t-i}) \qquad (8.4)$$

where $P(C^t = 1|C^{t-i} = 1)$ are probabilities estimated from ground-truth annotations, $P(C^t = 1|S^t)$ is the confidence score of a detector for the concept C in shot S^t, w_i is a concept-dependent weighting coefficient (with $\sum_i w_i = 1$) that measures the contribution from the shot that is temporally i shots apart from S^t, while d is the maximum temporal distance within which the shots are considered.

8.3.2 Rule learning example

As an example of rule learning using the FOILS algorithm, consider the event "*airplane take-off*" and a simple initial rule, such as:

$$Airplane(?a) \ \wedge \ Sky(?s) \ \wedge \ Ground(?g) \rightarrow AirplaneIsTakingOff(?a)$$

The algorithm enriches an initial rule with spatio-temporal relations, using a training set. The literal candidates considered by the algorithm are all the classes and properties defined in the ontology domain (e.g. *HasBoundingBox(s,Sbox)*), the temporal properties used to encode Allen's logic (e.g. $Temporal : before(?a, ?s)$) and the spatial properties used to encode the relative positions between concepts (e.g. *Spatial:BoxOverlapsBox(?tas, ?aBox,?sBox)*). At each step the most promising literal is added, considering the performance of the rules over the training data, until the recognition performance does not improve. Thus, the result of the FOILS algorithm is:

$$Airplane(?a) \ \wedge \ Sky(?s) \ \wedge \ Ground(?g) \ \wedge HasBoundingBox(?a, ?aBox) \wedge$$
$$HasBoundingBox(?s, ?sBox) \ \wedge \ HasBoundingBox(?g, ?gBox) \wedge$$
$$Spatial : BoxOverlapsBox(?tas, ?aBox, ?sBox) \wedge$$
$$Spatial : BoxIsInBox(?tag, ?aBox, ?gBox) \ \wedge \ Temporal : After(?tas, ?tag) \wedge$$
$$MovingObject(?a) \rightarrow AirplaneIsTakingOff(?a)$$

This rule can be translated in the following sentence: "`IF` an airplane, sky and ground instances (a, s, g) occur in a shot `AND` they have a bounding box ($aBox$, $sBox$, $gBox$, respectively) `AND` for a time interval tas the bounding

box of the airplane is on the bounding box of the sky `AND` for a time interval *tag* the bounding box of the airplane is on the bounding box of the ground `AND` the time interval *tas* is after of the interval *tag* `AND` the airplane is a moving object `THEN` that airplane is taking-off". In some cases we can observe that FOILS adds some literals that are not necessary for the event representation, however this does not affect negatively the performance of the rule. In this example, the "*moving object*" concept, that in our ontology is an hypernym of "*airplane*", is added to the rule even if it is not necessary.

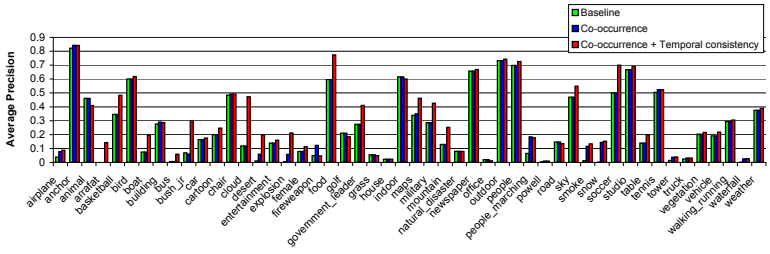
Once the rule is learned it is applied to the ontology, that contains the instances obtained by the semantic classifiers, to automatically extend the video annotation with instances of the "*airplane take-off*" event. In this case the ontology contains instances resulting from the detection of "*airplane*", "*sky*" and "*ground*" detectors. These detectors have been created using the Viola and Jones algorithm (provided by OpenCV) and color-based pixel classification with SVM, to detect and localize objects. Then, the spatio-temporal evolution of the appearance of concepts is determined using a tracker, based on an improved version of the particle filter [11]. Concept instances are associated with color and luminance histograms, that are used by the tracker to identify each instance in a video sequence. As an example, Fig. 8.1 shows a sequence of "*airplane take-off*" with results of concept detectors.

Figure 8.1: Examples of airplane, sky and ground detection and tracking in a TRECVid video sequence.

8.4 Experimental results

We evaluated how much our method improves the performance of individual concept detectors, exploiting concept co-occurrence and temporal consistency. We have built automatically an ontology from the MediaMill detectors thesaurus [208], following the method described in Sect. 8.3. The dataset used for this experiment is the training set of TRECVid 2005; it was divided using a 4-fold approach, maintaining groups of consecutive shots in the same fold, to be able to evaluate the effects of time consistency. The parameters of eq. 8.3 and eq. 8.4 (α and d, respectively) were chosen in preliminary experiments on the training data. In the training phase we also identified the concepts that took advantage of the use of the co-occurrence relation, and computed the refined confidence score only for them (based on eq. 8.3, with $\alpha = 0.1$). The Mean Average Precision (MAP) computed for all the concepts improved by 4.37%. After the co-occurrence refinement we computed, for all concepts, the temporal consistency refinement (setting $d = 15$ in eq. 8.4). The overall improvement of MAP, obtained by the combination of the two techniques, is 17.64%. Fig. 8.2 shows the performance of the baseline detectors and the results of the two refinement techniques, in terms of Average Precision. We report only the 50 concepts that obtained the largest variations. The concepts whose detectors have a very low performance, like "*airplane*", "*desert*", "*explosion*", "*people marching*" are improved by use of co-occurrence, that exploits the results of more robust detectors. The use of temporal consistency greatly improves the performance of certain concepts that are related to topics often shown in consecutive shots within news videos, like politics (e.g. "*Arrafat*", "*Bush Jr*", "*government leader*") or sports (e.g. "*soccer*", "*basketball*", "*boat*"). Small improvements are obtained for detectors with high performance, like for "*anchor*", "*people*" and "*outdoor*".

We also checked the capability of our system to detect composite concepts using semantic rules, automatically learned from concept instances, in two different video domains: broadcast news and surveillance. For the first domain, we considered four events selected from the LSCOM events/activities [106]: "*airplane flying*", "*airplane take-off*", "*airplane landing*", "*airplane taxiing*". The other set of events is related to the video surveillance of shopping malls: "*person enters in a shop*" and "*person exits a shop*". The dataset used for the news domain comprises 65 TRECVid 2005 videos



Concept	Baseline	Co-occurrence	Co-occurrence + Temporal consistency
airplane	0.04	0.08	**0.09**
anchor	0.82	**0.84**	**0.84**
animal	**0.46**	**0.46**	0.41
arrafat	0.00	0.00	**0.14**
basketball	0.35	0.35	**0.48**
bird	0.60	0.60	**0.62**
boat	0.07	0.07	**0.19**
building	0.28	**0.29**	**0.29**
bus	0.01	0.01	**0.06**
bush_jr	0.07	0.06	**0.30**
car	0.16	0.16	**0.18**
cartoon	0.20	0.20	**0.25**
chair	0.48	**0.49**	**0.49**
cloud	0.12	0.12	**0.47**
desert	0.01	0.06	**0.20**
entertainment	0.14	0.14	**0.16**
explosion	0.01	0.06	**0.21**
female	0.08	0.08	**0.11**
fireweapon	0.05	**0.12**	0.05
food	0.60	0.60	**0.77**
golf	**0.21**	**0.21**	0.19
government_leader	0.27	0.27	**0.41**
grass	0.05	0.05	0.05
house	0.02	0.02	0.02
indoor	**0.62**	**0.62**	0.60
maps	0.34	0.35	**0.46**
military	0.29	0.29	**0.43**
mountain	0.13	0.13	**0.25**
natural_disaster	0.08	0.08	0.08
newspaper	0.66	0.66	**0.67**
office	**0.02**	**0.02**	0.01
outdoor	0.73	0.73	**0.74**
people	0.70	0.70	**0.73**
people_marching	0.07	**0.18**	**0.18**
powell	0.00	**0.01**	**0.01**
road	**0.15**	**0.15**	0.13
sky	0.47	0.47	**0.55**
smoke	0.01	0.12	**0.13**
snow	0.00	0.14	**0.15**
soccer	0.50	0.50	**0.70**
studio	0.67	0.67	**0.69**
table	0.14	0.14	**0.20**
tennis	0.50	**0.52**	**0.52**
tower	0.01	**0.04**	**0.04**
truck	0.02	**0.03**	**0.03**
vegetation	0.20	0.20	**0.22**
vehicle	0.19	0.19	**0.22**
walking_running	0.29	0.29	**0.30**
waterfall	0.00	**0.03**	**0.03**
weather	0.38	0.38	**0.39**

Figure 8.2: Average precision of 50 concepts, selected from the 101 MediaMill thesaurus, showing comparison of baseline with the proposed refinement approaches: co-occurrence only and the combination of temporal consistency with co-occurrence (in this case the overall improvement for all the concepts is 17.64%).

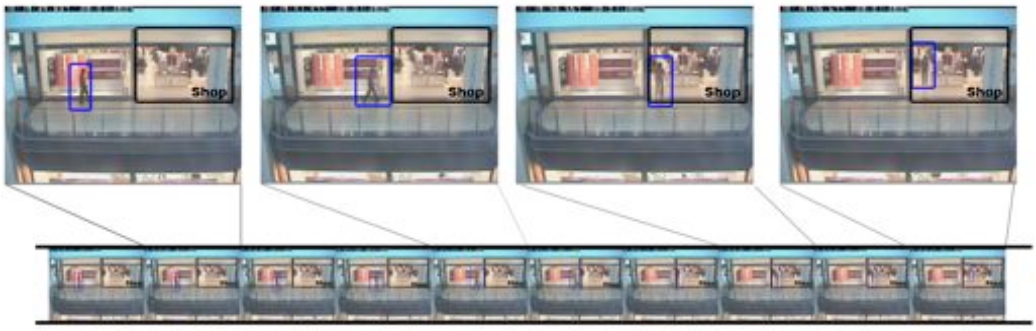
and 100 videos containing airplane events taken from the web[2] (called in the following Web Dataset). The TRECVid videos were selected from the TRECVid development set, considering those containing the LSCOM concepts "*airplane take-off*", "*airplane landing*" and "*airplane flying*". We inspected all the videos annotated with the "*airplane*" concept to select those that contain the "*airplane taxiing*" event, since this concept is not used in LSCOM. The videos used for the second domain are the CAVIAR[3] surveillance videos, selected from the front of view of the 2^{nd} set. These videos were filmed from a fixed position camera that frames a mall shop and the area in front of the shop. In the experiments, the scene framed was divided in four parts, as shown in Fig. 8.3, to determine when a person is in the shop, in front of it or in front of the showcase. The two datasets were divided using a 3-fold approach, to learn the rules.

We have then used these rules to annotate these videos, evaluating the results, in terms of precision and recall, as shown in Tab. 8.3. As it can be observed, the overall results for all the rules are extremely promising. The performance of "*airplane flying*" and "*airplane taxiing*" is better than that of "*airplane landing*" and "*airplane take-off*"; this is due to the fact that the rules modeling those events are simpler. The performance of the rules is dependent on the performance of the detectors and tracker. Investigation of the cases in which the rules fail, has shown that the main cause of failure is due to the performance of the sky and ground detectors. In particular, these detectors are affected by the low quality of the images and the presence of superimposed graphics. In a few cases the fault was the airplane detector, especially when superimposed graphics and text covered the appearance of the airplane, that occurred mostly in TRECVid videos. This fact is reflected by the different performance in the two datasets.

The results of the recognition of video surveillance actions show a good performance both in terms of precision and recall. The fixed camera and lighting conditions reduce the variability of the appearance of the observed events and objects; this leads to a good performance of the person detector and of the tracker. The performance of the rules is mainly dependent on the errors of the tracker, that happened sometimes when multiple persons' trajectories overlapped.

[2]This dataset is available on demand from the URL: http://www.micc.unifi.it/dome
[3]CAVIAR Dataset: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

(a)



(b)

Figure 8.3: a) CAVIAR Surveillance video dataset: view of the mall shop areas. b) Example of person detector and tracking in a video sequence.

Data Set	Action/Event	Precision	Recall
TRECVid 2005	Airplane flying	0.94	0.52
TRECVid 2005	Airplane take-off	0.32	0.40
TRECVid 2005	Airplane landing	0.69	0.69
TRECVid 2005	Airplane taxiing	0.92	0.78
Web Dataset	Airplane flying	0.93	0.92
Web Dataset	Airplane take-off	0.78	0.81
Web Dataset	Airplane landing	0.84	0.94
Web Dataset	Airplane taxiing	0.96	0.78
Web Dataset + TRECVid 2005	Airplane flying	0.93	0.72
Web Dataset + TRECVid 2005	Airplane take-off	0.55	0.60
Web Dataset + TRECVid 2005	Airplane landing	0.76	0.81
Web Dataset + TRECVid 2005	Airplane taxiing	0.94	0.78
CAVIAR	Person enters in the shop	0.96	0.76
CAVIAR	Person leaves from the shop	0.95	0.89

Table 8.3: Precision and recall of "*Airplane flying*", "*Airplane take-off*", "*Airplane landing*", "*Airplane taxiing*", "*Person enters in the shop*", "*Person leaves from the shop*" for different datasets.

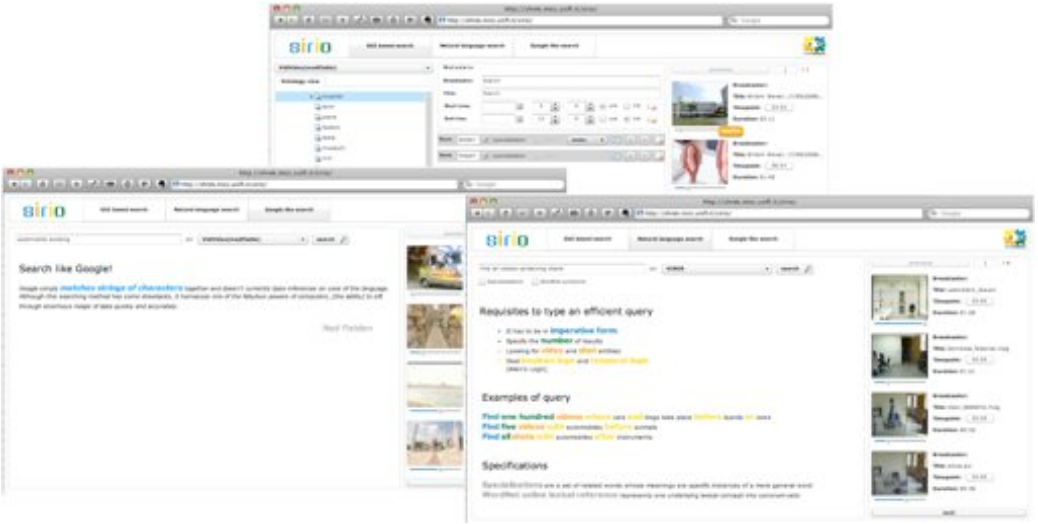
8.5 The Sirio web-based search engine

Browsing and searching video archives is performed exploiting the ontology described in Sect. 8.3. To this end we have developed a web-based prototype system, Sirio[4] [31], which provides integrated support for boolean-temporal, semantic, and query-by-example (QBE) queries. The system is based on the Rich Internet Application (RIA) paradigm. RIAs can avoid the usual slow and synchronous loop for user interactions, typical of web environments that use only the HTML widgets available to standard browsers. This has allowed to implement a visual querying mechanism that exhibits a look and feel approaching that of a desktop environment, with the fast response that is expected by users.

The search engine is a web application written in Java, executed in an Apache Tomcat application server, and supports multiple ontologies (for different video domains), ontology reasoning services and W3C SPARQL (SPARQL Protocol and RDF Query Language) queries. The GUI is a Flash application, written in Flex, that is executed in a client-side Flash Virtual Machine. Videos, returned as results of queries, are streamed using the

[4]http://www.micc.unifi.it/vidivideo - contact authors to obtain access passwords

(a)



(b)

Figure 8.4: (a) Browsing interface: the ontology graph view is used to explore parts of the full ontology, checking the instances of video clips annotated with the selected concept. All the instances of a concept are visible as streaming video clips. (b) Search interfaces: GUI query builder; natural language search; Google-like search.

RTMP protocol. To browse an archive, inspecting the annotated concept instances (i.e. video clips), the user navigates the ontology structure, presented as a graph. Fig. 8.4 (a) shows the browser interface. The user can select a concept from a tag cloud, that shows the concepts with the largest number

of instances, or navigate the ontology following the concept relations.

The prototype provides three different search modalities, as shown in Fig. 8.4 (b), aiming at different types of users: a GUI to build composite queries that include boolean-temporal operators (based on Allen's logic), visual prototypes for QBE and video metadata (like broadcaster and programme names, broadcast dates, etc.) has been developed for professional users. A free-text interface for Google-like searches and a natural language interface, that allow to compose queries with boolean-temporal operators, have been developed for novice users that do not require to specify complex queries or broadcast metadata. Using the ontology relations and reasoning it is possible to extend user's queries through subsumption and meronymy. The natural language and the Google-like interface, require another form of query expansion, using synonym relations based on WordNet, so that users can formulate their queries in a natural way, without being forced to select terms from a lexicon.

8.6 Conclusion

Ontologies are a source of a-priori knowledge that can be usefully exploited to complement classifiers and achieve higher performance, especially for concepts related to dynamic composite events. In this chapter we have presented an algorithm for automatic learning of ontology rules for video annotation and methods to refine the performance of automatic concept detection. We have also presented a prototype system to browse and search video archives. Our future work will deal with learning of rules that cope with uncertainty, to use fuzzy ontology reasoning that can exploit detector confidence scores. We will also investigate the use of fuzzy temporal Horn logic to overcome the expressivity limitations of SWRL.

Chapter 9

Conclusion

This chapter summarizes the contribution of the thesis and discusses avenues for future research.

9.1 Summary of contribution

This thesis makes a contribution to the field of multimedia understanding. We have proposed models and methods for effective visual search for objects and events in images and videos. We focused on retrieval of object instances (in particular trademarks) and event categories (such as human actions) and we provided a step-by-step methodology to reduce the semantic gap and to achieve automatic annotation and retrieval of visual content. The major contributions are summarized below:

- In Chapter 3, we presented a real system for logo recognition in large sports video archives. Trademark recognition and retrieval was performed by matching a set of local descriptors (SIFT) for each trademark instance against the set of features detected in each frame of the video. Accurate localization of trademarks was obtained through robust clustering of matched feature points in the video frame. Experimental results were shown, along with an analysis of the precision and recall, in a realistic application scenario in a variety of situations.

- In Chapter 4, we extended this approach by introducing a robust *context-dependent* similarity measure between local descriptors. This measure takes into account not only the intrinsic visual features but

also their context and spatial configuration. The strength of the proposed method resides in several aspects: *i*) the inclusion of the information about the spatial configuration in similarity design as well as visual features; *ii*) the ability to control the regularization of the solution via our energy function; *iii*) the invariance to many transformations including translation, scale, rotation and also partial occlusion; *iv*) the theoretical groundness of the matching framework which shows that under the hypothesis of existence of a reference logo into a test image, the probability of success of matching and detection is high while very low under background. The validity of the method was shown through extensive experiments on a challenging logo image dataset.

- In Chapter 5, we developed a novel approach for image forensics based on local visual features (the approach was built on ideas that came from Chapter 3). Given a suspected photo, our method allows to reliably detect if a certain region has been duplicated and, furthermore, to determine which geometric transformation was applied to perform such tampering. The technique has shown effectiveness with respect to diverse operative scenarios such as composite processing and multiple cloning.

- In Chapter 6, we introduced a method for event classification based on the popular bag-of-words model. The proposed approach used static visual features that represent the visual appearance of the scene; the dynamic progression of the event is modelled as a *string* composed by the temporal sequence of the bag-of-words histograms (*characters*). Strings are compared using the Needleman-Wunsch edit distance and SVMs with a string kernel have been used to deal with these feature vectors of variable length. Experiments have been performed on soccer videos and TRECVID 2005 news videos.

- Chapter 7 focused on categorization of human action classes from video collections. The novelty lies in: *i*) a novel 3D spatio-temporal gradient descriptor that, combined with optic flow, outperformed the state-of-the-art without requiring fine parameter tuning; *ii*) a more effective codebook model obtained by applying a radius-based clustering method and a soft assignment that considers the information of two or more relevant codeword candidates. The method was applied on

standard KTH and Weizmann datasets showing its validity and outperforming several recent approaches.

- In Chapter 8 we developed a novel rule-based methodology to describe and recognize composite concepts and events. It is able to automatically learn rules, expressed in Semantic Web Rules Language (SWRL), exploiting the knowledge embedded in a multimedia ontology. The relationship between concepts, their co-occurrence and the temporal consistency of video data are then used to improve the performance of individual concept detectors. We have also presented a prototype system to browse and search video archives.

9.2 Directions for future work

Nowadays social websites for media sharing (e.g. YouTube, Flickr and Facebook) have become more and more popular, allowing people to easily upload, share and annotate personal media content with keywords usually referred to as *tags*. These tags provide additional contextual and semantic information with which users can organize and access shared media content. Flickr hosts more than 2 billion images with about 3 million new uploads per day. YouTube reported in March 2010 more than 2 billion views per day, 24 hours of videos uploaded per minute, and also estimated that a common user spends, on average, 15 minutes each day on the site.

Because of this new scenario, the main directions for our future research are twofold. The first one is related to scalability. In fact, one direction of research is focused on how to extend visual search to such huge archives. To this end we are working on a compact codebook representation, in order to extend the work presented in Chapter 7 to more unconstrained videos, and also on a scalable string-based representation of video sequences (Chapter 6). A key aspect, that has to be adequately taken into consideration, is given by the fact that if hundreds of concepts need to be learned, overfitting to training domains must be overcome. This is the so-called *domain change* problem, which refers to the fact that concept detection is often applied on different domains of visual data than the ones it was trained on (it is a common problem within the TRECVID evaluation community). Therefore, visual search systems must become flexible enough to keep up with users' information needs. This requires new strategies for bootstrapping visual learning beyond

the manual annotation of limited datasets that constitutes the state of the art.

The second main research direction is also related to this problem. The availability of large archives of user-generated visual content provides a new opportunity to recover training data. But from a more general point of view, user-tagged images and videos introduce new challenges and opportunities. First of all, social visual search and analysis is very important for helping people organize and access the increasing amount of user-tagged multimedia. Since user tagging is known to be uncontrolled, ambiguous, and overly personalized, a fundamental problem is how to interpret the relevance of a user-contributed tag with respect to the visual content the tag is describing. For these reasons, several researchers are working on novel tag recommendation and re-ranking strategies, mainly based on tag co-occurrence, to ease the task of tagging visual data. The basic intuition is that, if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content [128]. The main research in this area is focused on the problem of how to integrate social tags and visual representation of the data. While research on image tagging has received a considerable attention in recent years [130], there are still very few works that address the problem of automatically assigning tags to videos and locating them temporally (or spatially) within the video sequence. This is a really interesting task since user tags are usually associated with the entire video and are not located temporally within the sequence. For this reason, users that search for a specific tag are forced to watch whole sequence of retrieved videos. Thus, our future research will deal (in particular) with video tag suggestion and temporal localization based on collective knowledge and visual similarity of frames. Other work will deal with exploitation of semantic relations between tags and the use of other sources of social knowledge to improve semantic relatedness of the suggested tags. Our idea is to integrate the appearance of visual data and the social knowledge given by user-tags with an a priori semantic knowledge represented by ontology models (Chapter 8).

Appendix A

Appendix

This appendix is related to context-dependent trademark matching and retrieval, previously presented in Chapter 4. Here we proof the convergence of our matching criterion (see Section 4.3.1).

A.1 Proof of proposition 2 in Section 4.3.1

Proof. Let $\mathcal{N}^{\theta,\rho}(X)$, $\mathcal{N}^{\theta,\rho}(Y)$ be two random variables standing for the number of interest points falling inside the context cell (θ, ρ) of respectively a reference logo and a test image. Here $\mathcal{N}^{\theta,\rho}(X) \rightarrow \mathcal{B}(n, 1/Q)$, $\mathcal{N}^{\theta,\rho}(Y) \rightarrow \mathcal{B}(m, 1/Q)$ and Q is the number of cells in the context ($Q = N_a \times N_r$, in practice $Q = 64$). Following the definition of our fixed point $\mathbf{K}_{X,Y}$ in (4.4), we have

$$\mathbf{K}_{Y|X} \quad \propto \quad \frac{1}{C} \exp\big(\mathcal{N}(X,Y)\big), \tag{A.1}$$

where $\mathcal{N}(X,Y)$, stands for the number of matching points in the context of X, Y

$$\mathcal{N}(X,Y) \;=\; \sum_{\theta,\rho}^{Q} \mathcal{N}^{\theta,\rho}(X)\, \mathcal{N}^{\theta,\rho}(Y). \tag{A.2}$$

Under $H_1 \longrightarrow \exists Y_J$ **s.t.** $(X, Y_J) \in H_1$

Since $\mathbf{K}_{Y_J|X} + \sum_{j\neq J}^{m} \mathbf{K}_{Y_j|X} = 1$, using (4.8), p_s, $q_s = 1 - p_s$ are respectively

$$\begin{aligned} &\mathbb{E}\big(\mathbf{K}_{Y_J|X}\big|(X,Y_J) \in H_1\big), \\ \text{and} \\ &\sum_{j\neq J}^{m} \mathbb{E}\big(\mathbf{K}_{Y_j|X}\big|(X,Y_j) \in H_0\big), \end{aligned} \tag{A.3}$$

here the expectation $\mathbb{E}$ is with respect to $\{X, Y_1, \ldots, Y_m\}$. Now, combining (A.1), (A.3), p_s and q_s may be respectively rewritten

$$\begin{aligned} & \frac{1}{C} \exp\Big(\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big)\Big) \\ \text{and} \qquad & \\ & \frac{1}{C}(m-1)\exp\Big(\mathbb{E}_{H_0}\big(\mathcal{N}(X,Y)\big)\Big), \end{aligned} \tag{A.4}$$

$\mathbb{E}_{\mathbb{H}_1}$ (resp. $\mathbb{E}_{\mathbb{H}}$) denotes the expectation w.r.t data in H_1 (resp. H_0) equal to

$$\begin{aligned} \mathbb{E}_{H_0}\big(\mathcal{N}(X,Y)\big) &= \mathbb{E}_{H_0}\left(\sum_{\theta,\rho}\mathcal{N}^{\theta,\rho}(X)\mathcal{N}^{\theta,\rho}(Y)\right) \\ &= \sum_{\theta,\rho}\mathbb{E}_{H_0}\Big(\mathcal{N}^{\theta,\rho}(X)\mathcal{N}^{\theta,\rho}(Y)\Big) \\ &= \sum_{\theta,\rho}\mathbb{E}_{H_0}\Big(\mathcal{N}^{\theta,\rho}(X)\Big)\,\mathbb{E}_{H_0}\Big(\mathcal{N}^{\theta,\rho}(Y)\Big), \\ &\quad \mathcal{N}^{\theta,\rho}(X), \mathcal{N}^{\theta,\rho}(Y) \overset{i.i.d}{\to} \mathcal{B}(n, 1/Q) \\ &= n^2\ (1/Q)^2 Q \\ &= n^2/Q. \end{aligned} \tag{A.5}$$

$$\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big) = \mathbb{E}_{H_1}\left(\sum_{\theta,\rho}\mathcal{N}^{\theta,\rho}(X)\mathcal{N}^{\theta,\rho}(X)\right). \tag{A.6}$$

Under H_1, $\mathcal{N}^{\theta,\rho}(X) \simeq \mathcal{N}^{\theta,\rho}(Y)$ and

$$\begin{aligned} \mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big) &\simeq \mathbb{E}_{H_1}\left(\sum_{\theta,\rho}\mathcal{N}^{\theta,\rho}(X)^2\right) \\ &= \sum_{\theta,\rho}\mathbb{E}_{H_1}\Big(\mathcal{N}^{\theta,\rho}(X)^2\Big) \\ &= \sum_{\theta,\rho}\mathbb{E}_{H_1}\left((\sum_{i}^{n} Z_{\theta,\rho,i})^2\right), \\ &\quad Z_{\theta,\rho,i} \to \mathcal{B}(1, 1/Q) \\ &= \sum_{\theta,\rho}\mathbb{E}_{H_1}\left(\sum_{i,j}^{n} Z_{\theta,\rho,i}Z_{\theta,\rho,j}\right) \\ &= \sum_{\theta,\rho}\mathbb{E}_{H_1}\left(\sum_{i}^{n} Z^2_{\theta,\rho,i}\right) \\ &\quad + \quad \mathbb{E}_{H_1}\left(\sum_{i,j\neq i}^{n} Z_{\theta,\rho,i}Z_{\theta,\rho,j}\right). \end{aligned} \tag{A.7}$$

Since $Z_{\theta,\rho,i}, Z_{\theta,\rho,j} \overset{i.i.d}{\rightarrow} \mathcal{B}(1, 1/Q)$

$$\begin{aligned}\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big) &\simeq \sum_{\theta,\rho}\sum_{i}^{n}\mathbb{E}_{H_1}\left(Z^2_{\theta,\rho,i}\right)\\ &+ \sum_{i,j\neq i}^{n}\mathbb{E}_{H_1}\left(Z_{\theta,\rho,i}\right)\mathbb{E}_{H_1}\left(Z_{\theta,\rho,j}\right)\\ &= Q\big(n/Q + n(n-1)(1/Q)^2\big)\\ &= n^2/Q + n(1-1/Q),\end{aligned} \tag{A.8}$$

therefore,

$$\begin{aligned}p_s &\propto \exp\big(n^2/Q + n(1-1/Q)\big)\\ q_s = 1-p_s &\propto (m-1)\exp\big(n^2/Q\big).\end{aligned} \tag{A.9}$$

Now, we consider a normalization factor C (in A.4) which guarantees $p_s + q_s = 1$, accordingly p_s is

$$\frac{\exp\big(n(1-1/Q)\big)}{\exp\big(n(1-1/Q)\big) + (m-1)} \tag{A.10}$$

Under $H_0 \longrightarrow \nexists\ Y_J$ s.t. $(X, Y_J) \in H_1$

Equations (A.3) are updated as

$$\begin{aligned}&\mathbb{E}\big(\mathbf{K}_{Y_J|X}\big|(X,Y_J)\in H_0\big),\\ &\text{and}\\ &\sum_{j\neq J}^{m}\mathbb{E}\big(\mathbf{K}_{Y_j|X}\big|(X,Y_j)\in H_0\big)\end{aligned} \tag{A.11}$$

$$\begin{aligned}p_s &\propto \exp\Big(\mathbb{E}_{H_0}\big(\mathcal{N}(X,Y)\big)\Big)\\ q_s = 1-p_s &\propto (m-1)\exp\Big(\mathbb{E}_{H_0}\big(\mathcal{N}(X,Y)\big)\Big),\end{aligned} \tag{A.12}$$

Combining (A.5), (A.12), p_s is $1/m$. □

Appendix B

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below.[1]

International Journals

1. **L. Ballan**, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra. "Event Detection and Recognition for Semantic Annotation of Video", *Multimedia Tools and Applications*, vol. in press, 2011. (Special Issue: Survey Papers in Multimedia by World Experts) [DOI:10.1007/s11042-010-0643-7]

2. **L. Ballan**, M. Bertini, A. Del Bimbo, G. Serra. "Video Annotation and Retrieval using Ontologies and Rule Learning", *IEEE Multimedia*, vol. 17, iss. 4, pp. 80-88, 2010. [DOI: 10.1109/MMUL.2010.4]

3. **L. Ballan**, M. Bertini, A. Del Bimbo, G. Serra. "Semantic Annotation of Soccer Videos by Visual Instance Clustering and Spatial/Temporal Reasoning in Ontologies", *Multimedia Tools and Applications*, vol. 48, iss. 2, pp. 313-337, 2010. [DOI: 10.1007/s11042-009-0342-4]

4. **L. Ballan**, M. Bertini, A. Del Bimbo, G. Serra. "Video Event Classification using String Kernels", *Multimedia Tools and Applications*, vol. 48, iss. 1, pp. 69-87, 2010. (Special Issue on Content Based Multimedia Indexing) [DOI: 10.1007/s11042-009-0351-3]

[1] The author's bibliometric indices are the following: H-index = 4, total number of citations = 65 (source: Google Scholar on December 16, 2010).

Submitted

1. **L. Ballan**, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra. "Human Action Localization and Recognition using Spatio-Temporal Interest Points and Tracking", *Expert Systems: The Journal of Knowledge Engineering*, 2010. (Submitted after major revision)

2. I. Amerini, **L. Ballan**, R. Caldelli, A. Del Bimbo, G. Serra, "A SIFT-based forensic method for copy-move attack detection and transformation recovery", *IEEE Transactions on Information Forensics & Security*, 2010. (Submitted after major revision)

International Conferences and Workshops

1. **L. Ballan**, M. Bertini, A. Del Bimbo, M. Meoni, G. Serra. "Tag suggestion and localization in user-generated videos based on social knowledge", in *Proc. of ACM Multimedia International Workshop on Social Media (WSM)*, Firenze (Italy), 2010. **(Best paper award)**

2. I. Amerini, **L. Ballan**, R. Caldelli, A. Del Bimbo, G. Serra, "Geometric tampering estimation by means of a SIFT-based forensic analysis", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas (USA), 2010.

3. **L. Ballan**, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra. "Recognizing Human Actions by Fusing Spatio-temporal Appearance and Motion Descriptors", in *Proc. of IEEE International Conference on Image Processing (ICIP)*, Cairo (Egypt), 2009.

4. **L. Ballan**, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra. "Effective Codebooks for Human Action Categorization", in *Proc. of ICCV International Workshop on Video-oriented Object and Event Classification (VOEC)*, Kyoto (Japan), 2009.

5. **L. Ballan**, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra. "Human Action Recognition and Localization using Spatio-temporal Descriptors and Tracking", in *Proc. of AI*IA International Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAI*HBA)*, Reggio Emilia (Italy), 2009.

6. **L. Ballan**, M. Bertini, A. Del Bimbo, G. Serra. "Action Categorization in Soccer Videos using String Kernels", in *Proc. of IEEE International*

Workshop on Content-Based Multimedia Indexing (CBMI), Chania (Crete), 2009.

7. **L. Ballan**, M. Bertini, A. Del Bimbo, G. Serra. "Video Event Classification Using Bag of Words and String Kernels", in *Proc. of International Conference on Image Analysis and Processing (ICIAP)*, Salerno (Italy), 2009.

8. **L. Ballan**, A. Bazzica, M. Bertini, A. Del Bimbo, G. Serra. "Deep Networks for Audio Event Classification in Soccer Videos", in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, New York (USA), 2009.

9. **L. Ballan**, M. Bertini, A. Jain. "A System for Automatic Detection and Recognition of Advertising Trademarks in Sports Videos", in *Proc. of ACM International Conference on Multimedia (MM)*, Vancouver, (Canada), 2008.

10. **L. Ballan**, M. Bertini, A. Del Bimbo, A. Jain. "Automatic Trademark Detection and Recognition in Sport Videos", in *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, Hannover (Germany), 2008.

National Conferences

1. **L. Ballan**, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra. "Robust space-time features combination for human action recognition", in *Proc. of GIRPR National Conference*, Marina di Ascea (SA), Italy, 2010.

2. **L. Ballan**, M. Bertini, A. Del Bimbo, A. Jain. "Automatic Detection of Advertising Trademarks in Sport Video", in *Proc. of Italian Research Conference on Digital Library Systems (IRCDL)*, Padova, Italy, 2008.

Technical Reports

1. H. Sahbi, **L. Ballan**, G. Serra, and A. Del Bimbo. "Context-Dependent Logo Matching and Retrieval", TELECOM ParisTech, Technical Report, 2010D009, 2010.

Bibliography

[1] "Dublin Core Metadata Initiative." [Online]. Available: http://dublincore.org/

[2] "TV Anytime Forum." [Online]. Available: http://www.tv-anytime.org/

[3] U. Akdemir, P. Turaga, and R. Chellappa, "An ontology based approach for activity recognition from video," in *Proc. of ACM Multimedia (MM)*, 2008.

[4] F. Aldershoff and T. Gevers, "Visual tracking and localisation of billboards in streamed soccer matches," in *Proc. of SPIE Electronic Imaging*, San Jose, CA, USA, 2004.

[5] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "Geometric tampering estimation by means of a sift-based forensic analysis," in *Proc. of IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010.

[6] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura, "Comm: Designing a well-founded multimedia ontology for the web," in *Proc. of Int'l Semantic Web Conference (ISWC)*, 2007.

[7] A. Artikis, M. Sergot, and G. Paliouras, "A logic programming approach to activity recognition," in *Proc. of ACM Int'l Workshop on Events in Multimedia*, 2010.

[8] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 285–305, November-December 2003.

[9] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, 2002.

[10] A. D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, "Trademark matching and retrieval in sports video databases," in *Proc. of ACM Multimedia Information Retrieval (MIR)*, Augsburg, Germany, 2007.

[11] A. D. Bagdanov, A. Del Bimbo, F. Dini, and W. Nunziati, "Improving the robustness of particle filter-based visual trackers using online parameter adaptation," in *Proc. of IEEE Int'l Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2007.

[12] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line handwriting recognition with support vector machines - a kernel approach," in *Proc. of Int'l Workshop on Frontiers in Handwriting Recognition*, 2002.

[13] L. Bai, S. Lao, G. J. F. Jones, and A. F. Smeaton, "Video semantic content analysis based on ontology," in *Proc. of Int'l Machine Vision and Image Processing Conference*, 2007.

[14] L. Bai, S. Lao, W. Zhang, G. J. F. Jones, and A. F. Smeaton, "A semantic event detection approach for soccer video based on perception concepts and finite state machines," in *Proc. Intl'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2007.

[15] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action categorization," in *Proc. of ICCV Int'l Workshop on Video-oriented Object and Event Classification (VOEC)*, Kyoto, Japan, September 2009.

[16] ——, "Recognizing human actions by fusing spatio-temporal appearance and motion descriptors," in *Proc. of IEEE Int'l Conference on Image Processing (ICIP)*, Cairo, Egypt, November 2009.

[17] ——, "Event detection and recognition for semantic annotation of video," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 279–302, January 2011, (Special Issue: Survey Papers in Multimedia by World Experts).

[18] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies," *Multimedia Tools and Applications*, vol. 48, no. 2, pp. 313–337, June 2010.

[19] ——, "Video annotation and retrieval using ontologies and rule learning," *IEEE MultiMedia*, vol. 17, no. 4, pp. 80–88, 2010.

[20] ——, "Video event classification using string kernels," *Multimedia Tools and Applications*, vol. 48, no. 1, pp. 69–87, 2010.

[21] M. Barni and F. Bartolini, *Watermarking Systems Engineering: Enabling Digital Assets Security and Other Applications*. Marcel Dekker, 2004.

[22] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[23] S. Bayram, I. Avcibas, B. Sankur, and N. Memon, "Image manipulation detection with binary similarity measures," in *Proc. of European Signal Processing Conference (ESPC)*, 2005.

[24] S. Bayram, H. T. Sencar, and N. Memon, "A survey of copy-move forgery detection techniques," in *Proc. IEEE Western New York Image Processing Workshop*, 2008.

[25] ——, "An efficient and robust method for detecting copy-move forgery," in *Proc. of IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC, USA, 2009.

[26] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[27] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*. Berlin, Germany: Springer, 1984.

[28] M. Bertini, A. Del Bimbo, and G. Serra, "Learning rules for semantic video event annotation," in *Proc. of Int'l Conference on Visual Information Systems (VISUAL)*, 2008.

[29] M. Bertini, A. Del Bimbo, G. Serra, C. Torniai, R. Cucchiara, C. Grana, and R. Vezzani, "Dynamic pictorially enriched ontologies for digital video libraries," *IEEE MultiMedia*, vol. 16, no. 2, pp. 42–51, Apr/Jun 2009.

[30] M. Bertini, A. Del Bimbo, C. Torniai, R. Cucchiara, and C. Grana, "Dynamic pictorial ontologies for video digital libraries annotation," in *Proc. of ACM Int'l Workshop on Many Faces of Multimedia Semantics (MS)*, 2007.

[31] M. Bertini, G. D'Amico, A. Ferracani, M. Meoni, and G. Serra, "Sirio, Orione and Pan: an integrated web system for ontology-based video search and annotation," in *Proc. of ACM International Conference on Multimedia (MM) - DEMO Session*, Firenze, Italy, October 2010.

[32] M. Bertini, A. Del Bimbo, and W. Nunziati, "Common visual cues for sports highlights modeling," *Multimedia Tools and Applications*, vol. 27, no. 2, pp. 215–218, Nov 2005.

[33] M. Bertini, A. Del Bimbo, and G. Serra, "Learning ontology rules for semantic video annotation," in *Proc. of ACM Int'l Workshop on Many Faces of Multimedia Semantics (MS)*, 2008.

[34] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M. Strintzis, "Semantic annotation of images and videos for multimedia analysis," in *Proc. of European Semantic Web Conference (ESWC)*, 2005.

[35] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[36] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007.

[37] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[38] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. of ACM Int'l Workshop on Computational Learning Theory*, 1992.

[39] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, Aug 2000.

[40] S. Bravo-Solorio and A. K. Nandi, "Passive method for detecting duplicated regions affected by reflection, rotation and scaling," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2009.

[41] D. Brezeale and D. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 38, no. 3, pp. 416–430, May 2008.

[42] R. Caldelli, I. Amerini, and F. Picchioni, "A DFT-based analysis to discern between camera and scanned images," *International Journal of Digital Crime and Forensics (IJDCF), e-Forensics 2009 Special Edition*, vol. 2, no. 1, 2010.

[43] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[44] C. Chao, H.-C. Shih, and C.-L. Huang, "Semantics-based highlight extraction of soccer program using DBN," in *Proc. of IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[45] D. Chen, J. Yang, and H. D. Wactlar, "Towards automatic analysis of social interaction patterns in a nursing home environment from video," in *Proc. of ACM Multimedia Information Retrieval (MIR)*, 2004.

[46] J. Chen and J. Ye, "Training svm with indefinite kernels," in *Proc. of Int'l Conference on Machine Learning (ICML)*, 2008.

[47] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008.

[48] M. Chen, A. G. Hauptmann, and H. Li, "Informedia @ TRECVID2009: Analyzing video motions," in *Proc. of the TRECVID Workshop*, 2009.

[49] V. Christlein, C. Riess, and E. Angelopoulou, "A study on features for the detection of copy-move forgeries," in *Information Security Solutions Europe*, 2010.

[50] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[51] T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers,*, vol. 14, no. 3, pp. 326–334, June 1965.

[52] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital watermarking*. San Francisco, CA: Morgan Kaufmann, 2002.

[53] 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, no. 10, pp. 1210–1224, 2005.

[54] S. Dasiopoulou, C. Saathoff, P. Mylonas, Y. Avrithis, Y. Kompatsiaris, S. Staab, and M. Strintzis, *Semantic Multimedia and Ontologies Theory and Applications*. Springer, 2008, ch. Introducing Context and Reasoning in Visual Content Analysis: An Ontology-Based Framework, pp. 99–122.

[55] 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, no. 2, pp. 1–60, 2008.

[56] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[57] R. J. M. Den Hollander and A. Hanjalic, "Logo detection in video by string matching," in *Proc. of IEEE Int'l Conference on Image Processing (ICIP)*, Barcelona, ES, 2003.

[58] ——, "Logo recognition in video by line profile classification," in *Proc. of SPIE Electronic Imaging*, San Jose, CA, USA, 2004.

[59] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 2005.

[60] C. Dousson and P. Le Maigat, "Chronicle recognition improvement using temporal focusing and hierarchization," in *Proc. of Int'l Joint Conference on Artificial Intelligence*, 2007.

[61] B. Dybala, B. Jennings, and D. Letscher, "Detecting filtered cloning in digital images," in *Proc. of ACM Int'l Workshop on Multimedia & Security (MM&Sec)*, New York, NY, USA, 2007.

[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] J. P. Eakins, K. J. Riley, and J. D. Edwards, "Shape Feature Matching for Trademark Image Retrieval," in *Lecture Notes in Computer Science*, vol. 2728, 2003, pp. 439–443.

[64] S. Ebadollahi, X. L., S.-F. Chang, and J. R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proc. of IEEE Int'l Conference on Multimedia and Expo (ICME)*, 2006.

[65] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, 2006.

[66] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. of International Conference on Computer Vision (ICCV)*, 2003.

[67] H. Farid, "Photo fakery and forensics," *Advances in Computers*, vol. 77, 2009.

[68] ——, "A survey of image forgery detection," *IEEE Signal Processing Magazine*, vol. 2, no. 26, pp. 16–25, 2009.

[69] H. Farid and S. Lyu, "Higher-order wavelet statistics and their application to digital forensics," in *Proc. of IEEE CVPR Workshop on Statistical Analysis in Computer Vision*, Madison, WI, USA, 2003.

[70] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[71] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998, ch. A semantic network of English verbs.

[72] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[73] ——, "A sparse object category model for efficient learning and exhaustive recognition," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[74] P. Fihl, M. Holte, and T. Moeslund, "Motion primitives for action recognition," in *Proc. of Int'l Workshop on Gesture in Human-Computer Interaction and Simulation*, 2007.

[75] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[76] A. R. J. Francois, R. Nevatia, J. Hobbs, R. C. Bolles, and J. R. Smith, "VERL: an ontology framework for representing and annotating video events," *IEEE MultiMedia*, vol. 12, no. 4, pp. 76–86, Oct-Dec. 2005.

[77] J. Fridrich, D. Soukal, and J. Lukás, "Detection of copy-move forgery in digital images," in *Proc. of DFRWS*, 2003.

[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 Int'l Conference on Multimedia & Expo (ICME)*, New York, USA, 2009.

[79] R. Garcia and O. Celma, "Semantic integration and retrieval of multimedia metadata," in *Proc. of the Knowledge Markup and Semantic Annotation Workshop*, 2005.

[80] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: A texture classification example," in *Proc. of International Conference on Computer Vision (ICCV)*, 2001.

[81] B. Georis, M. Mazière, F. Brémond, and M. Thonnat, "A video interpretation platform applied to bank agency monitoring," in *Proc. of Intelligent Distributed Surveillance Systems Workshop*, 2004.

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

[83] P. E. Gill, W. Murray, and M. H. Wright, *Practical optimization.* London, United Kingdom: Academic Press., 1981.

[84] L. Gorelick, M. Blank, E. Schechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[85] T. Gruber, "Principles for the design of ontologies used for knowledge sharing," *International Journal of Human-Computer Studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[86] B. Haasdonk, "Feature space interpretation of svms with indefinite kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 482–492, 2005.

[87] A. Hakeem and M. Shah, "Ontology and taxonomy collaborated framework for meeting classification," in *Proc. of Int'l Conference on Pattern Recognition (ICPR)*, 2004.

[88] N. Harte, D. Lennon, and A. Kokaram, "On parsing visual sequences with the hidden Markov model," *EURASIP Journal on Image and Video Processing*, vol. 2009, pp. 1–13, 2009.

[89] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2004.

[90] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, 2003.

[91] A. Haubold and M. Naphade, "Classification of video events using 4-dimensional time-compressed motion features," in *Proc. of ACM International Conference on Image and Video Retrieval (CIVR)*, 2007, pp. 178–185.

[92] ——, "Classification of video events using 4-dimensional time-compressed motion features," in *Proc. of ACM Int'l Conference on Image and Video Retrieval (CIVR)*, 2007.

[93] A. G. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. D. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 958–966, Aug 2007.

[94] L. Hollink, S. Little, and J. Hunter, "Evaluating the application of semantic inferencing rules to image annotation," in *Proc. of Int'l Conference on Knowledge Capture*, 2005.

[95] H. Huang, W. Guo, and Y. Zhang, "Detection of copy-move forgery in digital images using sift algorithm," in *Proc. of IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, 2008.

[96] H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. Steele, and T. Serre, "Automated home-cage behavioral phenotyping of mice," *Nature communications*, September 2010.

[97] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: a comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[98] Y. Jing and S. Baluja, "Pagerank for product image search," in *Proc. of WWW*, Beijing, China, 2008.

[99] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. of ACM Multimedia (MM)*, Beijing, China, 2009.

[100] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. of International Conference on Computer Vision (ICCV)*, 2005.

[101] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[102] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 9, pp. 1163–1173, 2004.

[103] T. Kato, "Database architecture for content-based image retrieval," *Proc. of SPIE Image Storage and Retrieval Systems,*, vol. 1662, pp. 112–123, 1992.

[104] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors." in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[105] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. of International Conference on Computer Vision (ICCV)*, 2005.

[106] L. Kennedy, "Revision of LSCOM event/activity annotations, DTO challenge workshop on large scale concept ontology for multimedia," Columbia University, ADVENT Technical Report N.221-2006-7, 2006.

[107] N. Khanna, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, "Forensic techniques for classifying scanner, computer generated and digital camera images," in *Proc. of IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, USA, 2008.

[108] G. Kienast, H. Stiegler, W. Bailer, H. Rehatschek, S. Busemann, and T. Declerck, "Sponsorship tracking using distributed multi-modal analysis (DIRECT-INFO)," in *Proc. of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, 2005, pp. 341–348.

[109] W. Kienzle, B. Scholkopf, F. Wichmann, and M. O. Franz, "How to find interesting locations in video: A spatiotemporal interest point detector learned from human eye movements," in *Proc. of* 29^{th} *Annual Symposium of the German Association for Pattern Recognition.* Springer, 09 2007.

[110] Y. S. Kim and W. Y. Kim, "Content-based trademark retrieval system using visually salient features," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997, pp. 307–312.

[111] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," *Proc. of BMVC*, 2008.

[112] J. Kleban, X. Xie, and W.-Y. Ma, "Spatial pyramid mining for logo detection in natural scenes," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, Hannover, Germany, 2008.

[113] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," *Applied Image Pattern Recognition Workshop,*, pp. 1–8, 2008.

[114] Y. Kompatsiaris and P. Hobson, *Semantic Multimedia and Ontologies: Theory and Applications*. Springer, 2008.

[115] B. Kovar and A. Hanjalic, "Logo appearance statistics in a sport video: Video indexing for sponsorship revenue control," in *Proc. of SPIE Electronic Imaging*, San Jose, CA, USA, 2002.

[116] R. Kowalski and M. Sergot, "A logic-based calculus of events," *New Generation Computing*, vol. 4, no. 1, pp. 67–95, 1986.

[117] Y. Lamdan, J. T. Schwartz, and H. J. Wolfson, "Object recognition by affine invariant matching," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, Ann Arbor, MI, USA, 1988, pp. 335–344.

[118] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[119] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. of Int'l Conference on Computer Vision (ICCV)*, 2003.

[120] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[121] I. Laptev and P. Perez, "Retrieving actions in movies," in *Proc. of Int'l Conference on Computer Vision (ICCV)*, 2007.

[122] G. Lavee, A. Borzin, E. Rivlin, and M. Rudzsky, "Building Petri nets from video event ontologies," in *Proc. of International Symposium on Visual Computing (ISVC)*, ser. LNCS, vol. 4841. Springer Verlag, 2007, pp. 442–451.

[123] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, no. 5, pp. 489–504, Jun 2009.

[124] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[125] C. Leslie, E. Eskin, J. Weston, and W. S. Noble, "Mismatch string kernels for SVM protein classification," in *Proc. of Int'l Conference on Neural Information Processing Systems (NIPS)*, 2003.

[126] L. Leslie, T.-S. Chua, and R. Jain, "Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation," in *Proc. of ACM Multimedia (MM)*, 2007.

[127] G. Li, Q. Wu, D. Tu, and S. J. Sun, "A sorted neighborhood approach for detecting duplicated regions in image forgeries based on DWT and SVD," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, Beijing, China, 2007.

[128] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1310–1322, 2009.

[129] H.-J. Lin, C.-W. Wang, and Y.-T. Kao, "Fast copy-move forgery detection," *WSEAS Trans. Sig. Proc.*, vol. 5, no. 5, pp. 188–197, 2009.

[130] D. Liu, X.-S. Hua, and H.-J. Zhang, "Content-based tag processing for internet social images," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 723–738, 2011.

[131] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[132] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[133] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[134] 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, no. 2, pp. 240–251, Feb. 2008.

[135] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *Journal of Machine Learning Research*, 2002.

[136] D. G. Lowe, "The viewpoint consistency constraint," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 57–72, 1988.

[137] ——, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[138] W. Lu, A. L. Varna, and M. Wu, "Forensic hash for multimedia information," in *Proc. of SPIE Media Forensics and Security*, 2010.

[139] M. Luo, Y.-F. Ma, and H.-J. Zhang, "Pyramidwise structuring for soccer highlight extraction," in *Proc. of ICICS-PCM*, 2003.

[140] W. Luo, J. Huang, and G. Qiu, "Robust detection of region-duplication forgery in digital image," in *Proc. of ICPR*, Washington, D.C., USA, 2006.

[141] R. Luss and A. D'Aspremont, "Support vector machine classification with indefinite kernels," *Proc. of Int'l Conference on Neural Information Processing Systems (NIPS)*, 2008.

[142] S. Lyu and H. Farid, "How realistic is photorealistic?" *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 845–850, 2005.

[143] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[144] B. Mahdian and S. Saic, "Detection of copy-move forgery using a method based on blur moment invariants," *Forensic Science International*, vol. 171, no. 2-3, pp. 180–189, 2007.

[145] N. Maillot and M. Thonnat, "Ontology based complex object recognition," *Image and Vision Computing*, vol. 26, no. 1, pp. 102–113, 2008.

[146] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[147] R. Mehran, B. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. of European Conference on Computer Vision (ECCV)*, 2010.

[148] M. Merler, C. Galleguillos, and S. Belongie, "Recognizing groceries in situ using in vitro training data," in *Proc. of CVPR Workshop on Semantic Learning Applications in Multimedia*, Minneapolis, MN, USA, June 2007.

[149] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[150] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.

[151] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[152] J. A. Miller and G. Baramidze, "Simulation and the semantic web," in *Proc. of the Winter Simulation Conference (WSC)*, December 2005.

[153] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Proc. of Int'l Conference on Neural Information Processing Systems (NIPS)*, 2003.

[154] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, L. Kennedy, A. G. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, July-Sept. 2006.

[155] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[156] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[157] N. Negroponte, *Being digital.* Knopf, 1995.

[158] M. Neuhaus and H. Bunke, "Edit distance-based kernel functions for structural pattern classification," *Pattern Recognition*, vol. 39, no. 10, pp. 1852–1863, October 2006.

[159] B. Neumann and R. Moeller, "On scene interpretation with description logics," in *Cognitive Vision Systems: Sampling the Spectrum of Approaches*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 3948, pp. 247–278.

[160] R. Nevatia, J. Hobbs, and B. Bolles, "An ontology for video event representation," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[161] T.-T. Ng, S.-F. Chang, J. Hsu, and M. Pepeljugoski, "Columbia photographic images and photorealistic computer graphics dataset," ADVENT, Columbia University, Tech. Rep., 2004.

[162] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification." in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[163] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[164] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[165] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. of European Conference on Computer Vision (ECCV)*, 2006.

[166] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 36, p. 719, 2005.

[167] P. Over, G. Awad, J. Fiscus, M. Michel, A. F. Smeaton, and W. Kraaij, "TRECVid 2009–goals, tasks, data, evaluation mechanisms and metrics," in *Proc. of the TRECVID Workshop*, Gaithersburg, USA, 2009.

[168] X. Pan and S. Lyu, "Region duplication detection using image feature matching," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 857–867, 2010.

[169] A. Paschke and M. Bichler, "Knowledge representation concepts for automated SLA management," *Decision Support Systems*, vol. 46, no. 1, pp. 187–205, 2008.

[170] N. Pattanasri, A. Jatowt, and K. Tanaka, "Enhancing comprehension of events in video through explanation-on-demand hypervideo," in *Advances in Multimedia Modeling*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 4351, pp. 535–544.

[171] F. Pelisson, D. Hall, O. Riff, and J. L. Crowley, "Brand identification using gaussian derivative histograms," in *Lecture Notes in Computer Science*, vol. 2623, 2003, pp. 492–501.

[172] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[173] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 758–767, 2005.

[174] A. Popescu and H. Farid, "Exposing digital forgeries by detecting duplicated image regions," Dartmouth College, Computer Science, Tech. Rep. TR2004-515, 2004.

[175] ——, "Statistical tools for digital forensics," in *Proc. of International Workshop on Information Hiding*, Toronto, Canada, 2005.

[176] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[177] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[178] J. R. Quinlan, "Learning logical definitions from relations," *Machine Learning*, vol. 5, no. 3, pp. 239–266, 1990.

[179] J. A. Redi, W. Taktak, and J.-L. Dugelay, "Digital image forensics: a booklet for beginners," in *Multimedia Tools and Applications, S.I. on Survey Papers in Multimedia by World Experts*, vol. 51, no. 1, 2011, pp. 133–162.

[180] D. E. Riedel, S. Venkatesh, and W. Liu, "Recognising online spatial activities using a bioinformatics inspired sequence alignment approach," *Pattern Recognition*, vol. 41, no. 11, pp. 3481–3492, 2008.

[181] L. G. Roberts, "Machine perception of three-dimensional solids," Lincoln Laboratory, Tech. Rep. N.15, 1963.

[182] S.-J. Ryu, M.-J. Lee, and H.-K. Lee, "Detection of copy-rotate-move forgery using zernike moments," in *International Workshop on Information Hiding*, 2010.

[183] D. Sadlier and N. O'Connor, "Event detection in field sports video using audio–visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225–1233, Oct 2005.

[184] H. Sahbi, J.-Y. Audibert, and R. Keriven, "Incorporating context and geometry in kernel design," TELECOM ParisTech, Tech. Rep. N.2009D002, January 2009.

[185] H. Sahbi, J.-Y. Audibert, J. Rabarisoa, and R. Kerivan, "Context-dependent kernel design for object matching and recognition," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, 2008.

[186] H. Sahbi, L. Ballan, G. Serra, and A. Del Bimbo, "Context dependent logo matching and retrieval," TELECOM ParisTech, Tech. Rep. N.2010D009, March 2010.

[187] J. SanMiguel, J. Martinez, and A. Garcia, "An ontology for event detection and its application in surveillance video," in *Proc. of IEEE Int'l Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2009.

[188] S. Savarese, A. Del Pozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *Proc. of Workshop on Motion and Video Computing*, 2008.

[189] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[190] A. Scherp, T. Franz, C. Saathoff, and S. Staab, "F–a model of events based on the foundational ontology DOLCE+DnS ultralight," in *Proc. of Int'l Conference on Knowledge Capture (K-CAP)*, 2009.

[191] 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)*, Amsterdam, NL, 2007.

[192] C. Schmid and R. Mohr, "Local Grayvalue Invariants for Image Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.

[193] H. Schneiderman and T. Kanade, "A statistical method for 3D object detection applied to faces and cars," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[194] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. of Int'l Conference on Pattern Recognition (ICPR)*, 2004.

[195] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT descriptor and its application to action recognition," in *Proc. of ACM Multimedia (MM)*, 2007.

[196] L. Seidenari and M. Bertini, "Non-parametric anomaly detection exploiting space-time features," in *Proc. of ACM Multimedia (MM)*, 2010.

[197] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press, 2004.

[198] V. Shet, D. Harwood, and L. Davis, "VidMAP: Video monitoring of activity with prolog," in *Proc. of IEEE Int'l Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2005.

[199] X. Shuai, C. Zhang, and P. Hao, "Fingerprint indexing based on composite set of reduced sift features," in *Proc. of ICPR*, 2008.

[200] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, Feb. 2008.

[201] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. of Int'l Conference on Computer Vision (ICCV)*, 2003.

[202] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. of ACM Multimedia Information Retrieval (MIR)*, 2006.

[203] 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, no. 12, pp. 1349–1380, 2000.

[204] L. Snidaro, M. Belluz, and G. Foresti, "Domain knowledge for surveillance applications," in *Proc. of Int'l Conference on Information Fusion*, 2007.

[205] 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, no. 5, pp. 975–986, Aug. 2007.

[206] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[207] ——, "Are concept detector lexicons effective for video search?" in *Proc. of IEEE Int'l Conference on Multimedia & Expo*, Beijing, China, July 2007, pp. 1966–1969.

[208] 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 *Proc. of ACM Multimedia (MM)*, 2006.

[209] H. Su, A. Bouridane, and M. Gueham, "Local image features for shoeprint image retrieval," in *Proc. of BMVC*, 2007.

[210] A. Swaminathan, M. Wu, and K. Liu, "Digital image forensics via intrinsic fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 101–117, 2008.

[211] Q. Tian, S. Zhang, W. Zhou, R. Ji, B. Ni, and N. Sebe, "Building descriptive and discriminative visual codebook for large-scale image applications," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 441–477, 2011.

[212] S. D. Tran and L. S. Davis, "Event modeling and recognition using Markov logic networks," in *Proc. of European Conference on Computer Vision (ECCV)*, 2008.

[213] C. Tsinaraki, P. Polydoros, F. Kazasis, and S. Christodoulakis, "Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content," *Multimedia Tools and Applications*, vol. 26, no. 3, pp. 299–325, Aug. 2005.

[214] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Empowering visual categorization with the GPU," *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp. 60–70, 2011.

[215] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[216] R. H. Van Leuken, M. F. Demirci, V. J. Hodge, J. Austin, and R. C. Veltkamp, "Layout indexing of trademark images," in *Proc. of ACM International Conference on Image and Video Retrieval (CIVR)*, Amsterdam, NL, 2007.

[217] R. Vezzani and R. Cucchiara, "Video surveillance online repository (ViSOR): an integrated framework," *Multimedia Tools and Applications*, vol. 50, no. 2, pp. 359–380, 2010.

[218] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[219] ——, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[220] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[221] F. Wang, Y.-G. Jiang, and C.-W. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. of ACM Multimedia (MM)*, 2008.

[222] A. Watve and S. Sural, "Soccer video processing for the detection of advertisement billboards," *Pattern Recognition Letters*, vol. 29, no. 7, 2008.

[223] 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*, 2009.

[224] W. Wei, S. Wang, X. Zhang, and Z. Tang, "Estimation of image rotation angle using interpolation-related spectral signatures with application to blind detection of image forgery," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 507–517, 2010.

[225] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang, "Selection of concept detectors using ontology-enriched semantic space," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1085–1096, 2008.

[226] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of European Conference on Computer Vision (ECCV)*, 2008.

[227] S. A. J. Winder, G. Hua, and M. Brown, "Picking the best DAISY," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[228] S.-F. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. of Int'l Conference on Computer Vision (ICCV)*, 2007.

[229] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[230] T. Xiang and S. Gong, "Incremental and adaptive abnormal behaviour detectionq incremental and adaptive abnormal behaviour detection," *Computer Vision and Image Understanding*, vol. 111, pp. 59–73, 2008.

[231] D. Xu and S.-F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1985–1997, 2008.

[232] G. Xu, Y.-F. Ma, H.-J. Zhang, and S. Yang, "A HMM based semantic analysis framework for sports game event detection," in *Proc. of IEEE Int'l Conference on Image Processing (ICIP)*, Barcelona, Spain, September 2003.

[233] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, 2001.

[234] J. Yang and A. G. Hauptmann, "Exploring temporal consistency for video analysis and retrieval," in *Proc. of ACM Multimedia Information Retrieval (MIR)*, 2006.

[235] ——, "(Un)reliability of video concept detection," in *Proc. of ACM Int'l Conference on Image and Video Retrieval*, 2008.

[236] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. of ACM Multimedia Information Retrieval (MIR)*, 2007.

[237] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[238] P. Y. Yin and C. C. Yeh, "Content-based retrieval from trademark databases," *Pattern Recognition Letters*, vol. 23, no. 1-3, pp. 113 – 126, 2002.

[239] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua, "Building a comprehensive ontology to refine video concept detection," in *Proc. of ACM Multimedia Information Retrieval (MIR)*, 2007.

[240] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, pp. 345–357, 2008.

[241] D. Zhang, D. G. Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted HMMs for unusual event detection," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[242] J. Zhang, M. Marszałek, 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, no. 2, pp. 213–238, 2007.

[243] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. Huang, "SIFT-bag kernel for video event analysis," in *Proc. of ACM Multimedia (MM)*, 2008, pp. 229–238.