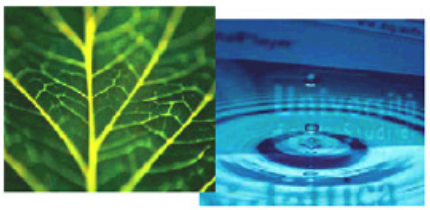
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Multimedia Content Analysis for Event Detection

Andrea Rosani

Advisor:

Prof. Francesco G.B. De Natale

Università degli Studi di Trento

March 2015

... to my wife Maria Stella and my daughter Letizia

Abstract

The wide diffusion of multimedia contents of different type and format led to the need of effective methods to efficiently handle such huge amount of information, opening interesting research challenges in the media community. In particular, the definition of suitable content understanding methodologies is attracting the effort of a large number of researchers worldwide, who proposed various tools for automatic content organization, retrieval, search, annotation and summarization. In this thesis, we will focus on an important concept, that is the inherent link between "media" and the "events" that such media are depicting. We will present two different methodologies related to such problem, and in particular to the automatic discovery of event-semantics from media contents. The two methodologies address this general problem at two different levels of abstraction. In the first approach we will be concerned with the detection of activities and behaviors of people from a video sequence (i.e., what a person is doing and how), while in the second we will face the more general problem of understanding a class of events from a set visual media (i.e., the situation and context). Both problems will be addressed trying to avoid making strong a-priori assumptions, i.e., considering the largely unstructured and variable nature of events.

As to the first methodology, we will discuss about events related to the behavior of a person living in a home environment. The automatic understanding of human activity is still an open problems in the scientific community, although several solutions have been proposed so far, and may provide important breakthroughs in many application domains such as context-aware computing, area monitoring and surveillance, assistive technologies

for the elderly or disabled, and more. An innovative approach is presented in this thesis, providing (i) a compact representation of human activities, and (ii) an effective tool to reliably measure the similarity between activity instances. In particular, the activity pattern is modeled with a signature obtained through a symbolic abstraction of its spatio-temporal trace, allowing the application of high-level reasoning through context-free grammars for activity classification.

As far as the second methodology is concerned, we will address the problem of identifying an event from single image. If event discovery from media is already a complex problem, detection from a single still picture is still considered out-of-reach for current methodologies, as demonstrated by recent results of international benchmarks in the field. In this work we will focus on a solution that may open new perspectives in this area, by providing better knowledge on the link between visual perception and event semantics. In fact, what we propose is a framework that identifies image details that allow human beings identifying an event from single image that depicts it. These details are called "event saliency", and are detected by exploiting the power of human computation through a gamification procedure. The resulting event saliency is a map of event-related image areas containing sufficient evidence of the underlying event, which could be used to learn the visual essence of the event itself, to enable improved automatic discovery techniques.

Both methodologies will be demonstrated through extensive tests using publicly available datasets, as well as additional data created ad-hoc for the specific problems under analysis.

Keywords [event detection, activity recognition, behavior analysis, anomaly detection, context-free grammars, regular expressions, gaming, saliency]

Acknowledgment

There are many persons that I would like to thank, for different reasons.
I would like to express my gratitude to the staff of the Multimedia Signal Processing and Understanding research group, in particular to Prof. Francesco De Natale, Dr. Giulia Boato and Dr. Nicola Conci, who helped me a lot in these past years, making it possible to restart my study activity after some years of work, opening new possibilities for the future of my job.
A particular thanks goes to the people I met, who helped me with their comments and suggestion, or just sharing the time of this experience, in particular Duc Tien, Valentina, Paolo, Cecilia, Krishna, Mattia, Alfredo, and Nicola.
A huge hug goes to my family, to my wife Maria Stella and my daughter Letizia, for their unconditioned support and love. To my Dad, who watches over me from the "high mountains". To my Mum, my sister Laura and her family. To Margherita who helped us a lot.
Last but not least I want to say thank you to all the people I met during these years, because everyone gave some contribution to this trip, making it somehow more interesting.

Contents

List of Tables

List of Figures

Introduction

Multimedia data gained an important role in recent years, thanks also to the advances in communications, computing and storage technology. The potential of multimedia became fundamental in improving the processes in different fields like surveillance, wearable computing, biometrics, and remote sensing, but also in advertising and marketing, education and training, entertainment, medicine.

The wide diffusion of multimedia content has evidenced new requirements for more effective access to global information repositories. Content analysis, indexing, and retrieval of multimedia data are one of the most challenging and fastest growing research areas. A consequence of the increasing consumer demand for multimedia information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data. In particular, there is the need of robust techniques to index/retrieve and compress multimedia information, new scalable browsing algorithms allowing access to very large multimedia databases, and semantic visual interfaces integrating the above components into unified multimedia browsing and retrieval systems.

As an example, the wide diffusion of video surveillance systems generated a huge amount of video data to be processed, with different application spanning from security to ambient assisted living. Automatic recognition of human activities and behaviors is still a challenging problem due to many reasons, including limited accuracy of the data acquired by sensing devices,

high variability of human behaviors, gap between visual appearance and scene semantics.

One of the major concept shared by different multimedia analysis frameworks is the description of events represented in such data, in order to provide information about the content of the media. In fact, whatever they are personal experiences, such as a wedding or a birthday, or large social happenings, such as a concert or a football match, events mark our lives and memories. Moreover, event recognition is a crucial task to provide high-level semantic description of the video content. The bag-of-words (BoW) approach has proven to be successful for the categorization of objects and scenes in images, but it is unable to model temporal information between consecutive frames.

In this work we will discuss about the problem multimedia content analysis addressing in particular two sub-problems: the modeling and matching of activities for human behavior understanding from video; the content mining for social event detection from still images.

1.1 Activity modeling and matching for human behavior understanding from video

Being able to understand human activities and behaviors is a key feature in the field of ambient intelligence [43, 50]. *Activities* can be defined as the concatenation of atomic actions that produce voluntary human body motion patterns of arbitrary complexity, describing *what* elements compose an event [99]. *Behaviors* instead relate human activities with the surrounding environment (people, objects, situations), inferring *how* and *why* a certain situation is occurring [9]. A *behavior* can be seen as the response of a human to the internal, external, conscious, or unconscious stimula he receives [20]. While the recognition of *activities* is *syntactic*,

as it can be typically associated to a sequence of characteristic elements, *behaviors* imply a joint analysis of *content* and *context*, thus providing a semantically richer description of the event. For example the activity of "running" is considered a natural behavior on a soccer field, while it would be reported as suspicious if detected inside a bank office.

Among the different approaches proposed in literature to detect activities and behaviors [26], video analysis is often preferred because of its limited cost of installation and maintenance, and lower obtrusiveness. Video data contain a lot of significant information to infer human behaviors, including location, posture, motion, as well as interaction with objects, other people, and the environment [10]. In this context, motion patterns are probably the most popular descriptor for many reasons: they are rather easy and fast to calculate, robust, and especially, they can be captured even in far range and from different perspectives, where posture analysis or object detection may fail [84] [42].

The motion pattern of a moving object (often associated to its trajectory) is defined as the spatio-temporal evolution of one or more feature points extracted from the visual sequence. In particular, when tracking humans, a convenient representation of the motion trajectory is the one that maps the centroid of the bounding box retrieved by the tracking algorithm on the ground plane, which becomes the reference system.

The research in this area has been very active in the past decades, and very efficient detectors and tracking algorithms have been proposed [69]. Usually, the output consists of a trajectory T, namely a raw set of coordinates (Fig. 1.1 associated to a temporal reference for each moving target in the scene, as recalled in (1.1), where $P_i = (x_i, y_i)$ and i is a frame counter.

$$T = \{P_i, t_i\};\ i = 0...I \tag{1.1}$$

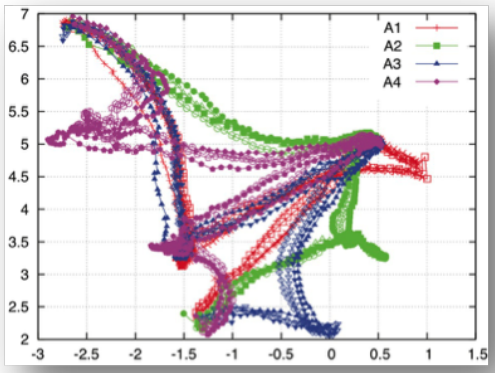


Figure 1.1: Examples of variability in tracks.

Although very basic, this representation makes it possible to perform a simple yet effective low-level classification of the incoming samples by matching them with a set of pre-stored templates [83] [95]. These approaches provide good results especially when the environment is known and motion patterns are rather constrained, such as in vehicular traffic monitoring. Instead, when the variety and diversity of patterns is higher, the performance of low-level analysis drops [55], due to the increasing noise and uncertainty in the numerical representation that hinders a reliable matching. Human behaviors fall into this category of events, because of their high variability and largely unconstrained nature.

In such situation, it is necessary to improve the raw samples description by introducing a symbolic representation [85][23]. This can be achieved in many different ways, including trajectory segmentation, down-sampling, quantization or approximation, where the common underlying objective is to achieve a new representation, in which symbols are more expressive and less noisy. Furthermore, symbolic representations can be easily complemented with additional attributes, including motion features (e.g., direction, speed, acceleration), and environmental information (e.g., proximity

to key points or objects). They can be easily made invariant to translation, rotation, and scaling [69]. This makes symbolic approaches suitable for both activity detection, where the syntax of symbol chains is considered, and behavior analysis, where additional attributes may be exploited to interpret more sophisticated scenarios.

In this work we propose a framework for human behavior analysis in indoor environments using Context-Free Grammars (CFGs). Activities are modeled considering the prominent areas of the environment visited by the subject. Compared to other existing approaches that use CFGs for activity modeling and matching, our method provides several important novelties. First of all, it classifies the events considering both positive and negative samples, thus ensuring a better separation of the classes while maintaining good generalization properties. Second, we introduce a retraining procedure, in order to update the grammar rules in presence of changes in the environmental setup or in the users' habits. Finally, we introduce the capability of dealing with concatenated or nested actions. According to the definition provided at the beginning of the section, we will use the term *behavioral* analysis not only to refer to the mere detection of the activity performed by the individual (what), but also to consider the semantic connotation of such activity (why and how).

1.2 Content mining for social event detection

Recent studies demonstrate that users find it easier to search and browse media archives when they are organized according to underlying events [108]. Many works propose the use of event models to enable efficient media indexing and retrieval (see, e.g., [66][64][96]), and some interesting prototypes have also appeared [1][3].

In this context, an interesting yet still open problem is how to capture

the relationships between media and events: is it possible to automatically discover events from media? What visual cues allow humans understanding which event a media depicts? Answering these questions would open great opportunities for improving media archiving systems by enabling faceted search, event-media networks, event summarization, storytelling, etc. As an example, given a photo collection a system could recognize the underlying event (e.g., a wedding), cluster pictures according to the relevant event structure (e.g., ceremony, party, cake cutting, etc.), associate appropriate tags to each picture, select representative images for the event, and so on.

Although the scientific literature reports several interesting ideas, many problems remain unsolved mostly due to the heterogeneity, multi-modality and unstructured nature of the data [82]. The report on the Social Event Detection (SED) task of the MediaEval benchmarking initiative [2] suggests that current technologies for media event detection, although interesting from a scientific viewpoint, are still inadequate for potential commercial exploitation. Current research efforts are mainly focused on defining the best possible features to describe an image, the most appropriate strategies to learn such representations from groundtruth data, and adequate matching procedures to find revealing patterns in unknown data. Little attention has been put on understanding which are the key elements that allow a human observer recognizing an event when looking at a set of media or even a single picture that depicts it. The question is: would it be possible to mimic such human comprehension mechanisms on a computer? And especially, would it be possible to understand which distinctive patterns enable such comprehension? In this work we introduce a new concept that we call *event saliency*. Event saliency refers to the event-specific visual contents of an image, i.e., the parts of the image that allow an observer associating it to a given event category with high confidence. It is to be pointed out that event saliency is a very different concept from the tra-

Figure 1.2: Which is the event associated to this picture? Visual elements that allow answering the question have been removed, making it almost impossible to guess. We define such information *event saliency*, to distinguish it from the classical *visual saliency*.

ditional *visual saliency*. In fact, visual saliency typically highlights the part of a picture that grabs users' attention at a first glance [34]. This is generally connected to the brightness of colors, the contrast, the position, the prominence and, in general, the image syntax. On the contrary, event saliency captures the picture areas that are related to the event independently of their visual prominence, and is therefore concerned with image semantics. Event saliency may include part of the background, a peripheral image area, or a small but revealing detail. The idea is illustrated in Fig. 1.2, where the black box exactly hides the detail that would make possible recognizing the event.

In order to demonstrate the concept of event saliency, we address the problem of detecting it from images. To this purpose, we propose a tool to extract event saliency maps from event-related images by exploiting crowd intelligence through games. The ultimate goal is to generate a groundtruth to be used for further studies, where the application of the event saliency concept may empower existing event detection approaches or allow defining innovative approaches.

The choice of performing this task through *gamification* (i.e., the use of game mechanics and game design techniques in non-game contexts) was driven by the consideration that this strategy has been proven to be successful in solving complex problems that require human intervention [104]. With respect to crowdsourcing, a well-studied gamification approach is more engaging and entertaining, thus attracting more users with higher commitment and less bias. In this work we show how a carefully designed game allowed creating a significant collection of accurate event saliency maps out of a dataset of representative images associated to a set of common events. To achieve this goal, we involved a large community of users in an adversarial game, where the real objective (producing the maps) was hidden.

In Ch. 3 we introduce the *event saliency*, as the collection of perceptual elements contained in an image that allow humans recognizing the depicted event. Furthermore, we propose *EventMask*, a GWAP conceived to detect event saliency in event-related pictures.

1.3 Structure of the Thesis

The remainder of this work is the following. In Ch. 2 we will introduce and detail our solution for activity modeling and matching, based on the use of Context Free Grammars, while in Ch. 3 we will introduce and analyze a gamification able to provide the relevant part inside an image, related to events. Finally in Ch. 4 conclusions will follow.

Context-Free Grammars for Activity Modeling and Matching

Automatic recognition of human activities and behaviors is still a challenging problem due to many reasons, including limited accuracy of the data acquired by sensing devices, high variability of human behaviors, gap between visual appearance and scene semantics. Symbolic approaches can significantly simplify the analysis, turning raw data into chains of meaningful patterns. This allows getting rid of most of the clutter produced by low-level processing operations, embedding significant contextual information into the data, as well as using simple syntactic approaches to perform the matching between incoming sequences and models. In this context we propose a symbolic approach to learn and detect complex activities through sequences of atomic actions. Compared to previous methods based on Context Free Grammars (CFGs), we introduce several important novelties, such as the capability to learn actions based on both positive and negative samples, the possibility of efficiently re-training the system in the presence of misclassified or unrecognized events, the use of a parsing procedure that allows correctly detecting the activities also when they are concatenated and/or nested one with each other. Experimental validation on three datasets with different characteristics demonstrates the robustness of the approach in classifying complex human behaviors.

The objective of this chapter is to define how to exploit a symbolic representation of the motion patterns associated to a person moving in a

known indoor environment, in order to acquire knowledge about his/her behavior. This information is very important in situation where there is the need to monitor person's activity, like in home care (e.g. fall detection), as well as when the system should be able to rise up an alert depending on a substantial difference in users habits, like in video-surveillance (e.g. robber detected using abnormal behavior). We will achieve this goal by modeling human motion patterns through Context-Free Grammars (CFGs). It will be demonstrated that the proposed strategy allows not only to acquire and recognize the examples provided during the training phase, but also to generalize them, thus being able to detect instances of the activities that have not been included in the training set.

2.1 Background

High-level reasoning based on symbolic representations of the scene under investigation could provide effective results in behavioral analysis. Some of the most significant approaches proposed in this context are summarized in the following paragraphs.

A typical way of introducing higher-level interpretation is to extract features from low-level data and feed them into a probabilistic model that can statistically describe the event structure. The work proposed by Duong et al. in [35] implements a strategy to learn and recognize human activities through a Switching Hidden Semi-Markov Model. The authors propose to adopt a two-layer representation, in which the bottom layer defines the atomic activities through a sequence of concatenated Hidden Semi-Markov Model; the upper layer is then used to handle the temporal structure of activities, composing the event by means of a sequence of switching variables. Similarly, the authors of [76] propose a variant of a Hidden Markov Model (HMM) to exploit both the hierarchical structure and the shared seman-

tics contained in the motion trajectories, introducing a Rao-Blackwellised particle filter in the recognition process to achieve real-time performances. Through this approach, the actions of a subject are learned from an unsegmented training set.

The authors in [59] propose a scalable approach that includes two major modules: a low-level action detector to process low-level data using a Dynamic Bayesian Network (DBN), and a Viterbi-based inference algorithm used to maintain the most likely activity given the DBN status and the output of the low-level detectors.

The main advantage of these methods is in the capability of handling the uncertainties generated during the low-level processing. On the other hand, as the event complexity increases, the recognition performance dramatically drops, due to a combination of factors including insufficient training data, semantic ambiguity in the model, or temporal ambiguity in competing hypothesis. Although some methods for unsupervised parameter estimation of the graphical model have been proposed (see [19]), the major problem remains the definition of the network topology, which is usually too complex to be learned automatically, requiring the help of human operators.

Another category of approaches performs activity recognition in a symbolic domain, introducing an intermediate layer between low-level feature extraction and high-level reasoning. Low-level primitives are processed using HMM-like approaches, while high-level behavior modeling is based on the Context-Free Grammars formalism [47]. Ivanov and Bobick [48] proposed a two-stages strategy. In the first phase, candidate features for low-level temporal domain are extracted and considered as "signatures". As far as the matching strategy is concerned, Stochastic Context-Free Grammars (SCFG) are adopted, providing longer range temporal constraints, disambiguating uncertain low-level detections, and allowing the inclusion

of a-priori knowledge about the temporal structure of events.

In [67] a system is proposed to generate detailed annotations of complex human behaviors performing the Towers of Hanoi through a parameterized and manually-defined stochastic grammar. In [68] the authors also use SCFGs to extract high-level behaviors from video sequences, in which multiple subjects can perform different separable activities. An alternative approach is proposed in [93]. Here, the so-called attribute grammars [56] are employed as descriptors for features that can not be easily represented by finite symbols.
A common drawback of the systems relying on formal grammars is in the definition and update of the production rules. In fact, an exhaustive formalization and structuring of the observable activities that a person can perform in everyday life is not practical [74]. For this reason, in [45] a computational framework is proposed, able to recognize behaviors in a minimally supervised manner, relying on the assumption that everyday activities can be encoded through their local event subsequences, and assuming that this encoding is sufficient for activity discovery and classification.

Another major limitation of SCFG-based systems is that the parsing strategy can handle only sequential relations between sub-events, being unable to capture the parallel temporal relations that often exist in complex events. To overcome this issue, the authors of [112] propose to derive the terminal symbols of a SCFG from motion trajectories. In particular, they transform them into a set of basic motion patterns (*primitives*) taken as terminals for the grammar. Then, a rule induction algorithm based on the Minimum Description Length (MDL) derives the spatio-temporal structure of the event from the primitive stream.

In a recent work in this area [28], the authors employed an induction algorithm called *EMILE* [6], originally used for Natural Language Processing (NLP) applications. Here, each sentence (i.e., each symbolic sequence) is

iteratively decomposed in *expressions* and *contexts*. Intuitively, given the entire set of training sentences, the algorithm searches for frequent combinations of *expressions* and *contexts*, and interprets them as a grammatical type.

However, the main drawback of this approach is that the generalization properties of the grammar cannot be controlled during training. The more diverse are the examples proposed in input, the larger becomes the final set of patterns that satisfies the grammar. In fact, part of these patterns do not belong to the training, but arise from generalization. It is then possible that unwanted expressions satisfy the resulting grammars .

Moreover, given two grammars generated from disjoint training sets, it is not guaranteed that their overlap is null, implying that, due to generalization, it is not possible to guarantee the separation of languages. If the grammars are used to classify the symbol strings, this means that there will be a subset of strings that will fit multiple classes.
A very detailed overview about the literature in the field can be found in [7]. For the sake of completeness, we report hereafter a short summary about the most relevant benefits of CFG-based approaches in activity and behavior modeling compared to other competing algorithms [110]:

- ability to model the hierarchical structure of events, which is difficult to capture with graphical models such as HMM;
- ability to take into account temporal relationships, so that long-term activities can be considered;
- capability of describing sequential features, resulting in a more efficient representation if compared to that obtained via bag of features;
- richness in semantics;

- simpler understanding compared to other knowledge driven models based on ontology (VERL [39], VEML [75]).

2.2 Motivations

The proposed framework stems from a recent work in this area [28] and extends it by introducing a more sophisticated learning strategy. In fact, behavior classes are in general not well separated, especially in the case of indoor or home monitoring, due to the high variability of human behaviors. Fig. 2.1 provides an example, referring to two different home activities, each one associated to an ideal model (the trajectory described by the solid line). Dashed lines report the actual behavior of two subjects performing the same actions in real life. It is possible to observe that these behaviors show non-negligible spatio-temporal differences compared to the model, leading to potential errors when using simple approaches based on the matching with pre-stored templates.

Results have been presented in [88] [89], [90], exploiting the properties of a Context Free Grammar originally developed for Natural Language Processing [98]. In these cases, grammars tend to overlap, and, for a single activity, multiple grammar rules may return a positive match. In the current approach, instead, we define a methodology to overcome the drift problem, by adopting a learning strategy that considers both positive and negative examples, and introducing a re-training stage, so as to improve the accuracy of the detection.

It will be shown that this procedure can avoid the overlapping of classes while learning the models, allowing a better generalization, and maintaining a good separation among them.
Furthermore, the proposed method is well suited to incremental learning. In fact, it does not require the storage of the original training set, but can

simply extend the knowledge of the system by feeding additional samples validated by the user into the learning procedure. In this way, false and missed alarms can be progressively learned in order to increase the accuracy of the detector, as well as adapting to changes in both environmental conditions and users' habits.
Finally, the proposed method allows processing the video stream in real time, as soon as the motion patterns are available, as it behaves like a symbolic parser [73], [11]. Thanks to the implemented parsing strategy, the method is also able to handle complex situations that typically degrade the performances of traditional matching tools, such as the presence of concatenated or nested activities, namely, when an action is partially or totally executed within another one [36].
It is worth noting that the algorithm does not impose any specific technology for data acquisition, which can be performed though various positioning devices (video tracking, sensor networks [101], RFIDs, etc.), thus providing a completely customizable solution for indoor monitoring.

2.3 Proposed Framework

To cope with the issues mentioned in the previous section, the proposed method operates a significant simplification of the observed domain, associating symbols only to a limited set of points of interest in the environment, called *Hot Spots*. Human actions are then described in terms of time-ordered sequences of such symbols. The obtained sequences are learned and recognized through Context-Free Grammars.

The most important steps of the proposed method can be summarized in the following points:

1. *pre-processing* of the incoming paths and conversion into the symbolic domain;

2. *learning* of the grammar sets that encode the rules for each set of training patterns;

3. *classification* of the incoming trajectories into the available rules, performed through parsing;

4. *update* of the grammar rules according to user feedback.

In this section, after briefly introducing the CFG formalism, we will provide a detailed description for each of the above mentioned items.

2.3.1 Context Free Grammar formalism

According to grammars theory, a set of strings over a finite set of symbols is defined as a *language*. A grammar is a tool that allows specifying which strings belong to a specific language.
A Context-Free Grammar (CFG) is defined as [72]:

$$G = (N, T, P, S) \tag{2.1}$$

where N is a finite set of *non-terminal* symbols, T is a finite set of *terminal* symbols ($N \cap T = 0$), P is a finite grammar of the form $A \rightarrow u$ ($A \in N$ and $u \in (N \cup T)^+$), and S is the starting symbol ($S \in N$).

The set P derives a string of terminal labels w from a *non-terminal* symbol A, if there is a derivation tree $A \rightarrow w$ with root A [72]. A language $L(G)$ of a CFG G is the set of all strings derived from the starting symbol S. $L(G)$ is called *ambiguous* if there are two or more derivations of the same string. In the proposed framework we will use *unambiguous* CFGs, i.e., there will be a unique derivation for the considered string.

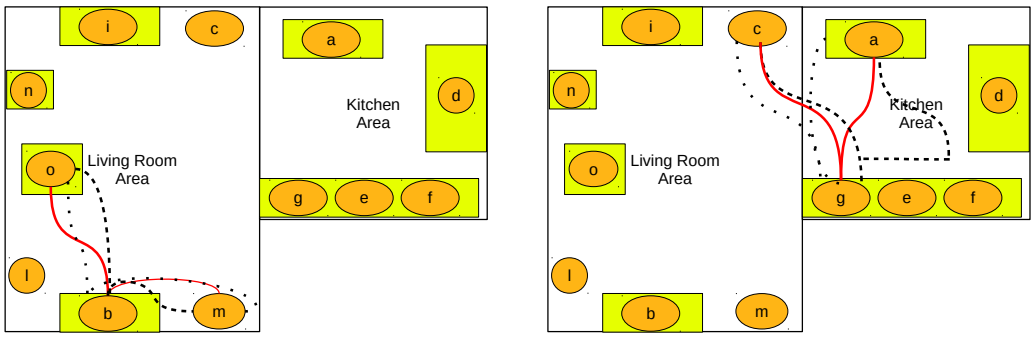


Figure 2.1: Examples of variability in human actions for two different activities: left image represent a sample for the action "Have a rest" and the right a sample of "Cooking". Solid lines indicate the reference path, while dashed lines refer to different subjects performing that action in real-life. a, b, c, etc. represent the labels for Hot Spots.

2.3.2 Activity representation

In our approach we convert human motion patterns into temporal concatenations of *Hot Spots* that the user has visited in a given time frame. We say that the user has visited a *Hot Spot* if he/she has been in the proximity of it for a predefined temporal interval. Both proximity and visit time of each Hot Spot are application-driven and should be defined according to the environment and based on the user's habits. As an example, operating a given appliance in a kitchen may request some time; a shorter stop at that location may therefore have a different meaning. Furthermore, older people may need longer time to perform the same action. Similarly, more stringent spatial requirements are necessary when moving in smaller rooms, compared, for example, to large exhibition areas. This allows simplifying the representation of (1.1) to a stream of symbols, each one associated to a pair (region-index, time-stamp) as in:

$$T' = \{R_j, t_j\};\ j = 0...J \tag{2.2}$$

where R_j is the index of the *Hot Spot* and t_j is the temporal reference. Describing the path in terms of a sequence of *Hot Spots* rather than sampling it at fixed time intervals provides a twofold advantage: (i) it reduces noise and outliers in the trajectory caused by limited accuracy in acquisition and/or tracking, and (ii), it generates a simpler representation that preserves the significant spatio-temporal evolution of the activity, while making more tractable the next processing steps.

2.3.3 CFG Rules Discovery

Grammatical inference is a discipline related to a large number of fields, including machine learning and pattern recognition. It basically consists in feeding data into an entity, the *learner*, which returns a grammar capable of *explaining* it [32]. If we want our grammar to learn a particular concept associated to a given set of symbolic patterns, we should provide it as input to the *learner*, which will return as output a set of grammar rules that generate a language. If the input patterns are characterized by a certain level of diversity, i.e., diverse instances of the same concept, the grammar will learn all these possible variations, but will also provide a certain degree of generalization. As an example, let us consider the following input patterns:

- *(ab)*;
- *(abab)*;
- *(aabb)*.

These patterns, can be exhaustively synthesized using the following grammar rules:

$P_{example_1} = S \rightarrow ab; S \rightarrow aS; S \rightarrow bb; S \rightarrow SS;$

It can be easily seen that all three original strings satisfy this grammar. For instance, also the strings *(aab)*, *(ababab)*, and *(abbb)* are part of the language, although not present in the initial set of strings.

This problem is known in literature [28], and can be partially solved by introducing the possibility of providing counterexamples during the training phase, by feeding the *learner* with both positive and negative examples, where negative examples can be used to separate the languages and to eliminate false positives.

To achieve this goal, we use a different learning algorithm, able to: (i) exploit both positive and negative samples as well as additional production rules in training, and (ii) allow the incremental learning of Context-Free Grammars.

For the first grammar generation our algorithm receives in input a labeled set of positive and negative samples, and builds a grammar P such that all the strings labeled as positive, and no string labeled as negative, can be derived from P. Given a set of behaviors to be classified b_j, $j = 1, ..., J$, and a set of observed behavioral patterns t_k, $k = 1, ..., K$, $K \gg J$, we create J clusters I_j such that:

$$I_j = I_j^{PS} \cup I_j^{NS} \tag{2.3}$$

where $I_j^{PS} = [t_k : t_k \in b_j]$; $I_j^{NS} = [t_k : t_k \notin b_j]$ are the positive and negative samples of the j-th behavior as classified by a supervisor. The grammar generation produces therefore a set of J grammars P_j, such that each grammar will fulfill I_j^{PS} and not I_j^{NS}.

The procedure described above for grammar generation allows sharply reducing the overlap of the generated languages. However, it does not guarantee that all strings resulting from the language (and not included in the training set) belong to a unique language. In such situation the pattern is not classified and will be possibly used for the re-training procedure.

At any point of the operation of the classifier, when a certain number of patterns t_k^* have been stored, for which the classification was not successful (either satisfying more than one grammar or none of them), the user may decide to run a re-training. In this case, the supervisor is again requested to manually classify the critical samples. Then, a new set of clusters I_j^* is produced, and a new set of grammars P_j^* is generated based on I_j^*. It is to be observed that the re-training starts from the previous grammars P_j and processes the new samples only. Therefore the updated grammars P_j^* will respect the rule $P_j \in P_j^*$.
In summary, the key features of the proposed system, as compared to other grammar-based behavior learning tools, are:

- CFGs are generated from positive and negative samples (possibility to limit grammar generalization and overlap);
- incremental learning of CFGs (possibility to easily do re-training, adding false positives and missed alarms in the training set as soon as an expert recognizes them).

The learning process described so far is illustrated by the following example, where we represent the positive and negative samples by the unit clauses of the form $ps(w)$ and $ns(w)$, where w is a string represented by an atom or a list of atoms [72]. The terminal symbols in the list are restricted to atoms other than $p, q, r, ...z$, which are used for non-terminal symbols. The symbol S is the starting symbol and the generated grammar is an unambiguous CFG in extended Chomsky normal form.

Let us consider two sets of *positive* samples only:

$Set_1 = ps(ab); ps(aa); ps(abab); ps(aabb); ps(ababab); ps(aababb); ps(aaab); ps(abaabb);$
$ps(aabbabab); ps(aabbaabbab); ps(aabaabbabb);$
$Set_2 = ps(ba); ps(aaa); ps(abb); ps(abba)$

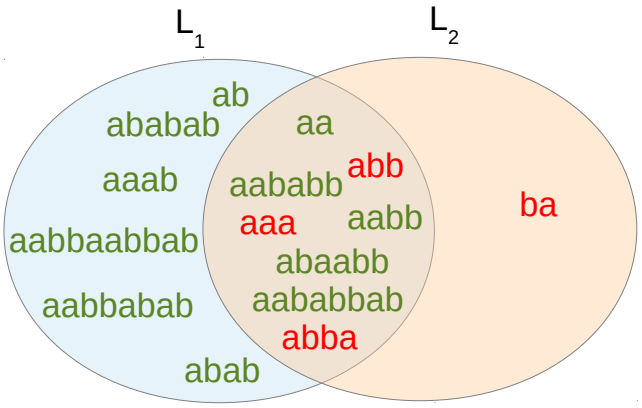


Figure 2.2: Intersection between languages $L_1 = L(P_1)$ and $L_2 = L(P_2)$ generated by positive samples only.

Our algorithm, according to the description provided in the previous paragraphs, generates the following grammars:

$P_1 =$	$S \to aa; S \to ab; S \to Sa; S \to Sb; S \to bS$
$P_2 =$	$S \to aa; S \to aS; S \to ba; S \to bb; S \to bS$

As can be seen in Fig. 2.2, the two grammar rules P_1 and P_2 generate two languages such that $L_1 \cap L_2 \neq 0$ and, in particular, some of the training samples belong to both languages. By introducing the use of *negative samples*, we can overcome this situation, forcing the separation of the training samples. To this purpose we built Set'_1 containing Set_1 as positive samples and Set_2 as negative, and viceversa for Set'_2.

The new grammar rules P'_1 and P'_2 have been derived accordingly.

Still, we are not sure that the intersection between the two new languages L'_1 and L'_2 is empty. For example the new atom $(aaaa)$ satisfies both grammars, thus resulting in an ambiguity (Fig. 2.3). To cope with

$Set'_1 = ps(aa); ps(ab); ps(abab); ps(aabb); ps(ababab); ps(aababb); ps(aaab); ps(abaabb);$
$ps(aabbabab); ps(aabbaabbab); ps(aabaabbabb);$
$ns(ba); ns(aaa); ns(abb); ns(abba)$
$Set'_2 = ps(ba); ps(aaa); ps(abb); ps(abba);$
$ns(aa); ns(ab); ns(abab); ns(aabb); ns(ababab); ns(aababb); ns(aaab); ns(abaabb);$
$ns(aabbabab); ns(aabbaabbab); ns(aabaabbabb);$

$P'_1 = \; p \to aS; p \to bS; S \to aa; S \to ab; S \to ap; S \to bb; S \to bp$
$P'_2 = \; p \to aa; p \to bb; S \to ap; S \to ba; S \to Sa$

this, we apply the update procedure described above. The misclassification of the atom $(aaaa)$ will be considered initially as an error. In a second stage, it will be prompted to an evaluator that classifies it. The grammars are updated accordingly. Adding *(aaaa)* as a positive sample of Set_2, will lead, for example, a the new set of rules.

$P''_1 = \; p \to bS; q \to aS; S \to aa; S \to ab; S \to ap; S \to qb; S \to qp$
$P''_2 = \; p \to aa; p \to bb; p \to ab; S \to ba; S \to Sa$

2.3.4 Parsing the CFG

An appropriate parsing procedure has been defined in order to check the compliance of the input strings with the generated grammar P. The parser receives in input the symbols corresponding to the *Hot Spots* visited by the person in the monitored environment. Symbols are progressively stored in a buffer, which initial length is equal to the maximum length (K) of the grammar words used in the learning stage. The parser reads the symbols in the buffer, calculating all possible combinations, without repetition, of the considered string, until they reach the minimum possible dimension.

The number of combinations N_k to be considered for a string of length $1 \leq k \leq K$ can be computed as:

$$N_k = \binom{K}{k} = \frac{K!}{k!(K-k)!} \tag{2.4}$$

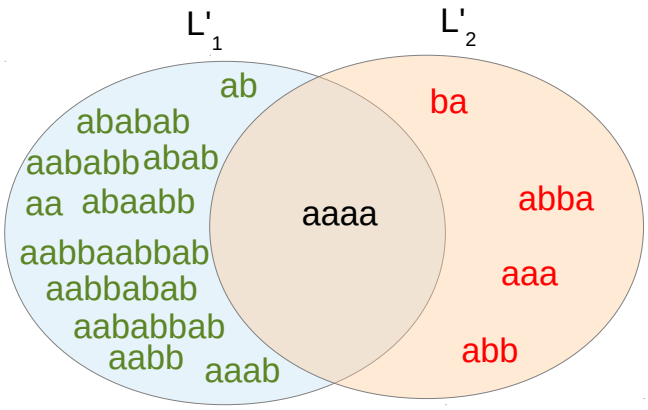


Figure 2.3: Intersection between languages $L'_1 = L(P'_1)$ and $L'_2 = L(P'_2)$ generated by positive and negative samples

In this way, we ensure that in presence of complex actions, the parser is able to detect also nested and concatenated subsequences, which are removed from the input pattern as soon as they are associated to an action. Then, the parsing can proceed on the remaining symbols in the stack.

When an activity is detected, the associated symbols are removed from the buffer, new symbols are added to the parsing string until the buffer is filled, and the process iterates.

The symbols remaining in the buffer can represent either actions that have not been learned by the system (for instance, a new or a rare behavior) or anomalous patterns, possibly generated by noise. They can be signaled as errors or anomalies, or can be stored for successive learning phases (e.g., personalizing on a given user's behavior).

The update procedure is needed to maintain a coherent model for the learned activities, given that potential modifications in terms of scene arrangements or users' habits may occur.

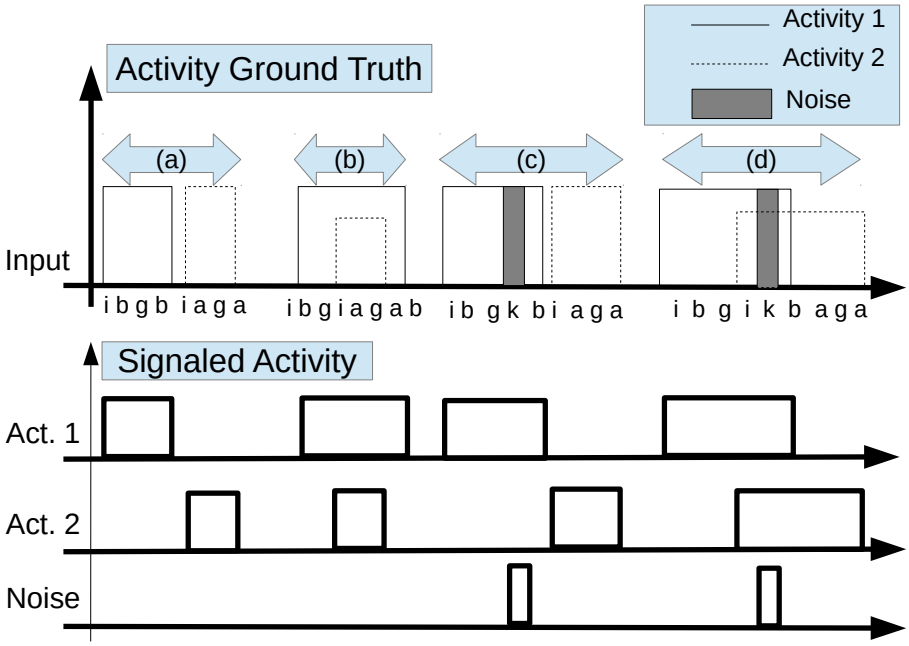


Figure 2.4: Activity spotting examples: (a) 2 consecutive sequences; (b) hierarchy between two activities; (c) two nested activities with noisy symbols; (d) two overlapping activities with noisy symbols.

For a better understanding of the parsing procedure we present a test to demonstrate the capability of the parsing strategy in spotting known activity patterns from a continuous event stream. In particular, we show how the proposed engine can recognize activities also in a concatenated and nested form. To this aim, we randomly selected some activity instances from our database and composed them in different configurations, as shown in Fig. 2.4. Activity 1 (Act. 1 in the figure) is represented by the string *(ibgb)* and Activity 2 by *(iaga)*. Noise is represented by the symbol *(k)* and represent an outlier in the sequence of *Hot Spots*. We consider the following situations: consecutive activities (a); nested activities (b); consecutive activities with noise (c); and overlapping (interleaved) activities in presence of noise (d). From the top: (i) the ground truth for the activ-

ity stream with the corresponding sequence of hot spots; (ii) the signaled activities; and (iii) detected noise patterns. As can be seen from the figure, and thanks to the parsing strategy, the system is able to disclose chunks of activities even if the incoming data stream is corrupted by noise.

2.4 Datasets for Human behavior analysis

Given that the proposed method uses the motion trajectory just to detect the proximity of the subject to the hot-spots identified in the environment, in principle any sensor providing such information is viable for our purposes. Examples of devices that can provide the requested information include video cameras, but also active and passive RFIDs, WSNs, acoustic sensors, etc. Adopting any combination of such sensors would also make it possible to provide more reliable estimates, reducing problems caused by occlusions, presence of multiple subjects, and so on.
In our experimental validation we have considered three different datasets, based on different sensor systems and application scenarios.

The first dataset, known as the "Ubicomp dataset" [101], is considered a benchmark in the area. It is composed by a set of action-related data collected by 14 state-change sensors installed in a home environment where a single person lives. Data have been acquired over 28 days, and annotation has been manually provided by the person living in the environment, distinguishing among seven activities (see Table 2.2), chosen on the basis of the so-called Katz ADL index [53]. The outcome of this process is a set of 2120 sensor events with 245 activity instances. Results presented in [101] and [100] use standard probabilistic graphical models for action recognition, in particular Hidden Markov Models and Conditional Random Fields, thus allowing comparison with state of art action recognition methodologies.

The second and third datasets have been build in our research labs, using the facilities available in two domains: assisted living and video-surveillance. Both datasets use visual information to track the subjects and extract the positioning information. The former refers to the "Home Dataset" and has been collected in a realistic domestic environment designed for testing ambient-assisted-living technologies (Fig. 2.5). The dataset includes a total of 81 trajectories equally divided into 3 classes: A) *Cooking*, B) *Eating*, C) *Taking a break*, and executed by 9 volunteers performing the same activity for 3 times in slightly different ways.
The latter, called "Office Dataset", has been recorded in an office environment, using multiple cameras (Fig. 2.6). The dataset includes 120 paths divided into 4 classes: A) *Arrival*, B) *WorkTime*, C) *Have a break*, D) *Print*.

Similarly to the previous one, it has been performed by 10 volunteers doing the same activity for 3 times in slightly different ways.
In both cases, the video streams have been processed by a motion tracker to extract the top-view trajectories of the moving subjects.

2.5 A Context-Free Grammar behavior analysis tool

In this section we present and discuss the results obtained by our method on the datasets described in Section 2.4.

Considering that the tracker is out of the scope of this work, the experimental validation only concerns the behavior analysis module.

In the following sub-sections we present first the results achieved on the Ubicomp dataset, introducing a comparison of our methodology with most of the common state of art learning strategies in the field; then, we present the results on the two visual datasets, where additional tests are proposed to show the specific features of the proposed method.

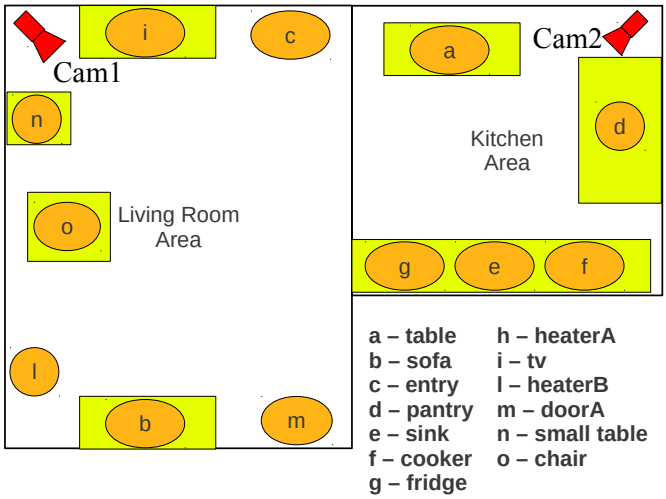


Figure 2.5: Top: map of the home environment and camera positions. *Hot Spots* are provided with the corresponding legend; Bottom: views of the environment from the two installed cameras (Cam1 left, Cam2 right).

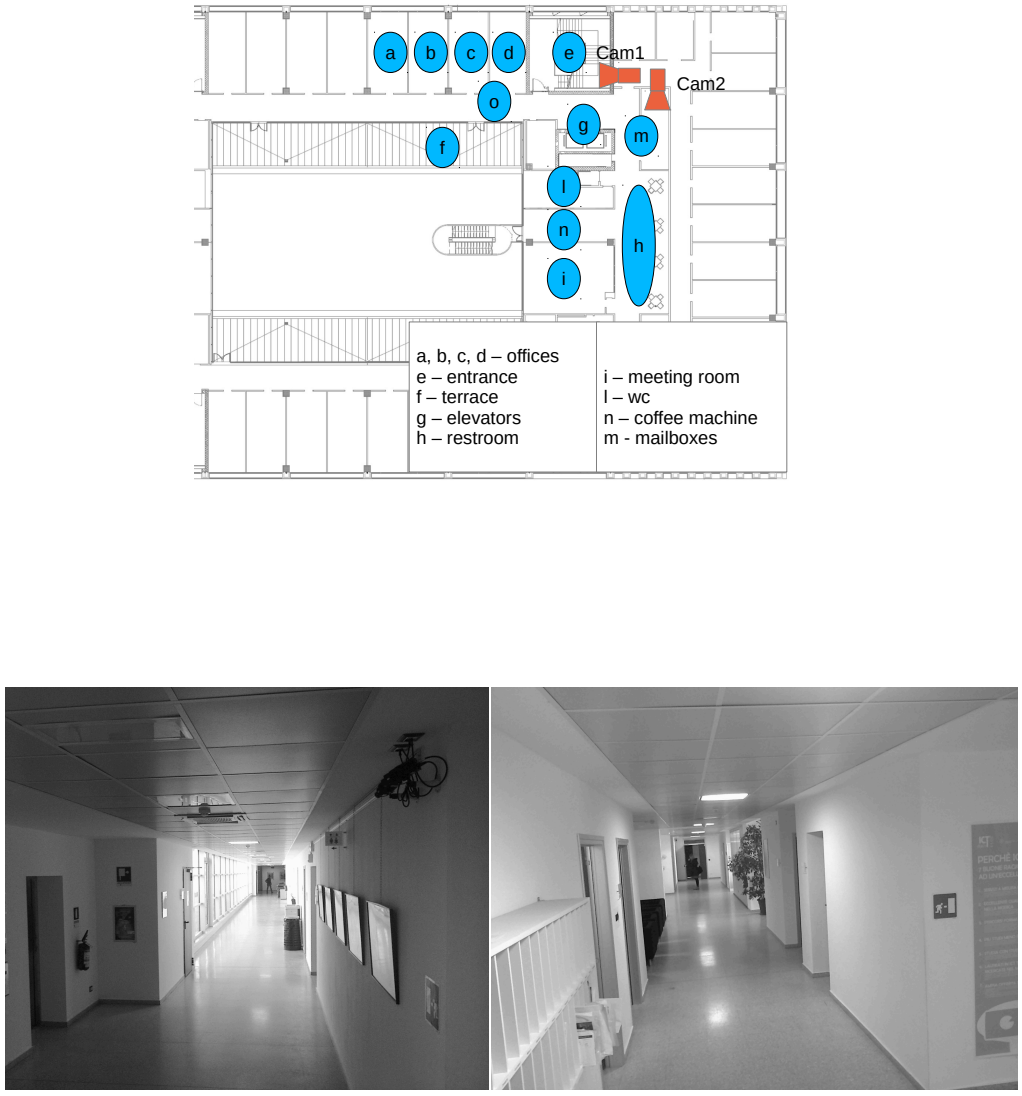


Figure 2.6: Top: map of the office environment and camera positions. *Hot Spots* are provided with the corresponding legend; Bottom: views of the environment from the two installed cameras (Cam1 left, Cam2 right).

2.5.1 Ubicomp dataset

In order to make it suitable to be fed into our framework, we had first to characterize each activity as a sequence of sensor events.

To this purpose, we mapped the 14 sensors assigning an identifier to each of them (Fig. 2.7) according to the indication in Table 2.1, so that each action is described by the sequence of sensor events generated within the action time-slot.

As an example, we report hereafter the sequence of sensor events for the action "Sleeping", according to the notation used in our framework:

$$
\begin{aligned}
Set_{Sleeping} = {} & ps(nn); ps(nnn); ps(bnni); ps(nninnn); \\
& ps(nncbbn); ps(bnccnc); ps(bnnibnn); \\
& ps(innccnn); ps(bncicnn); ps(cnnccnn); \\
& ps(bnncicnn); ps(nnncicnn); ps(nnnncicnn); \\
& ps(nnccnccnin); ps(nncnincicn); \\
& ps(nnncicnncicnn); ps(bnncicinncnicn); \\
& ps(nnbnncibbiciiicbn); \\
& ps(nbccbnnnbnbbbbibbbnncicncnncnn);
\end{aligned}
$$

Starting from this description, using as positive samples the strings for the considered action and as negative samples the ones of all the other actions, we can generate the set of grammars required to classify each action. As an example, we show in the following the grammar for the action "Sleeping":

$$
\begin{aligned}
P_{Sleeping} = {} & S \rightarrow bn; S \rightarrow bS; S \rightarrow cn\ S \rightarrow cS; S \rightarrow iS; \\
& S \rightarrow nc; S \rightarrow ni; S \rightarrow nn; S \rightarrow nS
\end{aligned}
$$

To make the results comparable to the ones presented in [101] and [100], we adopted the "leave one day out" validation proposed by the authors.

According to this rule, we have separated the test and training sets using

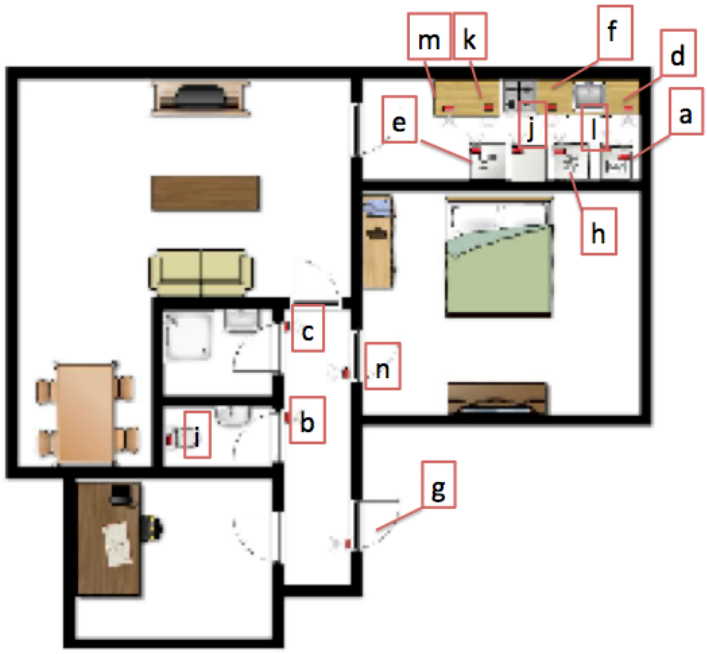


Figure 2.7: Ubicomp dataset. Map of the considered environment, with sensor labels.

Table 2.1: Sensor remapping.

Position	Letter assigned
Microwave	a
Hall-Toilet door	b
Hall-Bathroom door	c
Cups cupboard	d
Fridge	e
Plates cupboard	f
Frontdoor	g
Dishwasher	h
ToiletFlush	i
Freezer	j
Pans Cupboard	k
Washingmachine	l
Croceries Cupboard	m
Hall-Bedroom door	n

for testing one full day of sensor readings and for training the remaining days. We have iterated this process so that each of the days contained in the dataset has been considered for testing.

Table 2.2: Confusion Matrix using the proposed framework on the *"Ubicomp dataset"*. The values are percentages

	Leaving	Toilet.	Shower.	Sleeping	Breakfast	Dinner	Drink	N.C.
Leaving	87.88	9.09	0.00	0.00	0.00	0.00	0.00	3.03
Toilet.	1.77	89.38	0.00	1.77	0.00	0.00	4.42	2.65
Shower.	0.00	4.55	95.45	0.00	0.00	0.00	0.00	0.00
Sleeping	4.76	0.00	4.76	80.95	0.00	0.00	0.00	9.52
Breakfast	0.00	0.00	0.00	0.00	89.47	0.00	5.26	5.26
Dinner	0.00	11.11	0.00	0.00	0.00	33.33	11.11	44.44
Drink	0.00	11.11	5.56	0.00	0.00	0.00	72.22	11.11

The obtained averaged results are presented in Table 2.2. Comparing the above results with the ones shown in [101], it is possible to observe that our framework provides a similar accuracy in classifying the actions "Leaving", "Toileting", "Showering", "Sleeping" and achieves better results for the actions "Breakfast", and "Drink". The action "Dinner", instead, shows a limited accuracy, mostly due to the fact that it has the lowest number of occurrences, thus resulting in a very limited training set for rules generation. Besides this very appreciable result, we have also to stress that the proposed method presents the advantage of real-time operation and very low hardware requirements, since the run-time process only consists of a simple symbolic parsing. Moreover in Table 2.3 we compare our framework to the results shown in [100], where the same authors propose a set of measures considering different action recognition models that can be taken as a benchmark for new action recognition algorithms. In particular Precision and Recall are presented along with F-Measure and Accuracy (please see [100] for the corresponding definitions). The models considered in the work span from a "Naive Bayes" approach to Hidden Markov Mod-

els (HMM), Hidden Semi-Markov Models (HSMM), Conditional Random Fields (CRF). For comparison purposes our approach is reported as the last one.

Table 2.3: Comparison of obtained results respect [100].

Model	Precision	Recall	F-Measure	Accuracy
Naive Bayes	67.3 ± 17.2	64.8 ± 14.6	65.8 ± 15.5	95.3 ± 2.8
HMM	54.6 ± 17.0	69.5 ± 12.7	60.8 ± 14.9	89.5 ± 8.4
HSMM	60.2 ± 15.4	73.8 ± 12.5	66.0 ± 13.7	91.0 ± 7.2
CRF	66.2 ± 15.8	65.8 ± 14.0	65.9 ± 14.6	96.4 ± 2.4
Our Framework	89.9 ± 10.9	78.3 ± 19.6	83.7 ± 14.0	85.5 ± 11.8

As can be seen, the proposed framework obtained significant improvements in terms of Precision, Recall and F-Measure. The accuracy is still high, although slightly lower compared to the other models. This is mainly due to the fact that our approach produces some unrecognized actions, which limit the accuracy parameter. This fact can be partially recovered by the re-training procedure, as will be shown in the next paragraphs.

2.5.2 Home environment

As far as the "Home Dataset" is concerned, we first randomly divide each set of examples in the in 2 parts, a training and a test set, each one containing one half of the samples provided for each behavior. Based on the training set we generate 3 grammars, one for each action. Then, we apply the grammar to classify the test set. In order to cross-validate the results, every experiment is repeated 10 times with different random partitions, and the classification results are averaged. Cross-validation is applied to all the tests reported in the following.

In the first test we consider positive samples only in the grammar generation. Results are reported in Table 2.4.

From the confusion matrix we can observe that part of the patterns are

Table 2.4: Classification accuracy using positive samples only.

In/Out	A	B	C	$Unknown$
A	0.54	0.15	0	0.31
B	0.08	0.54	0	0.38
C	0	0.15	0.62	0.23

misclassified and about one third of the patterns are classified as unknown, meaning that they are not recognized as valid patterns by the parser. In the second experiment we introduce negative samples in the grammar generation, where for each class, the positive samples are the same as above, while the negative ones consist of the training of the other classes. Consequently, each of the 3 grammars is generated from 14 positive and 28 negative examples. Results are reported in Table 2.5.

Table 2.5: Classification accuracy considering positive and negative samples.

In/Out	A	B	C	$Unknown$
A	0.69	0	0	0.31
B	0	0.62	0	0.38
C	0	0	0.77	0.23

It can be observed that negative samples allow a better discrimination, i.e., removing the overlap among the three classes, as clearly shown by the new confusion matrix. On the contrary, the number of unrecognized samples remains unchanged, as the negative samples just restrict the region associated to each class.

In order to improve the detection of unrecognized samples, incremental learning can be used. For this test, the dataset is divided into 3 subsets of 9 instances each. We generate the grammars from the first subset using positive and negative samples (9 and 18, respectively). Then, we classify the second subset, and we select the faulty patterns (misclassified + non

recognized samples). These patterns are then used as additional training samples to update the grammars according to the proposed re-training procedure. Finally, the last subset is used for testing. The average results obtained after the first and second learning stage are reported in Tables 2.6 and 2.7, respectively.

Table 2.6: Average classification accuracy after one learning stage.

In/Out	A	B	C	$Unknown$
A	0.63	0	0	0.37
B	0	0.55	0	0.45
C	0	0	0.55	0.45

Table 2.7: Average classification accuracy after re-training.

In/Out	A	B	C	$Unknown$
A	0.74	0	0	0.26
B	0	0.69	0	0.31
C	0	0	0.67	0.33

Comparing the two tables one can observe that the performance considerably improves after re-training, leading to an average 70% of correct classification. It is also to be noted that the last result is better than the one shown in Table 2.5, although the total number of samples presented to the grammar generation tools is slightly lower on average. In fact, in the former, one half of the samples were used for training (14 per class), while in the latter, the average number was 13 (9 initial + 4 in re-training).

2.5.3 Office environment

The experiments for this last dataset have been performed in the same way as for the previous case. In this case we have 4 behavior classes and 30 samples per class. The first test is performed using 15 patterns per class

for training, based on positive samples only. The test was performed on the remaining 15 samples per class. The confusion matrix is reported in Table 2.8.

Table 2.8: Classification accuracy using positive samples only.

In/Out	A	B	C	D	$Unknown$
A	0.46	0.07	0	0	0.47
B	0	0.40	0	0.13	0.47
C	0	0	0.53	0	0.47
D	0.07	0	0	0.40	0.53

Also in this case, we tested the grammar generation tool adding negative examples, thus using 15 positive and 45 negative samples per class. Results are reported in Table 2.9.

Table 2.9: Classification accuracy considering positive and negative samples.

In/Out	A	B	C	D	$Unknown$
A	0.53	0	0	0	0.47
B	0	0.53	0	0	0.47
C	0	0	0.53	0	0.47
D	0	0	0	0.47	0.53

Finally, re-training is simulated splitting the dataset in 3 equal parts (10 samples per class), and using the first subset for initial grammar generation, the second subset for the first test and the re-training, and the last subset for final test. The intermediate and final results are reported in Table 2.10 and Table 2.11, respectively.

It can be observed that the figures are consistent with what has been presented in the previous case, with a slightly larger performance gap between the first test (single training, positive samples only) to the final one (after re-training).

Table 2.10: Average classification accuracy after one learning stage.

In/Out	A	B	C	D	$Unknown$
A	0.55	0	0	0	0.45
B	0	0.50	0	0	0.50
C	0	0	0.48	0	0.52
D	0	0	0	0.58	0.42

Table 2.11: Average classification accuracy after re-training.

In/Out	A	B	C	D	$Unknown$
A	0.73	0	0	0	0.27
B	0	0.71	0	0	0.29
C	0	0	0.82	0	0.18
D	0	0	0	0.78	0.22

2.6 Real time behavior analysis in compressed domain

In the context of human behavior analysis applied to a real scenario, an innovative solution based on a real time analysis of video with application in the field of fall detection for elderly care is presented in the following paragraph. The system performs anomaly detection and proposes the automatic reconfiguration of the camera network for better monitoring of the ongoing event. The developed framework is tested on a publicly available dataset and has also been deployed and evaluated in a real environment. Algorithms for fall detection operate in many cases in the pixel domain, whereas most of the surveillance cameras only provide the video in the compressed domain. In order for these algorithms to be applied, the video has to be decoded, introducing an additional processing layer. Furthermore, most algorithms are not operating in real time, barely reaching 20-25 frames per second on a PC-based platform, which hampers the ability

of their deployment in real scenarios. In order to respond to this need, especially in case of elderly care, it is necessary to develop low-complexity algorithms, which can be deployed directly in the DSP (Digital Signal Processor) onboard of the camera and possibly in the compressed domain, thus dropping the need for decoding. In this work we present an algorithm which completely operates in the compressed H.264 [109] domain and that requires a negligible complexity, hence it can be deployed on DSP (or similar) processor. Fall detection and reconfiguration is achieved by proposing a generic entropy measure derived using the distribution of the motion field extracted from the compressed video bit stream [57].

2.6.1 Evaluation

In order to demonstrate the utility and robustness of the algorithm, we first evaluate the performance of the fall detection algorithm by testing it against the reference fall detection dataset published by the University of Montreal [8], widely used to validate algorithms in this field.

To show the reconfiguration capability, we deployed a set up in a real environment and observe its performance during the occurrence of fall. To this extent we used two cameras "Sony SNC-EP521 indoor", day/night, with PTZ. These IP cameras are equipped with a 36x optical zoom allowing operators to cover large, open areas and zoom in for detailed close-up shots. Panning can span from 0 to 340 degrees, with max 105 degrees tilt, and their configuration can change using built in network commands. The cameras have been installed in our Department facility, and falling events have been recorded thanks to the collaboration of volunteers.

Fall detection

Since the algorithm operates in the compressed domain, we had to convert all the videos in the dataset [8] into the H.264 format using the JM H.264

reference encoder [46], at the frame rate of 25 frames per second. The thresholds necessary for a proper operation of the algorithm are learned for each camera and are maintained constant for that particular camera for all scenarios. Fall is defined as an event lasting 5-10 seconds, starting from the momentary stop by the subject just before the fall and ending with a motion less layover of the subject. The total number of correct fall detections, as compared to the ground truth, are deemed as true positives (TP), while false detections are termed as false positives (FP). Finally, true falls which have been skipped by the detector are termed as false negatives (FN). The results obtained for the video dataset are given in Table 2.12 in terms of Precision, Recall and F-Measure. A comparison with respect to the state of the art techniques is provided in 2.13. As can be seen, the fall detection algorithm performs reasonably well especially given the fact that it operates in real time. The algorithm fails to detect the falls, when the subject is very far away from the camera and subsequently the motion entropy generated by the subject is very low. In such scenario noise becomes dominant thereby causing false detections. Another scenario where the algorithm fails is in case of actions, which correspond to bending down on the floor etc. However, since we also took into consideration the momentary fall entropy, just after the fall most of such false detections have been resolved.

Table 2.12: Performance of the algorithm on the dataset [8].

Precision	Recall	F-Measure
0.89	0.86	0.88

Table 2.13: Comparison to the state of the art approaches described in[37].

	Our method	K-NN	C4.5	SVM	Bayes	Feng et. all
Sensitivity	0.86	0.75	0.85	0.95	0.80	0.98

Comparison

Our algorithm completely operates in the compressed domain. Hence it has the advantage of being very light in terms of computational and memory requirements. Nevertheless it compares very well with the other pixel domain state of the art fall detection methods as we can see from the table 2.12. Our method also provides a significant improvement with respect to other compressed domain methods like [25]. Most of these methods rely on the segmentation of moving object and the trajectory of its centroid, and also include other features like velocity of centroid. Present algorithm also uses these aspects, but it turns out to be more robust as it also exploits the motion disorder as one of the factors to determine fall detection. Furthermore, the compressed domain method presented in [25] uses AC and DC coefficients along with motion vectors to achieve object segmentation, which are heavily dependent on the quantization parameter used for encoding the video bit stream. The proposed method, instead is entirely based on motion vectors, which are independent with respect to changes in QP. In terms of complexity our solution offers the lowest complexity of all compressed domain methods as it operates at the level of 32×32 blocks, and the number of operations required for processing one frame are 5.2K, 16K, 48K, 106K computations for CIF, VGA, HD, full HD resolutions, respectively.

Reconfiguration

In case of real evaluation the video stream obtained from the camera has a resolution of 720×576 pixels and a frame rate of 25 frames per second. The H.264 bit stream obtained from the camera is encoded in the baseline profile. In order to access the Network Abstraction Layer (NAL) packets from the camera we have used the functions available in the *ffmpeg* library[79]. Fall detection and moving object segmentation are implemented using the

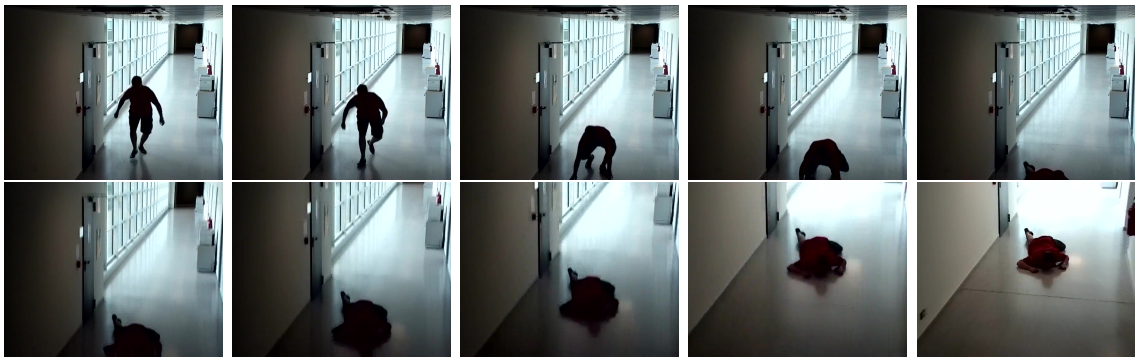


Figure 2.8: Fall detection and subsequent reconfiguration of the camera for better view.

motion vectors extracted from the H.264 (JM 18.6 version) decoder [46]. In order to control the camera automatically the *curl* library functions [78] are adopted. The whole set up is implemented on an Intel i5 processor, 3.10 GHz.

Fall detection and subsequent reconfiguration is shown in Fig. 2.8. As we can see from the images, fall of the person occurs towards the end of the image in one of the frames. However, camera instantly reconfigures to bring back the view of the fallen person. This shows that the algorithm works in real time and is robust enough to work in tricky illumination conditions.

Event identification using Games With a Purpose

The concept of "event" emerged in the last years as a key feature to efficiently index and retrieve media. Several approaches have been proposed to analyze the relationship between events and related media, enable event discovery, perform event-based media tagging, indexing, and retrieval. Despite the outstanding work done by several researchers in this area, a major problem that still remains open is how to infer the inherent link between visual concepts and events. In particular, the possibility of understanding which perceptual elements allow a human recognizing the event depicted by an image, would for sure open new directions in event media discovery. In this work we introduce the concept of Event Saliency to define the above event-revealing perceptual elements, and we propose an original method to detect it by exploiting crowd knowledge through gamification. We propose an adversarial game with a hidden purpose, where users are engaged in two competitive roles: on one side they should mask a photo collection to prevent the competitors recognizing the related event; on the other side they should discover events masked by other players. A set of rules and an adequate score system avoid cheating and force players to focus on details that really matter, thus allowing the accurate detection of event-related contents in media. A suitable algorithm composes the masks produced by different players on the same media, taking into account the results of the game. The final result is a saliency map that, differently from the

traditional concept of saliency, does not focus on perceptual prominence but rather on event-related semantics of media. Results of EventMask are collected in a publicly available dataset which can be exploited for further research in this domain. In this chapter we will introduce the analysis of images, and in particular we will address the problem of content mining for social event detection, starting from the information coming from another multimedia content respect to the previous section, i.e. galleries of still images. The objective is to provide a framework able to create the necessary groundtruth to automaticaly classify a single image respect to the represented event.

3.1 Background

In this section we review the state of the art in two areas that are strictly connected to our work: event-based media analysis and gamification for media analysis. A specific sub-section is dedicated to each of these aspects.

3.1.1 Event-based media analysis

Events provide a rich source of contextual information that can be exploited to address a number of different tasks in multimedia signal processing and analysis. Since the pioneering work of Ramesh Jain [108], many researchers have investigated the relationships between events and associated media, to solve problems like event discovery, automated media event clustering and summarization, event-based media retrieval and event networks. For a thorough review on the subject please look at [62] and references thereby. These researches prove that events provide a rich contextual information, which can be exploited to achieve better media indexing and retrieval. Also, media contain important traces of the underlying events that can be exploited for classification purposes. For instance, in [31] a

photo collection is clustered based on the implicit event structures and the emerging event fingerprints are extracted to eventually discover the type of the related event. [41] proposes a similar idea but with the ambitious objective of determining the event on the basis of a single media item, using a visual concept vector. In practice, event classes are learned with a Mixture Subclass Discriminant Analysis and a nearest neighbor criterion is used to associate the media to an event class. The main problem with the above method, as well as with other approaches that use visual information only, is the dependency on visual concept detectors, which still perform poorly. Multi-concept detection can partially solve this problem by finding evidence of a high-level concept, e.g., an event, from the joint presence of multiple visual concepts even with low individual accuracy. In this case, the joint weak detection of multiple related concepts reinforces the higher-level classification [97].

Another common solution is to use additional data, whenever available, to complement visual information. Several works attempt to exploit image annotations and tags to collect information about time, spatial coordinates, keywords in a multimodal analysis framework. For instance, in [60] events are detected from photo collections by analyzing user-supplied tags. In [29] event taxonomies are automatically extracted from annotated media, while personal photo galleries are organized via event clustering techniques in [64] and [27]. Wider information is elaborated in [86], by focusing on the domain of pictures and extracting event and place semantics from tags assigned to Flickr photos. In [80] a set of pictures is used to produce two image similarity graphs, one using visual features and the other using textual features, and then combine them in a single hybrid similarity graph. A number of features including time, location and textual information are exploited in [14], to face this issue as an unsupervised clustering problem. In [15] the same authors designed a two-step method to first cluster a

Twitter stream and then perform event vs. non-event clustering. Event detection in social media (e.g., Facebook and Twitter) is also studied in [16] and [54].

Also social ties can be considered an important source of information. A new concept of social interaction is defined in [107], where social affinity is computed via a random-walk on a social interaction graph to determine similarity between two pictures. In [38] the authors propose to use the social information produced by users in the form of tags, titles and photo descriptions for classifying photos in event categories. In [21] various informations (e.g., time, location, textual and visual features) are combined within a framework that incorporates external data sources from datasets and online web services. In [61], the authors exploit geo-tagging information retrieved from online sources to determine the bounding box for a set of venues, while using time information to determine the set of events that can be compared to those occurred at the examined venue. In [82] a multimodal clustering approach is proposed, which predicts the same cluster relationship by exploiting pairwise similarities for all different modalities and achieving supervised fusion of the heterogeneous features. The social event detection is transformed into a watershed-based image segmentation in [30] and [77], where visual and non-visual information are jointly exploited. A fully automated system for event recognition from an image gallery has been recently proposed in [111], by exploiting metadata information. In [22] authors identify, retrieve and classify photos in collaborative web photo collections associated with social events, by using contextual cues and spatio-temporal constraints.

In recent years, the importance of event-based media analysis and social media in general has been witnessed by the large participation of the research community to international challenges proposed around the event detection tasks by TRECVID [4] and MediaEval [2], which made also avail-

able to the research community various annotated datasets, suitable for training, testing and comparing different methodologies.

With special reference to the area of event media analysis, the main contribution of this paper is in the definition of the concept of event saliency, as a way to bring some light on the perception mechanisms that allow human beings understanding events when looking at representative images. In particular, what we aim to do is to highlight on an image the revealing visual contents with respect to the underlying event. It is easy to imagine how this result could open new directions in the framework of the above referenced event media analysis techniques, making it possible to focus the attention on important visual concepts and their relationships, while getting rid of irrelevant visual information that may mislead the analysis.

3.1.2 Gamification in media analysis

Human computing is becoming a common solution to face complex or extensive problems, where traditional computer-based approaches fail. Crowdsourcing is an interesting solution in this domain, based on fragmenting a task into a large number of sub-tasks and involving large communities of workers to solve every micro-task for a relatively small individual reward. Although very interesting in many respects, current crowdsourcing technologies suffer a number of problems. In particular, it is widely accepted that crowdsourced tasks have to be rather simple and fast to perform, there should be no dependency on each other, and there's a risk connected to the reliability of the work performed. This implies a careful design of the tasks and not all the problems are suitable to be defined in an appropriate way.

More recently, gamification emerged as an alternative modality of crowd-based problem solution. Luis von Ahn coined the term Games With A Purpose (GWAP) [104] to define a particular type of crowdsourced game

that is fun and engaging, and includes a task that can only be completed by humans. Many different game mechanisms comply with the above characteristics. A first class is based on *output agreement*, where users have to collaborate to reach a consensus. An example is the ESP Game for collaborative labeling of images [103], where players try to label a given image while playing in pairs: the goal of the game is to agree on as many tags as possible that describe the given image; the two players that convene on the maximum number of keywords win. The hidden goal in this case is to tag images according to the content, for further use in image storage and retrieval. A second class is based on *input agreement*. An example is TagATune [58], where players are given inputs and are prompted to produce descriptive outputs, so that their partners can asses whether their inputs are the same or different. Another example is WhoKnows? [106], a game whose purpose is to detect inconsistencies in Linked Data and score properties to rank them for sophisticated semantic-search scenarios. A third class of GWAPs introduces the so called *inversion problem*, where the problem is posed indirectly through a "double negation" criterion. Peekaboom [105] is a nice example of this type of games, aimed at supporting the creation of metadata associated to visual objects contained in images. It is played in couples in an adversarial way: the first player is given an image and a keyword related to it, and progressively reveals the image part related to the keyword until the other player guesses the object. Recently, a game called Bubbles [33] has been proposed for selecting discriminative features for fine-grained categorization, where users can reveal small circular areas of blurred pictures to inspect details and guess the represented subject.

Still pictures are not the only target of media analysis gamification. Yahoo's Video Tag Game [102] evolves previous approaches for collaborative tagging by introducing the time variable. Users are not trying to tag the

video as a whole, but different fragments of it, building a larger and richer set of metadata for that video content. In the context of event recognition in videos, a very recent work [18] proposes the new concept of minimally needed evidence to predict the presence or absence of an event in a video, which allowed improving event retrieval performance on two challenging datasets released under TRECVID [52][51]. Notice that video saliency is in general quite different from image saliency, since motion patterns tend to assume a prominent role in the perception.

InfoGarden [63] is a casual game, transforming document tagging into an activity like weeding a garden and protecting plants from gophers, designed to extend the willingness to maintain personal archives by enhancing the experience of personal archive management. Games used for the evaluation and curation of the underlying data are effectively incentives-driven tools, as discussed in [94], employing ease of use, fun and competition as incentives for users to perform what would otherwise be an unrewarding and demanding manual tasks. An overview about the application of game mechanics in information retrieval can be found in [40].

The social dimension is also very important in games. *Social gaming* [5] has been one of the emerging trends in the last years, attracting both research and industrial efforts. As an example, the evolution of social networks such as Facebook or Twitter has introduced content generation as a mainstream concept, forcing users to continuously produce and consume contents. Geopositioning games, such as Foursquare[1], Gowalla[2], Buzzd[3] or Facebook Places[4] encourage users to do check-ins indicating where they are, and thus letting the system extract information about their location, top visited places, typical routes. Collabio [17] is a Facebook application

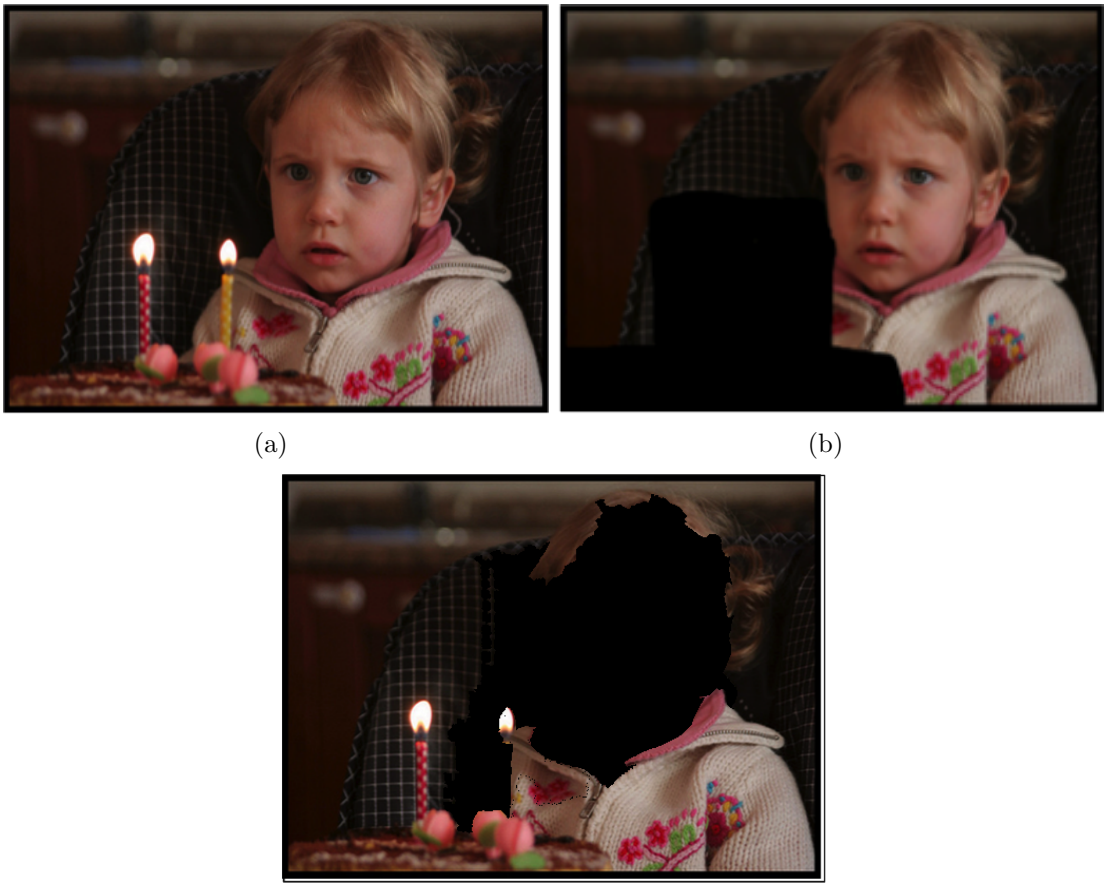
[1]http://foursquare.com/
[2]http://gowalla.com/
[3]http://www.buzzd.com/
[4]http://www.facebook.com/places/

that allows friends to tag each other with descriptive terms through a game. In GuessWho [44] users enter knowledge about their peers to enrich the organizational social network: each player is prompted with the name of a person and is asked to provide either names of people who are related to that person or tags that describe him/her. CityExplorer [65] combines social gaming with geospatial data gathering to set particular location of places and landmarks. Pirates! [12] encourages users to collaborate in the mapping process of WiFi networks available in the surroundings. Recently, a gamification approach that moves away from thinking of gamification as an "additive" process towards a more "holistic" paradigm was proposed in [49]. In this context, a novel definition was proposed that addresses gamification as a complete system in itself, positioning it as the process of adding an actionable layer of context.

For what concerns gamification, this work introduces a novel adversarial game concept, which allows obtaining cross-validated event saliency maps by combining the results of several users playing in different roles on the same images. Several interesting strategies have been considered into the game, from the "target inversion", to the introduction of negative scores originated by inter-user competition, to the timing of the game. The combined effect of such mechanisms was to ensure accurate completion of the hidden task, avoiding cheating and bias, and incentivizing users to keep on playing. Besides the specific application of the game in the event saliency context, we believe that EventMask could provide significant clues on how to design a game with a hidden purpose to gain knowledge about complex analysis tasks in multimedia.

3.2 Event Saliency

Let us consider the image in panel (c) of Fig. 3.1, which shows the same photo of Fig. 1.2 covered with a different mask. It is now rather easy to understand which event it depicts. In fact, a revealing element (the typical birthday pie with the candles) is now clearly visible. Moreover, this visual element suggests a number of additional information that in some way disclose also the covered parts. For instance, we can assume



(a) (b)

(c)

Figure 3.1: Comparison between event saliency and visual saliency. Panel (a) represents the reference image; panel (b) has the same image where event salient areas are covered by a mask, resulting in a not intuitive association of the image to a specific event; panel (c) shows a second version of the masked image, where visual salient areas are covered: despite this, details crucial for the event recognition are still visible.

that the baby is a girl, because of the color of the candles, and that it is her second birthday. Incidentally, the mask used in this example has been obtained with a well-known saliency detector [71][70], thus underlining the low relatedness of visual saliency with event semantics. If we are able to cover all the image areas that are revealing the event and only those, so that the concealed image cannot be anymore associated to the correct event while minimizing the coverage, we have defined what we call an *event saliency map*. The event saliency is largely independent of the color and contrast of the relevant visual elements, and can be made of any number of disjoint elements, with arbitrary shape and in any position over the image. In fact, revealing information can be hidden everywhere in the picture: in the background, in a small detail, or even in a low contrast or blurred area. For this reason, it is difficult to imagine automatic techniques able to extract it from a picture with the currently available technologies. On the contrary, the ability of human beings at recognizing events even from a small detail is incredibly high. For this reason we decided to use human interaction to gain knowledge about this complex problem. It is important however to notice that the overall purpose of the proposed method is not to directly involve people in event detection, while to learn from people the event semantics contained in visual media, i.e., what really matters in images to detect the subjacent event. In this sense, our method provides as output a groundtruth that can be used as a basis to implement better event detection and event-based image classification algorithms, or to improve existing ones.

3.2.1 EventMask

Among the various possibilities of involving people in this task, we decided to use gamification for various reasons. With respect to expert-based approaches, gamification provides the possibility of involving large numbers

of users, thus covering cultural, personal and social diversities, and preventing biases due to the expert's individual experience. With respect to crowdsourcing-based approaches, we reduce cheating by exploiting the intrinsic incentive of game playing, and we have the possibility of better hiding the real purpose of the task, thus avoiding biases.

EventMask is designed as a GWAP where the real purpose of the game is disjoint from the apparent goal, and hidden to the players. In the game each user can play multiple roles, without the need of another user on-line, thus resulting in more simple usage and easier user engagement. The game is competitive, in the sense that each player has to compete off-line with others to both increase their score and cut the scores of the competitors. This provides also the reward mechanism, which incentivizes using the game. Another characteristic of EventMask is that it is formulated as an *inversion problem*: we do not ask people to do what we expect as a result of the game, but to negate the contrary. In the specific context of EventMask, this means that we never ask a player to highlight what is important for him/her to recognize an event, but rather to conceal it to other people, so that they cannot recognize the event itself. This is very important for two reasons: (i) it prevents users' bias due to their personal experience (if I have to indicate what is important for me, I will probably focus on a limited number of details that I believe are the most significant due to my own experience of that event); and (ii) it obliges a user to detect all the revealing details (if I have to hide an event, I will proceed initially with prominent details, but after covering those, I will focus the attention of the remaining parts, thus discovering additional traces, until all the revealing details are covered).

The following sections provide the description of the game mechanics and of the processing applied to the data generated by the players. This includes the description of the overall structure of the game, the players

roles and rules, the scoring system, and the production of results. It is important to point out that all these mechanisms have been carefully defined in order to stimulate fair and correct playing, avoid cheating, introduce mutual control to avoid external moderation, reveal the performance and reputation of players, motivate users.

Game Mechanics

The basic idea of the game is simple. EventMask is an adversarial game, in which users are alternately engaged into two competitive roles: masking and discovery. In the *masking* role the user is presented a random image related to an event, and is requested to hide (cover with uniform color using a simple painting tool) the minimum part of it that makes the event unrecognizable. An honest player will correctly hide the event, while leaving the highest possible area of the image visible (see example in Fig. 3.2). In *discovery* role, the user is presented an image masked by another player, randomly selected among the images he never masked, and is requested to classify it into an event class chosen from a list. The list contains a set of event labels much larger than the set of events represented in the dataset and showing some degree of similarity, so as to make difficult to "guess" an event just based, e.g., on the environment or the background information. An honest player will select the correct class if some significant clue remained in the masked image, or will return a "no-choice".

Each player can act in either role as much as he wants, but in order to guarantee enough data for the discovery role, the possibility of playing as discoverer is bounded by the number of previously masked images for that player. Players may gain points when acting in both roles, to incentivize them playing honestly and performing well in each of the two tasks.

While discovery is rather straightforward, masking implies more com-

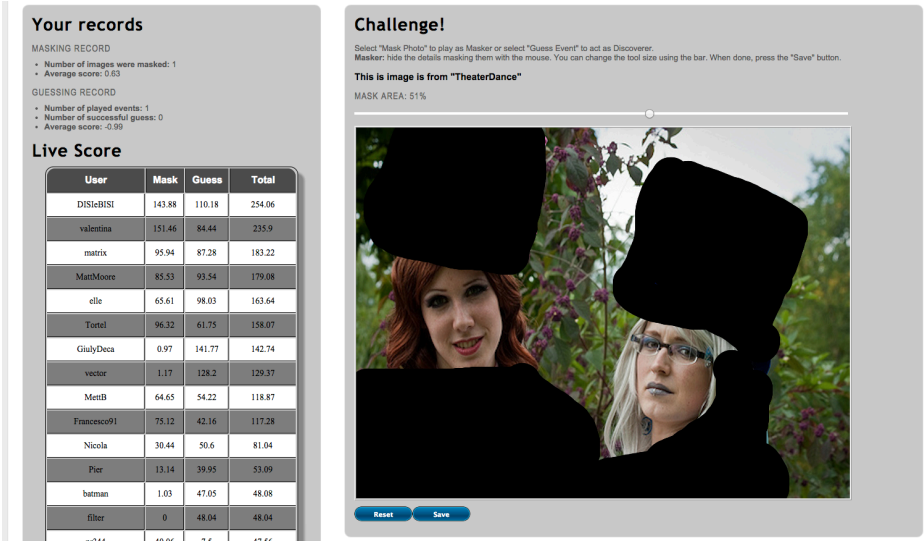


Figure 3.2: Screenshot of the application showing an example of masking procedure. Images are prompted to players and they should hide with a tool the most relevant parts related to the event corresponding to that photo.

plex reasoning. A good player will try covering the lowest possible area to maximize the score (see section 3.2.1), but sometimes this will imply leaving some revealing traces in the shape and position of the masked area. These aspects are further elaborated in the results section.

Access to the game is provided via web at this URL:
http://mmlab.science.unitn.it/eventmask

Scoring system

A critical aspect of gamification is how to reward the players for their performance, while at the same time preventing cheating. In EventMask this is achieved thanks to a well conceived scoring system. The assignment of points is performed in such a way to incentivize honest playing, while rewarding top performances.

Points are assigned when playing both in masking and discovery roles.

In masking role, a provisional score is assigned to the player according to the percentage of the image area left uncovered. In this sense, covering the whole image will result in a zero score, while leaving it completely visible gives maximum score. In the first case, the user is sure that the image will be unrecognizable, but it will get no reward out of this. In the second case, the image will be recognized almost for sure by every discoverer, then the player did a bad job, but the provisional score is maximum. The masking score should then be adjusted by some verification mechanism to avoid cheating. This mechanism is provided by discovery. A discoverer is presented a masked image: if he can recognize the event, the score of the masking player for that image is halved, while he is assigned points proportionally to the masked area. In the above example, a player that left the image completely visible will rapidly lose all the points earned.

Also discovery role may lead to some abnormal behaviors. The typical case is guessing. In order to discourage guessing, players loose points for giving incorrect classification of events. Accordingly, when the event is difficult to recognize with sufficient confidence, it is left open the possibility to skip it, with no point loss. The joint rewarding mechanism between masking and discovery guarantees that players cannot gain points with incorrect behaviors, and they ensure by themselves the validation of the results by playing competitively in both roles (discoverers act as evaluators of the work carried out by maskers). Points rules are summarized in Table 3.1.

Finally, a global score system is provided to recognize the reputation of the various users, prompting it in home page, so that the result is publicly available to the community of players, thus motivating people to improve their performance (see example in Fig. 3.2).

It is to be pointed out that, besides providing a score system for the players, the above rules allow validating the quality of the masks generated

Table 3.1: Game scoring

Role	**Scoring rules**
Masker	Gains A^I provisional points for each image I, where $A^I = 100 - C^I$ and C^I is proportional to the covered areas. Looses $A^I/2$ points when a Discoverer is able to detect the event represented beside the presence of the mask generated.
Discoverer	Gains A^I points with a successful recognition of the event connected to an image I. Looses A^I points for each wrong guess on image I. Do not loose any point if he decide to pass the turn.

by the users. All this information is fundamental to build the final event saliency maps, as explained in the following section.

3.2.2 Event-saliency map generation

Although the game is designed to incentivize effective playing, event saliency detection is not trivial even for humans. Users can forget important parts, roughly segment objects, in particular at their boundaries, and sometimes cheat. Since however users act independently of each other, we can imagine that such behaviors will be largely uncorrelated, while the peak of correlation will be concentrated on the really salient areas. Accordingly, to achieve the final map, we fuse the masks produced by different players on the same input image, each one validated by multiple discoverers. In the following we explain how such fusion process is achieved.

The input information is made of the various masks and their relevant scores, as generated by discoverers. The ideal output is made of all the

image segments that are sufficient to discover the event, and only those. It is to be pointed out that a mask that was never discovered not necessarily is a good map, as it may be over-complete. At the same time, two complementary masks that individually failed with all discoverers may jointly produce a good map. For this reason a simple weighted sum is not an effective fusion strategy. The proposed composition algorithm is a mix of Boolean and weighting rules that keeps into account all the above considerations.

We start from a set of binary masks M_j^I associated to the image I, one for each masking player j, defined as follows:

$$M_j^I(n,m) = \begin{cases} 1 & \text{if pixel } (n,m) \text{ is masked} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Then, based on the results of discovery, we substitute the zeros of each mask with a value proportional to the average number of times the event was recognized for that mask. The rational behind this operation is that, if the event was recognized, the unmasked pixels should contain some evidence of it. At the end of this process, each binary mask M_j^I is transformed into a real-valued mask $\hat{M}_j^I$, as follows:

$$\hat{M}_j^I(n,m) = \begin{cases} P/T & \text{if } M_j^I(n,m) = 0 \text{ AND event recognized by } P \text{ out of } T \text{ discoverers} \\ 1 & \text{if } M_j^I(n,m) = 1 \end{cases} \tag{3.2}$$

Finally, we intersect all the weighted masks associated to image I by applying a pixel-wise product. For a binary image this will have the effect of a AND operation, with the result of achieving an intersection, i.e., the minimum set of points for which there is agreement among users. Since the modified masks are no more binary, the product operation is not exactly a

(a) $P/T = 0.80$

(b) $P/T = 0.40$

(c) $P/T = 0$

(d)

Figure 3.3: Example of event saliency map generation. Masked images are represented along with the corresponding result on the global map, according to the procedure of weighting and fusion described in Equations (2) and (3), respectively. Panel (a) shows an example of starting masked image (left) with the first generated map (right), panel (b) shows the second mask considered (left) with the corresponding result on the global map (right), panel (c) reports an example (taken among those generated) of "art" mask (left) and its reduced influence on the map (right), and panel (d) shows the original image (left) with the final event saliency map (right).

AND, but nevertheless tends to sharply decrease the values in the presence of areas with values near to zero in at least one weighted mask.

The product operation also allows making the saliency map evolve in time as soon as new validated masks are made available by the game, by simply multiplying the current map times the incoming modified mask (and normalizing with respect to the number of masks). The final *event saliency map* is therefore defined as follows:

$$\bar{M}^I = (\prod_j \hat{M}_j^I)^{1/j} \tag{3.3}$$

Fig. 3.3 shows step-by-step the creation of an event saliency map associated to the test image already presented in Figures 1 and 2. Panel (a) left shows the masked image provided by the first player. Here, an important detail is still visible, i.e., part of a cake. This important detail leaded many players to correctly guess the event ($P/T = 0.80$). The resulting weighted map $\hat{M}_j^I(n, m)$ reported on panel (a) right, shows a full black region corresponding to the mask, and a dark gray region in the remaining area, related to the presence of important details in an unknown region outside the mask. Panel (b) left is the outcome of a mask created with care: the cake is fully covered and the candles are concealed as well. For this reason only a couple of players were able to discover the event represented, thus leading to a lighter value of the background ($P/T = 0.40$). The right side of panel (b) shows the combination of this map with the previous one. Since users are playing, it may happen that sometimes a player simply wants to make a joke, thus producing some funny results (see, e.g., the mask in panel (c) left) as well as some anomalous masks to create false leads or simply cheating. The example demonstrates (panel (c) right) that our methodology is resilient, as the combination of the weighting procedure and the mask fusion procedure allows removing the effect

of malicious inputs, even for a very low number of players. In fact these are largely incoherent among different players, and they are filtered out by the mask composition procedure. The fusion of all the masks, according to Equation (3.3), results in a event saliency map highlighting the most informative regions inside the considered image (both reported in panel (d), right and left, respectively).

3.3 Results

In this section we present a set of results generated during a series of game sessions launched within different communities, involving several hundred people. The main gaming sessions were organized within our labs, involving more than 300 volunteers in different stages, selected among Master and PhD students in different disciplines, not involved in specific activities related to image retrieval (to avoid bias). Further sessions where organized at various conferences and research workshops [92], including the Show&Tell demo session at IEEE ICASSP 2014 [91]. The objective of these tests was to demonstrate that the proposed game-based approach can provide new insights on event saliency, which can be used as a basis for further studies in content-based event detection from media. In particular, we aim at demonstrating (i) that the concept of event-saliency exists and it is different for the classical visual saliency, and (ii) that the EventMask game is properly designed to produce effective event-saliency maps. Moreover, the extensive use of the game made it possible to generate a dataset of event-saliency annotated images which are made available to the research community.

3.3.1 Datasets and Experiments

In this section we introduce the results achieved on a set of images taken from two publicly available image datasets. The complete event-saliency set produced by our work was also made available to the research community.[5] In the following we will show a set of selected examples for visual evaluation, to allow perceiving the meaningfulness of event saliency and comparing it to traditional visual saliency.

As far as the datasets are concerned, the system has been tested using two different sets of event-related images. The former is a dataset related to large-scale social events, and includes a selection of photos extracted from the "MediaEval SED competition" [87], referred to the following event types:

- Concert
- Conference
- Exhibition
- Fashion
- Protest
- Sport
- Theater - Dance

The latter consists of a set of images collected from the "EiMM Dataset" [64], which encompasses events related to the personal sphere, and in particular the following categories:

- Concert

[5]The EventMask dataset, with all original images, their corresponding event saliency maps, and visual objects associated to the 15 event types, is available, after registering and playing, at the URL: http://mmlab.science.unitn.it/EventMaskDataset

- Graduation
- Meeting - Conference
- Mountaintrip
- Pic Nic
- Sea Holiday
- Ski Holiday
- Wedding

For both datasets we randomly selected 35 images per each event category, for a total of 525 images.

Every user was asked to play in both masking and discovery roles. During each session a player masked on average 25 images and discovered more than 100 images, with a return rate of 76%. When designing a game with a purpose, one should not forget that the primary incentive for user is engagement (being the objective of game supplier hidden). In our case, we pointed on the adversarial nature of the game, and in some sessions we also proposed some symbolic prizes for winners. The reaction was surprisingly positive, with most involved users continuing the game (70%) well beyond the minimum requested time (average time spent 15 minutes per session). Furthermore, they involved additional users in the game, although we did not pursue any viral mechanism. This promising trend indicates that GWAPs can provide a significant alternative to crowdsourcing to involve humans in multimedia-related computational tasks. The final maps have been generated considering to have at least 10 masks per each image and at least 5 validations for each mask. These numbers where defined by analyzing the progressive composition of the map, which resulted almost steady when 7 or more user masks have been processed. In this way each

image has been seen by users at least 60 times, for a total of more than 30,000 image elaborations made by the crowd to process the entire dataset.

During the discovering phase, we prompted the users with different events, allowing them to select their guess within a list. This list has been designed in order to avoid trivial solutions, thus, we included all categories available in either dataset plus additional classes not present in the data and partially overlapped from the visual viewpoint. As an example, in a "Skiholiday" picture it is possible to cover person and equipment, but probably not snow and winter mountains, which occupy a significant portion of the image. The presence in the list of other events connected to the *winter mountain* context like "Avalanche" prevents easy guessing. Moreover, the class "Unknown-Pass" was added to allow users refusing to choose. For the sake of completeness we report hereafter the list of additional events that are available for the discoverer role:

- Birthday
- Halloween
- Baptism
- Funeral
- CityTour
- CarAccident
- Avalanche
- Unknown-Pass

After collecting and validating the user masks, the information was passed to the procedure described in section III.C, to construct the event saliency maps of the relevant images. In Figures 3.4 and 3.5 we present

some representative results. The first column shows the original images, the second contains the *event saliency maps*, and the third shows the traditional visual saliency, calculated by the technique in [70]. In saliency maps, darker gray levels are associated to higher saliency (i.e., details masked by the majority of successful players for event saliency, and areas with greater perceptual impact in traditional saliency), while lighter grey levels refer to progressively less important areas.

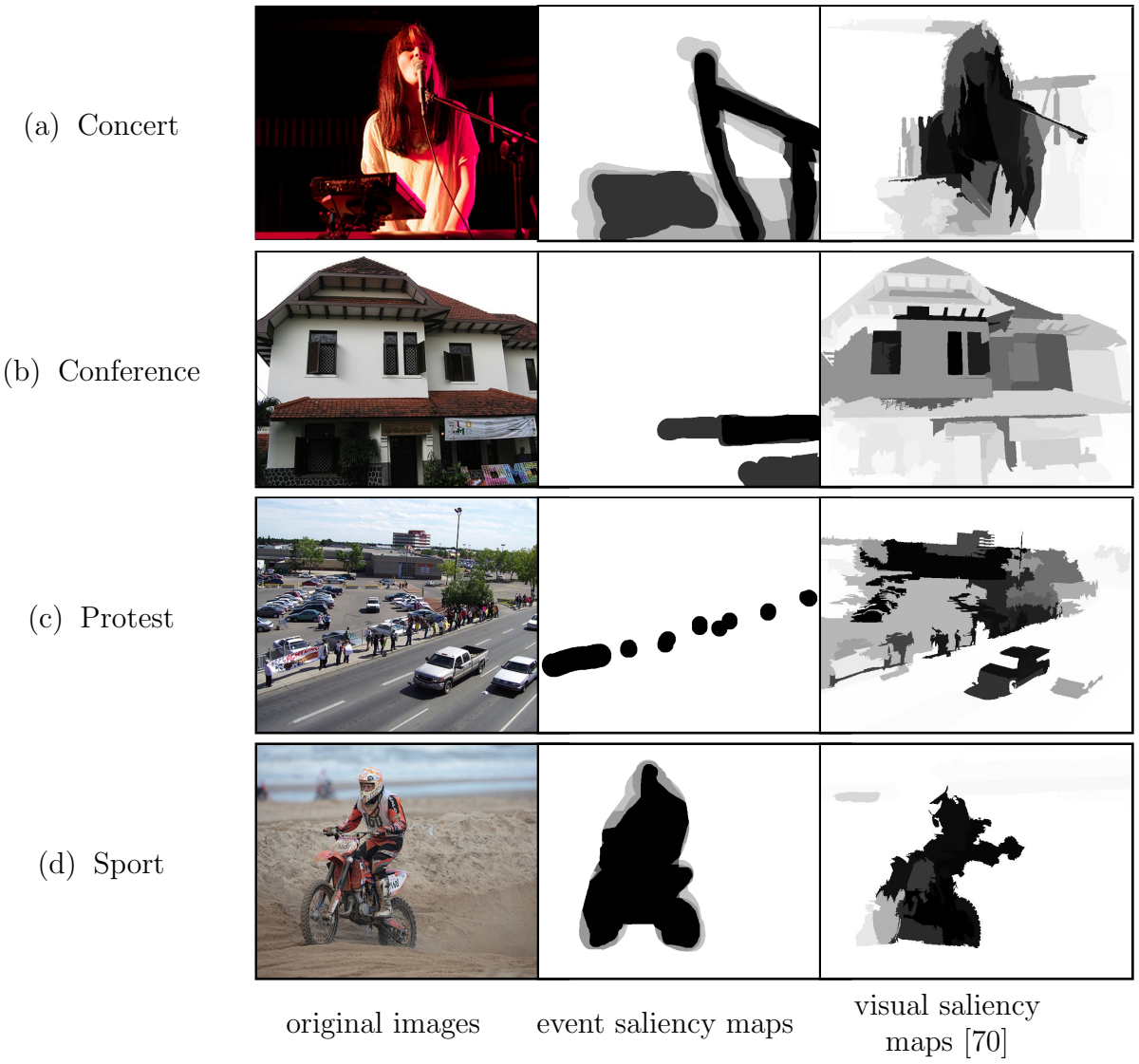


Figure 3.4: Results on MediaEval SED dataset [87].

The first set (Fig. 3.4) is related to the MediaEval dataset. It is interesting to notice that in most cases the event saliency covers a smaller area with respect to the visual saliency, focusing on the visual concepts that contain the highest amount of information in an event perspective. Such details may appear in any part of the images (see, e.g., image (b)) while visual saliency is typically in the center, they are not necessarily big, bright colored or high-contrast. Most often, significant event-related details are not concerned with people (who instead are visually attractive) but to objects or the environment. This is the case for example of images (a) and (c) where objects such as the microphone or the panels carried by people clearly reveal the nature of the depicted events. Only seldom visual and event saliency present a similar distribution, and this typically happens when the attention of the photographer is attracted by a subject strongly related to the event. This is the case for instance of image (d), where the background is rather irrelevant, while the foreground is both visually attractive and event-related. In this case the overlap between the two maps is significantly high. The second set of images confirms the above considerations also when applying the proposed approach to another type of events, more related to the personal dimension. Results are reported in Fig. 3.5 again referring to a subset of representative images distributed across the different classes. Again, event saliency is mostly concentrated on the details of the images, like in pictures (a), (b), and (e), with clear difference with respect to visual saliency, mostly connected to the foreground. In particular in image (a) we have a situation where the event (Meeting-Conference) is mostly revealed by the presence of badges on the persons, so we have a very compact representation of the event, described by that particular visual concept. A similar situation holds in image (e), where the spouse in the background represents the only distinctive element of the event (Wedding) on that picture.

Sometimes photos contain event-related details that are not in the foreground. This is the case of image (b) where the background of the image results more informative in a event-related perspective, since it contains the only hint about the nature of the event (Mountain Trip).

Figure 3.5: Results on EiMM dataset [64].

The most surprising case is the one depicted in image (c), where event and visual maps are almost complementary, as the event-revealing component is the sand and see area in the background, while the visual attention is attracted by the colored group of people in the foreground. Also in this dataset, it may happen that visual and event saliency show a good overlap (see, e.g., image (d)) when the main subject is also representative of the event.

3.3.2 Assessment of the Event-Saliency Maps

In order to assess the results of the game-based event-saliency detector, we introduce here two steps of analysis. First, we evaluate the performance of players by analyzing the users' masks and their relationships with the final map. Then, we evaluate the significance of the saliency maps obtained at the output of the process.

Concerning the first assessment, three parameters have been defined that put into relation the individual users' masks with each other and with the final one, as follows:

$$R(M_j^I) = \frac{card\left(M_j^I \cap \bar{M}^I\right)}{card\left(\bar{M}^I\right)} \tag{3.4}$$

$$P(M_j^I) = \frac{card\left(M_j^I \cap \bar{M}^I\right)}{card\left(\bar{M}_j^I\right)} \tag{3.5}$$

$$D(M_j^I, M_k^I) = \frac{card\left(M_j^I \cup M_k^I - M_j^I \cap M_k^I\right)}{card\left(M_j^I \cup M_k^I\right)} \tag{3.6}$$

where *card* denotes the number of non-null elements in the mask (after binarization), R is the *recall* of j-th user on I-th image, measured as the percent coverage of the final mask; P is the corresponding *precision*, i.e., the percentage of pixels of the user mask that belong to the final map; and

Table 3.2: Event-saliency individual users' masks evaluation

	Average	Standard dev
Recall	56	12
Precision	78	17
Diversity	36	37

D represents the users' *diversity*, calculated as the number of non-matching points between two different user masks, normalized over the total covered area. Table 3.2 reports the average R, P, D values obtained over all users and images in the datasets, as well as their respective standard deviations.

It is worth noticing that the average user covers something more than half final map. This means that users tend to disregard some salient details. The average precision however is significant, meaning that users in general focus on important parts.

The third parameter shows that, although the single user is not fully reliable, the crowd provides complementary information, which can be exploited to achieve a reliable final result. As a matter of fact, the composition method proposed in our work gets rid of the over-concealed areas, while maintaining the minimum evidence agreed among all users.

This is also demonstrated by a further experiment performed with real users. The objective of this test was to understand if the areas covered by the map are really the ones that contain all the event-related information. To this purpose, we prepared a photo collection where each image in the dataset was covered with the corresponding binarized event-saliency mask. The images where then presented to a panel of volunteers including more than 100 people, never involved in the game before. Each user had to guess the event represented in each masked image, choosing among a list and without any penalty for the errors. The user had also the possibility to

(a) Concert $w = 0.33$ (b) Conference $w = 0.16$ (c) Exibition $w = 0.0$

(d) Fashion $w = 0.0$ (e) Protest $w = 0.5$ (f) Sport $w = 0.16$

(g) SeaHoliday $w = 0.16$ (h) SkiHoliday $w = 0.83$ (i) Wedding $w = 0.16$

Figure 3.6: Example of masked images, using the binary event saliency map superimposed on the original images, considering the same events represented in Figures 3.4 and 3.5. w represent the percentage of recognition of the event during the evaluation carried out.

mark the image as unknown. Each image was shown to at least 10 different volunteers.

The average result achieved on the whole dataset was of 71% failure in recognizing the event, with about 46% faulty detections. As a counter-test, the same images were successively shown to the same users unmasked, obtaining in that case a percentage of failure of 10%. Fig. 3.6 reports some examples referred to the images shown in Figures 3.4 and 3.5, along with the percentage of successful recognition of the event by the users, displayed by w in a range from 0 to 1.

It is clear that for some particular images it is much more difficult to hide the event. For instance, in image (h) the background is so particular that it

Table 3.3: Event-saliency maps evaluation: average percentage of event recognition on masked images for each event class.

MediaEval SED dataset [87]	**EiMM dataset [64]**
Concert $\hat{w} = 0.16$	Concert $\hat{w} = 0.16$
Conference $\hat{w} = 0.14$	Graduation $\hat{w} = 0.14$
Exibition $\hat{w} = 0.16$	Meeting - Conference $\hat{w} = 0.12$
Fashion $\hat{w} = 0.14$	Mountaintrip $\hat{w} = 0.83$
Protest $\hat{w} = 0.33$	Pic Nic $\hat{w} = 0.57$
Sport $\hat{w} = 0.14$	Sea Holiday $\hat{w} = 0.14$
Theater - Dance $\hat{w} = 0.33$	Ski Holiday $\hat{w} = 0.86$
	Wedding $\hat{w} = 0.14$

is rather easy to guess the relevant event category, even if some important details are covered. Of course, a larger variety of events associated to the same environment (e.g., different types of winter sports) would make the background much less informative. Other particular situations are those in which the shape of the mask results very similar to the shape of the covered object, thus providing some hint to guess the event. This is the case for example of image (f), where it is rather easy to understand that the concealed object is a motorbike, exploiting the capability of our brain to auto-complete shapes, connecting them to known objects [81]. Although these considerations may be interesting to reveal some characteristics of our cognition system, it is to be pointed out that, even in the above particular situations, the masks represent anyhow the most significant event-related contents of the image, thus making possible to use the relevant information to learn more about the distinguishing visual characteristics of media with respect to the underlying events.

In this respect, the few images that cannot be masked effectively by any player can be considered as a special class where the whole image (thus, probably, the environment) is representative enough to detect the correct event, which is a precious information anyway. In Table 3.3 the average

percentage $\hat{w}$ of event recognition achieved per each event is shown.

3.3.3 Discussion

The concept of event-saliency introduced by this work could provide a powerful source of information to improve the performance of a number of different applications, including image classification, event discovery from media, event-subevent characterization, automatic annotation, ontology definition, media framing/refocusing, summarization and storytelling. Although it is out of the scope of this work producing a thorough analysis of such envisaged applications, we would like to briefly introduce a couple of additional experiments that we conducted (i) to verify if event saliency detection can be potentially performed in an automatic way; and (ii) to test its potential in the framework of a very open problem such as event discovery from single images. Both experiments have been carried out by exploiting state of the art technologies for visual content description and matching, and in particular exploiting the well-known SURF descriptors [13], which provide fast and robust local feature detection, widely adopted in object detection, as well as image classification and retrieval [51].

In the first test, the goal was to use the knowledge acquired with our game to extend the event-saliency annotation to new images. To this purpose, we iteratively took out an image from the set, and removed the relevant data from the dataset. Then, we learned the remaining visual concepts and matched them to the current picture. Every detected concept was then used to create a heat-map, with a heat level proportional to the confidence of the detection. Finally, the resulting heat map was binarized and compared to the groundtruth (the saliency map produced by the game) in terms of precision and recall, as defined above. Table 3.4 presents the average results achieved over the datasets. It is possible to observe that the average recall value (67%) is just slightly better than the one of the average

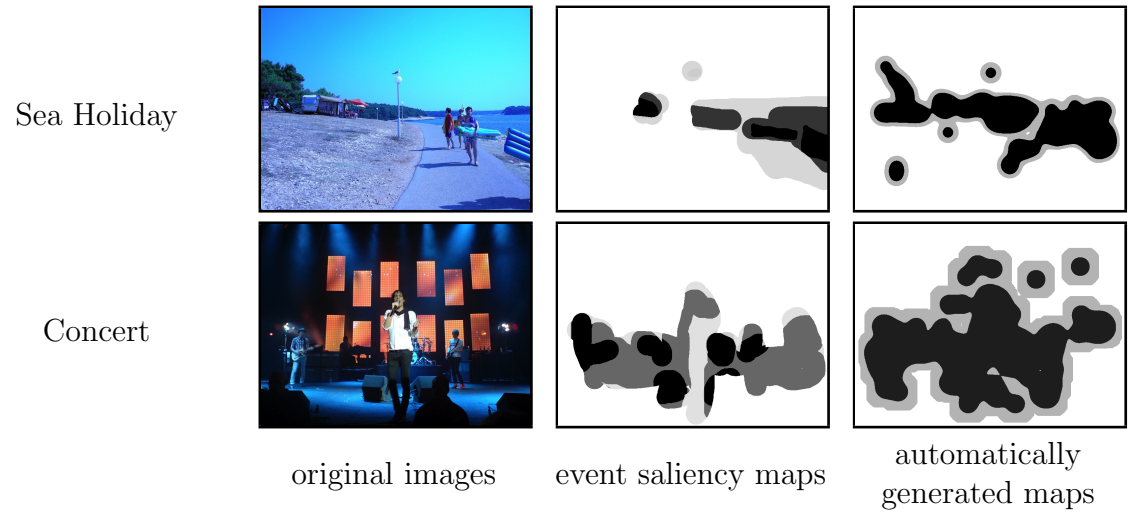


Figure 3.7: Examples of automatically generated maps compared with the relevant results produced with EventMask.

player of the game, while the precision is much worse (60%). Moreover, the standard deviation of these estimates is rather high. This confirms that the generation of accurate event-saliency maps is a non-trivial issue, which would require adequate tools able to tackle with the semantic nature of the problem. Fig. 3.7 shows two examples of automatically generated maps, compared with the relevant maps produced by the game.

In the second test, we wanted to verify if the use of event-saliency information could influence the behavior of an event detector. The cascade of SURF descriptors or their many variances and bag-of-visual concepts (BoW), followed by statistical classification, is considered a de-facto standard in image retrieval [10]. In our tests we performed event classification from single image using this classical approach. The key points were extracted using SURF with a minimum Hessian equal to 800 and then collected into a massive matrix and clustered by a K-Means algorithm, setting k=20. BoW creates an histogram that describes the image and a codebook: the relevant histogram-label pairs are used to train an SVM with RBF kernel for the final classification [24]. The training of the clas-

Table 3.4: Comparison between automatically generated maps and event-saliency maps produced by the game

	Average	Standard dev
Recall	67	24
Precision	60	30

Table 3.5: Percentage accuracy of event-detection from single image

	MediaEval SED dataset	EiMM dataset
Complete images	31.15	38.80
Event-saliency areas	45.95	41.54
Non event-saliency areas	29.52	20.20

sifier was performed on 3 different settings of the event-saliency dataset: (A) the complete set of whole images; (B) the event-saliency areas only; (C) the non event-saliency areas only. Then, we selected a new set of images containing 105 new images for each event class (corresponding to three times the dimension of the training set), extracted from the original datasets (MediaEval SED and EiMM, respectively), and applied the three trained event classifiers to such new set. The performance was then measured in terms of average accuracy. Table 3.5 shows the results obtained for the two datasets. It is possible to observe that the results achieved by the classifier trained with salient objects only is much better in the first dataset and comparable in the second dataset with those related to the whole images. It is also interesting to notice that the training performed on the background only (whole image with salient part removed) produces the worst results in both datasets, with a strong performance loss on the EiMM dataset. This is a further proof that event-saliency contains significant evidence of the event-related information contained in the image. Clearly, more sophisticated uses of the saliency information can be thought and remain an open problem for future research.

Conclusions

In this thesis we presented two different methodologies related to the problem of multimedia content analysis, and in particular to the automatic discovery of event-semantics from media contents. The two methodologies addressed this general problem at two different levels of abstraction. The first approach was related to the detection of activities and behaviors of people from a video sequence, identifying what a person is doing and how, while in the second faced the more general problem of understanding a class of events from a set visual media, considering the situation and context. Both problems have been addressed trying to avoid making strong a-priori assumptions, exploiting the largely unstructured and variable nature of events.

As for the first methodology is concerned, we have proposed a framework for human behavior analysis in a known scenario based on context-free grammars. The algorithm takes as input a set of sample trajectories associated to the activities to be detected, represents them in a symbolic form according to the sequences of hot spots visited during the action, and generates a corresponding set of grammars describing the relevant behaviors. Activity detection is then performed in real time. The major contributions of the proposed approach, as compared to other symbolic approaches for activity recognition, consists first in the possibility of using both positive and negative samples, thus allowing better discrimination capabilities; second, the ability to easily perform a re-training procedure to adapt to

changes in the environment and to achieve better personalization; finally, we have included the capability of effectively dealing with both concatenated and nested actions. The algorithm has been validated in different experimental scenarios, both using visual data and positioning sensors, targeted at monitoring daily activities of people in different environments and compared with state of the art recognition models.

The second methodology proposed has been devoted to the study of events in still images. We introduced the concept of event saliency, defining it as the set of visual concepts that are able to reveal the nature of an event from a set of related media, as opposed to visual saliency, traditionally connected to perceptual prominence. We have explained the importance of such concept from the viewpoint of event-based media analysis methodologies, which could provide powerful tools for media indexing, retrieval, clustering, and summarization. Furthermore, we have defined a strategy, based on gamification, to extract reliable event saliency maps from images by exploiting human interaction. Extensive tests performed with a mixed user population of several hundred individuals playing on two datasets, showed the potential of the proposed approach in building a representative grouhdtruth on a significant set of personal and social events. Moreover, we provided a tool to interact with the system allowing users to add new event categories and related images to the developed framework, thus extending the EventMask dataset and resulting event-saliency groundtruth. The results of such process can be exploited to gain knowledge about the inherent link between visual concepts and events, thus supporting automatic learning systems for concept recognition and media event detection.

Bibliography

[1] Glocal project: Event-based retrieval of networked media, 2012. http://glocal-project.eu/.

[2] Mediaeval benchmarking initiative for multimedia evaluation, 2014. http://www.multimediaeval.org/.

[3] Social sensor project: Sensing user generated input for improved media discovery and experience, 2014. http://www.socialsensor.eu/.

[4] Trec video retrieval evaluation, 2014. http://trecvid.nist.gov.

[5] G. Abramson and M. Kuperman. Social games in a social network. *Physical Review E*, 63(3):0309011–0309014, 2001.

[6] P. W. Adriaans and M. Vervoort. The EMILE 4.1 grammar induction toolbox. In *International Colloquium on Grammatical Inference*, pages 293–295, 2002.

[7] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: a review. *ACM Computing Surveys*, 43(3):16:1–16:43, 2011.

[8] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Multiple cameras fall dataset. *DIRO-Université de Montréal, Tech. Rep*, 1350, 2010.

[9] J. A. Baird and D. A. Baldwin. Making sense of human behavior: action parsing and intentional inference. *Intentions and intentionality: Foundations of social cognition*, pages 193–206, 2001.

[10] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51:279–302, 2011.

[11] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE MultiMedia*, 17(4):80–88, 2010.

[12] L. Barkhuus, M. Chalmers, P. Tennent, M. Hall, M. Bell, S. Sherwood, and B. Brown. Picking pockets on the lawn: The development of tactics and strategies in a mobile game. In *Proceedings of UbiComp*, pages 358–374, 2005.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[14] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 291–300, 2010.

[15] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the ACM International Conference on Weblogs and Social Media*, pages 438–441, 2011.

[16] E. Benson, A. Haghighi, and R. Barzilay. Event discovery in social media feeds. In *Proceedings of the Annual Meeting of the Associa-*

tion for Computational Linguistics: Human Language Technologies - Volume 1, pages 389–398, 2011.

[17] M. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Personalization via friendsourcing. *ACM Transactions Computer-Human Interaction*, 17(2):1–28, 2010.

[18] S. Bhattacharya, F. Yu, and S.F. Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 105–112, 2014.

[19] J. A Bilmes et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. *International Computer Science Institute*, 4(510):126, 1998.

[20] P. V. K. Borges, N. Conci, and A. Cavallaro. Video-based human behavior understanding: a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(11):1993–2008, 2013.

[21] M. Brenner and E. Izquierdo. Social event detection and retrieval in collaborative photo collections. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 21:1–21:8, 2012.

[22] M. Brenner and E. Izquierdo. Multimodal detection, retrieval and classification of social events in web photo collections. In *Proceedings of the Workshop on Social Events in Web Multimedia - ACM International Conference on Multimedia Retrieval*, 2014.

[23] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby. Detecting anomalies in peoples trajectories using spectral graph anal-

ysis. *Computer Vision and Image Understanding*, 115(8):1099 – 1111, 2011.

[24] C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27, 2011.

[25] L. Chia-Wen and L. Zhi-Hong. Automatic fall incident detection in compressed video for intelligent homecare. In *Proceedings of the International Conference on Computer Communications and Networks*, pages 1172–1177, 2007.

[26] A. A. Climent-Pérez, P. Flórez-Revuelta, and F. Chaaraoui. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[27] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing Communications and Applications*, 1(3):269–288, 2005.

[28] M. Daldoss, N. Piotto, N. Conci, and F. G. B. De Natale. Activity detection using regular expressions. *Lecture Notes in Electrical Engineering*, 158:91–106, 2013.

[29] M. S. Dao, G. Boato, and F. G. B. De Natale. Discovering inherent event taxonomies from social media collections. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 48:1–48:8, 2012.

[30] M. S. Dao, G. Boato, F. G. B. De Natale, and T. V. Nguyen. Jointly exploiting visual and non-visual information for event-related social

media retrieval. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 159–166, 2013.

[31] M. S. Dao, D. T. Dang-Nguyen, and F. G. B. De Natale. Robust event discovery from photo collections using signature image bases (sibs). *Multimedia Tools and Applications*, 70(1):1–29, 2012.

[32] C. De La Higuera. A bibliographical study of grammatical inference. *Pattern recognition*, 38(9):1332–1348, 2005.

[33] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2013.

[34] K. Duncan and S. Sarkar. Saliency in images and video: a brief survey. *IET Computer Vision*, 6(6):514–523, 2012.

[35] T. V. Duong, H. Bui, D. Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching Hidden Semi-Markov Model. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 838–845, 2005.

[36] K. Eunju, S. Helal, and D. Cook. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, 9(1):48–53, 2010.

[37] W. Feng, . Liu, and M. Zhu. Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera. *Signal, Image and Video Processing*, pages 1–10, 2014.

[38] C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Bringing order to your photos: Event-driven classification of flickr images based on so-

cial knowledge. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 189–198, 2010.

[39] A. R. J. Francois, R. Nevatia, J. Hobbs, R. C. Bolles, and J. R. Smith. VERL: an ontology framework for representing and annotating video events. *IEEE Multimedia*, 12(4):76 – 86, 2005.

[40] L. Galli, P. Fraternali, and A. Bozzon. On the application of game mechanics in information retrieval. In *Proceedings of the ACM International Workshop on Gamification for Information Retrieval*, pages 7–11, 2014.

[41] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing*, pages 85–90, 2011.

[42] M. I. Gonzalez Duarte and S. Chacon Murguia. An adaptive neural-fuzzy approach for object detection in dynamic backgrounds for surveillance systems. *IEEE Transactions on Industrial Electronics*, 59(8):3286–3298, 2012.

[43] D. Gowsikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4):747–765, 2014.

[44] I. Guy, A. Perer, T. Daniel, O. Greenshpan, and I. Turbahn. Guess who?: enriching the social graph through a crowdsourcing game. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1373–1382, 2011.

[45] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell. A novel sequence representation for unsupervised analysis of human activities. *Journal of Artificial Intelligence*, 173(14):1221–1244, 2009.

[46] HHI. H.264 reference decoder from heinrich hertz institute, January 2014. http://iphome.hhi.de/suehring/tml/.

[47] B. Hu, W. Wang, and H. Jin. Human interaction recognition based on transformation of spatial semantics. *IEEE Signal Processing Letters*, 19(3):139–142, 2012.

[48] Y. A. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, 2000.

[49] M. Jacobs. Gamification: Moving from addition to creation. In *Proceedings of the ACM CHI Workshop on Designing Gamification: Creating Gameful and Playful Experiences*, 2013.

[50] A. Jaimes and N. Sebe. Multimodal humancomputer interaction: A survey. *Computer Vision and Image Understanding*, 108(12):116 – 134, 2007.

[51] Y. G. Jiang, S. Bhattacharya, S. F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[52] Y. G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S. F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *Proceedings of NIST TRECVID Workshop*, 2010.

[53] S. Katz, T. D. Downs, H. R. Cash, and R. C. Grotz. Progress in development of the index of ADL. *The Gerontologist*, 10(1):20–30, 1970.

[54] W. Kazufumi, O. Masanao, O. Makoto, and O. Rikio. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2541–2544, 2011.

[55] K. M. Kitani, Y. Sato, and A. Sugimoto. Recovering the basic structure of human activities from a video-based symbol string. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pages 9–9, 2007.

[56] D. E. Knuth. Semantics of context-free languages. *Theory of Computing Systems*, 2(2):127–145, 1968.

[57] K. R. Konda, A. Rosani, N. Conci, and F. G. B. De Natale. Smart camera reconfiguration in assistend home environments for elderly care. In *Proceedings of the European Conference on Computer Vision*, page To appear, 2015.

[58] E. Law and L. von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206, 2009.

[59] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[60] C. Ling and R. Abhishek. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 523–532, 2009.

[61] X. Liu, R. Troncy, and B. Huet. Using social media to identify events. In *Proceedings of the ACM SIGMM International Workshop on Social Media*, pages 3–8, 2011.

[62] Z. Ma, Y. Yang, N. Sebe, and A.G. Hauptmann. Knowledge adaptation with partially shared features for event detection using few exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1789–1802, 2014.

[63] C. Maltzahn, A. Jhala, M. Mateas, and J. Whitehead. Gamification of private digital data archive management. In *Proceedings of the ACM International Workshop on Gamification for Information Retrieval*, pages 33–37, 2014.

[64] R. Mattivi, G. Boato, and F. G. B. De Natale. Event-based media organization and indexing. *Infocommunications Journal*, 3(3):9–18, 2011.

[65] S. Matyas, C. Matyas, C. Schlieder, P. Kiefer, H. Mitarai, and M. Kamata. Designing location-based mobile games with a purpose: collecting geospatial data with CityExplorer. In *Proceedings of the ACM International Conference on Advances in Computer Entertainment Technology*, pages 244–247, 2008.

[66] V. Mezaris, A. Scherp, R. Jain, and M. S. Kankanhalli. Real-life events in multimedia: detection, representation, retrieval, and applications. *Multimedia Tools and Applications*, 70(1):1–6, 2013.

[67] D. Minnen, I. Essa, and T. Starner. Expectation grammars: leveraging high-level expectations for activity recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 626–632, 2003.

[68] D. Moore and I. Essa. Recognizing multitasked activities using stochastic context-free grammar. In *Proceedings of AAAI Conference*, pages 770–776, 2001.

[69] B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1114 –1127, 2008.

[70] O. Muratov, G. Boato, and F. G. B. De Natale. Diversification of visual media retrieval results using saliency detection. In *Proceedings of IS&T/SPIE Electronic Imaging*, pages 86670I–86670I, 2013.

[71] O. Muratov, P. Zontone, G. Boato, and F. G. B. De Natale. A segment-based image saliency detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1217–1220, 2011.

[72] K. Nakamura. Incremental learning of context free grammars by bridging rule generation and search for semi-optimum rule sets. In *Grammatical Inference: Algorithms and Applications*, volume 4201 of *Lecture Notes in Computer Science*, pages 72–83. 2006.

[73] K. Nakamura and M. Matsumoto. Incremental learning of context free grammars. In *Grammatical Inference: Algorithms and Applications*, volume 2484 of *Lecture Notes in Computer Science*, pages 174–184. 2002.

[74] F. Nater, T. Tommasi, H. Grabner, L. Van Gool, and B. Caputo. Transferring activities: Updating human behavior analysis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1737–1744, 2011.

[75] R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, volume 4, page 39, june 2003.

[76] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui. Learning and detecting activities from movement trajectories using the hierarchical Hidden Markov Models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 955–960, 2005.

[77] T. V. Nguyen, M. S. Dao, R. Mattivi, and F. G. B. De Natale. Event detection from social media: User-centric parallel split-n-merge and composite kernel. In *Proceedings of the Workshop on Social Events in Web Multimedia - ACM International Conference on Multimedia Retrieval*, 2014.

[78] Open Source. Open source multiple contributions. Command line tool for transferring data with url syntax, 2014. http://curl.haxx.se/.

[79] Open Source. Open source multiple contributions. Trans standard multimedia framework for media manipulation., 2014. http://www.ffmpeg.org/.

[80] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-based landmark and event detection for tagged photo collections. *IEEE MultiMedia*, 18(1):52–63, 2011.

[81] L. Pessoa, E. Thompson, and A. Noë. Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, 21(06):781–796, 1998.

[82] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 23:1–23:8, 2012.

[83] C. Piciarelli, C. Micheloni, and G.L. Foresti. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1544–1554, 2008.

[84] N. Piotto, G. Boato, N. Conci, F. G. B. De Natale, and M. Broilo. Object trajectory analysis in video indexing and retrieval applications. *Studies in Computational Intelligence*, 287:3–32, 2010.

[85] N. Piotto, N. Conci, and F. G. B. De Natale. Syntactic matching of trajectories for ambient intelligence applications. *IEEE Transactions on Multimedia*, 11(7):1266 –1275, 2009.

[86] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 103–110, 2007.

[87] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval Multimedia Benchmark Workshop*, 2013.

[88] A. Rosani, G. Boato, and F. G. B. De Natale. Weighted symbolic analysis of human behavior for event detection. In *Proceedings of the IS&T/SPIE Electronic Imaging*, 2013.

[89] A. Rosani, N. Conci, and F. G. B. De Natale. Human behavior understanding for assisted living by means of hierarchical context free grammars. In *Proceedings of the IS&T/SPIE Electronic Imaging*, 2014.

[90] A. Rosani, N. Conci, and F.G.B. De Natale. Human behavior recognition using a context-free grammar. *Journal of Electronic Imaging*, 3(3), 2014.

[91] A. Rosani, D. T. Dang-Nguyen, G. Boato, and F. G. B. De Natale. Eventmask: a game-based analysis of event-related saliency in photo galleries. In *Proceedings of the IEEE ICASSP, Show and Tell*, 2014.

[92] A. Rosani, D. T. Dang-Nguyen, G. Boato, and F. G. B. De Natale. Eventmask: a gamification for image processing. In *Proceedings of the Annual GTTI Meeting on Multimedia Signal Processing*, 2014.

[93] J. Seong-Wook and R. Chellappa. Attribute grammar-based event recognition and anomaly detection. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition Workshop*, pages 107 – 107, 2006.

[94] K. Siorpaes and E. Simperl. Human intelligence in the process of semantic content creation. *World Wide Web*, 13(1-2):33–59, 2010.

[95] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747 –757, 2000.

[96] R. Troncy, B. Malocha, and A. T. S. Fialho. Linking events with media. In *Proceedings of the ACM International Conference on Semantic Systems*, pages 42:1–42:4, 2010.

[97] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 68:1–68:2, 2011.

[98] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine learning*, pages 104–, 2004.

[99] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[100] T. Van Kasteren, G. Englebienne, and B. J. A. Krse. Human activity recognition from wireless sensor network data: Benchmark and software. In *Activity Recognition in Pervasive Intelligent Environments*, volume 4, pages 165–186. Atlantis Press, 2011.

[101] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate activity recognition in a home setting. In *Proceedings of the International Conference on Ubiquitous computing*, pages 1–9, 2008.

[102] R. van Zwol, Lluis Garcia, G. Ramirez, B. Sigurbjornsson, and M. Labad. Video tag game. In *Proceedings of the International World Wide Web Conference*, 2008.

[103] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004.

[104] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[105] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 55–64, 2006.

[106] J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. Whoknows? evaluating linked data heuristics with a quiz that cleans up dbpedia. *Interactive Technology and Smart Education*, 8(4):236–248, 2011.

[107] Y. Wang, H. Sundaram, and L. Xie. Social event detection with interaction graph modeling. In *Proceedings of the ACM International Conference on Multimedia*, pages 865–868, 2012.

[108] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, 14(1):19–29, 2007.

[109] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003.

[110] W. Xu, J. Lu, Y. Zhang, and J. Wang. An unsupervised framework of video event analysis. In *Software Engineering and Knowledge Engineering: Theory and Practice*, volume 114, pages 329–337. 2012.

[111] M. Zaharieva, M. Zeppelzauer, and C. Breiteneder. Automated social event detection in large photo collections. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 167–174, 2013.

[112] Z. Zhang, T. Tan, and K. Huang. An extended grammar system for learning and recognizing complex visual events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2), February 2011.