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

Supervised and Semi-supervised Event Detection with Local Spatio-Temporal Features

Candidate
Lorenzo Seidenari

Supervisors
Prof. Alberto Del Bimbo

Dr. Marco Bertini

PhD Coordinator
Prof. Giacomo Bucci

ciclo XXIV, 2009-2011

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.

Ai miei genitori.

Acknowledgements

First I would like to thank my supervisor, Prof. Alberto Del Bimbo, for guiding my research with his knowledge and experience and for constantly challenging me with new problems. I also thank all my colleagues of the Media Integration and Communication Center (MICC) who were of great help during my research. A special acknowledgement goes to Andrew D. Bagdanov, Lamberto Ballan, Marco Bertini, Luca Costantini, and Giuseppe Serra who directly collaborated with my research. Finally, thanks go to everyone who read or helped in any way to the realisation of this thesis.

Contents

Chapter 1

Introduction

Imaging sensors have become extremely affordable recently. Most of them are carried in people pockets embedded in phones or cameras. It is estimated that more than 4 billions of people own a mobile phone. Smartphones can now even perform some basic automatic analysis of video content; this capability, augmented with the use of cloud-based services, certainly encourages the acquisition and production of more video content.

The amount of videos produced daily in digital format grows astonishingly fast; in fact, as an example, on Youtube *"60 hours of video are uploaded every minute, resulting in nearly 10 years of content uploaded every day"* [119] with a 25% of increase with respect to the 8 hours per minute users were uploading as of May 2011 [174]. This data is often poorly annotated and user comments provide little or no information about the true content of the video. Several social networks base their existence and appeal on the possibility for their users to easily upload and share personal or publicly available videos. On the one hand existing video archiving services have added a social dimension, providing the possibility to comment and tag content, while on the other hand more traditional social networks constantly enhance media sharing features.

At the same time, the affordability of cameras and the growing perceived need for security in city streets and public buildings pushes the video-surveillance market. As examples of this trend, the city of Chicago is building a pervasive network, combining police and non-police cameras [24] and the city of London has installed around ten thousand cameras as a crime deterrent [37].

We are therefore observing two main phenomena that are responsible for the massive production of digital video data: the first is related to the production and sharing of user generated content portraying events of interest for the user, the latter is related to the continuously growing presence of cameras and camera networks that municipalities and other public institutions are installing in cities, roads and buildings.

1.1 Motivation

It appears obvious that this huge amount of visual information needs to be processed with some kind of machinery. As we know, machines have impressive computational power but their ability to derive a meaning from analysed data is comparatively very poor. This distance between the meaning immediately available to the human mind and the lack of for computers is often referred as the *semantic gap* [139]. The goal of most automated video analysis systems is to somehow reduce this gap. The goal of this work is to implement effective methods for searching archives of generic videos, addressing in particular the detection and recognition of events. We want to overcome the current video search paradigm that is based on textual information co-located with the video data. Our aim is to enable users to retrieve video from a database with semantic queries; for example, Figure 1.1 shows the result of the query: "Retrieve all video with people hugging". In a different application, a system could provide the annotation of unseen

Figure 1.1: Result for query: "people hugging".

or novel data; this is specifically useful for video surveillance operators. In this case the query to be satisfied is no longer related to a specific event

Figure 1.2: Examples of normal (left column) and abnormal (right column) events for two common scenarios.

but covers a more broad spectrum of visual data regarded as abnormal with respect to the known/past data. Figure 1.2 shows the output of a system able to recognise visual information it has never processed before but that contradicts the learnt model.

1.2 Contributions

This thesis deals with various aspects of the analysis of video sequences. In order to extract valuable semantic information from video data we provide several tools to represent and automatically recognise the content in videos.

The first problem this work deals with is the detection of known categories of events; we refer to this problem as supervised detection of events. We train classifiers on videos from a finite and human annotated set of classes of events. The main building block of the approaches proposed in this thesis is the description of video content with local spatio-temporal features. The local descriptors we define are used to sparsely represent objects and their motion in the analysed sequences. The proposed approach is suitable for the extraction of semantic information from videos shot "in the wild", since

it can cope with occlusions, self-occlusions, illumination, scale and point-of-view variations.

In the quest for systems able to automatically process huge amounts of data, specifically for security and surveillance purposes, the definition of a finite set of events to be recognised poses some hard-to-ovecome limitations. The second problem this work addresses is the retrieval of events of interest regardless of their specific nature. We cast the detection of these events of interest as anomaly detection. This approach eases the challenge since it requires the gathering of just the normal data, that we assume is available in quantity.

Finally, since video data ultimately must be stored or transmitted, we propose a method that can leverage both semantic high level features and low level image features to reduce the disk space and bandwidth needed.

The rest of the thesis is organised as follows. We start with a review of the state of the art in the analysis and annotation of events in Chapter 2. This chapter serves to build a background for the subsequent chapters and its emphasis is on approaches that exploit the temporal nature of videos; in particular, we extensively review techniques based on local spatio-temporal features that represent the foundation of most of the work presented in the following chapters.

Chapters 3 and 4 deal with the classification of events belonging to specific known classes. We concentrate on the recognition of human actions and activities. In Chapter 3 we define a novel descriptor based on the local variations of appearance. The proposed descriptor is efficient and does not require any parameter tuning. We adopt radius-based clustering and smooth assignment of feature descriptors to create effective codebooks. We also deal with the reduction of dimensionality by applying a novel technique based on deep learning. Chapter 4 aims at defining a more compact representation for local space-time features to alleviate the main drawbacks of space-time descriptors: computational efficiency and high dimensionality. We use 3D Zernike moments to compute a representation of space-time patches that is not redundant by construction. Furthermore, given the hierarchical structure of the proposed descriptor, we formulate a novel measure of similarity based on pyramidal matching of features. The proposed similarity is a valid Mercer kernel.

Chapter 5 is a shift in the event retrieval paradigm with respect to previous chapters. We advocate the use of anomaly detection techniques when

the classes of interests are too many or unknown in advance. We exploit the features presented in Chapter 3 and a non-parametric technique in order to build a model of the dynamic appearance of the scene. Our model uses multiple scales and the context of local patterns to detect unusual events. The system runs in real-time with no specific hardware or parallelisation of any sort.

Finally in Chapter 6 we propose an application of semantic cue extraction in videos that greatly reduces the storage and bandwidth requirements of streaming video. The proposed method is based on the selective removal of uninteresting areas with Gaussian blur and the successive compression with the H.264 codec. Areas of interest can be detected with a detector or, in case of limited available computing power, with a combination of cheap-to-compute local features like image corners and edges.

Chapter 2

Literature review

This chapter gives a brief survey of related work on event recognition using local visual features. [1]

2.1 Introduction

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. Typically videos are automatically segmented in shots and a representative keyframe of each shot is analysed to recognise 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 [157] even for a human annotator, as shown in Figure 2.1. 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 [65]. An event filmed in a video is related to the temporal aspect of the video itself and to some

[1] Part of this chapter 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* [8].

changes in the properties of the entities and scenes represented; therefore there is need of representing and modelling time and properties' variations, using appropriate detectors, feature descriptors and models.

Several surveys on semantic video annotation have been recently presented. A review of multi-modal video indexing was presented in [142], considering entertainment and informative video domains. Multi-modal approaches for video classification have been surveyed in [23]. A survey on event detection has been presented in [82], focusing on modelling techniques; our work 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 behaviour recognition in surveillance applications has been provided in [73], while in [124] are reported the most recent works on human action recognition. A survey of crowd analysis methods was reported in [177]. In this chapter we survey methods that have been applied to different video domains, considering edited videos (i.e. videos that have been created from a collection of video material, selecting what elements to retain, delete, or combine, like movies) and unedited videos (i.e. videos that have not been processed and are simply the result of video recording, like surveillance videos).

(a) (b) (c)

Figure 2.1: 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* ?

The problem of semantic video annotation is strictly related to the problem of generic visual categorisation, like classification of objects or scenes, rather than that of recognising a specific class of objects. Recently it has been shown that part-based approaches are effective methods for scene and object recognition [43, 137, 170, 179] 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 therefore achieving robustness against occlusion and clutter. To deal effectively with changes of viewing conditions the features should be invariant to geometrical transformations such as translation, rotation, scaling and also affine transformations. SIFT [93] and SURF [10] features have become the de facto standards, because of their good performance and (relatively) low computational cost. In this field, a solution that recently has become very popular is the Bag-of-Words (BoW) approach. It has been originally proposed for information retrieval, where it is used for document categorisation 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 quantised invariant local descriptors, called *visual-words*. The main reason for the success of this approach is that it provides methods that are sufficiently generic to cope with many object types simultaneously. 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 [138] challenges.

The problem of the detection and recognition of events and activities is recently getting a larger attention, also within the TRECVid evaluation: the high-level concept detection task of TRECVid 2009 [120] considered the problem of event detection, with 7 out of 20 high-level concepts to be detected that were related to events and actions [32]. The most recent approaches proposed in this task have started to cope with the problem of representing videos considering the temporal aspects of it, analysing more than one keyframe per shot and introducing some representation of the context [120, 169]. 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).

2.2 Events and Actions

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 [141] which can be exploited in the event modelling as explained in Sections 2.3.1. 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 [96, 132]. 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*, [14]) 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 [44]. This kind of human events are usually recognised 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 [64].

Events involving multiple people or happening in longer timeframes can be referred as *activities* [124]. 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* [177]: persons are no more considered as individuals but the global motion of a crowd is modelled [101]. 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 provided by static features or scene classifiers may improve event recognition performance [51, 91, 98].

2.3 Features for Actions and Events

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 generalisation 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 [93, 104]; image regions can be normalised [106] to obtain invariance to deformations due to viewpoint change and another normalisation can be applied to obtain rotation and partial photometric invariance [93].

This kind of description has been extended in the object and scene categorisation scenario exploiting the bag-of-words framework [137]. Through the use of an intermediate description, the codebook, images are compactly

represented. The codebook is usually obtained with a vector quantisation procedure exploiting some clustering algorithm such as k-means. This intermediate description allows both fast data access, by building an inverted index [114, 137], and generalisation over category of objects by representing each instance as a composition of common parts [43]. As in the textual counterpart the bag of visual words does not retain any structural information: 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 generalisation 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 [129] can recover some image structure without loss of generalisation power. A global approach has been proposed by Lazebnik *et al.* [83]; 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.

Given the success of bag of keypoints representations in static concept classification, efforts have been made to introduce this framework in event categorisation. The first attempt in video annotation has been made by Zhou *et al.* [180], 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. 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.

2.3.1 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 [78, 118, 162, 164] while the approaches presented in [32, 38] are explicitly designed for space-time features. Laptev extended Harris corners keypoints to the space-time domain [78]; 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 [10] 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 [63] have been extended by Oikonomopoulos *et al.* [118]. The detector proposed by Dollár *et al.* [38] 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.* [32], 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 [78, 162, 164] both in space and time or may simply be a parameter of it [38, 80]; moreover scale for space and time can be fixed as in [38] or a dense sampling can be performed to enrich the representation [6, 80]. Figure 2.2 shows an example of the response of the detectors presented in [6], 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 [68] or extending boosting and Haar features used for object detection [156]. 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 [116]. 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 [91, 98, 105].

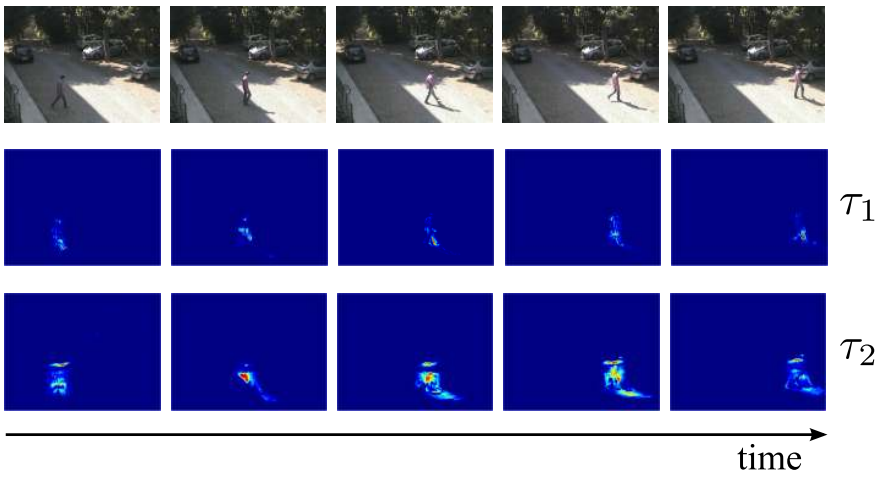
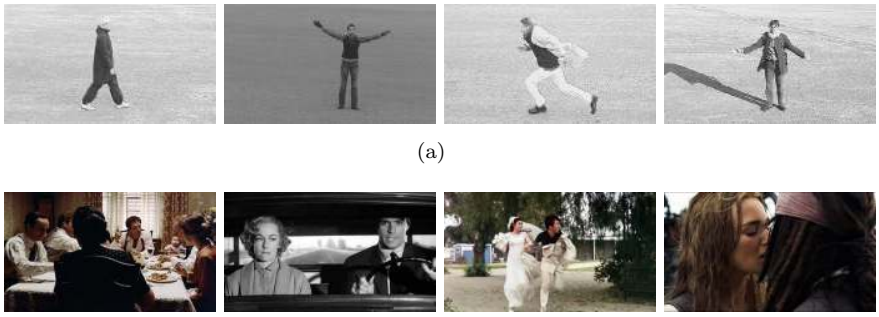


Figure 2.2: Spatio-temporal interest point detector [6] running at different temporal scales (blue low response, red high response); first row: original video frames, second row detector response at temporal scale τ_1 (mostly due to the limbs), third row: detector response temporal scale τ_2 (mostly due to the torso), with $\tau_1 < \tau_2$. Frames taken from the ViSOR video repository [154].

2.3.2 Descriptors

The regions extracted by detectors need to be represented compactly. Descriptors are usually computed using a common pipeline as outlined in [163] for static features and, partially, in [79] 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 [71, 78]. In order to obtain more robust descriptors a region normalisation can be applied [78]; 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 [38], image gradients [78], spatio-temporal gradients [6,38,71,131] and optical flow [6,38,78]. 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 motion information these two descriptions have been found to be complementary [6] 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 [38, 79] or local [6, 71, 131] 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 [38, 79]. A different technique is to compute higher order derivatives of image intensity values [78]. Finally, following the approach of SIFT a descriptor normalisation and clipping can be applied to obtain robustness w.r.t. contrast change [71]. As shown in [163], for static feature descriptors, parameters can be learnt instead of "handcrafted"; Marszalek *et al.* performed such an optimisation by training on datasets [98]. 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. Figure 2.3 shows sample frames of these two datasets.

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



(a)

(b)

Figure 2.3: Sample frames from actions in KTH (a) and Hollywood (b) datasets.

2.3.3 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 [93] 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 [92,131] or encoded in the final action representation [105,111,128,165]. Scovanner *et al.* [131] 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.* [111] and Wong *et al.* [165] exploited graphical models to introduce a structural representation of the human action

by modelling relations among body parts and their motion. Savarese *et al.* [128] augmented the action descriptor by computing visual words spatio-temporal correlograms instead of a flat word-count. Finally Mikolajczyk and Uemura [105] 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 [41, 80, 81]. 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 [41, 81] or non-uniform layout [80], and by binning features in space and time, position dependent feature statistics is computed.

2.4 Classification of complex 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. As briefly discussed in Section 2.3, many methods proposed recently 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 [168] 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 [157], BoW is extended constructing relative motion histograms between visual words (ERMH-BoW) in order to employ motion relativity and visual relatedness. Zhou *et al.* [180] presented a SIFT-

Bag based generative-to-discriminative framework for video event detection, providing improvements on the best results of [168] 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.* [9] modelled events as a sequence composed of histograms of visual features, computed from each frame using the traditional bag-of-words (see Figure 2.4). 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 [58] to classify the behaviour of caged mice.

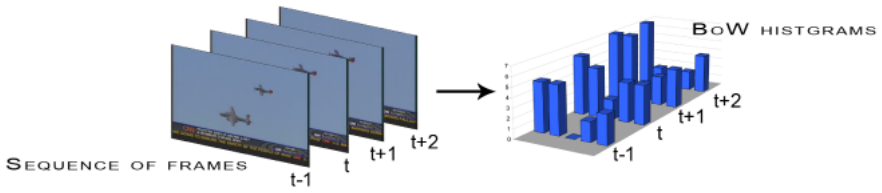


Figure 2.4: Shots are represented as a sequence of BoW histogram; Events are so described by concatenation of histograms of variable size, depending on the clip length. Example taken from [9].

2.4.1 Detection of unknown or rare events

Building models for activities performed by multiple persons interacting with each other and possibly with the objects present in the scene is often extremely complex and requires the detection and tracking of the subjects involved [33]. This difficulties are accentuated if the behaviors we want to recognise have a long temporal extent. One of the toughest challenges in this set of problems is the definition of a finite set of human activities for which to train classifiers; moreover even if the delineation of all the behaviours of interest was possible, the gathering of sufficient data is still required. In this particular setting may be convenient to attempt to solve the problem of

detecting interesting behaviour as a semi-supervised problem. Indeed it is often plausible to be able to collect a great amount of data to be regarded as normal; robust models can be then trained with this data in order to detect novel or anomalous patterns that are not respecting the known normal distribution. In visual analysis anomalous pattern detection has been performed mainly based on the analysis of trajectories [2,25,56] and more recently with the use of local image descriptions [1, 70, 96, 102], based on optical flow or appearance.

Chapter 3

Event Detection with Spatio-Temporal Features

Recognition and classification of human actions for annotation of unconstrained video sequences has proven to be challenging because of the variations in the environment, appearance of actors, modalities in which the same action is performed by different persons, speed and duration and points of view from which the event is observed. This variability reflects in the difficulty of defining effective descriptors and deriving appropriate and effective codebooks for action categorisation. In this chapter we propose a novel and effective solution to classify human actions in unconstrained videos. It improves on previous contributions defining a novel local descriptor that uses image gradient and optic flow to respectively model the appearance and motion of human actions at interest point regions. In the formation of the codebook we employ radius-based clustering with soft assignment in order to create a rich vocabulary that may account for the high variability of human actions. We obtain a strong reduction of computation time by applying codebook size reduction with Deep Belief Networks with little loss of accuracy.[1]

[1]The work presented in this chapter is based on the preliminary work published as "Effective Codebooks for Human Action Categorization" in Proc. of ICCV-WS Video and Object Event Categorization (VOEC), 2009 [7]

3.1 Introduction

With the continuous growth of video production and archiving, the need for automatic annotation tools that enable effective retrieval by content has accordingly gained increasing importance. In particular, action recognition is a very active research topic with many important applications such as human-computer interaction, video indexing and video-surveillance. Existing approaches for human action recognition can be classified as using holistic or part-based information [8, 135]. Most of the holistic-based methods usually perform better in a controlled environment and are also computationally expensive due to the requirement of pre-processing the input data. Moreover, these representations can be influenced by motions of multiple objects, variations in the background and occlusions. Instead, part-based representations that exploit interest point detectors combined with robust feature descriptors, have been used very successfully for object and scene classification tasks in images [43, 179]. As a result, nowadays most video annotation solutions have exploited the bag-of-features approach to generate textual labels that represent the categories of the main and easiest to detect entities (such as objects and persons) in the video sequence [52, 140].

The definition of effective descriptors that are able to capture both spatial and temporal features has opened the possibility of recognizing dynamic concepts in video sequences. In particular, interesting results have been obtained in the definition of solutions to automatically recognise human body movements, which usually represent a relevant part of video content [107, 123, 124, 149]. However, the recognition and classification of such dynamic concepts for annotation of generic video sequences has proven to be very challenging because of the very many variations in environment, people and occurrences that may be observed. These can be caused by cluttered or moving background, camera motion and illumination changes; people may have different size, shape and posture appearance; semantically equivalent actions can manifest differently or partially, due to speed, duration or self-occlusions; the same action can be performed in different modes by different persons. This great variability on the one hand reflects in the difficulty of defining effective descriptors and on the other makes it hard to obtain a visual representation that may describe such dynamic concepts appropriately and efficiently.

3.1.1 Effective Spatio-Temporal Descriptors

Holistic descriptors of body movements have been proposed by a few authors. Among the most notable solutions, Bobick *et al.* [17] proposed motion history images and their low-order moments 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. Efros *et al.* [39] created stabilised spatio-temporal volumes for each action video segment and extracted a smoothed dense optic flow field for each volume. They have proved that this representation is particularly suited for distant objects, where the detailed information of the appearance is not available. Yilmaz and Shah [172] used a spatio-temporal volume, built stacking object regions; descriptors encoding direction, speed and local shape of the resulting 3D surface were generated by measuring local differential geometrical properties. Gorelick *et al.* [48] 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. Global descriptors that jointly encode shape and motion were suggested in Lin *et al.* [88]; Wang *et al.* [158] exploited global histograms of optic flow together with hidden conditional random fields. Although encoding much of the visual information, these solutions have shown to be highly sensitive to occlusions, noise and change in viewpoint. Most of them have also proven to be computationally expensive due to the fact that some pre-processing of the input data is needed, such as background subtraction, segmentation and object tracking. All these aspects make these solutions only suited for representation of body movements in videos taken in controlled contexts.

Local descriptors have shown better performance and are in principle better suited for videos taken in both constrained and unconstrained contexts. They are less sensitive to partial occlusions and clutter and overcome some of the limitations of the holistic models, such as the need of background subtraction and target tracking. In this approach, local patches at spatio-temporal interest points are used to extract robust descriptors of local moving parts and the bag-of-features approach is employed to have distinctive representations of body movements. Laptev [78] and Dollár [38] approaches have been among the first solutions. Laptev [78, 130] proposed an extension to the Harris-Förstner corner detector for the spatio-temporal case; interesting parts were extracted from voxels surrounding local maxima

of spatio-temporal corners, i.e. locations of videos that exhibit strong variations of intensity both in spatial and temporal directions. The extension of the scale-space theory to the temporal dimension permitted to define a method for automatic scale-selection. Dollár *et al.* [38] proposed a different descriptor than Laptev's, by looking for locally periodic motion. While this method produces a denser sampling of the spatio-temporal volume, it does not provide automatic scale-selection. Despite of it, experimental results have shown that it improves with respect to [130].

Following these works, other authors have extended the definition of local interest point detectors and descriptors to incorporate time or combined static local features with other descriptors so to model the temporal evolution of local patches. Sun *et al.* [145] have fused spatio-temporal SIFT points with holistic features based on Zernike moments. In [162], Willems *et al.* extended SURF feature to time and defined a new scale-invariant spatio-temporal detector and descriptor that showed high efficiency. Scovanner *et al.* [131], have proposed to use grouping of 3D SIFT, based on co-occurrence, to represent actions. Kläser *et al.* [71] have proposed a descriptor based on histograms of oriented 3D gradients, quantised using platonic solids. Gao *et al.* [47] presented MoSIFT, an approach that extend the SIFT algorithm to find visually distinctive elements in the spatial domain. It detects spatio-temporal points with a high amount of optical flow around the distinctive points motion constraints. More recently, Laptev *et al.* [80] proposed a structural representation based on dense temporal and spatial scale sampling, inspired by the spatial pyramid approach of [83] with interesting classification results in generic video scenes. Kovashka *et al.* [75] extended this work by defining a hierarchy of discriminative neighbourhoods instead of using spatio-temporal pyramids. Liu *et al.* [91] combined MSER and Harris-Affine [106] regions with Dollár's space-time features and used AdaBoost to classify YouTube videos. Shao *et al.* [134] applied transformation based techniques (i.e. Discrete Fourier Transform, Discrete Cosine Transform and Discrete Wavelet Transform) on the local patches and used the transformed coefficients as descriptors. Yu *et al.* [175] presented good results using the Dollar's descriptor and random forest-based template matching. Niebles *et al.* [112] trained an unsupervised probabilistic topic model using the same spatio-temporal features, while Cao *et al.* [27] suggested to perform model adaptation in order to reduce the amount of labeled data needed to detect actions in videos of uncontrolled scenes. Comparative evaluations of the per-

formance of the most notable approaches were recently reported by Wang *et al.* [158] and Shao *et al.* [135].

3.1.2 Suitable Visual Codebooks

According to the bag-of-features model actions are defined as sets of codewords obtained from the clustering of local spatio-temporal descriptors. Most of the methods have used the k-means algorithm for clustering because of its simplicity and speed of convergence [43, 112, 137, 170]. However, both the intrinsic weakness of k-means to outliers and the need of some empirical preevaluation of the number of clusters hardly fit with the nature of the problem at hand. Moreover, with k-means the fact that cluster centres are selected almost exclusively around the most dense regions in the descriptor space results into ineffective codewords of action primitives. To overcome the limitations of the basic approach, Liu *et al.* [92] suggested a method to automatically find the optimal number of visual word clusters through maximisation of mutual information (MMI) between words and actions. MMI clustering is used after k-means to discover a compact representation from the initial codebook of words. They showed some performance improvement. Recently Kong *et al.* [74] have proposed a framework that unifies reduction of descriptor dimensionality and codebook creation, to learn compact codebooks for action recognition optimizing class separability. Differently, Uemura and Mikolajczyk [105] explored the idea of using a large number of features represented in many vocabulary trees instead of a single flat vocabulary. Yao *et al.* [171] recently proposed a similar framework using a training procedure based on a Hough voting forest. Both these methods require higher efforts in the training phase.

3.1.3 Our Contribution

In this chapter we propose a novel and effective solution to classify human actions in unconstrained videos. It improves on previous contributions in the literature through the definition of a novel local descriptor and the adoption of a more effective solution for the codebook formation. We use image gradient and optic flow to respectively model the appearance and motion of human actions at regions in the neighbourhood of local interest points and consider multiple spatial and temporal scales. These two descriptors are used in combination to model local features of human actions and activities.

Unlike similar related works [71, 131], no parameter tuning is required.

In the formation of the codebook we recognise that the clusters of spatio-temporal descriptors should be both in a sufficiently large number and sufficiently distinguished from each other so to represent the augmented variability of dynamic content with respect to the static case. To this end radius-based clustering [62] with soft assignment has been used. In fact, with radius-based clustering cluster centers are allocated at the modes corresponding to the maximal density regions, so resulting into a statistics of the codewords that better fits with the variability of human actions with respect to k-means clustering. Experiments carried on standard datasets show that the approach followed outperforms the current state of the art methods. To avoid too large codebooks we performed codebook compression with Deep Belief Networks. The solution proposed shows good accuracy even with very small codebooks. Finally, we provide several experiments on the Hollywood2 dataset [80] and on a new surveillance dataset (MICC-Surveillance), to demonstrate the effectiveness and generality of our method for action recognition in unconstrained video domains.

The rest of the chapter is organised as follows: the descriptor is presented in Section 3.2. Action representation and categorisation is presented in Section 3.3. The experimental results, with an extensive comparison with the state-of-the-art approaches, are hence discussed in Section 3.5. Here we also included experiments on unconstrained videos to demonstrate the effectiveness of the approach also in this case. Conclusions are drawn in Section 3.6.

3.2 Spatio-temporal Local Descriptors of Appearance and Motion

Spatio-temporal interest points are detected at video local maxima of the Dollár's detector [38] applied over a set of spatial and temporal scales. Using multiple scales is fundamental to capture the essence of human activity. To this end, linear filters are separately applied to the spatial and temporal dimension: on the one hand, the spatial scale permits to detect visual features of high and low detail; on the other, the temporal scale allows to detect *action primitives* at different temporal resolutions. The filter response

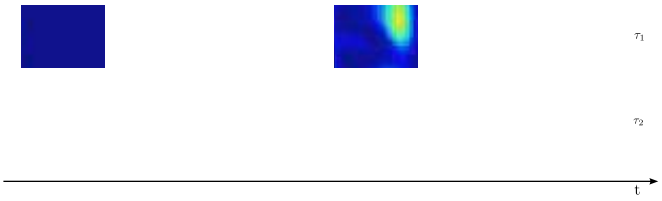


Figure 3.1: Response of the spatio-temporal interest point detector at two temporal scales $\tau_1 < \tau_2$ (low response in blue, high response in red); first row: original video frames, second row detector response at temporal scale τ_1 (mostly due to motion of human limbs); third row: detector response temporal scale τ_2 (mostly due to motion of human torso).

function is defined as:

$$R = \Big(I * g_\sigma * h_{ev}\Big)^2 + \Big(I * g_\sigma * h_{od}\Big)^2 \tag{3.1}$$

where $I(x, y, t)$ is the image sequence, $g_\sigma(x, y)$ is a spatial Gaussian filter with scale σ, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters that provide a strong response to temporal intensity changes for periodic motion patterns, respectively defined as:

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2} \tag{3.2}$$
$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2} \tag{3.3}$$

where $\omega = 4/\tau$. In the experiments we used $\sigma = \{2, 4\}$ as spatial scales and $\tau = \{2, 4\}$ as temporal scales. Figure 3.1 shows an example of temporal scaling of human body parts activity during walking: torso has high response at high temporal scale, while limbs respond at the lower scale.

Three-dimensional regions of size proportional to the detector scale (6x) are considered at each spatio-temporal interest point, and divided into equally sized sub-regions (three for each spatial dimensions along the x and y, and two for the temporal dimension t), as shown in Figure 3.2.

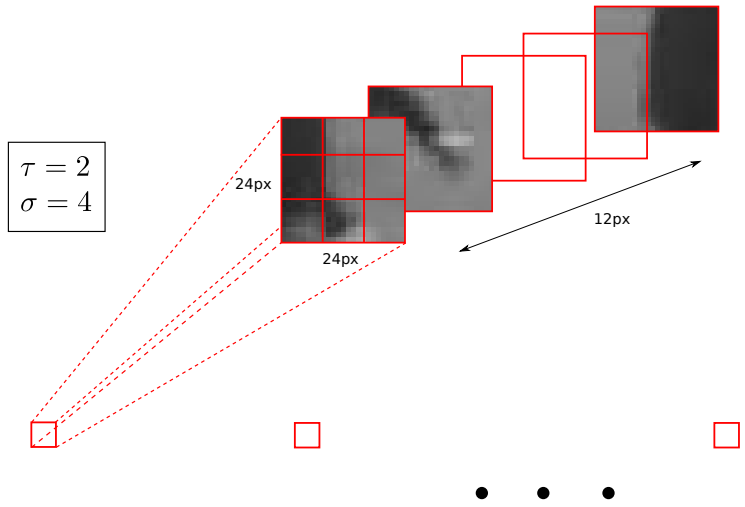


Figure 3.2: Three dimensional region at the spatio-temporal interest point corresponding to a swinging arm.

For each sub-region, image gradients on x, y and t are computed as:

$$G_x = I(x+1, y, t) - I(x-1, y, t) \tag{3.4}$$
$$G_y = I(x, y+1, t) - I(x, y-1, t) \tag{3.5}$$
$$G_t = I(x, y, t+1) - I(x, y, t-1) \tag{3.6}$$

and the optic flow with relative apparent velocity V_x, V_y is estimated according to [95].

Orientations of gradients and optical flow are computed for each pixel as:

$$\phi = \tan^{-1}\left(G_t / \sqrt{G_x^2 + G_y^2}\right) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \tag{3.7}$$
$$\theta = \tan^{-1}\left(G_y / G_x\right) \in [-\pi, \pi] \tag{3.8}$$
$$\psi = \tan^{-1}\left(V_y / V_x\right) \in [-\pi, \pi] \tag{3.9}$$

where ϕ and the θ are quantised in four and eight bins, respectively.

The local descriptor obtained by concatenating ϕ and θ histograms (H3DGrad) has therefore size $3 \times 3 \times 2 \times (8+4) = 216$. There is no need to re-orient the 3D neighbourhood, since rotational invariance, typically required in object detection and recognition, is not desirable in the action classification context.

This approach is much simpler to compute than those proposed in [131] and [71]. In particular, in [131] the histogram is normalised by the solid angle value to avoid distortions due to the polar coordinate representation (instead of quantizing separately the two orientations as in our approach), moreover the size of the descriptor is 2048; in [71] the 3D gradient vector is projected on the faces of a platonic solid. This latter approach requires additional parameter tuning, to optimise the selection of the solid used for the histogram computation and whether to consider the orientations of its faces or not. Differently from [80] our 12-bin H3DGrad descriptor models the dynamic appearance of the three-dimensional region used for its computation, instead of being a 4-bin 2D histogram cumulated over time.

The ψ is quantised in eight bins with an extra "no-motion" bin added to improve performance. The local descriptor of ψ (HOF) has size $3 \times 3 \times 2 \times (8+1) = 162$. Histograms of ϕ, θ and ψ are respectively derived by weighting pixel contributions respectively with the gradient magnitude $M_G = \sqrt{G_x^2 + G_y^2 + G_t^2}$ (for ϕ and θ), and the optic flow magnitude $M_O = \sqrt{V_x^2 + V_y^2}$ (for ψ).

In order to obtain an effective codebook for human actions these two descriptors can be combined according to either early or late fusion. In the former case the two descriptors are first concatenated and the combined descriptor is hence used for the definition of the human action codebook. In the latter a codebook is obtained from each descriptor separately; then the histograms of codewords are concatenated to form the representation (see Figure 3.3).

Figure 3.4 shows the classification accuracy measured with the KTH dataset, using codebooks based on the H3DGrad descriptor (a), HOF descriptor (b), and early (c) and late fusion (d), with 4000 codewords. Each action, is represented by an histogram H of codewords w obtained according to k-means clustering with hard assignment:

$$H(w) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 1 & \text{if } w = \underset{v \in V}{\operatorname{argmin}}(D(v, f_i)); \\ 0 & \text{otherwise;} \end{cases} \tag{3.10}$$

where n is the number of the spatio-temporal features, f_i is the i-th spatio-temporal feature, and $D(v, f_i)$ is the Euclidean distance between the codeword v of the vocabulary V and f_i.

We present in Table 3.1 the average accuracy obtained by H3DGrad and

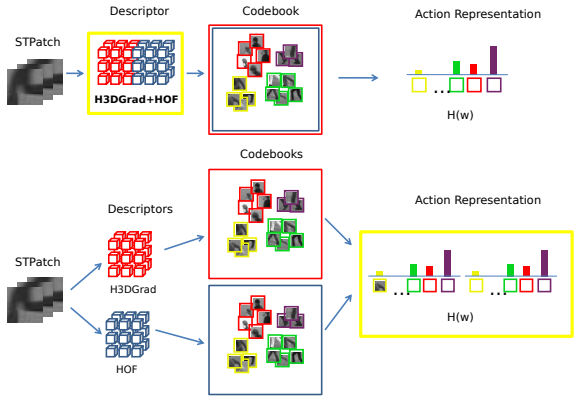


Figure 3.3: Two fusion strategies: early-fusion (at the descriptor level) and late-fusion (at the codebook level).

Descriptor	**KTH**	**Weizmann**
H3DGrad	90.38	92.30
HOF	88.04	89.74
H3DGrad + HOF (early fusion)	91.09	92.38
H3DGrad + HOF (late fusion)	**92.10**	**92.41**

Table 3.1: Average class accuracy of our descriptors, alone and combined, on the KTH and Weizmann datasets.

HOF respectively, and by the early and late fusion. From the figures, it appears clearly that late fusion provides the best performance. This can be explained with the fact that H3DGrad and HOF descriptors have quite complementary roles (for example the *boxing* action is better recognised when using H3DGrad descriptor while *hand-clapping* action is better recognised by HOF, as shown in Figure 3.4 (a),(b)). Late fusion improves recognition performance for all the classes except one. A similar behaviour was observed with the Weizmann dataset, although in this case the improvement was not so significant mainly due to the limited size and intra-class variability of the dataset (see Table 3.1).

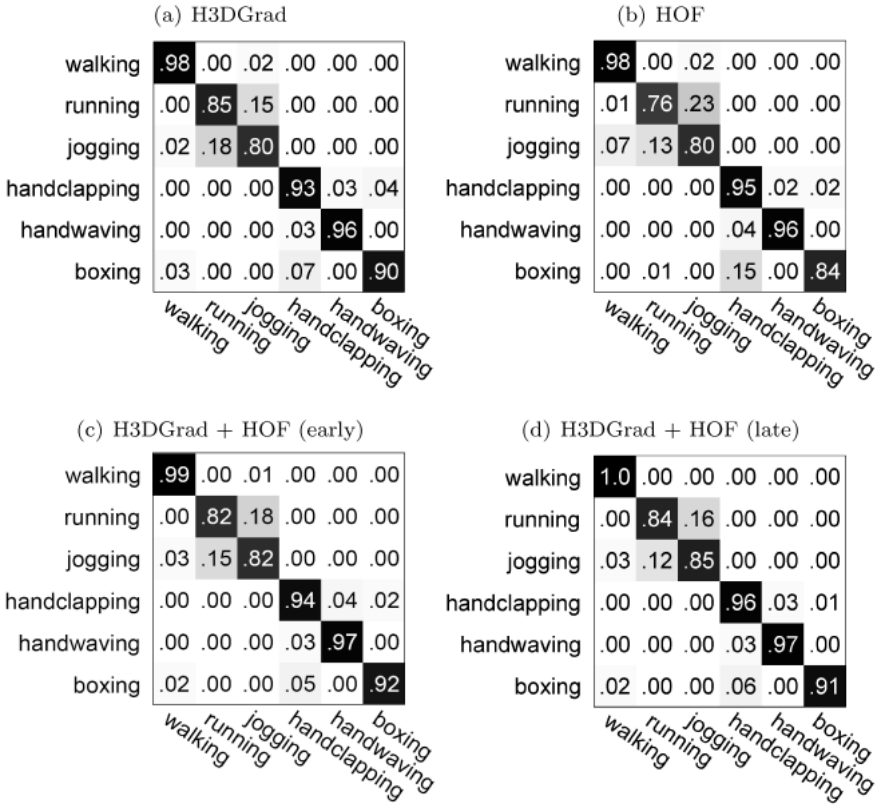


Figure 3.4: Classification accuracy on the KTH dataset using k-means clustering, hard assignment and different descriptors combination strategies (i.e. early or late fusion).

3.3 Action Representation and Classification

In order to improve with respect to k-means and to account for the high variability of human actions in terms of appearance or motion we used radius-based clustering for codebook formation.

Figure 3.5 shows the codeword frequency of radius-based clustering and k-means with hard quantisation on the KTH dataset. It is interesting to note that with k-means most of the codewords have similar probability of occurrence, so making it difficult to identify a set of words that have at the

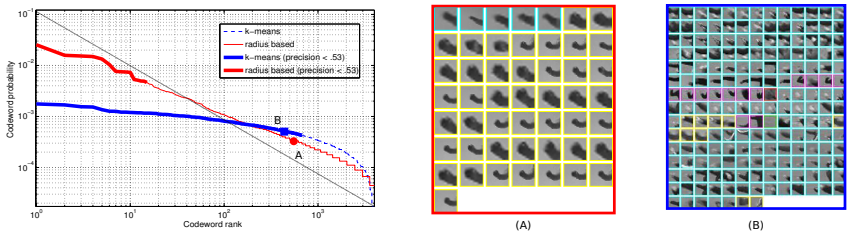


Figure 3.5: Log-log plots of codeword frequency using k-means and radius-based clustering with hard assignment. Bold lines indicate regions where the average cluster precision [103] is below 0.53. The dotted diagonal line represents the Zipfian distribution. Two sample clusters are shown at near frequencies, respectively obtained with radius-based clustering (A) (most of the features in the cluster represent spatio-temporal patches of the same action) and with k-means (B) (features in the cluster represent patches of several actions). Patches of actions have different colors: *boxing* (cyan), *hand-waving* (magenta), *hand-clapping* (yellow), *running* (green), *walking* (red), *jogging* (blue).

same time high discrimination capability and good probability of occurrence. In contrast radius-based shows a much less uniform frequency distribution. Interestingly, with radius-based clustering, the codeword distribution of the human action vocabulary is similar to the Zipf's law for textual corpora. It seems therefore reasonable to assume that codewords at intermediate frequencies are the most informative also for human action classification, and the best candidates for the formation of the codebook.

Due to the high dimensionality of the descriptor, codebooks for human actions usually have cluster centres that are spread in the feature space, so that two or more codewords are equally relevant for a feature point (codeword *uncertainty*); moreover cluster centres are often too far from feature points so that they are not any more representative (codeword *plausibility*). With radius-based clustering, codeword *uncertainty* is critical because it frequently happens that feature points are close to the codewords boundaries [152]. Instead, codeword *plausibility* is naturally relaxed due to the fact that clusters are more uniformly distributed in the feature space. To reduce the *uncertainty* in codeword assignment, we therefore performed radius-based clustering with soft assignment by Gaussian kernel density estimation

smoothing. In this case, the histogram H is computed as:

$$H(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_\sigma(w, f_i)}{\sum_{j=1}^{|V|} K_\sigma(v_j, f_i)} \tag{3.11}$$

where K_σ is the Gaussian kernel: $K_\sigma(\cdot,\cdot) = \frac{1}{\sqrt{2\pi}\sigma} e^{(-\frac{d(\cdot,\cdot)^2}{2\sigma^2})}$ being σ the scale parameter tuned on the training set, and $d(\cdot,\cdot)$ is the Euclidean distance.

Figure 3.6 compares the classification accuracy with codebooks obtained with k-means clustering with both hard and soft assignment, and radius-based clustering with soft assignment, respectively for the KTH and Weizmann dataset. The plots have been obtained by progressively adding less frequent codewords to the codebooks (respectively up to 4000 and 1000 codewords for the two datasets). The performance of k-means is improved by the use of soft assignment. With a small number of words radius-based clustering with soft assignment has lower performance than k-means due to the fact that the codewords used have higher frequency than those used by k-means (see Figure 3.5). As the number of codewords in the codebook increases, radius-based clustering outperforms k-means, whether with hard or soft assignment. This reflects the fact that in this case radius-based clustering permits to have also sparse regions being represented in the codebook. Besides, soft assignment helps to reduce *uncertainty* in the dense regions. Figure 3.7 shows the confusion matrix for different human actions on KTH and Weizmann datasets with radius-based soft assignment. The average accuracy is respectively 92.66% and 95.41% for the two datasets.

3.4 Person tracking and data association

When multiple persons, possibly not interacting and performing different and separate actions, there is need of a segmentation procedure to map space-time interest points detected in the video to each subject. Person tracking is used to assign the detected spatio-temporal interest points to each person present in a video, to localise both in space and time each recognised action. The tracker adopted in our system implements a particle filter based tracking algorithm, presented by [5], that tracks position, size and speed of the target, describing the target appearance with its colour histogram (using hue and saturation channels). The tracker is initiated using the human detector of [36], implemented in OpenCV. The detector is run

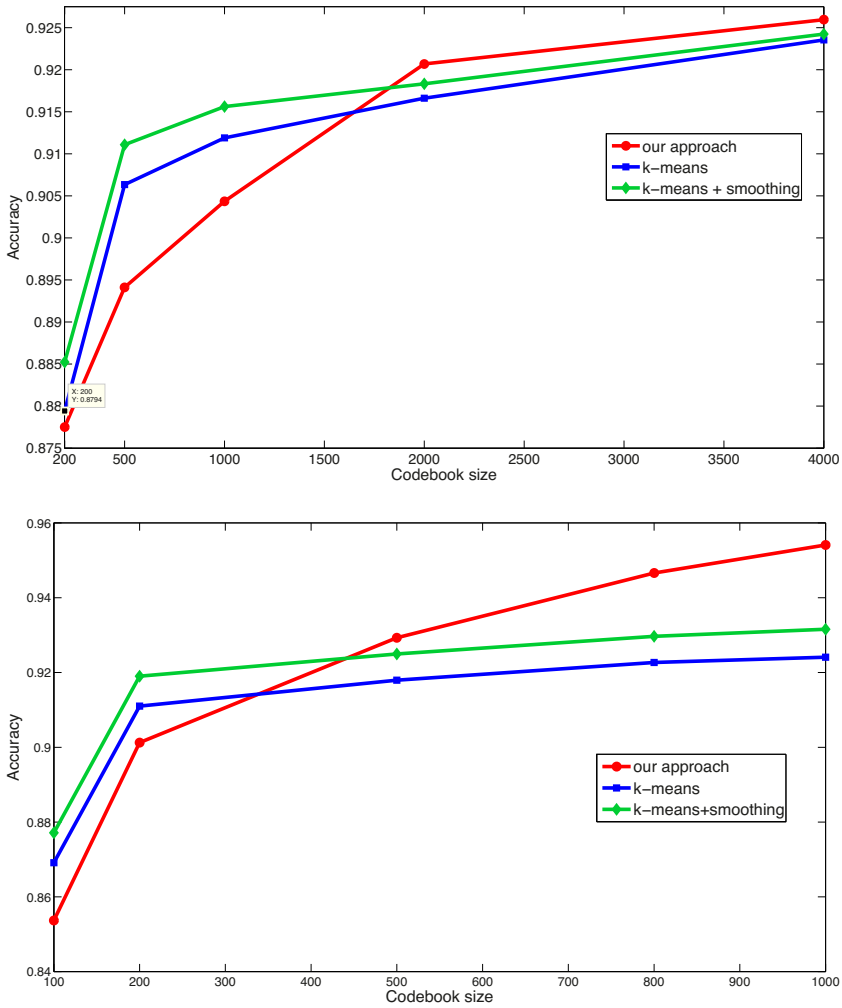


Figure 3.6: Classification accuracy on KTH (top) and Weizmann (bottom) datasets with codebooks created with k-means with hard assignment, k-means with soft assignment and radius-based with soft assignment.

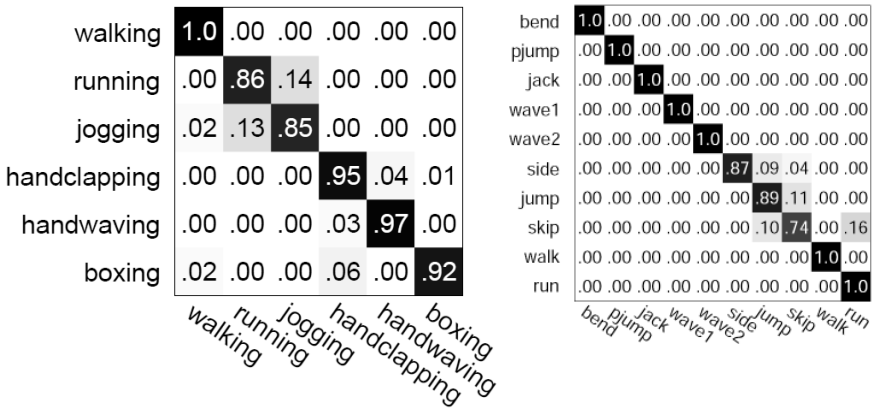


Figure 3.7: Classification accuracy on KTH (left) and Weizmann (right) datasets using radius-based clustering with soft assignment.

frame-wise to obtain both new targets to follow and measures for existing tracks. Measures obtained from the people detector are associated to targets by solving a data association problem, using a fast greedy algorithm that has a much lower complexity than the optimal solution obtainable with the Hungarian algorithm [166]. This greedy algorithm can be executed in real-time, as needed in video-surveillance applications, and works as follows: a matrix M that contains all the matching scores $m_{i,j}$ between the i_{th} target and the j_{th} measure of the person detector is computed. The matching score is computed as:

$$m_{i,j} = e^{-\frac{d_{i,j}^2}{D}} \tag{3.12}$$

where $d_{i,j}$ is the Euclidean distance between the static part of the model (position and size) of the target and the position and size of the detected person (represented using top-left and bottom-right coordinates of the bounding boxes) and D is adaptively chosen based on the target size.

The maximum $m_{i,j}$ are iteratively selected, and the i rows and j columns belonging to target and detector in M are deleted. This is repeated until no further valid $m_{i,j}$ is available. Two approaches are followed to avoid the erroneous association of a detection to a target: *i)* only the associated detections with a matching score $m_{i,j}$ above a threshold are used, to avoid that a detection that is far from a target is matched; *ii)* if a detection overlaps

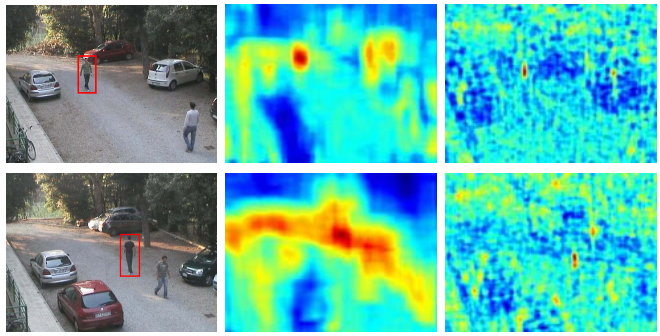


Figure 3.8: Original frame, hue/saturation histogram and person detector generated likelihood computed for the farthest target (highlighted with red bounding box). In this example the pedestrian detector is run at a single scale; histogram likelihood is generated using the values of the Batthacharya distance between the template histogram and a corresponding (same scale and aspect ratio) window. In both cases scale and aspect ratio variations are not considered, for the sake of visualisation.

more than one target no association is performed. If a detection is not associated to any target and does not overlap any existing target then it is used to start a new track.

The template of the target appearance is updated every time a new detection is associated to the track. In this way we prevent template drift and we allow the color histogram to adapt with respect to illumination changes and maneuvers which can change the target appearance. The state update equation, defined over the 8-dimensional state vector x_k (composed by 4 components for position and size and 4 components for their velocities), realises a 1^{st}-order dynamic model:

$$x_k = Ax_{k-1} + v_{k-1}, \quad A = \begin{bmatrix} I_4 & I_4\Delta t \\ 0 & I_4 \end{bmatrix}, \tag{3.13}$$

where I_4 is an 4×4 identity matrix, Δt is the time step and v_{k-1} is an additive, zero mean, isotropic Gaussian uncertainty term that represents the uncertainty in the state update. This uncertainty is parametrised in terms of the standard deviation on each component of the state vector. The measurement model exploits the results of the person detector whenever they are available.

The person detector likelihood is strongly peaked in presence of a target, as shown in the third column of Figure 3.8. This behavior allows to detect as distinct objects even very close pedestrians, but is not suitable to use it as likelihood of the target [3] since in particle weight computation it could assign very high weights to a few or no particles, and almost uniform low weights to the remaining population, leading thus to a degeneracy problem. To deal with this issue the target model of the particle filter is based on the color histogram of the tracked object, aiming at robustness against non-rigidity, rotation and partial occlusion [117]; after updating the template histogram with the new measure histogram, weights are computed according to the Batthacharya distance between the particle and the template histograms. On the other hand the color histogram is too weak to be used as an aspect model in a real-world video-surveillance scenario and should not be used as a sole measurement provider, as shown in the second column of Figure 3.8; this is due to background pixels contaminating the template and the lack of discriminativity of the histogram caused also by its subsampling (we used eight hue bins and eight saturation bins, to reduce sensitivity to light conditions).

To improve the particle filter capability to effectively track the target, even if its appearance is not strongly characterised, the tracking method implements a particular technique, based on the use of the similarity of the current estimate with the original target histogram as an index of tracking quality, to manage the uncertainty in the state update equation by means of on-line adaptation of the error v_{k-1}. In particular, let us consider the case where the variances of position and size of the target are set to very high values. In this case the filter samples over a wide enough area to maximise the possibility of capturing the target in case of erratic changes in direction or velocity. The pitfall in this strategy, however, is that it also increases the likelihood that the particle filter will become distracted by spurious similar patches in the background. Considering also the variances of the velocities the problem is even worse: from equation 3.4, in the update equation for propagating a particle from time $k-1$ to k, the uncertainty in the dynamic component is propagated to the static component. To reduce this effect a *blindness* value is computed by passing the similarity of estimate and original target histogram through a sigmoid; this *blindness* value is used to adjust the variances in such a way that the noise in the static component of the state observations is never amplified by the noise in the dynamic components.

This allows the tracker to switch between two different behaviors: one that relies on the predicted motion of the target and one that behaves like a random-walk model.

Figure 3.9: Example of multiple person tracking, spatio-temporal interest point detection and their association to the tracks.

3.4.1 Action classification and track annotation

By mapping the features associated to each tracked person in a video to the vocabulary, we can represent it by the frequency histogram of visual words. In order to reduce outliers, histograms of tracks that contain too few interest points, are discarded. Then, the remaining histograms are fed to a classifier to predict the action category.

3.5 Experimental Results

We have assessed our approach for categorisation of human actions in different conditions. Particularly, it has been tested on the KTH and Weizmann datasets that show staged actions performed by an individual in a constrained non-cluttered environment. Moreover, in order to have a more complete assessment of the performance of the proposed solution even in real world scenes with high variability and unconstrained videos, we also carried out experiments on the Hollywood2 and MICC-UNIFI Surveillance datasets. This latter, made publicly available at `http://www.openvisor.org` [155], includes real world video surveillance sequences containing actions performed by individuals with cluttering and varying filming conditions. Experiments were performed using non-linear SVMs with the χ^2 kernel [179].

(a) Walking

(b) Running

(c) Pickup object

(d) Enter car

(e) Enter car (from a different view point)

(f) Exit car

(g) Handshake

(h) Give object

Figure 3.10: Sample frames of sequences from the MICC-UNIFI Surveillance dataset.

3.5.1 Experiments on KTH and Weizmann datasets

The KTH dataset, currently the most common dataset used for the evaluations of action recognition methods [158], contains 2391 short video sequences showing six basic actions: *walking, running, jogging, hand-clapping, hand-waving, boxing*. They are performed by 25 actors under four different scenarios with illumination, appearance and scale changes. They have been filmed with a hand-held camera at 160×120 pixel resolution. The Weizmann dataset contains 93 short video sequences showing nine different persons, each performing ten actions: *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*. They have been filmed with a fixed camera, at 180×144 pixel resolution, under the same lighting condition.

Table 3.2 reports the average accuracy of our method in comparison with the most notable research results published in the literature. The performance figures reported are those published in their respective papers. For a fair comparison, our experiments have been performed with the setup suggested by the creators of the KTH and Weizmann datasets [48, 130], that has been used in [47, 71, 80, 92, 126, 130, 131, 145, 158, 162, 176]. In particular, with the KTH dataset, SVM classifiers have been trained on sequences of 16 actors and performance was evaluated for the sequences of the remaining 9 actors according to 5-fold cross-validation. With the Weizmann dataset SVM classifiers have been trained on the videos of 8 actors and tested on the one remaining, following leave-one-out cross-validation.

While showing the best performance, our solution has also the nice property that it does not require any adaptation to the context under observation. Instead other solutions require some tuning of the descriptor to the specific context. Namely, Laptev *et al.* [80] perform different spatio-temporal sampling of video frames and define a set of descriptors; hence they represent each action with the best combination of sampling and descriptors; Kläser *et al.* [71] use a parametrised 3D gradient descriptor; parameter values are optimised for the dataset used; Liu *et al.* [90] use both local and global descriptors and select the best combination of them according to an optimisation procedure; Scovanner *et al.* [131] optimise the codebook by associating co-occurrent visual words.

Other researchers have claimed higher performance on the KTH datasets: 94.2% Bregonzio *et al.* [21]; 93.2% Liu and Shah [92]; 93.43% Lin *et al.* [88]. However, these results were obtained with classifiers trained on larger sets

of data. Therefore, for the sake of fairness, they have not been included in Table 3.2. An exhaustive list of the different experimental setups and results has been recently published by Gao *et al.* [47].

Method	KTH	Weizmann	Features	Optimisations
Our method	**92.66**	**95.41**	H3DGrad + HOF	-
Yu *et al.* [176]	91.8	-	HoG + HOF	-
Wang *et al.* [158]	92.1	-	HOF	-
Gao *et al.* [47]	91.14	-	MoSIFT	-
Sun *et al.* [145]	89.8	90.3	2D SIFT + 3D SIFT + Zernike	-
Rapantzikos *et al.* [126]	88.3	-	PCA-Gradient	-
Laptev *et al.* [80]	91.8	-	HoG + HOF	codebook, sampling
Wong and Cipolla [164]	86.62	-	PCA-Gradient	-
Scovanner *et al.* [131]	-	82.6	3D SIFT	codebook
Liu *et al.* [92]	-	90.4	PCA-Gradient + Spin images	codebook -
Kläser *et al.* [71]	91.4	84.3	3D HoG	descriptor
Willems *et al.* [162]	84.26	-	3D SURF	-
Schüldt *et al.* [130]	71.7	-	ST-Jets	-

Table 3.2: Comparison of classification accuracy with some state-of-the-art methods on KTH and Weizmann datasets.

3.5.2 Tracker evaluation

We evaluate our tracking module quality by measuring multiple object tracking accuracy (MOTA) as defined by [11]. MOTA is an intuitive performance metric for multiple object trackers and measures a tracker performance at keeping accurate trajectories. For each frame processed a tracker should produce a set of object hypotheses, each of which should ideally correspond to a real visible object. In order to compute MOTA a consistent hypothesis-object mapping over time must be produced; the complete procedure to obtain this mapping is specified in detail in [11]. MOTA takes into account all possible errors that a multi-object tracker makes: false positives, missed objects and identity switches. False positives (fp) arise when, for example, the tracker is initiated on a false detection or when an object is missed and consequently a wrong pattern replaces the correct object hypothesis. Misses or false negatives (fn) arise whenever an object is not mapped to any of the hypotheses proposed by the tracker; finally identity switches (sw) happen

whenever an object hypothesis is mapped to the wrong object, for example after an occlusion or when an object tracker fails and another tracker is reinitialised. Errors are normalised by the number of objects present (gt) with respect to the whole sequence.

MOTA is defined as follows:

$$MOTA = 1 - \frac{\sum_t fp_t + fn_t + sw_t}{\sum_t gt_t} \tag{3.14}$$

We represent persons as bounding boxes and we consider a mapping correct if $\frac{O \cap H}{O \cup H} \geq 0.5$, where O and H are the areas of the object and the hypothesis bounding boxes mapped. We measured MOTA for all five sequences in which our final recognition experiments were performed and another sequence. The last sequence is recorded with a PTZ camera, panning tilting and zooming on targets and targets are instructed to produce overlapping trajectories in order to create difficult situations for a multiple object tracker. In the first five test sequences most of the errors are caused by false alarms of the pedestrian detector that cause instantiation of trackers; in the classification stage this empty tracks can be filtered since they usually do not contain enough detected space-time interest points. In the last sequence most of the errors are due to identity switches since target manoeuvres are more complex. MOTA is quite satisfying in all sequences, considering also that, in order to attain real-time performance, our appearance model is weak and no online classifier is used to perform data association or learn the template.

Sequence	FPR	FNR	SWITCH	MOTA
1	27.92	2.92	0	68.35
2	38.56	12.40	2	49.82
3	13.15	32.16	0	54,67
4	23.65	9.18	0	67.20
5	15.02	27.48	0	57.74
6	14.59	3.82	52	79.38

Table 3.3: Multiple object tracking accuracy(MOTA) together with false positive rate (FPR), false negative rate (FPR) and amount of identity switches (SWITCH).

Figure 3.11: Sample frames from a challenging sequence. First and second rows: the tracker is able to handle occlusions without losing the track or switching object identities. Third row: occlusion is correctly handled between yellow target and orange target but a false positive arise (red target) due to a false alarm of the pedestrian detector. Fourth row: after successfully handling the first occlusion the tracker lose the yellow target; a new track is initiated (magenta) afterwards, and correctly tracked until the end of the sequence.

3.5.3 Experiments on MICC-UNIFI Surveillance dataset

The MICC-UNIFI Surveillance dataset is composed by 175 real world video sequences of human actions with durations ranging from 3 to 20 seconds. The videos have been taken from wall mounted Sony SNC RZ30 cameras at 640×480 pixel resolution, in a parking lot. The scenes are captured from different viewpoints, at different degrees of zooming, with different shadowing and unpredictable occlusions, at different duration, speed and illumination conditions. Eight subjects perform seven everyday actions: *walking, running, pick-up object, enter car, exit car, handshake* and *give object*. A few examples are shown in Figure 3.10. We followed a repeated stratified random

sub-sampling validation, using 80% of the videos of each class as training set. Experiments were performed using a 2000 codeword codebook. The confusion matrix of classification accuracy is reported in Figure 3.12: the average accuracy is 86.28%. Most of the misclassifications observed with our method occurred with the *give object* and *handshake* actions. They are both characterised by a very fast motion pattern and small motion of the human limbs. Figure 3.13 reports sample sequences of these actions with evidence of details. In Table 3.4, we report a comparison of our method with other codebook creation approaches (k-means with hard and soft assignment) and with other state-of-the-art descriptors that publicly make their implementation available: MoSIFT[2] [47] and Dollár *et al.* [3] [38]. The results show that the proposed method outperforms the other approaches, and that the proposed codebook creation approach performs better than the typical k-means clustering whether with hard and soft assignment.

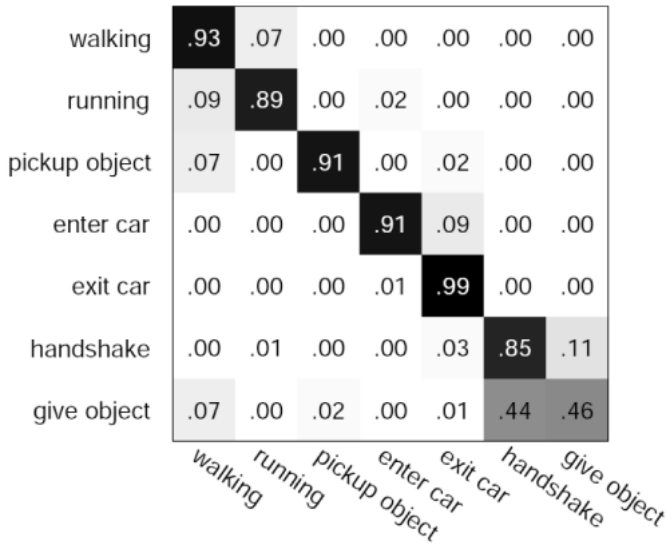


Figure 3.12: Classification accuracy on the MICC-Surveillance dataset using radius-based clustering with soft assignment.

[2] http://lastlaugh.inf.cs.cmu.edu/libscom/downloads.htm
[3] http://vision.ucsd.edu/~pdollar/research.html

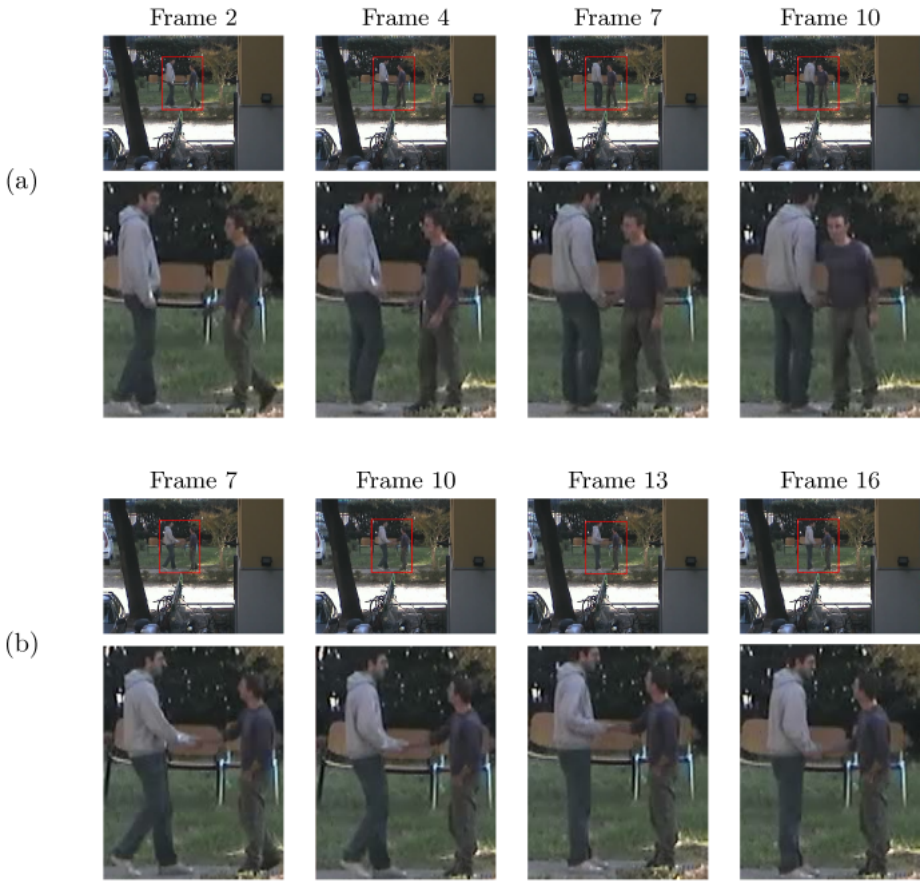


Figure 3.13: Sample frames of *give object* (a) and *handshake* (b) action sequences in the MICC-Surveillance dataset. For each sequence the second row shows the detail indicated in red in the first row.

Method	MICC-Surveillance
Our method	**86.28**
k-means + soft	83.74
k-means	82.90
Dollár's *et al.* [38]	72.50
MoSIFT [47]	75.88

Table 3.4: Comparison of classification accuracy on MICC-Surveillance dataset with our method, k-means with soft assignment, k-means with hard assignment, and with the descriptors proposed in [38] and [47].

3.5.4 Experiments on Hollywood2 dataset

The Hollywood2 dataset [99] is composed by sequences extracted from DVDs of 69 Hollywood movies, showing 12 different actions in realistic and challenging settings: *answer phone, drive car, eat, fight person, get out of car, handshake, hug person, kiss, run, sit down, sit up, stand up.* We performed our experiments with the same setup of [80, 158] using the "clean" training dataset, containing scenes that have been manually verified. This dataset is composed by 1707 sequences divided in training set (823) and test set (884), with different frame size and frame rate; train and test set videos have been selected from different movies. To be comparable with other experimental results the performance has been evaluated computing the average precision (AP) for each class and reporting also the mean AP over all classes. Codebooks have been created using 4000 codewords, as in [158]. We have compared our codebook creation approach with k-means clustering using both soft and hard assignments, and with an implementation of the method proposed in [80] using the provided descriptor and detector[4]. Results are reported in Table 3.5, showing that the proposed method outperforms the other approaches in the majority of action classes and in terms of mean AP.

3.5.5 Reducing the Codebook Size

Large codebooks, although being able to exploit the most informative codewords as illustrated in Figure 3.5, imply high time and space complexity. Reduction of codebook size with preservation of descriptive capability is therefore desirable. Linear dimensionality reduction techniques such as Principal

[4] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

Action	*k-means*	*k- means + soft*	*Our method*	Laptev *et al.* [80]
Answer phone	0.178	0.186	**0.195**	0.134
Drive car	0.864	**0.865**	0.863	0.861
Eat	0.552	0.564	0.564	**0.596**
Fight person	0.564	0.557	0.578	**0.643**
Get put of car	0.362	0.364	**0.362**	0.297
Handshake	0.142	0.143	0.167	**0.179**
Hug person	0.251	0.257	0.275	**0.345**
Kiss	0.494	**0.510**	0.503	0.467
Run	0.631	0.636	**0.659**	0.619
Sit down	0.483	0.493	**0.509**	0.505
Sit up	0.215	0.231	**0.227**	0.143
Stand up	0.511	0.513	**0.514**	0.485
mean AP	0.437	0.443	**0.451**	0.439

Table 3.5: Comparison of per-class AP performance on Hollywood2 dataset with codebooks created with our method, k-means with soft assignment, k-means with hard assignment and with the detector+descriptor proposed by Laptev *et al.* [80].

Component Analysis or Latent Semantic Analysis, are not suited to this end because they are not able to handle high order correlations between codewords that are present in human action representation [151]. We have therefore applied nonlinear dimensionality reduction with Deep Belief Networks (DBNs) [53,151]. A DBN is composed of several Restricted Boltzmann Machines (RBM) building blocks that encode levels of non-linear relationships of the input vectors. It is pre-trained by learning layers incrementally using contrastive divergence [28]. After pre-training, the auto-encoder is built by reversing the network and connecting the top layer of the network to the bottom layer of its reversed version. The auto-encoder is then used to fine-tune the network using a standard back-propagation algorithm.

Since the action representation $H(w)$ can be considered as a coarse probability density estimation of the features of a human action (see equation 3.11), given a set of space-time features $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, the value of the i-th bin of H can be considered as the probability that a space-time

descriptor $f \in \mathcal{F}$ is represented by the codeword w_i. This probability can hence be used as an input for an RBM according to [53].

Figure 3.14 reports plots of accuracy measured at different codebook sizes, with PCA, LSA and DBN codebook reduction and radius-based clustering with soft assignment, on the KTH dataset. Codebook reduction was applied to a 4000 codewords codebook. The dimension of the input layer is equal to the size of the uncompressed codebook and the dimension of the output layer is the compressed codebook size. Each hidden layer is one half the dimension of its input layer. The network depth ranges between five and eight depending on the size of the output codebook. The performance of our approach outperforms that of the method recently proposed in [74], especially for the smaller codebook sizes.

Figure 3.15 reports plots of mean computation times for a KTH video sequence as a function of codebook size for radius-based clustering with soft assignment. The accuracy values of Figure 3.14 have been reported on the plot for the sake of completeness. It can be noticed that strong codebook size reductions result into time improvements of more than two orders of magnitude. A compressed codebook with 100 codewords scores 89.57% recognition accuracy with respect to 92.66% of a 4000 codewords codebook.

Figure 3.16 shows that DBN-compressed codebooks on the one hand provide good accuracy even with very small codebook sizes, and on the other hand make radius-based clustering still competitive with respect to k-means clustering with 100 or less codewords.

Table 3.6 reports a comparison in terms of classification accuracy at different codebook sizes with DBN, PCA and LSA on the MICC-UNIFI surveillance dataset. Codebook reduction was applied to the 2000 codeword codebook obtained with radius-based clustering and soft assignment in the previous classification experiment. The smaller number of available training videos, with respect to KTH, is responsible for the reduction in classification accuracy, although the DBNs largely outperform the other methods. This experiment shows another advantage of the use of DBNs over PCA and LSA when the number of sequences available for training is relatively small, i.e. the possibility to create larger dictionaries that usually yield higher classification accuracy although maintaining a speed improvement of an order of magnitude. Table 3.7 reports a comparison of MAP performance obtained using compressed codebooks created with DBN, PCA and LSA on the Hol-

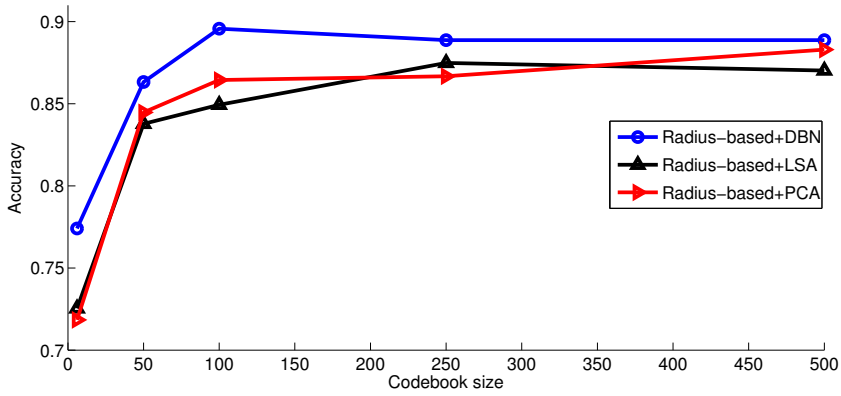


Figure 3.14: Classification accuracy on KTH dataset at different codebook sizes, with different codebook reduction techniques, for radius-based clustering with soft assignment.

lywood2 dataset. Codebook reduction was applied to the 4000 codeword codebook obtained with radius-based clustering and soft assignment used in the classification experiment. Despite the challenging dataset, the performance is still comparable with that obtained with full sized codebooks by several approaches reported in [158].

Codebook size	6	50	100	250	500
DBN	0.386	0.397	0.412	0.431	0.474
PCA	0.333	0.378	0.405	-	-
LSA	0.330	0.346	0.335	-	-

Table 3.6: Classification accuracy on MICC-UNIFI dataset at different codebook sizes, with different codebook reduction techniques, for radius-based clustering with soft assignment. Using PCA and LSA it is not possible to create codebooks larger than the number of training videos; using DBNs this issue is not present.

Codebook size	6	50	100	250	500
DBN	0.281	0.372	0.383	0.375	0.374
PCA	0.191	0.323	0.329	0.337	0.338
LSA	0.204	0.322	0.316	0.311	0.314

Table 3.7: Classification of MAP performance on Hollywood2 dataset at different codebook sizes, with different codebook reduction techniques, for radius-based clustering with soft assignment.

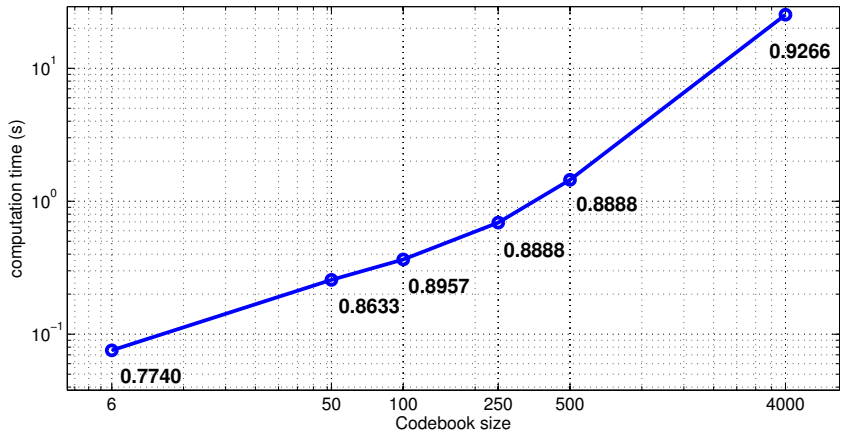


Figure 3.15: Mean computation times for a KTH video sequence at different codebook sizes with radius-based clustering and DBNs. The numbers associated to the markers indicate the classification accuracy.

3.6 Conclusions

In this chapter we have presented a novel method for human action categorisation that exploits a new descriptor for spatio-temporal interest points that combines appearance (3D gradient descriptor) and motion (optic flow descriptor), and effective codebook creation based on radius-based clustering and a soft assignment of feature descriptors to codewords. The approach was validated on KTH and Weizmann datasets, on the Hollywood2 dataset and on a new surveillance dataset that contain unconstrained video sequences that include more realistic and complex actions. Results outperform the state-of-the-art with no parameter tuning. We have also shown that a strong

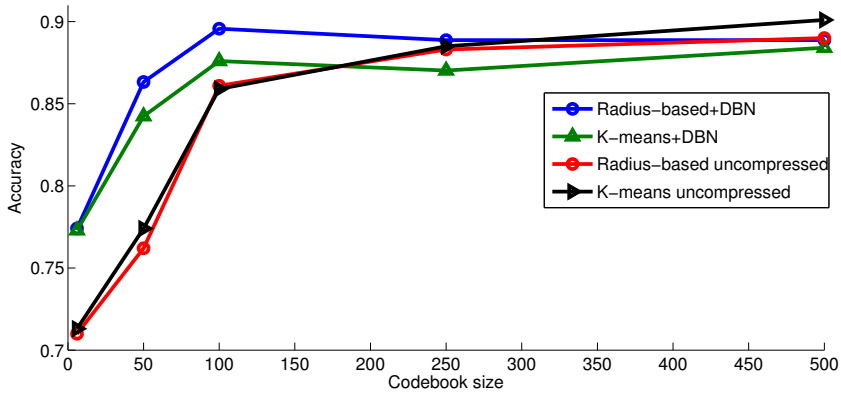


Figure 3.16: Classification accuracy as a function of codebook size, for DBN-compressed and uncompressed codebooks. Radius-based clustering with soft assignment is compared with k-means clustering with hard assignment.

reduction of computation time can be obtained by applying codebook size reduction with Deep Belief Networks, with small reduction of classification performance.

Chapter 4

Pyramid Kernel Descriptors Based on Space-time Zernike Moments

Local space-time descriptors are the main and most powerful tool for robust video representation and are the fundamental building block in event recognition algorithms. Space-time descriptors are usually carefully engineered in order to obtain feature invariance to photometric and geometric variations. The main drawback of these descriptors is high dimensionality and efficiency. In this chapter we propose a novel descriptor based on 3D Zernike moments computed for space-time patches. Moments are by construction not redundant and therefore optimal for compactness. Given the hierarchical structure of our descriptor we propose a novel similarity procedure that exploits this structure comparing features as pyramids. The approach is tested on a public dataset and compared with state-of-the art descriptors.[1]

[1]The work presented in this chapter has been published as "Space-time Zernike Moments and Pyramid Kernel Descriptors for Action Classification"" in *Proc. of International Conference on Image Analysis and Processing (ICIAP)*, [34]

4.1 Introduction

As shown in Chapter 3 one of the most powerful video representation for event recognition is obtained through the computation of a set of features describing local spatio-temporal regions. Indeed several techniques have been developed in the recent years mainly based on the use of local descriptions of the imagery. Following the success of SIFT [93] in object and scene recognition and classification [137], several space-time extensions of the local patch descriptors have been proposed. Similarly to local image features [104, 106] space-time features are localised through a detection step and then computed on the extracted patches; videos are represented as a collection of descriptors. Space-time descriptors represent the appearance and the motion of a local region and are engineered in order to retain invariance to geometric and photometric transformations. Laptev *et al.* [80] have defined a descriptor as a concatenation of histograms of oriented 2D gradients and histograms of optical flow. In order to reduce the computation burden an extension of SURF have been presented in [162]. Scovanner *et al.* [131] extended the SIFT to three-dimensional gradients normalizing 3D orientations bins by the respective solid angle in order to cope with the issue of the uneven quantisation of solid angles in a sphere. To solve this issue Kläser *et al.* [71] proposed to exploit 3D pixel gradients developing a technique based on Platonic solids. Finally Ballan *et al.* [7] developed an efficient descriptor decorrelating the spatial and temporal components and creating separated histograms of 3D gradient orientations. However all of these descriptors are extremely high-dimensional and often retain redundant information.

In the same time, researchers have exploited moments and invariant moments in pattern recognition [45]. Moments are scalar quantities used to characterise a function and to capture its significant features and they have been widely used for hundreds of years in statistics for description of the shape of probability density functions. Moments and in particular Zernike moments are a common choice in shape representation [86]. Zernike moments have been also proposed in action recognition as holistic features in [145] to describe the human silhouettes. Representations based on Zernike polynomials outperform other moment based descriptors in term of noise robustness, information redundancy and reconstruction error [146].

Despite the fact that feature matching is an important step in the recognition process few works have analysed it. Lowe [93] showed that in order to retrieve meaningful patches it is necessary to look at the distances of the

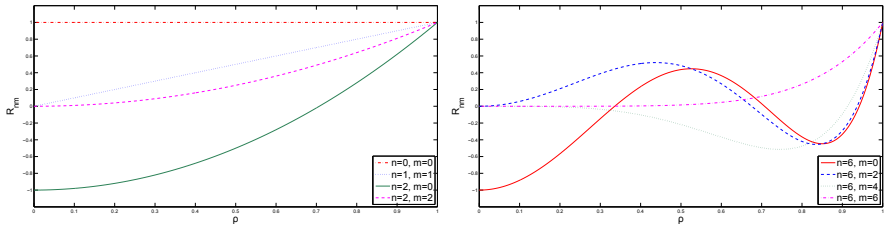
second nearest neighbour. More recently Bo *et al.* [16] provided a kernel view of the matching procedure between patches. Their work formulates the problem of similarity measurement between image patches as a definition of kernels between patches. Since these kernels are valid Mercer kernels it is straightforward to combine them or plug them into kernelised algorithms.

In this chapter we propose a new method for classification of human actions based on an extension of the Zernike moments to the spatio-temporal domain. Furthermore, we propose a kernel suitable for matching descriptors that can be hierarchically decomposed in order to obtain a multiple resolution representation. This kernel is inspired by multi-resolution matching of sets of features [49, 83], but instead of matching sets of features we match single space-time patches at multiple resolutions. To the best of our knowledge 3D Zernike moments have never been used as local space-time features and the pyramid matching scheme has never been used to define kernels between single features but only to match sets of features. Experimental results on KTH dataset shows that our system presents a low computational time maintaining comparable performance with respect to the state-of-the-art. The rest of the chapter is organised as follows. The generalisation of the Zernike moments to the three dimensions is presented in the next section. The Pyramid Kernel Descriptors are introduced in Section 4.3. The techniques for action representation and classification are presented in Section 4.4. Experimental results on the standard KTH dataset are discussed in Section 4.5. Finally, conclusions are drawn in Section 4.6

4.2 Space-time Zernike Moments

We first describe the formulation of the Zernike moments in two dimensions, and then introduce the generalisation to the space-temporal domain. Let $\mathbf{x} = [x_1, x_2]$ be the Cartesian coordinates in the real plane $\mathbb{R}^2$. Zernike polynomials are a set of orthogonal functions within the unit disk composed by a radial profile R_{nm} and a harmonic angular profile $H_m(\vartheta)$ defined as follows

$$V_{nm}(\rho, \vartheta) = R_{nm}(\rho) \cdot H_m(\vartheta) \tag{4.1}$$



(a) (b)

Figure 4.1: a) Radial profile up to the 2^{nd} order; b) Radial profile for the 6^{nd} order.

where $\rho = \sqrt{x_1^2 + x_2^2}$, $\vartheta = \tan^{-1}\left(\frac{x_2}{x_1}\right)$, $H_m(\vartheta) = e^{im\vartheta}$ and

$$R_{nm}(\rho) = \begin{cases} \sum_{s=0}^{(n-|m|)/2} \frac{(-1)^s (n-s)! \rho^{n-2s}}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} & \text{for } n-|m| \text{ even} \\ 0 & \text{for } n-|m| \text{ odd} \end{cases}. \quad (4.2)$$

The index n is named "order" and is a non-negative integer, and m is called "repetition" and it is an integer such that $n - |m|$ is even and non-negative. In Figure 4.1 some examples of the radial profile R_{nm} are shown. Both the Zernike polynomials and the radial profile $R_{nm}(\rho)$ satisfy the orthogonal condition

$$\int_0^{2\pi}\int_0^{1} V_{nm}^*(\rho,\vartheta) V_{n'm'}(\rho,\vartheta) \rho d\rho d\vartheta = \frac{\pi}{n+1}\delta_{nn'}\delta_{mm'} \quad (4.3)$$

and

$$\int_0^{1} R_{nm}(\rho) R_{n'm'}(\rho) \rho d\rho = \frac{1}{2(n+1)}\delta_{nn'}\delta_{mm'} \quad (4.4)$$

where δ indicates the Kronecker delta. Zernike polynomials are widely used to compute the Zernike moments [86, 110].

Let $f(\mathbf{x})$ be any continuous function, the Zernike moments are

$$A_{nm}(\mathbf{x_0}) = \frac{n+1}{\pi}\int\int_{\|\mathbf{x}-\mathbf{x_0}\|\leq 1} f(\mathbf{x}) V_{nm}^*(\mathbf{x}-\mathbf{x_0}) dx_1 dx_2 \quad (4.5)$$

where $\mathbf{x_0}$ denotes the point where the unit disk is centred. In this work we are interested in the computation of the Zernike moments for functions as

$f : \mathbb{R}^3 \mapsto \mathbb{R}$ where the third dimension is the time. To get the 3D Zernike polynomials [26, 115], the harmonic angular profile is substituted by the spherical harmonic functions

$$Y_m^l(\vartheta, \varphi) = N_m^l P_m^l(\cos\vartheta)\, e^{il\varphi} \tag{4.6}$$

where P_m^l denotes the Legendre function and N_m^l is a normalisation factor

$$N_m^l = \sqrt{\frac{2m+1}{4\pi}\frac{(m-l)!}{(m+l)!}}. \tag{4.7}$$

The spherical harmonic functions up to the 3^{rd} order are shown in Figure 4.2. In this case, given an order n, we use only the values of $m \geq 0$, and the index l is an integer such as $-m \leq l \leq m$. Then, the 3D Zernike polynomials are defined in spherical coordinates as follows

$$V_{nm}^l(\rho, \vartheta, \varphi) = R_{nm}(\rho) \cdot Y_m^l(\vartheta, \varphi) \tag{4.8}$$

and they satisfy the orthogonal condition within the unit sphere

$$\int_0^1 \int_0^{\pi} \int_0^{2\pi} \left[V_{nm}^l(\rho, \vartheta, \varphi)\right]^* V_{n'm'}^{l'}(\rho, \vartheta, \varphi) \sin(\vartheta)\, d\vartheta d\varphi d\rho = \delta_{nn'}\delta_{mm'}\delta^{ll'}. \tag{4.9}$$

Let $\boldsymbol{\xi} = [\mathbf{x}, t]$ be the generic point in the real plane $\mathbb{R}^2$ at the time t, the 3D Zernike moments are

$$A_{nm}^l(\boldsymbol{\xi_0}) = \frac{3}{4\pi} \int\limits_{\|\boldsymbol{\xi} - \boldsymbol{\xi_0} \leq 1\|} f(\boldsymbol{\xi}) \left[V_{nm}^l\left(\frac{\boldsymbol{\xi} - \boldsymbol{\xi_0}}{\boldsymbol{\sigma}}\right)\right]^* d\boldsymbol{\xi} \tag{4.10}$$

where $\boldsymbol{\xi_0}$ is the point where the unit sphere is centred, and $\boldsymbol{\sigma}$ tunes the size in pixel of the unit sphere for each coordinate. This is necessary because the patches, that we need to describe by using the 3D Zernike moments, can have different sizes in space and time. We use these space-time Zernike moments as descriptors for the local patches. The orthogonal condition (see equation 4.2) ensures that there is no redundant information in the descriptor, this allows to have a compact representation of the local features. Figure 4.4 show that we can obtain a rough but representative reconstruction of space-time cuboids from the 3D Zernike moments. In particular, we exploit the phase of these complex moments since from preliminary experiments proved to be more effective.

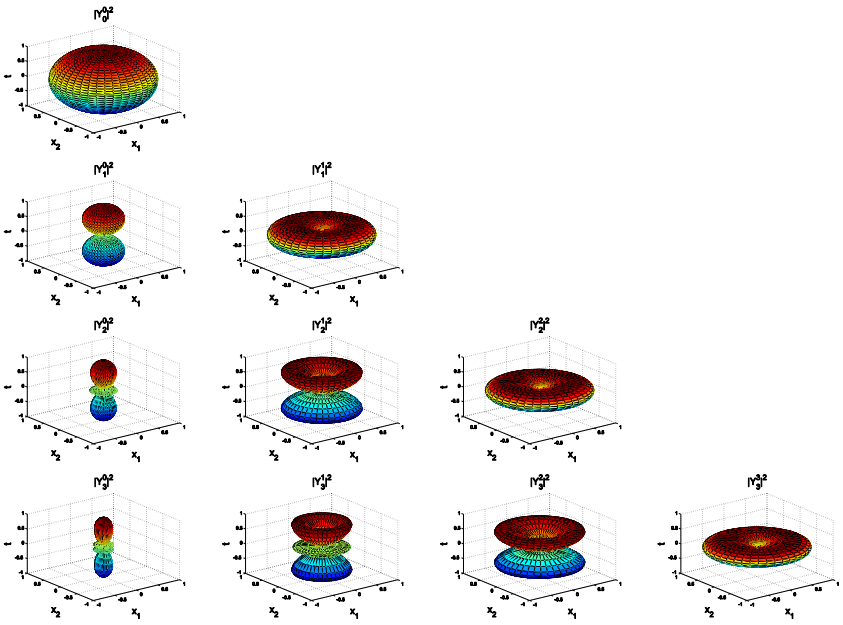


Figure 4.2: Spherical harmonic functions up to the 3^{rd} order.

4.3 Pyramid Kernel Descriptors

We introduce a descriptor matching kernel inspired by multi-resolution matching of sets of features [49, 83]; Grauman and Darrel [49] proposed the Pyramid Matching kernel to find an approximate correspondence between two sets of features points. Informally, their method takes a weighted sum of the number of matches that occur at each level of resolution, which are defined by placing a sequence of increasingly coarser grids over the features space. At any resolution, two feature points match if they fall into the same cell of the grid; number of matches computed at finer resolution are weighted more than those at coarser resolution. Later, Lazebnik *et al.* [83] introduced the Spatial Pyramid Matching kernel that work by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-regions.

Differently from these approaches our idea is to adapt the pyramid scheme for computing the similarity between two descriptor points. This allows to compute the similarity between two descriptors at multiple resolutions, exploiting a more distinctive representation when available and discarding it when at higher resolutions becomes noisy. We call our proposed approach "Pyramid Kernel Descriptors" because feature points are matched considering the descriptors as a multi-resolution set.

We consider a set of space-time interest points $X = \{\xi_1, \dots \xi_s\}$ and their descriptors $D = \{d_1, \dots, d_s\}$, where each descriptor can be organised in p sets $\{s^1, \dots, s^p\}$ hierarchically ordered. The pyramid kernel between d_i and d_j is defined as a weighted sum of the similarities of sets found at each level of the pyramid:

$$K(d_i, d_j) = \sum_{k=0}^{p} w_k k_c(s_i^k, s_j^k) \tag{4.11}$$

where w_k is the weight and $k_c(s_i^k, s_j^k)$ is a kernel to compute similarity between s_i^k and s_j^k. The similarity found at each level in the pyramid is weighted according to the description resolution: similarities made at a finer resolution, where features are most distinct, are weighted more than those found at a coarser level. Thus, if the p sets are arranged in ascending order the weight at level k can be defined as $w_k = 2^{k-p}$. In this case our proposed kernel is a valid Mercer kernel for the closure property of kernels since it is a weighted sum of valid kernels. As described in Section 4.2, our description based on space-time Zernike moments have a pyramid structure defined by the or-

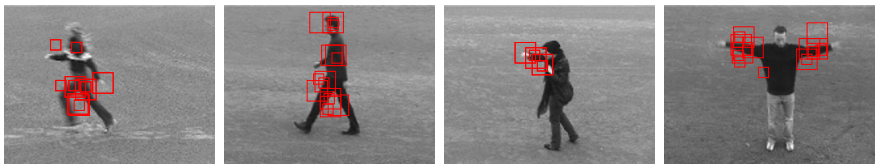


Figure 4.3: Examples of space-time interest points extracted at multiple scales for different actions. Clips are taken from the KTH dataset: running, walking, boxing and handwaving.

ders. In fact, lower order moments describe low frequencies of each cuboid while higher order moments encode higher frequencies. We define s^k as the concatenation of the phases of the complex Zernike moments for the first k orders: $s^k = \big(\arg(A^0_{00}), \ldots, \arg(A^l_{km})\big)$, where m and l are set according to Section 4.2. We use a normalised scalar product: $k_c(s^k_i, s^k_j) = \frac{s^k_i \cdot s^k_j}{\|s^k_i\| \|s^k_j\|}$, as a kernel between s^k_i and s^k_j, which is a valid Mercer kernel. Note that we normalise the scalar product computed at each level in order to have comparable values in the final sum.

For example, if we use a two level pyramid kernel descriptor $s_0 = \big(\arg(A^0_{00})\big)$ and $s_1 = \big(\arg(A^0_{00}), \arg(A^{-1}_{11}), \arg(A^0_{11}), \arg(A^1_{11})\big)$ and the corresponding weights $w_0 = 1, w_1 = \frac{1}{2}$. The final kernel between two space-time Zernike descriptors d_i, d_j computed up to the n^{th} order is:

$$K(d_i, d_j) = \sum_{k=0}^{n} 2^{k-n} \frac{s^k_i \cdot s^k_j}{\|s^k_i\| \|s^k_j\|}. \tag{4.12}$$

4.4 Action classification

We represent an action as a bag of space-time interest points detected by the adaptation described in Chapter 3 of the detector proposed by Dollár *et al.* [38].

Each point is described using Space-time Zernike moments and then a nearest-neighbor classifier based on the concept of instance-to-class similarity [19] is used for action categorisation. We chose not to employ descriptor codebooks (as in bag-of-words approaches) in order to better evaluate the effectiveness of our descriptor alone.

The instance-to-class nearest-neighbour classifier estimates the class pos-

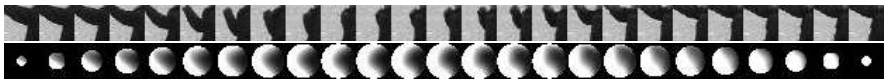


Figure 4.4: Frames of a cuboid (top). Reconstructed cuboid from complex 3D Zernike moments up to the 6^{th} order (bottom).

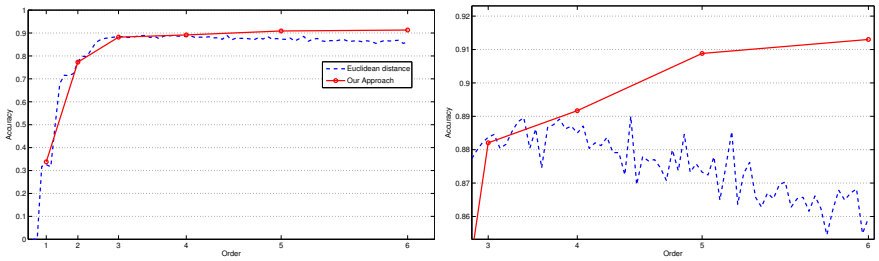


Figure 4.5: Comparison of the two similarity techniques; right) detail showing the effect of pyramid matching descriptors on high order moments.

terior probability given the query video clip with a non-parametric density estimation based on local Parzen windows centerd on descriptors belonging to the class. In [19] authors have shown that formulations based on more than one nearest neighbour per query descriptor do not significantly outperforms the simpler 1-NN formulation. Given this evidence, the implementation of this simple but effective classifier boils down to obtaining the most similar descriptor from the database for each feature extracted in a query clip (generally based on Euclidean distance between descriptors) and accumulating a vote for the class to which the database descriptor belongs to. The class with more votes is associated to the query clip. Instead of using Euclidean distance, we use our pyramid kernel descriptors (Section 4.3) to select the most similar descriptors which have, for each feature, the maximum kernel values.

4.5 Experimental Results

We tested our approach on the KTH action dataset. We used a leave-one-out procedure specifically we used the clips of 24 actors as a training set and the clips of the remaining actor as a test set. Performance is presented as the average accuracy of 25 runs, each with a different person. First we tested

our descriptor using the nearest-neighbour classifier based on the Euclidean distance and increasing the amount of moments (see Figure 4.5). With this approach the use of high order moments degrades the performance of the classifier. This is due to the fact that the high order filters response in small scale cuboids is mostly noisy. Then we used our pyramid similarity kernel increasing the levels of detail. As discussed in Section 4.3 levels with higher order moments are weighted more than levels with lower order moments. We can see that in this case we can exploit the higher details captured by high order moments without degrading the overall classifier performance. The confusion matrix reported in Figure 4.6 shows that as expected jogging

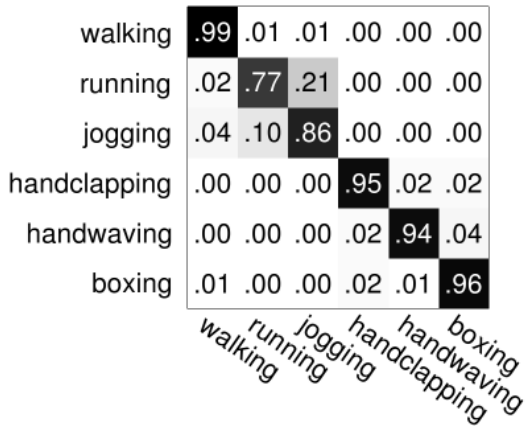


Figure 4.6: Confusion matrix for the KTH dataset.

and running are the most difficult actions to discriminate while for all other classes results are quite satisfying.

In Table 4.1 we compare our descriptor computation time, storage needs and accuracy on KTH dataset. Computation time is measured on our machine when the code was available while it is reported from the original publication if not. The accuracy is reported from the experiments reported in the original publication. We can see that Pyramid Zernike 3D descriptors are the smallest in terms of storage and are fast as other non-trivial implementations and C/C++ implementations; note that Gradient PCA requires only the projection of the concatenated pixel gradient values and that our descriptor is implemented without any optimisation in MATLAB.

Method	Size	Computation time	Accuracy
Pyramid Zernike 3D	84	0.0300 s	91.30%
Gradient + PCA [38]	100	0.0060 s	81.17%
3D SIFT [131]	640	0.8210 s	82.60%
Ext Grad LBP-TOP + PCA [100]	100	0.1000 s	91.25%
3DGrad [6]	432	0.0400 s	90.38%
HOG-HOF[2] [80]	162	0.0300 s	91.80%
HOG3D[2] [71]	380	0.0020 s	91.40%
SURF3D[2] [162]	384	0.0005 s	84.26%

Table 4.1: Descriptor complexity comparison together with accuracy.

4.6 Conclusions

In this chapter we have presented a method for action classification based on a new compact descriptor for spatio-temporal interest points. We introduce a new kernel suitable for matching descriptors that can be decomposed in multi-resolution sets. The approach was validated on the KTH dataset, showing results that have a low spatial and temporal computational complexity with comparable performance with the state-of-the-art. Our future work will deal with evaluation on more realistic datasets.

[2] c++ implementation

Chapter 5

Unsupervised event detection: anomaly detection

In this chapter we propose an approach for anomaly detection and localization, in video surveillance applications, based on spatio-temporal features that capture scene dynamic statistics together with appearance. Real-time anomaly detection is performed with an unsupervised approach using a non-parametric modelling, evaluating directly multi-scale local descriptor statistics. A method to update scene statistics is also proposed, to deal with the scene changes that typically occur in a real-world setting. The proposed approach has been tested on publicly available datasets, to evaluate anomaly detection and localisation, and outperforms other state-of-the-art real-time approaches. [1]

5.1 Introduction

The real-world surveillance systems currently deployed are primarily based on the performance of human operators that are expected to watch, often simultaneously, a large number of screens (up to 50 [147]) that show streams captured by different cameras. One of the main tasks of security personnel is to perform proactive surveillance to detect suspicious or unusual be-

[1] The work presented in this chapter has been published as "Multi-scale and real-time non-parametric approach for anomaly detection and localization" in *Computer Vision and Image Understanding (CVIU), 2012*, [12]

haviour and individuals [66] and to react appropriately. As the number of CCTV streams increases, the task of the operator becomes more and more difficult and tiring: after 20 minutes of work the attention of an operator degrades [50]. Operators usually take into account specific aspects of activity and human behaviour in order to predict possible perilous events [147], although often they can not explain their own criteria used to detect an unusual situation [66], or do not recognise unusual behaviours because they have not gathered enough knowledge of the environment and of the common behaviours they have to watch [143].

Video analytics techniques that automatically analyse video streams to warn, possibly in real-time, the operators that unusual activity is taking place, are receiving much attention from the scientific community in recent years. The detection of unusual events can be used also to guide other surveillance tasks such as human behavior and action recognition, target tracking, and person and car identification; in this latter case it is possible to use pan-tilt-zoom cameras to capture high resolution images of the subjects that caused the anomalous events.

Anomaly detection is the detection of patterns that are unusual with respect to an established normal behaviour in a given dataset, and is an important problem studied in several diverse fields [31]. Approaches to anomaly detection require the creation of a model of normal data, so to detect deviations from the model in the observed data. Three broad categories of anomaly detection techniques can be considered, depending on the approach used to learn the model: supervised [2,20,25,56,85,89,94,122], semi-supervised [136,178] or unsupervised [1,18,22,59–61,70,102,121,153,167,173]. In this work we follow an unsupervised approach, based on the consideration that anomalies are rare and differ amongst each other with unpredictable variations.

The model can be learnt off-line as in [20,25,56,76] or can be incrementally updated (as in [1,70,167,173]) to adapt itself to the changes that may occur over time in the context and appearance of a setting. Our approach continuously updates the model, to gather knowledge of common events and to deal with changes in "normal" behaviour, e.g. due to variations in lighting and scene setting.

Most of the methods for identifying unusual events in video sequences use trajectories [2,25,56,60,61,67,76,121,122,136,178] to represent the activities shown in a video. In these approaches objects and persons are tracked and

their motion is described by their spatial location. Blob features have been used in [29, 153, 167], without tracking the blobs. The main drawback of tracking-based approaches is the fact that only spatial deviations are considered anomalies, thus abnormal appearance or motion of a target that follows a "normal" track is not detected.

Optical flow has been used to model typical motion patterns in [1, 29, 70, 85, 102], but, as noted in [76], this measure also may become unreliable in presence of extremely crowded scenes; to solve this issue a dense local sampling of optical flow has been adopted in [1, 89]. Local spatio-temporal descriptors have been successfully proposed in [38, 78] to recognise human actions, while more simple descriptors based on spatio-temporal gradients have been used to model motion in [18, 76] for anomaly detection. Dynamic textures have been used to model multiple components of different appearance and dynamics in [60, 96].

Another issue that is common to both tracking and blob-based approaches is the fact that it is very difficult to cope with crowded scenes, where precise segmentation of a target is impossible. It is also important to consider that trajectory based methods rely on a long chain of algorithms (blob detection, data association, tracking, ground plane trajectory extraction) each of which may fail, leading to the failure of the whole anomaly detection system. Instead, approaches that are purely pixel-based, learning a scene representation independently of the explicit modelling of object motion, allow to skip the chain of intermediate decisions required by the sequence of algorithms, and detect an event directly from the representation of frames.

Some recent works consider the fact that, in some cases, an event can be regarded as anomalous if it happens in a specific context; for example the interaction of multiple objects may be an anomaly even if their individual behaviour, if considered separately, is normal. These works consider the scene [70, 76, 153], typically modelled with a grid of interest points, or the co-occurrence of behaviours and objects [60, 61, 94, 102] like persons and vehicles.

In this work we propose a multi-scale non-parametric approach that detects and localise anomalies, using dense local spatio-temporal features that model both appearance and motion of persons and objects. Real-time performance is achieved using a careful modelling of dense sampling of overlapping features. Using these features it is possible to cope with different types of anomalies and crowded scenes. The proposed approach addresses the

problem of high variability in unusual events and, using a model updating procedure, deals with scene changes that happen in real world settings. The spatial context of the spatio-temporal features is used to recognise contextual anomalies.

The rest of this chapter is structured as follows: scene representation, spatio-temporal descriptor and feature sampling are described in Section 5.2; in Section 5.3 is presented the real-time anomaly detection method, with multi-scale integration, context modeling and model updating procedure; finally experimental results, obtained using standard datasets are discussed in Section5.4. Conclusions are drawn in Section 5.5

5.2 Scene representation

Modeling crowd patterns is one of the most complex scenarios for detection of anomalies in video surveillance scenarios. Describing such statistics is extremely complex since, as stated in Section 5.1, the use of trajectories does not allow to capture all the possible anomalies that may occur, e.g. due to variations of scene appearance and the presence of unknown objects moving in the scene; this is due to the fact that object detection and tracking are often unfeasible both for computational issues and for occlusions. On the other hand, global crowd descriptors are not able to describe anomalous patterns which often occur locally (e.g. a cyclist or a person moving in an unusual direction among a crowd). The most suitable choice in this context is to observe and collect local space-time descriptors.

5.2.1 Feature sampling

Surveillance scenes are typically captured using low frame rate cameras or at a distance, leading to a short temporal extent of actions and movements (often just 5-10 frames). Therefore, it is necessary to sample these features densely in order to obtain as complete as possible coverage of the scene statistics. This approach is also motivated by the good performance obtained using dense sampling in object recognition [62] and human action recognition [159].

The solution adopted in the proposed method is to use spatio-temporal features that are densely sampled on a grid of cuboids that overlap in space and time. Figure 5.1 shows an example of spatial, temporal and spatio-

temporal overlaps of cuboids, and an example of application of overlapping spatio-temporal cuboids to a video. This approach permits localisation of an anomaly both in terms of position on the frame and in time, with a precision that depends on the size and overlap of cuboids; it also models the fact that certain parts of the scene are subject to different anomalies, illumination conditions, etc., and is well suited for the typical surveillance setup where a fixed camera is observing a scene over time. Considering the position of the cuboids on the grid it is also possible to evaluate the context of an anomaly, inspecting the nearby cuboids. Moreover, it makes it possible to reach real-time processing speed, since it does not require spatio-temporal interest point localisation. In our previous work [132] we have investigated how the overlap affects the performance of the system, and determined that a 50% spatial overlap provides the best performance, detecting more abnormal patterns without raising false positives, because spatial localisation of the anomaly is improved. On the other hand temporal overlap does not provide an improvement and, instead, may increase false detections.

5.2.2 Spatio-temporal descriptors

To compute the representation of each spatio-temporal volume extracted on the overlapping regular grid, we exploit the descriptor based on three-dimensional gradients computed using the luminance values of the pixels (Figure 5.1) described in Chapter 3.

It can be observed that if the overlap of cuboids precisely matches the subregions of nearby cuboids we can reuse the computations of these subregions for different cuboid descriptors (Figure 5.2). Using a number of spatial subregions that is a multiple of the overlap reduces the computational cost of the descriptors [150]: considering that a 50% overlap of cuboids is optimal then it is convenient to use an even number of spatial regions, since it is possible to reuse 50% or, depending on the position of the cuboid, 75% of the descriptors of nearby cuboids.

Therefore, we have divided the cuboid in 8 subregions, two along each spatial direction and two along the temporal direction. This choice increases the speed of the system of about 50%, with respect to a division of cuboids in $3 \times 3 \times 2$ regions [132].

This descriptor jointly represents motion and appearance, and it is robust to illumination and lighting changes, as required in a surveillance context in which a video might be recorded over a large extent of time. We do not

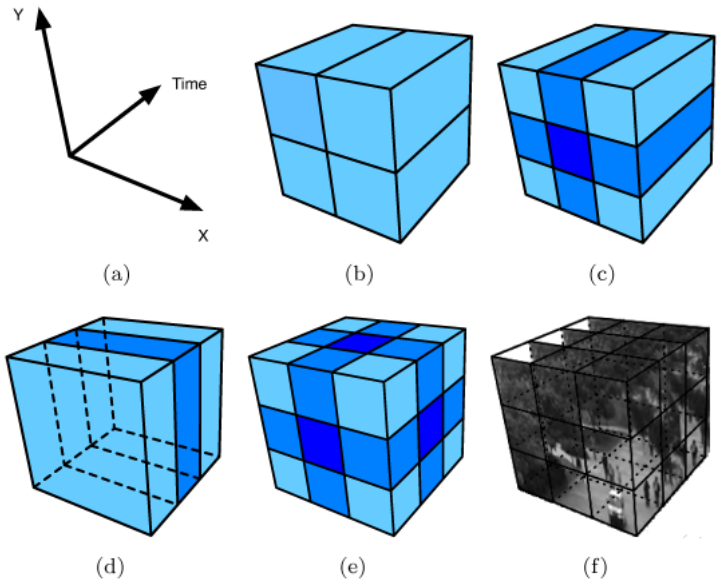


Figure 5.1: Examples of cuboids for spatio-temporal descriptors extraction, darker areas show the common parts due to the overlap of cuboids, if any. *(a)* spatial dimensions (X and Y) and temporal dimension (Time); *(b)* $2 \times 2 \times 1$ cuboids with no overlap; *(c)* $2 \times 2 \times 1$ cuboids with spatial overlap and no temporal overlap; *(d)* $1 \times 1 \times 2$ cuboid with temporal overlap and no spatial overlap; *(e)* $2 \times 2 \times 2$ cuboids with spatio-temporal overlap; *(f)* $2 \times 2 \times 2$ cuboids with spatio-temporal overlap, applied to a part of a frame of a surveillance video, to compute the spatio-temporal descriptors.

apply a re-orientation of the 3D neighborhood, since rotational invariance, otherwise useful in object detection and recognition tasks, is not desirable in a surveillance setting. The ϕ (with range $-\frac{\pi}{2}, \frac{\pi}{2}$) and θ ($-\pi$,π) are quantised in four and eight bins, respectively. The overall dimension of the descriptor is thus $2 \times 2 \times 2 \times (8 + 4) = 96$. Figure 5.3 shows three descriptors of cuboids containing a walking person, a cyclist and a moving cart. This construction of the three-dimensional histogram is inspired, in principle, by the approach proposed by Scovanner *et al.* [131], where they construct a weighted three-dimensional histogram normalised by the solid angle value (instead of separately quantizing the two orientations) to avoid distortions

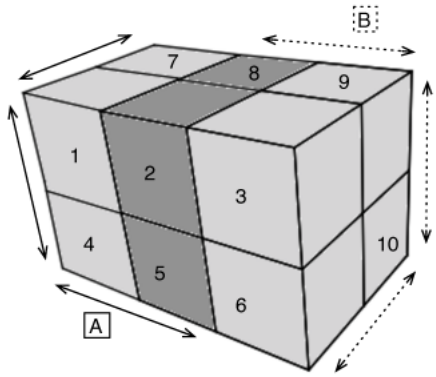


Figure 5.2: Example of two spatially overlapping cuboids: A and B. The subregions 2, 5, 8 (and another one below 8) are common to both cuboids, and their computation for cuboid B can skipped, once they have been computed for cuboid A.

due to the polar coordinate representation. However, we have found that our method is computationally less expensive, equally effective in describing motion information given by appearance variation, and shows an accuracy of human action recognition that is above or in line with other state-of-the-art descriptors [6], but without requiring tuning of descriptor parameters. In fact, we cannot afford any descriptor parameter learning since our setting is completely unsupervised.

5.3 Real-time anomaly detection

Our system is able to learn from a normal data distribution fed as a training set but can also start without any knowledge of the scene, learning and updating the "normal behaviour" profile dynamically, almost without any human intervention. The model can always be updated with a very simple procedure. Despite the simple formulation of this approach our system is able to model complex and crowded scenes, including both dynamic and static patterns.

Our technique is inspired by the one proposed in [22], where the proposed scene representation is global and static, based on global histograms of oriented gradients of single frames. Instead, in our approach, we use local

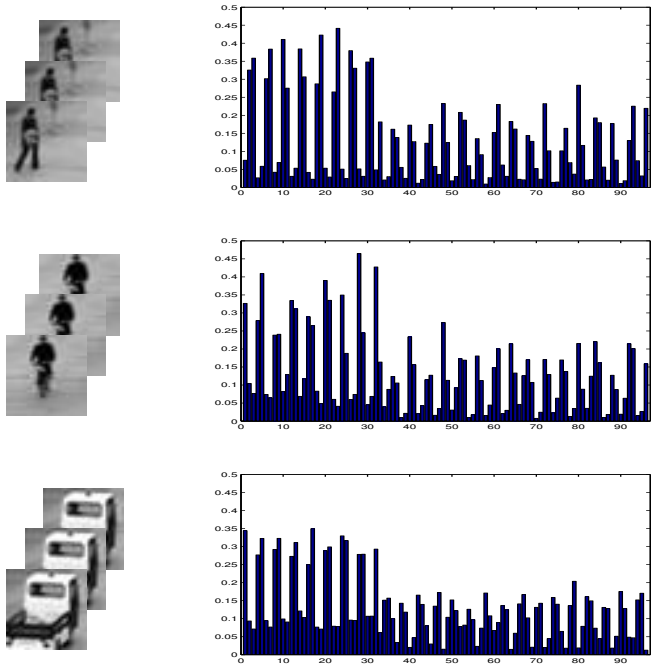


Figure 5.3: Example of three descriptors computed on cuboids containing a moving person, a cyclist and a moving cart

spatio-temporal features as a scene representation and we exploit the idea of the adaptive threshold in order to learn, over time, local models for different portions of the scene. Another significant difference with respect to [22] is the use of pure data instead of clusters. We do not perform clustering on data since we prefer not to corrupt data distribution in order to produce a more accurate estimation of the distance threshold used to detect anomalies. Also the model update procedure is different: since we are not applying any clustering procedure to data, our model update can be performed just by analyzing the detected anomalies stored over time; therefore it can be performed more frequently, without the need to operate either in detection mode or in maintenance mode.

As specified in Section 5.2 the use of local space-time gradients allows us to detect a wider range of anomalies while an appearance based method

restricts the anomalies that can be detected only to significant changes in a scene, e.g. a car parked in a wrong place, the presence of a fire truck or an unseen weather condition (rain, snow or fog).

5.3.1 Non-parametric model

In anomaly detection tasks a certain amount of normal data is usually available; our system can exploit this data as a training set to bootstrap itself and run in a semi-supervised fashion. Our system can also be run on-line with no previous knowledge of the scene, since a model update procedure is used. To jointly capture scene motion and appearance statistics we use the robust space-time descriptor, with dense sampling, described in Section 5.2. In order to decide if an event is anomalous we need a method to estimate normal descriptor statistics. Moreover, since no assumptions are made on the scene geometry or topology, it is important to define this normal descriptor distribution locally with respect to the frame.

Given a set of triples composed of descriptors d_q, their locations l_q and their scales s_q extracted from the past T frames, we would like to evaluate the likelihood of this data given the previously observed triples $\langle d, l, s \rangle$, i.e. $p\left(d_q, l_q, s_q | \mathbf{d}, \mathbf{l}, \mathbf{s}\right)$. The following assumptions are made: descriptors computed from neighbouring cells and from cells extracted at different scales are considered independent: this is a common Markovian assumption in low-level vision [46] that, even if may not hold for overlapping cells, allows to simplify the model and indeed proved to be effective, as reported in the experiments. We do not pose any prior on the locations, i.e. we do not consider any region of the frame more likely to generate anomalous descriptors. Since we consider a sequence of frames anomalous if at least a cell of the frame is considered as such, then the whole frame probability is obtained by marginalizing out the cell locations i. In the case of a single scale model we have:

$$p\left(d_q, l_q, s_q | \mathbf{d}, \mathbf{l}, \mathbf{s}\right) \propto \sum_i p\left(d_q^i, l_q^i | \mathbf{d^i}, \mathbf{l^i}\right). \tag{5.1}$$

For multi-scale models, we assume descriptors computed at different scales independent (even if overlapped), therefore we obtain:

$$p\left(d_q, l_q, s_q | \mathbf{d}, \mathbf{l}, \mathbf{s}\right) \propto \sum_i \prod_{j \in O^i} p\left(d_q^i, l_q^i, s_q^i | \mathbf{d^i}, \mathbf{l^i}, \mathbf{s^j}\right), \tag{5.2}$$

where O^i represent the set of patches overlapping region i.

To model the contextual anomalies, we need to compute the likelihood of a given descriptor with respect to its neighbouring observed cells; since we consider neighbouring models independent we obtain the following likelihood:

$$p\left(d_q, l_q, s_q | \mathbf{d}, \mathbf{l}, \mathbf{s}\right) \propto \sum_i \prod_{j \in O^i} \prod_{k \in N^{ij}} p\left(d_q^i, l_q^i, s_q^i | \mathbf{d^k}, \mathbf{l^k}, \mathbf{s^j}\right), \tag{5.3}$$

where N^{ij} represents the set of neighbouring locations at the same scale. The evaluation of probabilities in equation 5.1, 5.2, 5.3 are performed through non-parametric tests, as described in the following.

5.3.2 Implementation

Given a certain amount of training frames for each cell in our grid, space-time descriptors are collected and stored using a structure for fast nearest-neighbour search, providing local estimates of anomalies; an overview of this schema is shown in Figure 5.4. The training stage is very straightforward, since we do not use any parametric model to learn the local motion and appearance; instead we represent scene normality directly with descriptor instances.

A simple way to decide if an event happening at a certain time and location of the video stream should be considered anomalous, is to perform a range query on the training set data structure to look for neighbours. In this work we have used a fast approximate nearest-neighbour search over k-means trees, provided by the FLANN library [109]. A k-means tree is a hierarchical indexing data structure obtained by recursively splitting data. Once an optimal radius for each image location is learnt, all patterns for which the range query does not return any neighbour are considered anomalies. The problem with this technique is the intrinsic impossibility of selecting *a priori* a correct value for the radius. This happens for two reasons: firstly, each scene location undergoes different dynamics, for example a street will mostly contains fast unidirectional motion generated by cars and other vehicles, while a walkway will have less intense motion and more variations of the direction; moreover a static part of the scene, like the side of a parking lot, will mostly contain static information. Secondly, we want to be able to update our model dynamically by adding data which should be considered normal given the fact that we observed that kind of pattern for a sufficient

amount of time; therefore, since scene statistics must evolve over time, the optimal radius will evolve too. Finally, we also would like to select a value that encodes the system sensitivity, i.e. the probability that the observed pattern is not generated from the underlying scene descriptors distribution.

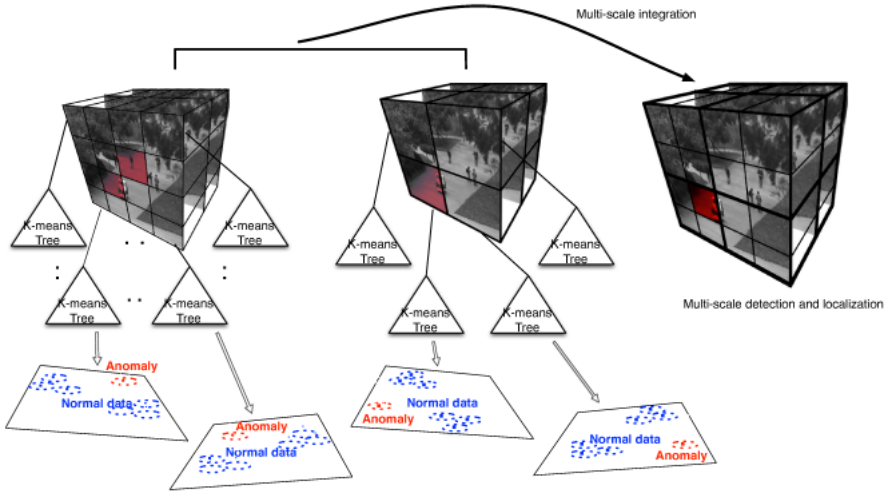


Figure 5.4: System overview. For each cell at each scale cuboids features are stored in efficient indexes based on k-means tree (fine on the left, coarse on the right). The planes underneath represent in a simplified view the high dimensional feature space. Anomaly detection may occur in different cells depending on the scale; the multi-scale integration mechanism reduces false alarms and provides a refined localisation of the anomaly (e.g. the cart on the walkway, see Figures 5.6 and 5.7).

To estimate the optimal radius for each data structure we exploit CDF_i^{-1}, the inverse of the empirical cumulative distribution of nearest-neighbour distances of all features in the structure of the cell i of the overlapping grid (Figure 5.5 shows an example for two grid cells). The estimate of the CDF of a random variable d for a value t is:

$$CDF(t) = \sum_{i=1}^{n} \mathbf{1}\{d_i \leq t\} \tag{5.4}$$

where $\mathbf{1}\{E\}$ is the indicator of event E and d_i are realisations of d. A practical and efficient procedure to directly estimate the inverse empirical

cumulative distribution CDF^{-1} of a set D of realisations of univariate random variables d_i, which share their density function, is the following: 1) sort d_i in ascending order, 2) remove duplicate values from the sorted list (usually needed for discrete variables) and store the sorted unique values in a vector D^{su}, 3) obtain $CDF^{-1}(p) = D^{su}_k$, where $k = \lfloor p \cdot |D^{su}| \rfloor$, p is a probability, $|D^{su}|$ is the cardinality of the set D^{su} and v_i denote the $i-th$ element of vector v.

Given a probability p_a below which we consider an event anomalous, we choose the radius $\hat{r}_i$ for cell i as:

$$\hat{r}_i = CDF_i^{-1}(1 - p_a)\,. \tag{5.5}$$

The anomaly probability p_a can be set to $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, \ldots$ depending on the user's need to obtain a more or less sensitive system. After setting such value p_a, optimal radii are estimated for each cell with likely different values. This optimal radius formulation allows easy data-driven parameter selection and model update.

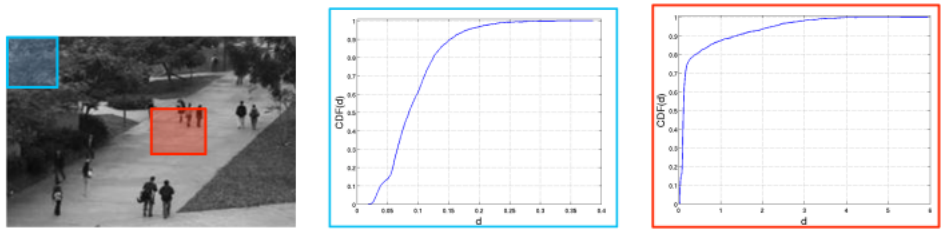


Figure 5.5: CDFs of different spatio-temporal cells: *left)* frame with highlighted positions of two cells, *center)* CDF of the blue (upper left) cell, *right)* CDF of the red (centered) cell.

5.3.3 Multi-scale integration

Anomalous events are generated by objects moving in different parts of the scene, therefore their scale can be subject to high variations due to the distance from the camera; moreover, we do not know the kind and the size of the objects that will generate anomalies. It is thus necessary to analyse video at multiple scales. In some initial experiments we observed that models based on smaller patches have a higher segmentation accuracy but suffer from false positives, while for models with bigger patches we observe

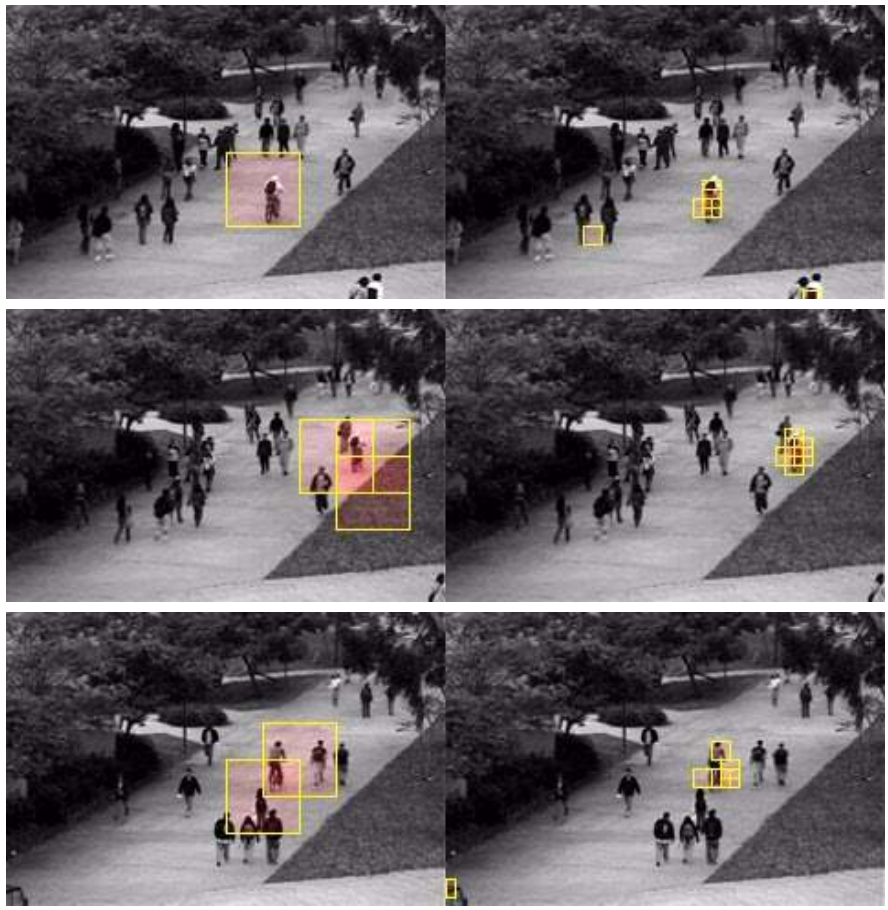


Figure 5.6: Single scale anomalies detections and localisations, before integration.

the opposite behaviour. We propose to improve our previous work [132] by exploiting a late fusion of the detection results of multiple models. This captures the abnormal patterns at different resolution. Since we aim at real-time performance, a dense patch sampling in scale is not computationally feasible; therefore, we limit the use of scales to two levels. Models are trained with patches of different size, with a factor of 4× difference. Anomaly detection is performed using the radius search with the optimal learnt distance and the

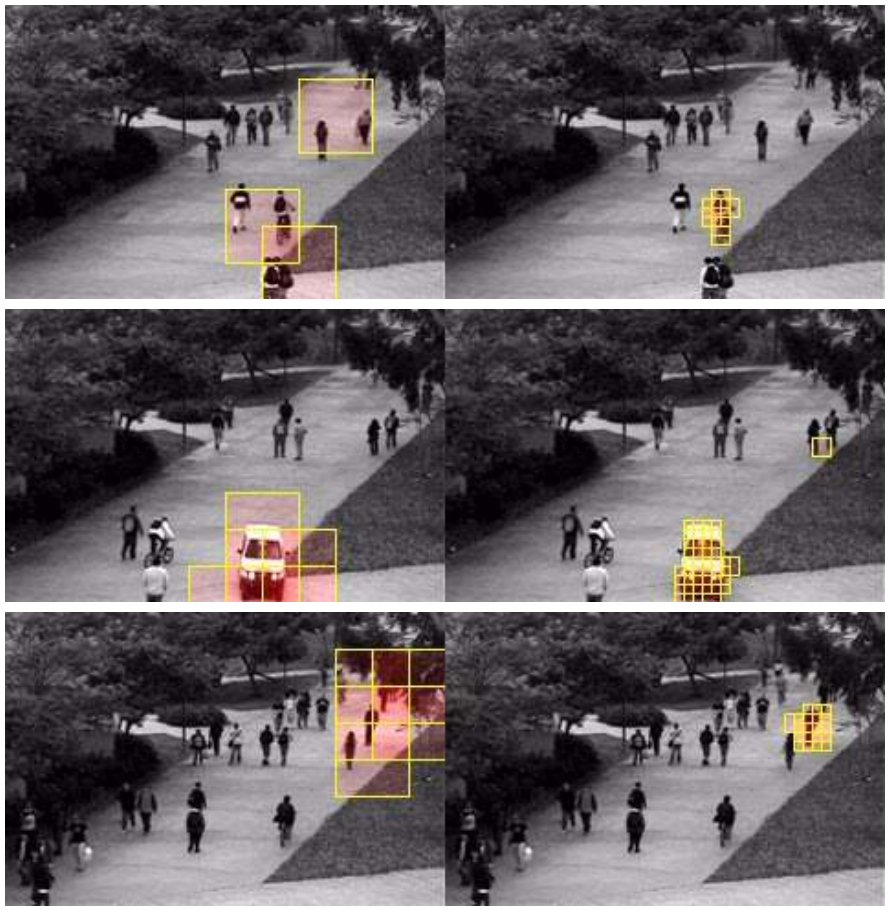


Figure 5.7: Single scale anomalies detections and localisations, before integration (continued).

final detection result is obtained from the intersection of all the detections. This allows the system to filter spurious small false positives and increases the capability of the system to accurately localise even smaller objects (i.e. pedestrians, cyclists). Moreover, space-time patterns that span more than one overlapping cell will be more likely considered anomalous, while a single isolated patch will be suppressed by the integration procedure. From a probabilistic point of view two likelihood maps are generated non-parametrically.

These maps represent how likely it is that a given space-time pattern it is an outlier for the observed statistic; the final likelihood map is generated via a product rule, resulting in the spatial intersection of the two detected areas, following equation 5.2. In our implementation, in order to keep the system executing in real-time, we used the following scales: 40×40 and 10×10, with a 50% overlap of 20 pixels and 5 pixels, respectively. Figures 5.6 and 5.7 shows different anomaly localisations at these scales.

5.3.4 Context modelling

A purely data-driven method, as the approach proposed in this chapter, can suffer from the lack of data in the case that statistics of patches from a region is too complex. This is a well known problem in all instance based methods like k-NN. To moderate this effect, we extend our model by considering the anomaly likelihood of a patch with respect to the observations of the nearby patches. Therefore we test the patch descriptor also against the models of the eight neighbouring cells. With this technique we increase the amount of data available for learning the local model of a part of the scene in a sensible way; in fact a patch that it is anomalous for a region but not for the neighbouring ones would not be considered as such, while patches that are outliers for all the neighbouring regions will be considered anomalies. The result of the detection is again obtained by product rule, therefore a patch is anomalous if and only if it is evaluated as such by all the models in its neighbourhood according to equation 5.3

5.3.5 Model update

Since applications for anomaly detection in video surveillance are designed to be executed for a long time, it is very likely that a scene will change its appearance over time; very simple examples are the event of a snowstorm, the cars that enter and exit a parking lot or the placement of temporary structures in a setting. It is therefore very important to provide a way to update our model. Again, we propose a very straightforward data-driven technique.

Together with the data-structure for each overlapping grid cell, we keep a list of anomalous patterns. We exploit the same range query approach presented in the previous subsection to look for normality in the abnormality list. This list is inspected on a regular basis, and new data is incorporated

by applying the following procedure. If an event happens very frequently it is likely that it will a have certain amount of neighbours in feature space, while truly anomalous event will still be outliers. After the estimation of an optimal radius for the anomalous pattern list, we discard all outliers in this list and incorporate all other data in the cell i training set. The optimal radius $\hat{r}_i$ for the updated cell is then recomputed.

Even if it is not required, since they can be used with default values, two parameters of the system can be tuned to adapt them to a particular scenario: grid density and overlap of cuboids. Reducing cuboid overlap can increase the detection performance, while using a more or less dense spatio-temporal grid can serve also as a system adaptation for a specific camera resolution or frame rate. These two parameters are directly bound to physical and technical system properties (e.g. camera resolution and computer processing speed) that the user can easily adjust to figure out a proper configuration. Instead, the system automatically computes the optimal radius parameter, that is a quantity that is extremely task, scene and time dependent.

5.4 Experimental results

We tested our approach on the UCSD[2] anomaly dataset presented in [96], which provides frame-by-frame local anomaly annotation. The dataset consists of two subsets, corresponding to different scenes using fixed cameras that overlook pedestrian walkways: one (called Peds1) contains videos of people moving towards and away from the camera, with some perspective distortion; the other (called Peds2) shows pedestrian movement parallel to the camera. Videos are recorded at 10 FPS with a resolution of 238×158 and 360×240, respectively. This dataset mostly contains sequences of pedestrians in walkways; annotated anomalies, that are not staged, are related to appearance and behaviour. In particular, they are non-pedestrian entities (cyclists, skaters, small carts) accessing the walkway and pedestrians moving in anomalous motion patterns or in non-walkway regions. The first subset contains 34 training video samples and 36 testing video samples, while the latter contains 16 training video samples and 12 testing video samples. Each sequence lasts around 200 frames, for a total dataset duration of ~ 33 minutes. 10 videos of the Peds1 subset have manually generated pixel-level binary masks, which identify the regions containing anomalies. We tested

[2] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

our approach on the whole UCSD dataset. Each anomalous frame in the testing set is annotated; for each cuboid classified as anomalous, we flag as anomalous each region of the frames from which it was created; frames that contain at least one anomalous region are considered anomalous. We follow the evaluation procedure of [96]: in the frame level evaluation an abnormal frame is considered correctly detected if at least one pixel of the frame is detected as anomalous; in the pixel level evaluation an abnormal frame is considered correctly detected only if at least the 40% of the anomalous pixels are detected correctly and considered a false positive otherwise. A "lucky guess" happens when a region different from the one that generated the anomaly is detected as anomalous in the same frame. The frame level detection evaluation does not takes into account this phenomenon. In our previous work [132] we evaluated the best parameters for dense sampling and overlapping of the spatio-temporal descriptors; the best results were obtained for cuboids of 40×40 pixels, with 8 frames of depth, a spatial overlap of 50% and no temporal overlap. In these experiments we used the same parameters.

We compare our system with results of other state-of-the-art approaches, as reported in [96]: MPCCA [70], Adam *et al.* [1], Mehran *et al.* [102] and Mahadevan *et al.* [96]. Results are reported using the ROC curve and the Equal Error Rate (EER) - that is the rate at which both false positives and misses are equal. Both multi-scale integration and contextual modeling help in lowering the EER, with respect to our previous work [132]. Our approach achieves a similar performance on both Peds1 and Peds2 datasets, showing the flexibility of the representation that is able to cope with diverse settings and types of anomalies. The other competing real-time approaches have a EER performance in the two datasets that varies between 6% to 11%; it can also be noted that these performance variations of the other systems are not uniform, thus there is no hint that a dataset is "more difficult" than the other. Figure 5.8, Figure 5.9 and Table 5.1 report the results for anomaly detection in Peds1 and Peds2. Figure 5.10 and Table 5.2 report results for anomaly localisation on Peds1.

Our approach, with the use of multiple scales and contextual queries, obtains the second best result both in temporal and spatial anomaly detection after the method proposed in [96], and is far superior to all the others in terms of spatial localisation and frame level localisation (except the close result of Social Force for Peds1). However, it has to be noted that the ap-

proach of [96] is not suitable for real-time processing since it takes 25 seconds to process a single frame on a computer with a computational power (3 GHz CPU with 2 GB of RAM) comparable to the one used in our experiments (2.6 GHz CPU with 3 GB of RAM). The good results in spatial anomaly localisation imply that we are not taking advantage of lucky guesses, but that we accurately localise the abnormal behaviours in space and time. Figure 5.11 shows a qualitative comparison of anomaly localisation of our approach with state-of-the-art off-line approach [96].

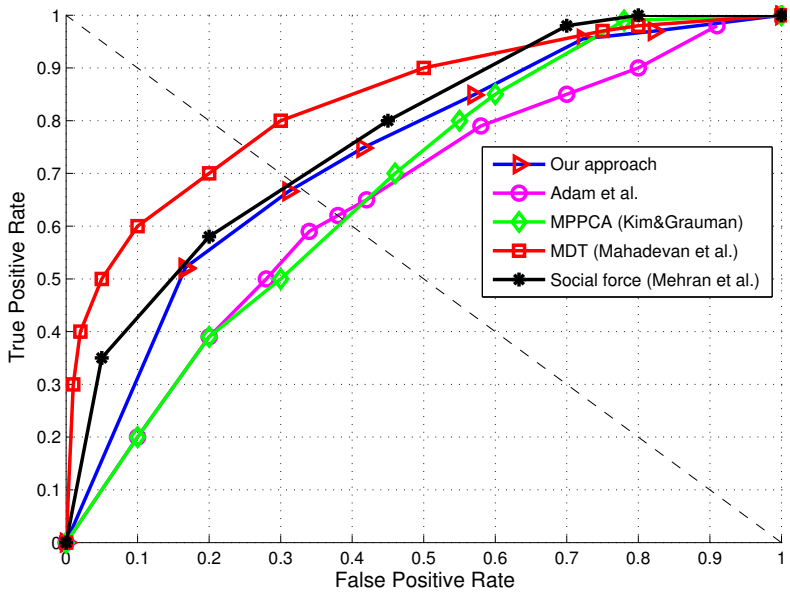


Figure 5.8: ROC curve to compare our method with state-of-the-art approaches on the Peds1 dataset. The dashed diagonal is the EER line.

Since our approach aims at real-time processing, we have evaluated the impact of the dense sampling of cuboids, computing the average number of processed frames per second while varying the spatial overlap of cuboids. The plot in Figure 5.12 shows how the steps of our method affect the performance. The use of multiple scales degrades the performance the most, almost halving the frame rate. The overhead of context modelling depends on the amount of features extracted, in particular it has little influence for the single scale algorithm but it strongly affects the multi-scale one since

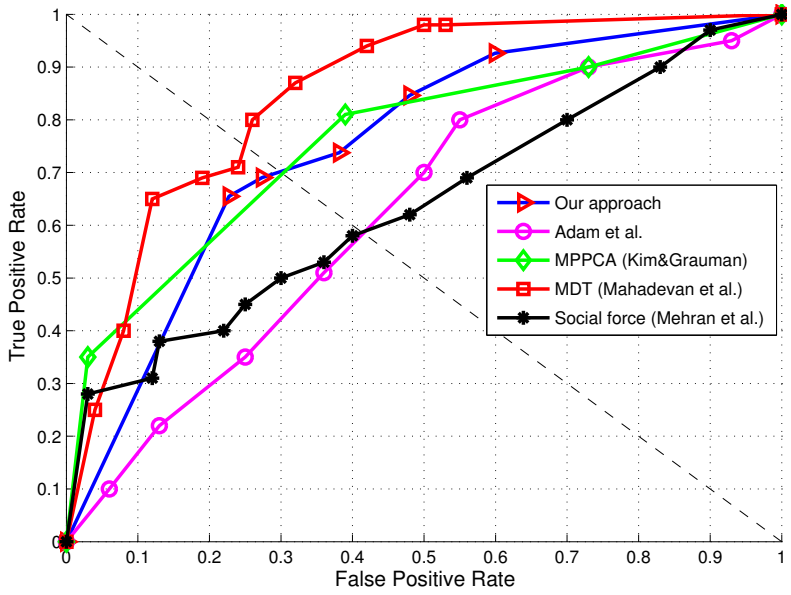


Figure 5.9: ROC curve to compare our method with state-of-the-art approaches on the Peds2 dataset. The dashed diagonal is the EER line.

the complexity of that step depends linearly on the amount of feature extracted. We also measured the anomaly detection overhead by computing the different frame rate in training (i.e. feature extraction and computation only) and testing and we found that for the single scale approach, without exploiting the context, it is only 3∼6% of the total computation time while using the context it increases to 11∼12% of the total computation time. For the multi-scale approach, the use of smaller patches (10×10) increases the burden of context modelling. Even with multiple scales and contextual neighbourhood queries our system is able to process 8 frames per second, with 50% patch overlap, and to obtain competitive results of detection and localisation with respect to non-realtime systems that require several seconds to process a single frame [96]. We expect that code optimisation exploiting modern multi-core CPUs will greatly reduce the computational gap between multi-scale and single-scale methods. Cuboid size does not affect the computation time since smaller cuboids imply an increased number of descriptors which are faster to compute while bigger cuboids generate fewer but slower

	UCSPed1	UCSPed2	Average
Single scale	34%	32%	33%
Multi-scale	32%	31%	32%
Context+Multi-scale	**31%**	**30%**	**30%**
MDT (Mahadevan *et al.* [96])	25%	25%	25%
MPPCA (Kim *et al.* [70])	40%	30%	35%
Social Force (Mehran *et al.* [102])	31%	42%	37%
Adam *et al.* [1]	38%	42%	40%

Table 5.1: Summary of quantitative system performance and comparison with state-of-the-art (lower values are better). EER is reported for frame level anomaly detection on Peds1 and Peds2 datasets together with the average over the two datasets.

Single scale	Multi-scale	**Context+Multi-scale**	MDT	MPPCA	SF	Adam
27%	28%	**29%**	45%	18%	21%	24%

Table 5.2: Detection rate on the anomaly localisation task (higher values are better).

to compute descriptors. The main reason for the decrease of computational performance when using the multi-scale approach is the increased number of model queries made when using smaller cuboids.

Since in video surveillance the precision of the alarms is important, because a human operator may be disturbed by a high number of false alarms, in Figure 5.13 we report the precision-recall curve for the UCSD dataset, created varying the p_a parameter from 10^{-5} to 10^{-2}, showing a good performance; considering low probabilities p_a for the anomalies the recall is reduced while raising the precision, and *vice versa*. In particular, the break-even point at 0.71 of precision and recall is obtained for $10^{-4} \leq p_a \leq 10^{-3}$ for Peds1 while for Peds2 the value of .87 is obtained for $10^{-5} \leq p_a \leq 10^{-4}$.

5.5 Conclusions

In this chapter we have presented a multi-scale non-parametric anomaly detection approach that can be executed in real-time in a completely unsupervised manner. The approach is capable of localizing anomalies in space

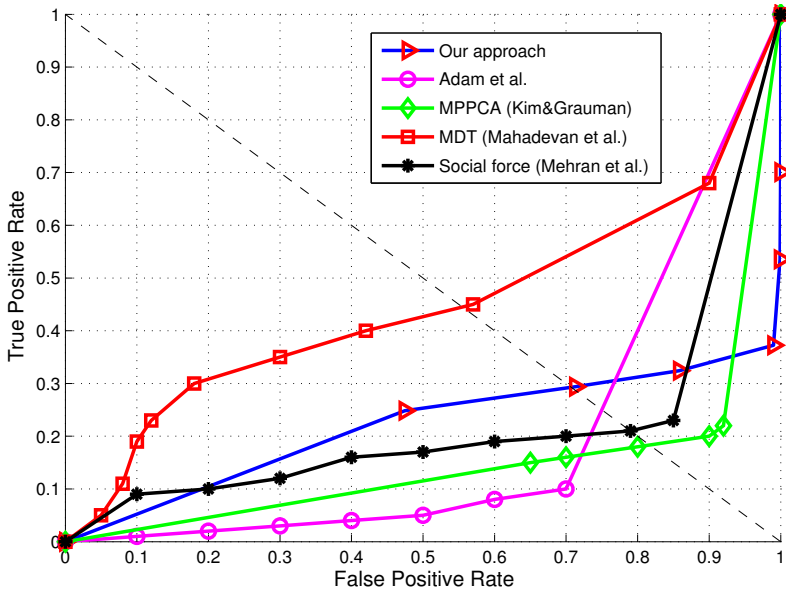


Figure 5.10: ROC curve to compare the localisation accuracy of our method with state-of-the-art approaches using Peds1 dataset. The dashed diagonal is the EER line (note that the plot of a random classifier is not diagonal in this case, but close to zero).

and time. We have also provided a straightforward procedure to dynamically update the learnt model, to deal with scene changes that happen in real-world surveillance scenarios. Dense and overlapping spatio-temporal features, that model appearance and motion information, have been used to capture the scene dynamics, allowing the detection of different types of anomalies. The proposed method is capable of handling challenging crowded scenes that cannot be modelled using trajectories or pure motion statistics (optical flow).

A comparison on a publicly available dataset shows that our method achieves the best performance with respect to existing state-of-the-art real-time solutions [1, 70, 102].

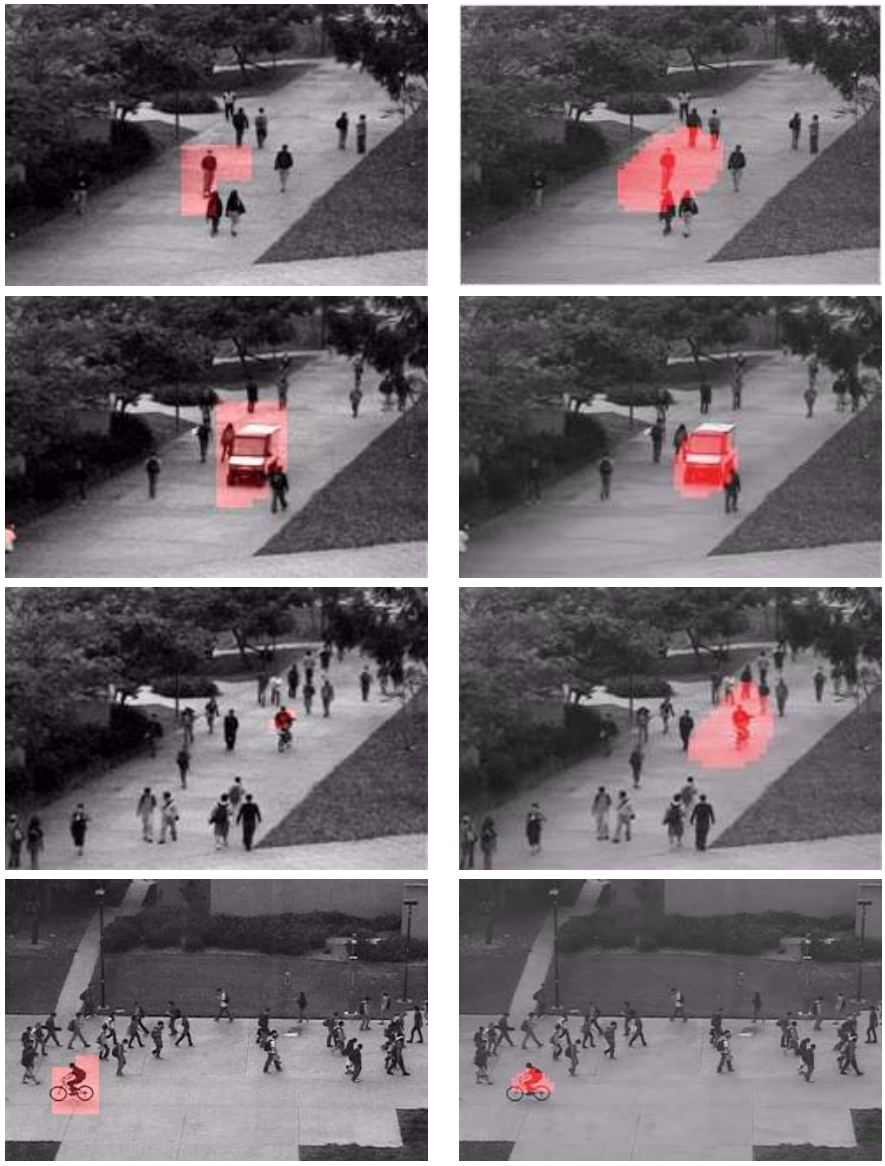


Figure 5.11: Anomaly localisation results (top) compared with the best performing method [102] (bottom) on the UCSD dataset.

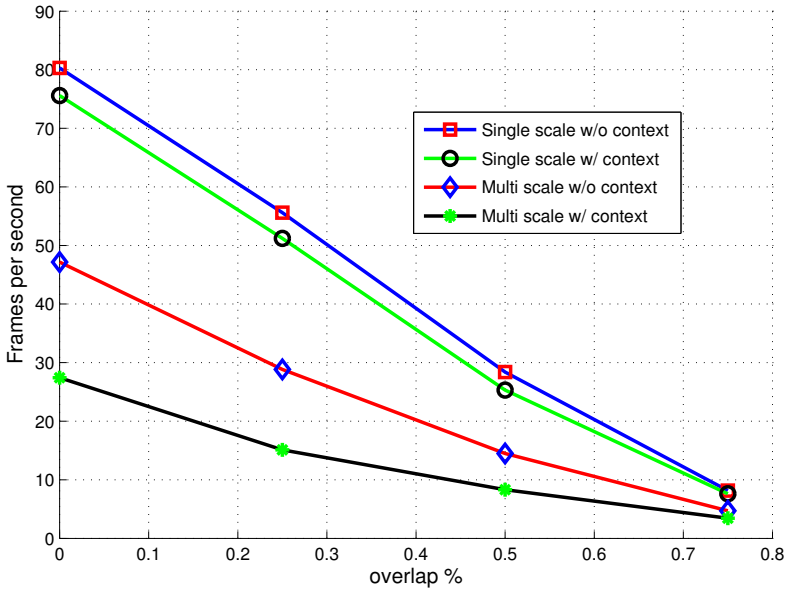


Figure 5.12: Comparison of the number of frames per second (FPS) processed while varying the spatial overlap of cuboids, using single-scale and multi-scale approach on the Peds1 dataset.

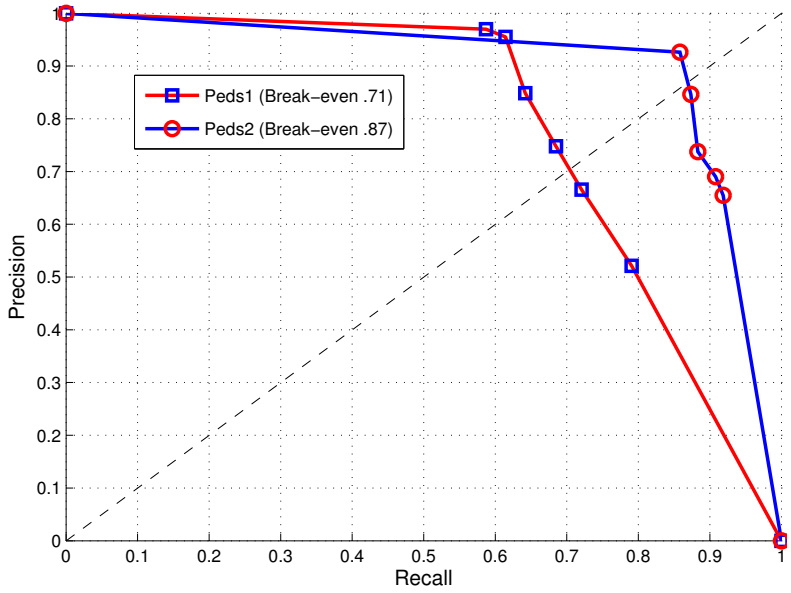


Figure 5.13: Precision/Recall curve of our approach for the two datasets. Break-even, i.e. the intersection with the dashed line, is reported in the legend.

Chapter 6

Semantic adaptive video coding for video surveillance applications

This chapter describes an approach to adaptive video coding for video surveillance applications. Using a combination of low-level features with low computational cost, we show how it is possible to control the quality of video compression so that semantically meaningful elements of the scene are encoded with higher fidelity, while background elements are allocated fewer bits in the transmitted representation. Our approach is based on adaptive smoothing of individual video frames so that image features highly correlated to semantically interesting objects are preserved. Using only low-level image features on individual frames, this adaptive smoothing can be seamlessly inserted into a video coding pipeline as a pre-processing state. Experiments show that our technique is efficient, outperforms standard H.264 encoding at comparable bitrates, and preserves features critical for downstream detection and recognition. [1] [2]

[1] This chapter has been published as "Adaptive Video Compression for Video Surveillance Applications" in *Proc. of International Symposium on Multimedia (ISM) 2011* [4].

[2] This work is partially supported by the EU EraSME ORUSSI Project and by SELEX Communications.

6.1 Introduction

Many critical video streaming applications require transmission of many streams over limited bandwidth. Two such examples are video surveillance networks and local UHF video streaming networks like those based on the ETSI TETRA standard used in emergency and security services [125]. These two example applications have several things in common, among them the need to deliver reasonably high-quality video from multiple cameras spread over large areas and to accomplish this using limited bandwidth [55]. One way to optimise such systems is to control the amount of redundant or irrelevant information transmitted by each camera. In this article we describe a system of adaptive video compression that automatically adjusts the amount of information transmitted according to how semantically "interesting" a part of a video is likely to be.

Consider the example of a video surveillance application, such as in a hospital or airport, where hundreds of cameras might be deployed to monitor tens of thousands of square meters. Systems of this type typically stream raw video feeds from all cameras to a central server for observation, analysis and possibly further processing. This creates a bottleneck at the central server, and bandwidth limitations become a critical issue in overall system efficiency. This bandwidth problem becomes even more acute when wireless IP cameras are deployed – an option that is becoming increasingly popular due to their rapid reconfigurability and lack of infrastructure requirements such as cabling. Note also that a large percentage of bandwidth is expended transmitting scenes of little or no interest because they do not contain objects of semantic interest (e.g., people, cars or aeroplanes). In such application scenarios selectively compressing video streams depending on the semantic content of each frame can result in significant bandwidth savings.

Another video streaming application that can benefit from this type of semantic adaptation are the UHF networks commonly used to stream video from dash-cams installed in state and local police cars. Many police departments require that dash-cams be used to record incidents and that they stream video back to a central headquarters for monitoring and archiving. At any one time, tens or even hundreds of cameras might be streaming video and in this application bandwidth is severely limited by the limitations imposed by using UHF radio frequencies for transmission. Again, significant amounts of bandwidth can be wasted transmitting irrelevant portions of the video frame that contain no semantically relevant information in the form of

faces, people, licence plates, etc.

In both of these examples, bandwidth is squandered by transmitting entire video frames at high bitrates. That is, the same number of bits is dedicated to encoding irrelevant portions of the frame, portions that have no intrinsic value to either application because they contain static and uninteresting objects, as is used to encode truly interesting parts of the frame that contain people, identifying details of cars or faces. Our approach to this problem is to detect interesting portions of video frames and allocate more bits in the compressed representation to them. The rest of the frame is allocated fewer bits in the compressed stream by smoothing away details that would otherwise be unnecessarily encoded in the transmitted video.

Robust and accurate object detectors have almost become a commodity technology in computer vision applications. Reliable, pre-trained detectors exist for pedestrians [36], for text [144], and for a broad variety of general object categories [40, 42]. Despite recent advances in efficient object detection [97], even single object detection still requires a significant amount of computational resources. Application of multiple detectors in order to detect semantically interesting scene elements (e.g., for cars, faces, people, text and licence plates) would require massive computational resources for each individual stream. As such, the detector approach is not feasible for our application scenarios. Note also that new detectors would have to be trained for each potentially interesting scene object, which limits the generality of the detector approach as well.

Most modern detectors are based on high-frequency image features in the form of edges, corner points or other salient image features. The two most popular features are the Histogram of Oriented Gradients (HOG) [36,42,97], which is based on a local histogram of image gradient directions, and SIFT descriptors calculated at interesting points in the image [40, 77]. Both of these descriptors are based on image derivatives calculated across a range of scales in the image. As such, in order to preserve such features in a compressed version of the video it is essential to preserve high-frequencies in each frame and transmit them with reasonable fidelity. If we selectively smooth a video frame, preserving regions containing many high-frequency features, we are more likely to preserve recognisability, or "detectability" using modern object detectors in the transmitted representation. At the same time, if we smooth regions that do not contain dense, high-frequency features we will can reduce the amount of information that must be encoded

and thus transmitted.

This paper is organised as follows: in the next section we discuss some of the related work on adaptive video encoding; the visual features used are described in Section 6.3; a description of our approach to adaptive video compression is provided in Section 6.4. Experimental results and a comparison of adaptive video coding with respect to standard H.264 coding are reported in Section 6.5; and conclusions are drawn in Section 6.6.

6.2 Related Work

Traditional adaptive video compression approaches do not consider the semantic content of video and instead adapt compression depending on the requirements of the network or device used to deliver video to the end user [148]. Semantic video compression, instead, alters the video by taking into account objects [30, 54, 69, 113] or a combination of objects and events [15], using pattern recognition techniques. Kim and Hwang proposed using video object planes (VOP) coding of MPEG-4 to encode differently the interesting objects in the scene [69]. However in [15] it was shown that this type of compression is less efficient than directly performing re-quantisation blocks containing an object that is relevant to users.

Adaptation in the compressed domain has been performed through re-quantisation [160], spatial resolution downscaling [133], temporal resolution downscaling [84], and by a combination of them [87].

Huang et al. [54] use smoothing to control the amount of bits allocated locally to encode each video frame. The more smoothing applied to a portion of a video frame, the fewer bits will be used to encode that region. Their approach is based on motion segmentation, however, and as such is highly sensitive to scene and camera motion. As such it is not directly applicable in cases where streams mobile or active cameras are used, or to detect static objects, like the licence plates of parked cars. Our approach is directly based on image features correlated with downstream detector features. As such, it is applicable to any type of stream, independent of motion characteristics.

Our approach to adaptive video encoding is based on the observation that the most useful image features for downstream object detection are based on edges and salient interest points. As such, by preserving these features we maximise the ability to detect semantically interesting scene objects after transmission. At the same time, by smoothing features that

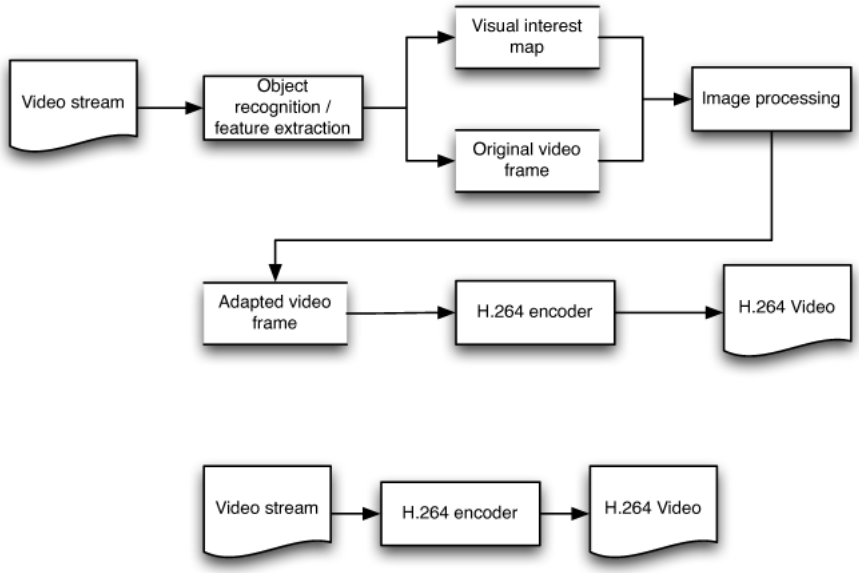


Figure 6.1: Our approach to adaptive video coding. (top) The schema for semantic video adaptation. (bottom) standard video coding.

are unlikely to contribute to positive detections we reduce the amount of irrelevant information transmitted. Figure 6.1 illustrates our system for adaptive video coding. The bottom diagram in Figure 6.1 illustrates the standard H.264 video coding pipeline. At the top of Figure 6.1 is shown our pipeline: before encoding each frame is passed through a sequence of low-level feature extraction (Visual Interest Map), followed by selective smoothing (Image Processing) which smooths details in uninteresting regions of the images before H.264 encoding.

6.3 Visual Features for Adaptive Video Compression

In most surveillance applications the most interesting objects are faces, people and cars. Face and people detectors are both often trained on features based on gradients [36, 156]. Other, more general object detectors are also based on similar features [42]. Moreover edge features are often exploited to estimate crowd density [35, 108] without resorting to object detection and tracking.

Since all MPEG coding standards perform an initial step of spatial color subsampling, as a form of lossy compression, the visual features we use for this work are based on the luminance of pixels. Another advantage of this is that the colour space used in MPEG and M-JPEG standards is YCbCr, so it is possible to extract directly the luminance information from the Y channel, without requiring any conversion.

As mentioned above, the features we use have been selected to be highly correlated with those used for object detection. In particular, corner points can be used to effectively detect text (useful in the case of license plates or identifying text on clothing) in video [13, 144], and edge features in the form of image gradients are used in many state-of-the-art object detectors [40, 42, 97]. Since our application scenario requires onboard adaptive encoding, we selected the features used in order to minimise the computational resources required.

For detecting corner features we use the FAST detector [127]. This detector has recently been used on mobile phones to augment reality using the limited computational resources of the mobile device [72]. The FAST detector is an approximation of the SUSAN detector in which a circular region around a candidate corner is analysed to determine if differences between

(a) (b) (c)

Figure 6.2: Examples of the features used for adaptive image encoding. (a) The original video frame. (b) Corners points detected using the FAST detector. (c) Edges detected using the Sobel gradient operator. Note how both the corner and edge features tend to be concentrated on and around the semantically relevant objects in the scene: people, cars, license plates, etc.

the central point and a pre-defined sequence of pixels in the region satisfy a learned contrast threshold. This detector has been shown to produce very stable features and is the most efficient and robust corner detection algorithm available.

Edge features are characterised by high image gradient values perpendicular to the edge itself. We use the Sobel gradient operator to detect pixels in the image corresponding to edges. The Sobel operator is very efficient, involving only two 3×3 convolutions of the image:

$$\mathbf{G}_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I \tag{6.1}$$

$$\mathbf{G}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I, \tag{6.2}$$

where I is the image to be processed (the video frame) and $*$ represents the 2D convolution operator. The Sobel edge response G, an estimate of the gradient magnitude at each point in the image, is then computed as:

$$\mathbf{G} = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}. \tag{6.3}$$

The FAST and Sobel edge features will be used in the next section to drive adaptive image compression. Essentially, regions of the image containing

a high density of FAST corner responses, or a high density of Sobel edge responses, should be preserved. Other regions can be smoothed in order to reduce detail encoded in transmission. See Figure 6.2 for an example of the extracted features on a typical video frame.

6.4 Adaptive Video Compression

MPEG video coding is based on two basic techniques [57, 161]: transform domain-based compression (intra-coding), where blocks of 8×8 pixels are processed to compute discrete cosine transform (DCT) of each, representing it as a matrix of coefficients; and block-based motion compensation performed on macroblocks, i.e. groups of 2×2 blocks (16×16 pixels), coding them with motion vectors and with the DCT coefficients of the "residual block" obtained from motion estimation. In both cases the DCT coefficients are quantised, as a lossy compression step, so that the high frequency coefficients go to zero in order to represent them with efficient Run Length Encoding (RLE) encoding and Variable Length Codes (VLC). The residual block typically contains high frequency components that have to be quantised differently from the intra-coded blocks.

In our approach we reduce the bandwidth needed for video streaming by selectively smoothing parts of each frame. We do not directly exploit the temporal structure of videos in order to reduce the need of buffering and to allow visual feature extraction even on moving cameras. This approach helps the encoder to more efficiently compress the DCT coefficients of both intra-coded and residual blocks since they will contain fewer high frequency components. The smoothing is defined by a set of semantic binary masks which are generated by collecting statistics of low-level visual features in a video frame. These masks could also be defined by a set of detectors for objects of interest or anomalous frame regions. The result would be a binary mask defined by the bounding boxes of objects detected in each frame as shown in Figure 6.4. This approach performs extremely well (see Table 6.2 in Section 6.5) but does not allow a sufficient frame-rate on low-end computational architectures and as discussed above does not generalise well when the number of objects of interest increases.

We instead design our masks by splitting each frame into square pixel regions. Region size is selected in order to optimally fit the DCT encoded pixel macroblocks used in H.264 video encoding. We tested 8×8, 16×16

and 32×32 regions. Smoothed regions will be assigned fewer bits by the encoding algorithm, allowing more bits to be assigned to non-smoothed ones. This approach will therefore decrease the bandwidth needed for streaming while maintaining a high quality of interesting frame regions.

The amount of smoothing applied to each region is determined by the density of Sobel and FAST features contained therein. Let I denote the current image to be encoded, and $F(\mathbf{x})$ and $S(\mathbf{x})$ the FAST and Sobel feature responses at pixel $\mathbf{x}$, respectively. Also, let B_i denote the i-th block of the image and let the function $B(\mathbf{x})$ map pixel $\mathbf{x}$ to the block containing it. We define the following feature threshold functions on local image blocks:

$$F(B_i) = \{x | x \in B_i \mid F(\mathbf{x}) > T_F\} \tag{6.4}$$

$$S(B_i) = \{x | x \in B_i \mid S(\mathbf{x}) > T_S\}, \tag{6.5}$$

where T_F and T_S are empirically determined thresholds on the FAST and Sobel feature responses.

Assuming there are n levels of smoothing, we will now define smoothing masks that are based on feature densities in each image block. The masks correspond to increasing feature densities. The i-th level mask corresponding to FAST feature density is defined as:

$$M_i^F(\mathbf{x}) = \begin{cases} 1 & \text{if } \tau_{i-1}^F \leq |F(B(\mathbf{x}))| < \tau_i^F \\ 0 & \text{otherwise} \end{cases}, \tag{6.6}$$

where τ_i^F for $i \in \{0, 1, \ldots, n\}$ is a strictly increasing series of thresholds used to determine which feature densities correspond to which mask. We require that $\tau_0^F = 0$ and $\tau_n^S = w \times h$, where w and h are the width and height of the image blocks used for encoding. These restrictions ensure that the sequence of image masks M_i completely partitions the image:

$$\bigcap_i \bigcup_{\mathbf{x}} M_i^F(\mathbf{x}) = \emptyset \tag{6.7}$$

$$\bigcup_i \bigcup_{\mathbf{x}} M_i^F(\mathbf{x}) \circ I = I \tag{6.8}$$

.

The i-th level mask corresponding to Sobel feature density is similarly defined:

$$M_i^S(\mathbf{x}) = \begin{cases} 1 & \text{if } \tau_{i-1}^S \leq |F(B(\mathbf{x}))| < \tau_i^S \\ 0 & \text{otherwise} \end{cases}, \tag{6.9}$$

with identical conditions on τ_i^S as for FAST features above.

The final smoothed image can now be written as:

$$I_s = \sum_{i=1}^{n} (M_i^S \circ M_i^F \circ I) * G_{\sigma_i}, \tag{6.10}$$

where $M_i^S \circ M_i^F \circ I$ represents the Hadamard (element-by-element) multiplication of the feature masks and image I, and G_{σ_i} is a Gaussian function with variance σ_i^2. I_s is an adaptively smoothed version of the original image I. Based on the density of FAST and Sobel features in each $w \times h$ block of the image, a variable amount of smoothing will be applied. The amount of smoothing applied is controlled by the sequence σ_i, while the density thresholds τ_i^S and τ_i^F are used to determine how interesting each block is in terms of each feature.

6.5 Experimental Results

In order to evaluate the performance of our approach to adaptive video compression, we acquired a set of three test videos using a real-world surveillance setup. Two Sony SNC RZ30 cameras recorded videos of a parking lot at 640×480 pixels and 25 FPS for a total of 2327 frames. The videos were been recorded in MPEG-4 format using the H.264 codec provided by the open source library x264. Each video is encoded with an average bitrate of 3532,44 Kbit/s.

We evaluate the performance of our algorithm by measuring the structural similarity index (SSIM) [158] and comparing the non-H.264 compressed videos and videos compressed with our approach, computing SSIM on the computed masks. SSIM is a visual quality assessment metric that models the perception of compression artefacts by the human visual system better than other standard quality measures based on peak signal-to-noise ratio (PSNR) or mean squared error (MSE). In fact MSE, and consequently PSNR, perform badly in predicting human perception of image fidelity and quality: MSE values of distorted images that present dramatically and visibly different visual qualities can be nearly identical [158]. For this reason its use has been proposed to drive the motion compensation coding of H.264 [170] and some encoders, like x264[3], have started to use it to drive the adaptation of

[3] http://www.x264.org

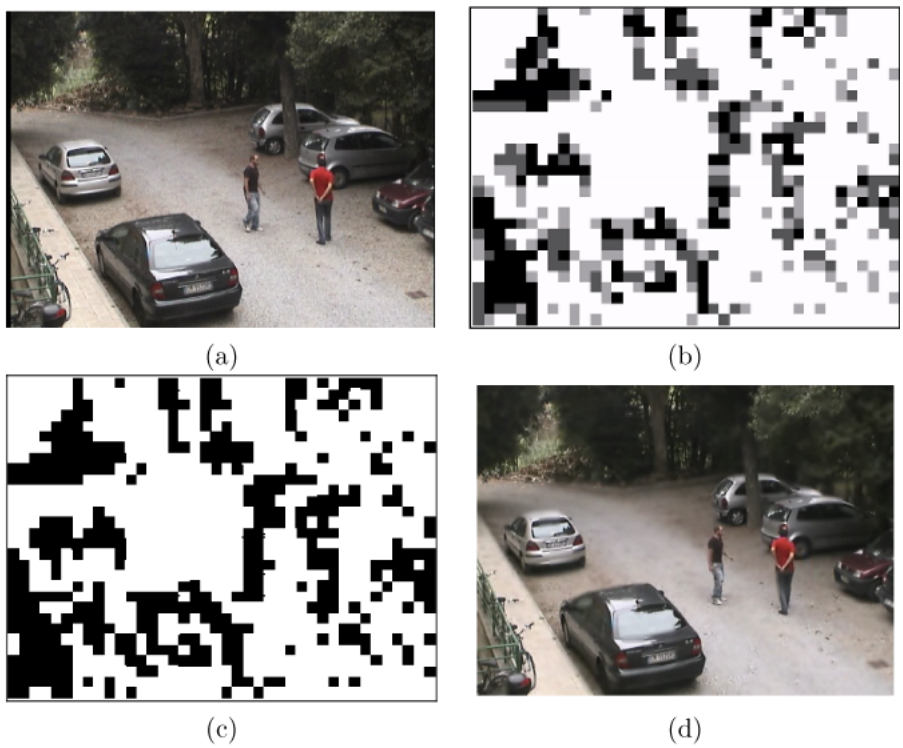
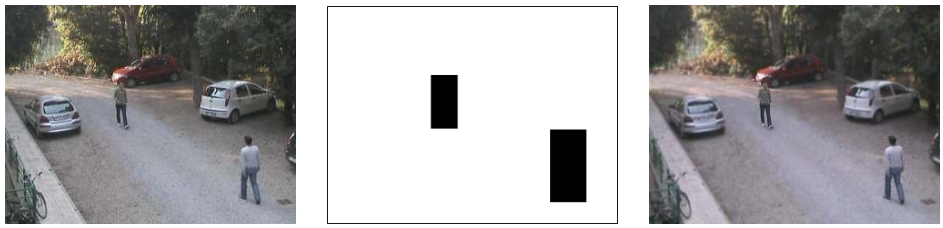


Figure 6.3: An example of feature-preserving adaptive compression. (a) Original video frame. (b) The multi-level mask computed from FAST and Sobel responses. (c) A single level mask from FAST and Sobel features. (d) The frame compressed using the single-level mask. Note how persons and license plates are encoded with high enough quality to preserve recognisability, while background details are smoothed away.

the quantisation coefficients during compression. SSIM is defined as:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}, \tag{6.11}$$

where X and Y are images, μ_X is the average of the luminance of X, μ_Y is the average of the luminance of Y, σ_{XY} is the covariance of the contrast of X and Y, σ_X^2 is the variance of the contrast of X, σ_Y^2 is the variance of the contrast of Y, $C_i = (K_i L)^2$ are constants used to avoid instability when $\mu_X^2 + \mu_Y^2$ is near 0, where $K_i \ll 1$ and L is the amplitude of the range of



(a) (b) (c)

Figure 6.4: Adaptive compression driven by pedestrian detector. A frame with two people (a), the masks built with the pedestrian detector (b) and the final adaptively encoded frame (c). All the scene but the two pedestrian is smoothed.

values that a pixel can have. SSIM is typically computed on 8×8 pixel windows, and can assume values between $[-1, 1]$, where 1 means that two images are identical. SSIM is measured in the non-smoothed regions only.

After an initial evaluation we found that the best performance is obtained by exploiting both low-level features (FAST and Sobel), using blocks of 16×16 and smoothing all regions without corners and with less than 3 non-zero pixels. Smoothing is performed using block filters approximating a Gaussian filter of the correct σ. In a preliminary set of experiments we also explored the possibility of using multiple levels of smoothing, in particular we used three levels of smoothing selected with three thresholds for both features. The performance of this approach is less appealing, see Table 6.1, with respect to plain binary mask driven smoothing. Figure 6.3 illustrates the performance of our approach, along with example masks derived from low-level features, on a typical video frame. Note how semantically meaningful features like the license plate and persons are preserved in the compressed representation.

6.5.1 Feature and Compression Evaluation

We evaluated low-level features, mask size and type for a set of sensible configurations. In particular, we tested the FAST and Sobel features separately and together. Results are reported for each feature or combination for the best window size. The average SSIM gain (ΔSSIM) is obtained with the following procedure: first videos are compressed with H.264 with Constant

Method	ΔSSIM (binary masks)	ΔSSIM multi-level masks
Sobel 8	0.013	-0.002
FAST 32	0.024	0.009
Sobel + Fast 16	0.027	0.01

Table 6.1: Comparison of different low-level features and mask generation strategies. Binary masks correspond to the situation where only two levels of smoothing are used ($n = 2$), whereas for multi-level masking three levels were used ($n = 3$). Average ΔSSIM is reported with respect to identical bitrate video encoded with standard H.264 compression. CRF is varied in order to obtain files of the same bitrate.

Rate Factor (CRF) in the range 25-20; for each of these files we compress the original video (V_o) with our adaptive technique tuning the CRF (lowering) in order to obtain approximately the same bitrate. We finally compute ΔSSIM as SSIM(V_{ac},V_o) - SSIM(V_c,V_o), where V_{ac} and V_c are the adaptive and H.264 coded videos, respectively. A negative value means that our technique degrades the video more than standard H.264 encoding, note that this happens only for Sobel features alone and with multi-level masks. For all other combinations we improve the SSIM without increasing the bitrate.

Apart from the increase of quality in regions of interest we are mainly interested in the reduction of bandwidth. To this end we measure how the file size decrease with the increase of CRF for adaptive encoded videos and H.264 encoded videos. As shown in Figure 6.5 we are able to spare between 40% to 10% of the bandwidth depending on the quality of the final encoding. It is also interesting to see how our encoder is able to retain the appearance of the original video; Figure 6.5 shows the average behavior of our algorithm.

6.5.2 Efficiency of Our Approach

We report in Table 6.2 the frame rate of our approach compared with the frame rate obtained using the pedestrian detector-based approach on the same machine. Low-level features allow our system to run at a very high frame rate, permitting our system to stream video in real-time. Though pedestrian detection allows us to reduce the video size by a higher factor, it should be noted that this is mainly due to the fact that masks generated by the pedestrian detector drive the algorithm to smooth most of the frame, reducing the bandwidth needed. Even if this seems a desirable property,

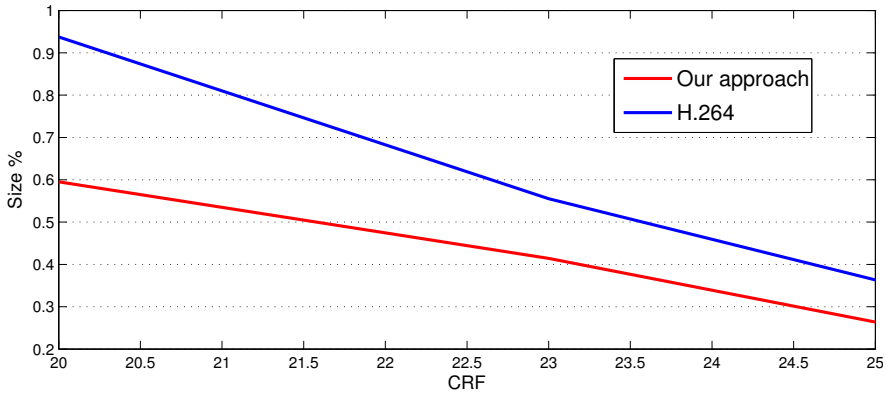


Figure 6.5: Average file size obtained by varying the CRF. File size is normalised with respect to the original (high quality) video.

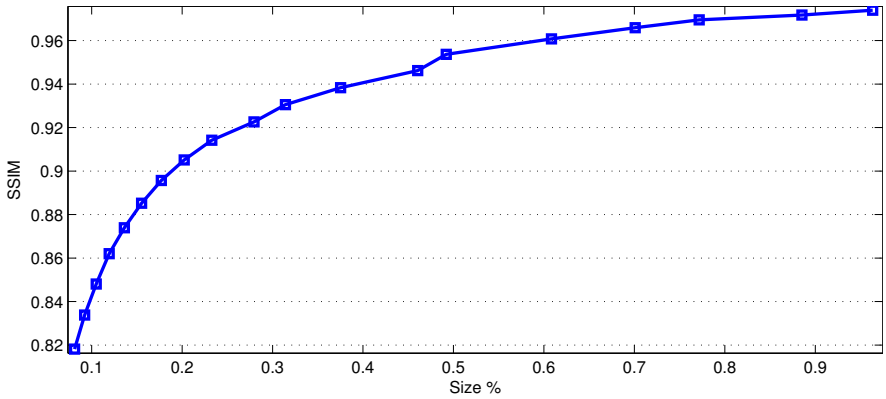


Figure 6.6: SSIM versus file size (normalised with respect to the original video size). For a video size of about 20% of the original, the SSIM computed on the interesting areas is above 0.9 and approaches 0.95. Even when video size decreases rapidly, below 20% of the original size, SSIM still remains above 0.8.

Video	Method	Frames/sec	Video Size (MB)
V1	Pedestrian Detector	1.3	2.2
V1	Sobel + Fast16	71.8	4.1
V2	Pedestrian Detector	1.6	3.0
V2	Sobel + Fast16	87.0	7.2
V3	Pedestrian Detector	1.9	15.5
V3	Sobel + Fast16	71.5	18.0

Table 6.2: Frame rate comparison for the two feature extraction approaches. File size is also reported for the same CRF (17).

Video (size)	Precision	Recall	ΔTP/P	ΔFP/P
Original (7.9 MB)	.92	.67	-	-
Compressed (1.8 MB)	.89	.81	.11	.09
Compressed (4.4 MB)	.93	.84	.13	.05
Compressed (7.2 MB)	.89	.84	.11	.09

Table 6.3: Performance of a pedestrian detector on the original and adaptively encoded frames. Precision is slightly reduced but recall is increased.

objects other than pedestrians are encoded with a lower quality. As an example, license plates are unreadable and other car details are consistently degraded.

6.5.3 Semantic Cue Preservation

Videos compressed and transmitted with our adaptive encoding will be subsequently inspected either by personnel or machines. In the following experiment we measure how encoded videos preserve the image features that allow high level object detectors to extract semantic information from videos. In particular, we processed a video with the Dalal&Triggs [36] pedestrian detector before and after the adaptive encoding for three levels of adaptive compression. From Table 6.3 it appears that the performance of the pedestrian detector is not substantially affected by our adaptive compression. In particular, the precision on the compressed video is reduced but the overall recall is improved. We also report the increase in true positives and false positives for compressed videos, which also explains the increase in recall.

6.6 Conclusions

In this paper we have presented a novel method for semantic video coding based on low-level features that require a very limited computational resources. The technique can be used as a pre-processing state before DCT encoding, and is able to reduce the size of videos down to half the original size, while maintaining the perceptive quality of the areas considered of interest. The approach has been shown to have also a beneficial effect on automatic analysis of the compressed video, improving the performance of the person detector based on the histogram of gradients.

Chapter 7

Conclusions

7.1 Summary of contribution

This thesis makes a contribution to the field of video understanding. In particular, we address the problem of detecting events in videos with a particular focus on the recognition of human actions. Our effort is dedicated to designing robust representations for local space-time patterns and to the definition of models that allow comparison between these patterns in order to build supervised and semi-supervised systems. We explore different aspects of video understanding, each requiring progressively less human supervision.

The first step to enable machines to automatically recognise video content is the design of a robust video representation. Indeed, we are not willing to narrow the domain of videos to analyse. For this reason, in Chapter 3 we propose a novel local spatio-temporal feature that is efficient to compute, adapts to various video domains without any tuning and represent appearance and motion. We also show how to correctly quantise these high dimensional features to obtain discriminative and rich codebooks of video words by applying density-based clustering. Finally, we apply a nonlinear dimensionality reduction technique to compress the codebooks, gaining in the process two orders of magnitude in training and testing speed. We propose a complete semantic video retrieval system that, addressing several issues in the recognition pipeline, improves over the state-of-the-art.

After this first step we recognise that local space-time feature design is still an open issue. We therefore proceed, using a different approach, with the definition of a novel feature in Chapter 4. We exploit Zernike moments to

benefit from the orthogonality condition of Zernike polynomials that ensures there is no redundancy in the representation. Unfortunately, high order moments are extremely sensitive to noise. We deal with this issue with a novel kernel for features that builds on the concept of pyramid matching. Our similarity metric allows the preservation of the information from higher order moments when discriminative and automatically discards it when it is not. The proposed descriptor is extremely low dimensional and, thanks to the pyramid kernel descriptor matching procedure, obtains state-of-the-art performance.

Chapters 3 and 4 present methods to detect events belonging to a finite set of known classes. To obtain a complete coverage of the video understanding problem, we must cope with scenarios for which annotation of events of interest is not available. This can happen due to the scarcity of annotated video data or for the too wide range of behaviours that need to be retrieved; finally, it is also possible that users do not know which are the events of interest but rather only those which are not. To cover all these situations we propose to shift the video annotation paradigm from supervised to semi-supervised, casting the retrieval of unknown events of interest as anomaly detection. In Chapter 5 we build on features presented in Chapter 3 to obtain scene representation. We define a model for anomaly detection with the real-time requirement in mind. Our model is based on non-parametric statistical tests and therefore requires almost no training time and very reduced testing time. The use of context together with a multi-scale representation greatly helps to suppress false positives. The scene model can be easily updated with a procedure derived from the anomaly detection algorithm. Our system runs in real-time on a modest platform and is the best among other real-time systems proposed in the the literature.

Chapter 6 presents an application of concept detection in video. We present an algorithm to reduce the bandwidth required for streaming video. The proposed technique is applicable to all DCT based video codecs as a simple preprocessing step. The approach is based on selective smoothing of image blocks. The choice of blocks is semantically driven. Smoothing masks can for example be generated with object, event or anomaly detectors. Since these detectors are often based on low-level features we propose also a technique based on the density of image structures like corners and edges to drive the adaptive compression. Dropping the use of detectors reduces the computational burden. This latter approach not only reduces the bitrate but

videos encoded with our algorithm preserve important image features. We performed further experiments with a pedestrian detector on videos encoded with our algorithm. The detector did not significantly loose precision and even improved recall.

7.2 Directions for future work

When it comes to scaling automatic video annotation to more data and inevitably more concepts we will face several challenges. First of all there is the need to obtain good annotation for a sufficient amount of video examples for each concept. This can be partly overcome using social media and, in the case of movies, the scripts. There is still a large amount of data for which these approaches are not feasible. For user generated content uploaded on social networks, the only "free" annotation available is the one we can extract from tags or user comments, which are an imprecise and noisy source of information. For video surveillance footage we do not have any means of easily obtaining annotation.

Moreover, when it comes to describing the interaction of a person with one or more other persons or objects, it is extremely complex to train discriminative models, unless we make strong assumptions (i.e. maximum number of people, maximum duration of the activity, strong prior on the point of view, etc.). The complexity is also increased by the need to classify activities exploiting context. Finally, the articulated nature of the human body is the main cause for the extreme variability of activities a human would define similar or the same.

We believe that in order to scale a human activity understanding system it is important to gather all the information available in order to build a rich representation. Such information resides in the single subject location and motion in the scene, the objects present in the scene (and their relative position to people) and the mutual location and distances of the people involved.

A possible approach is to start by densely sampling spatio-temporal features and subsequently add structure to the sequence representation. We can think of various way of adding structure to this bag-of-features: use a global structure, a spatial pyramid may be a solution, but we can think of adding the scene geometry if known; partition the scene not only globally but also using the detection of people and their parts; use the context of

persons to generate another partitioning of the scene.

Essentially, what we are describing is a progressive filtering of scene features into a more structured model of global, contextual and personal features that each correspond to a different aspect of the scene. We also propose dealing with this kind of data with unsupervised learning.

An advantage of moving to this type of generative model would be that it allows richer characterisation of agent behaviour in the scene, and would enable applications like retrieval and unsupervised characterisation of complex actions that discriminative models do not.

Appendix A

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. M. Bertini, A. Del Bimbo, **L. Seidenari**. "Multi-scale and real-time non-parametric approach for anomaly detection and localization", vol. in press, 2012. (S.I.: Semantic Understanding of Human Behaviours in Image Sequences) [DOI:10.1016/j.cviu.2011.09.009]

International Conferences and Workshops

1. 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.
2. 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.

[1] The author's bibliometric indices are the following: H-index = 3, total number of citations = 27 (source: Google Scholar on January 25, 2012).

3. 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.
4. **L. Seidenari** and M. Bertini, "Non-parametric Anomaly Detection Exploiting Space-time features", in *Proc. of ACM MultiMedia (ACMMM)*, Firenze, Italy, 2010.
5. **L. Seidenari** and Marco Bertini and Alberto Del Bimbo ,"Dense Spatio-temporal Features For Non-parametric Anomaly Detection And Localization", in *Proc. of ACM MultiMedia International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, Firenze, Italy, 2010.
6. A. D. Bagdanov, M. Bertini, A. Del Bimbo and **L. Seidenari**,"Adaptive Video Compression for Video Surveillance Applications" in *Proc. of ISM Int'l Symposium on Multimedia* ,Dana Point(CA),2011.
7. L. Costantini, **L. Seidenari**, G. Serra, A. Del Bimbo and L. Capodiferro,"Space-time Zernike Moments and Pyramid Kernel Descriptors for Action Classification", in *Proc. of International Conference on Image Analysis and Processing (ICIAP)*, 2011, Ravenna, Italy.

National Conferences

1. L. Ballan, M. Bertini,A. Del Bimbo,F. Dini,G. Lisanti,**L. Seidenari** and G. Serra, "RECENT RESEARCH ACTIVITIES IN VIDEOSURVEILLANCE AT UNIFI::MICC", in *Proc. of GIRPR National Conference*, Marina di Ascea (SA), Italy, 2010.
2. **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.

Bibliography

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[2] P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis, "Detecting abnormal human behaviour using multiple cameras," *Signal Processing*, vol. 89, no. 9, pp. 1723–1738, 2009.

[3] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174 – 188, 2002.

[4] A. D. Bagdanov, M. Bertini, A. Del Bimbo, and L. Seidenari, "Adaptive video compression for video surveillance applications," in *Proc. of IEEE Int'l Symposium on Multimedia (ISM)*, 2011.

[5] A. D. Bagdanov, F. Dini, A. Del Bimbo, 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.

[6] 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, 2009.

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

[8] ——, "Event detection and recognition for semantic annotation of video," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 279–302, 2011.

[9] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video event classification using string kernels," *Multimedia Tools and Applications*, vol. 48, no. 1, pp. 69–87, 2010.

[10] 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.

[11] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[12] M. Bertini, A. D. Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320–329, 2012.

[13] M. Bertini, C. Colombo, and A. Del Bimbo, "Automatic caption localization in videos using salient points," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, 2001, pp. 68–71.

[14] 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, 2005.

[15] M. Bertini, A. Del Bimbo, A. Prati, and R. Cucchiara, "Semantic adaptation of sport videos with user-centred performance analysis," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 433–443, 2006.

[16] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2010.

[17] 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.

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

[19] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification." in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[20] C. Brax, L. Niklasson, and M. Smedberg, "Finding behavioural anomalies in public areas using video surveillance data," in *Proc. of Int'l Conference on Information Fusion*, 2008.

[21] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[22] M. Breitenstein, H. Grabner, and L. Van Gool, "Hunting Nessie: Real time abnormality detection from webcams," in *Proc. of ICCV Int'l Workshop on Visual Surveillance*, 2009.

[23] 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, 2008.

[24] W. M. Bulkeley. (2009) Chicago's camera network is everywhere. [Online]. Available: http://online.wsj.com/article/SB10001424052748704538404574539910412824756.html

[25] S. Calderara, C. Alaimo, A. Prati, and R. Cucchiara, "A real-time system for abnormal path detection," in *Proc. of IET Int'l Conference on Imaging for Crime Detection and Prevention (ICDP)*, London, UK, 2009.

[26] N. Canterakis, "3d zernike moments and zernike affine invariants for 3d image analysis and recognition," in *Proc. of Scandinavian Conference on Image Analysis*, 1999.

[27] L. Cao, L. Zicheng, and T. Huang, "Cross-dataset action detection," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[28] M. A. Carreira Perpinan and G. E. Hinton, "On contrastive divergence learning," in *Proc. of Artificial Intelligence and Statistics (AISTATS)*, 2005.

[29] R. Castellanos, H. Kalva, O. Marques, and B. Furht, "Event detection in video using motion analysis," in *Proc. of ACM Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, 2010.

[30] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic segmentation and description for video transcoding," in *Proc. of IEEE Int'l Conference on Multimedia & Expo (ICME)*, 2003.

[31] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

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

[33] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[34] L. Costantini, L. Seidenari, G. Serra, A. Del Bimbo, and L. Capodiferro, "Space-time zernike moments and pyramid kernel descriptors for action classification," in *Proc. of IAPR Int'l Conference on Image Analysis and Processing (ICIAP)*, Ravenna, Italy, 2011.

[35] D. G. D. Kong and H. Tao, "Counting pedestrians in crowds using viewpoint invariant training," in *Proc. of IEEE British Machine Vision Conference (BMVC)*, 2005.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[37] J. Davenport. (2007, September) Tens of thousands of cctv cameras, yet 80% of crime unsolved. [Online]. Available: http://www.thisislondon.co.uk/news/article-23412867-tens-of-thousands-of-cctv-cameras-yet-80-of-crime-unsolved.do

[38] 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.

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

[40] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[41] 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.

[42] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[43] 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.

[44] 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.

[45] J. Flusser, B. Zitova, and T. Suk, *Moments and Moment Invariants in Pattern Recognition*. Wiley Publishing, 2009.

[46] W. T. Freeman and E. C. Pasztor, "Learning low-level vision," *International Journal of Computer Vision (IJCV)*, vol. 40, no. 1, pp. 25–47, 2000.

[47] Z. Gao, M.-Y. Chen, A. G. Hauptmann, and A. Cai, "Comparing evaluation protocols on the KTH dataset," in *Proc. of HBU Workshop*, 2010.

[48] 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.

[49] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. of IEEE Int'l Conference on Computer Vision (ICCV)*, 2005.

[50] N. Haering, P. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 279–290, 2008.

[51] A. Haubold and M. Naphade, "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, pp. 178–185.

[52] A. G. Hauptmann, M. G. Christel, and R. Yan, "Video retrieval based on semantic concepts," in *Proceedings of the IEEE*, vol. 96, no. 4, 2008.

[53] E. G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[54] H.-J. Huang, X.-M. Zhang, and Z.-W. Xu, "Semantic video adaptation using a preprocessing method for mobile environment," in *Proc. of IEEE Int'l Conference on Computer and Information Technology (CIT)*, 2010.

[55] S. Z. Hussain, "Performance evaluation of H.264/AVC encoded video over TETRA enhanced data service (TEDS)," Master's thesis, Helsinki University of Technology, 2009.

[56] I. Ivanov, F. Dufaux, T. M. Ha, and T. Ebrahimi, "Towards generic detection of unusual events in video surveillance," in *Proc. of IEEE Int'l Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2009.

[57] K. Jack, *Video Demystified*. LLH Publishing, 2001.

[58] H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. Steele, and T. Serre, "Automated home-cage behavioral phenotyping of mice," *Nature communications*, pp. 1–68, 2010.

[59] F. Jiang, Y. Wu, and A. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 907–913, 2009.

[60] ——, "Detecting contextual anomalies of crowd motion in surveillance video," in *Proc. of IEEE Int'l Conference on Image Processing (ICIP)*, 2009.

[61] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding (CVIU)*, vol. 115, no. 3, pp. 323–333, 2011, special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics.

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

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

[64] 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.

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

[66] H. Keval and M. Sasse, ""Not the Usual Suspects": A study of factors reducing the effectiveness of CCTV," *Security Journal*, vol. 23, no. 2, pp. 134–154, 2010.

[67] S. Khalid, "Activity classification and anomaly detection using m-mediods based modelling of motion patterns," *Pattern Recognition*, vol. 43, no. 10, pp. 3636–3647, 2010.

[68] 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 Annual Symposium of the German Association for Pattern Recognition*. Springer, 2007.

[69] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 122–129, 2002.

[70] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[71] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-Gradients," in *Proc. of IEEE British Machine Vision Conference (BMVC)*, 2008.

[72] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. Sixth IEEE and ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR)*, Nara, Japan, 2007.

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

[74] Y. Kong, X. Zhang, W. Hu, and Y. Jia, "Adaptive learning codebook for action recognition," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1178–1186, 2011.

[75] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[76] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[77] C. H. Lampert, "Detecting objects in large image collections and videos by efficient subimage retrieval," in *Proc. of IEEE Int'l Conference on Computer Vision (ICCV)*, 2009.

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

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

[80] 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.

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

[82] 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, 2009.

[83] 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.

[84] Z. Lei and N. Georganas, "H.263 video transcoding fo spatial resolution downscaling," in *Proc. of Conference on Information Technology: Coding and Computing*, 2002.

[85] J. Li, S. Gong, and T. Xiang, "Discovering multi-camera behaviour correlations for on-the-fly global activity prediction and anomaly detection," in *Proc. of IEEE Int'l Conference on Computer Vision Workshops (ICCV Workshops)*, 2009.

[86] S. Li, M.-C. Lee, and C.-M. Pun, "Complex zernike moments features for shape-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, no. 1, pp. 227–237, 2009.

[87] Y. Liang and Y.-P. Tan, "A new content-based hybrid video transcoding method," in *Proc. of IEEE Int'l Conference on Image Processing (ICIP)*, 2001.

[88] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. of IEEE Int'l Conference on Computer Vision (ICCV)*, 2009.

[89] C. Liu, G. Wang, W. Ning, X. Lin, L. Li, and Z. Liu, "Anomaly detection in surveillance video using motion direction statistics," in *Proc. of IEEE Int'l Conference on Image Processing (ICIP)*, 2010.

[90] 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.

[91] 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.

[92] 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.

[93] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[94] C. C. Loy, T. Xiang, and S. Gong, "Detecting and discriminating behavioural anomalies," *Pattern Recognition*, vol. 44, no. 1, pp. 117–132, 2011.

[95] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of Image Understanding Workshop*, 1981.

[96] 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)*, San Francisco, CA, USA, 2010.

[97] J. G. A. D. B. Marco Pedersoli and X. Roca, "Efficient discriminative multiresolution cascade for real-time human detection applications," *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1581–1587, 2011.

[98] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[99] ——, "Actions in context," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[100] R. Mattivi and L. Shao, "Human action recognition using lbp-top as sparse spatio-temporal feature descriptor," *Computer Analysis of Images and Patterns*, vol. 5702, pp. 740–747, 2009.

[101] R. Mehran, B. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. of IEEE European Conference on Computer Vision (ECCV)*, 2010.

[102] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[103] K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," in *Proc. of IEEE Int'l Conference on Computer Vision (ICCV)*, 2005.

[104] 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.

[105] 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.

[106] 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.

[107] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding (CVIU)*, vol. 104, no. 2-3, pp. 90–126, 2006.

[108] M. Morrow, A. B. Chan, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in *In IEEE Intl. Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2009)*, 2009.

[109] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. of Int'l Conference on Computer Vision Theory and Application (VISSAPP'09)*, 2009.

[110] A. Neri, M. Carli, V. Palma, and L. Costantini, "Image search based on quadtree zernike decomposition," *Journal of Electronic Imaging*, vol. 19, no. 4, p. 043023, 2010.

[111] J. 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.

[112] 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.

[113] T. Nishi and H. Fujiyoshi, "Object-based video coding using pixel state analysis," in *Proc. of IAPR Int'l Conference on Pattern Recognition (ICPR)*, 2004.

[114] 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.

[115] M. Novotni and R. Klein, "Shape retrieval using 3d zernike descriptors," *Computer-Aided Design*, vol. 36, no. 11, pp. 1047–1062, 2004.

[116] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. of IEEE European Conference on Computer Vision (ECCV)*, 2006.

[117] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, no. 1, pp. 99 – 110, 2003.

[118] 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.

[119] A. Oreskovic. (2012, January) Exclusive: Youtube hits 4 billion daily video views. [Online]. Available: http://www.reuters.com/article/2012/01/23/us-google-youtube-idUSTRE80M0TS20120123

[120] 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.

[121] C. Piciarelli and G. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835–1842, 2006, vision for Crime Detection and Prevention.

[122] C. Piciarelli and G. L. Foresti, "Surveillance-oriented event detection in video streams," *IEEE Intelligent Systems*, vol. PrePrints, no. 99, pp. 32–41, 2010.

[123] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding (CVIU)*, vol. 108, no. 1-2, pp. 4–18, 2007.

[124] ——, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[125] N. Qadri, M. Altaf, M. Fleury, and M. Ghanbari, "Robust video communication over an urban VANET," *Mobile Information Systems*, vol. 6, no. 3, pp. 259–280, 2010.

[126] K. Rapantzikos, Y. Avrithis, and S. Kollia, "Dense saliency-based spatiotemporal feature points for action recognition." in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[127] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. of IEEE European Conference on Computer Vision (ECCV)*, 2006.

[128] S. Savarese, A. Del Pozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *Proc. of IEEE Workshop on Motion and Video Computing*, 2008.

[129] 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.

[130] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. of IAPR Int'l Conference on Pattern Recognition (ICPR)*, 2004.

[131] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT descriptor and its application to action recognition," in *Proc. of ACM Int'l Conference on MultiMedia (MM)*, 2007.

[132] L. Seidenari, M. Bertini, and A. Del Bimbo, "Dense spatio-temporal features for non-parametric anomaly detection and localization," in *Proc. of ACM Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, 2010.

[133] T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatio-temporal resolution and different encoding formats," *IEEE Transactions on Multimedia*, vol. 2, no. 2, pp. 101–110, 2000.

[134] L. Shao, R. Gao, Y. Liu, and H. Zhang, "Transform based spatio-temporal descriptors for human action recognition," *Neurocomputing*, vol. 74, pp. 962–973, 2011.

[135] L. Shao and R. Mattivi, "Feature detector and descriptor evaluation in human action recognition," in *Proc. of ACM Int'l Conference on Image and Video Retrieval (CIVR)*, 2010.

[136] R. Sillito and R. Fisher, "Semi-supervised learning for anomalous trajectory detection," in *Proc. of IEEE British Machine Vision Conference (BMVC)*, 2008.

[137] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. of IEEE Int'l Conference on Computer Vision (ICCV)*, 2003.

[138] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. of ACM Int'l Workshop on Multimedia Information Retrieval (MIR)*, 2006.

[139] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1349–1380, 2000.

[140] 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 Int'l Conference on MultiMedia (MM)*, 2006.

[141] C. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Transactions on Multimedia*, vol. 7, no. 4, pp. 638–647, 2005.

[142] ——, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[143] A. Stedmon, S. Harris, and J. Wilson, "Simulated multiplexed CCTV: The effects of screen layout and task complexity on user performance and strategies," *Security Journal*, vol. -, no. 24, pp. 344–356, 2011.

[144] L. Sun, G. Liu, X. Qian, and D. Guo, "A novel text detection and localization method based on corner response," in *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, 2009.

[145] X. Sun, M. Chen, and A. G. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. of CVPR Workshop for Human communicative Behavior analysis (CVPR4HB)*, 2009.

[146] C. H. Teh and R. T. Chin, "On image analysis by the method of moments," vol. 10, no. 4, pp. 496–513, 1988.

[147] T. Troscianko, A. Holmes, J. Stillman, M. Mirmehdi, D. Wright, and A. Wilson, "What happens next? The predictability of natural behaviour viewed through CCTV cameras," *Perception*, vol. 33, no. 1, pp. 87–101, 2004.

[148] B. Tseng, C.-Y. Lin, and J. Smith, "Using MPEG-7 and MPEG-21 for personalizing video," *IEEE Multimedia*, vol. 11, no. 1, pp. 42–52, 2004.

[149] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 18, no. 11, pp. 1473–1488, 2008.

[150] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time bag of words, approximately," in *Proceeding of the ACM Int'l Conference on Image and Video Retrieval*, ser. CIVR '09, 2009.

[151] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Tilburg University, Tech. Rep. TiCC-TR 2009-005, 2009.

[152] 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.

[153] J. Varadarajan and J.-M. Odobez, "Topic models for scene analysis and abnormality detection," in *Proc. of IEEE Int'l Conference on Computer Vision Workshops (ICCV Workshops)*, 2009.

[154] 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. [Online]. Available: http://www.openvisor.org

[155] ——, "Video surveillance online repository (ViSOR): an integrated framework," *Multimedia Tools and Applications*, vol. 50, no. 2, pp. 359–380, 2010.

[156] P. A. 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.

[157] F. Wang, Y.-G. Jiang, and C.-W. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. of ACM Int'l Conference on MultiMedia (MM)*, 2008.

[158] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. of IEEE British Machine Vision Conference (BMVC)*, 2009.

[159] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition." in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[160] O. Werner, "Requantization for transcoding of MPEG-2 bit streams," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 179–191, 1999.

[161] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[162] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of IEEE European Conference on Computer Vision (ECCV)*, 2008.

[163] 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.

[164] S.-F. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. of IEEE Int'l Conference on Computer Vision (ICCV)*, 2007.

[165] 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.

[166] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[167] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893–908, 2008.

[168] 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.

[169] J. Yang and A. G. Hauptmann, "Exploring temporal consistency for video analysis and retrieval," in *Proc. of Int'l Workshop on Multimedia Information Retrieval (MIR)*, 2006.

[170] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. of Int'l Workshop on Multimedia Information Retrieval (MIR)*, 2007.

[171] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[172] 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.

[173] J. Yin and Y. Meng, "Abnormal behavior recognition using self-adaptive hidden markov models," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, vol. 5627, pp. 337–346.

[174] Youtube.com. (2011, May) Youtube press statistics. [Online]. Available: http://www.youtube.com/t/press_statistics

[175] G. Yu, N. Goussies, J. Yuan, and Z. Liu, "Fast action detection via discriminative random forest voting and top-k subvolume search," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 507–517, 2011.

[176] ——, "Fast action detection via discriminative random forest voting and top-k subvolume search," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 507–517, 2011.

[177] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345–357, 2008.

[178] D. Zhang, D. Gatica-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.

[179] 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.

[180] 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 Int'l Conference on MultiMedia (MM)*, 2008.