PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Event Based Media Indexing

Ivan Tankoyeu

Advisor:

Prof.Fausto Giunchiglia

Università degli Studi di Trento

February 2013

Abstract

Multimedia data, being multidimensional by its nature, requires appropriate approaches for its organizing and sorting. The growing number of sensors for capturing the environmental conditions in the moment of media creation enriches data with context-awareness. This unveils enormous potential for event-centred multimedia processing paradigm. The essence of this paradigm lies in using events as the primary means for multimedia integration, indexing and management.

Events have the ability to semantically encode relationships of different informational modalities. These modalities can include, but are not limited to: time, space, involved agents and objects. As a consequence, media processing based on events facilitates information perception by humans. This, in turn, decreases the individual's effort for annotation and organization processes. Moreover events can be used for reconstruction of missing data and for information enrichment.

The spatio-temporal component of events is a key to contextual analysis. A variety of techniques have recently been presented to leverage contextual information for event-based analysis in multimedia. The content-based approach has demonstrated its weakness in the field of event analysis, especially for the event detection task. However content-based media analysis is important for object detection and recognition and can therefore play a role which is complementary to that of event-driven context recognition.

The main contribution of the thesis lies in the investigation of a new event-

based paradigm for multimedia integration, indexing and management. In this dissertation we propose i) a novel model for event based multimedia representation, ii) a robust approach for mining events from multimedia and iii) exploitation of detected events for data reconstruction and knowledge enrichment.

Keywords

media indexing, event-based analysis, context processing, multimedia collection, event detection

Acknowledgements

I would like to express deep gratitude to my scientific advisor Professor Fausto Giunchiglia, for the chance that he gave me, for countless life and professional lessons and for his continuous help and support on different steps of my PhD life.

I wish to express sincere appreciation to Dr. Ilya Zaihrayeu and Dr. Julian Stöttinger. They played crucial roles in my research activity. I would like to thank them for patience and guidance without which this study would never have reached the current state.

I must also acknowledge my friends for their support. I would like to give special thanks to Siarhei Bykau, Maksim Khadkevich, Uladzimir Kharkevich, Heorhi Raik and Yury Zhauniarovich for their encouragement and assistance within tough periods in my life.

Finally, I thank my family: my mother Anastasia Tankoeva, my sister Nadezhda Tseliuk, her husband Dmitri Tseliuk and my future wife Viktoryia Tankoeva (Ambrushkevich) for their never ending love, great support and inspiration.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The storytelling being in a broad sense the process of transmitting information plays a vital role in mankind development. It is used not only as a way to entertain but also to share and transfer knowledge and experience. According to Wikipedia[1] the process of storytelling stands for “conveying of events in words, images, and sounds” . This definition allows us to distinguish main components of storytelling: *events* and corresponded *media*.

The story of our life memorized in autobiographical memory is compiled from a set of events experienced within our lifetime [88]. This signalized about the importance of events as “intellectual landmarks” for indexing one’s individual memory [61]. Furthermore, these “intellectual landmarks” can retrieve different types of related to an event information such as temporal, spatial, causal, etc. [9], [54]. Thus, we can say that events play a role of natural aggregators of contextual information. Therefore, events can be used to represent context in a semantically meaningful way that follows how we humans process our experiences. By the context we imply the information about environmental conditions of an event.

Due to the recent advances in computer technology and arrivals of new media capturing and storage devices the process of storytelling has made a major transformation. New media capturing and storage devices e.g. cameraphones

[1] `http://en.wikipedia.org/wiki/Storytelling`

have changed the traditional storytelling form to the digital one. Digital storytelling become even more popular with the developing of on-line facilities for sharing and publishing personal content in the Internet. Digital media that can be taken almost at no cost allows people to collect huge amount of personal data that can be used "to tell" a story of their life. This media is characterized by the heterogeneous nature, in the sense of diversity of content's form. However, the enormous amount of digital media requires appropriate techniques for its indexing and organizing.

To this end, we propose to use events as the primary means for media indexing and organization. Moreover, the ability of digital devices to capture the context (i.e. spatio-temporal information) along with the media content facilitates this task. In our research we use definition of [38] and refer events as "something that occurs in a certain place during a particular interval of time". This definition allows us to scope event mining task as the detection of spatio-temporal boundaries of an event.Therefore spatio-temporal data processing dominates in the event-based multimedia analysis.

Event-based multimedia analysis paradigm is a relatively young approach for indexing, management and organizing multimedia data. However, it demonstrates enormous potential for media analysis. To unveil this potential event-based techniques for semantic multimedia analysis require a lot of work.

This thesis touches different phases of event-based media analysis. On the initial phase we develop a model for event content representation with the focus on media components of this model. The second phase consists of the event detection part that describes an approach for mining events on the global and local scale. The phase of event detection is followed by the final phase of event exploitation, that clearly demonstrates the power of events for providing a new knowledge.

The reminder of this chapter consists as follows. **Section 1.1** demonstrates the necessity of events for media indexing; **Section 1.2** outlines the concrete goals of the thesis; **Section 1.3** exemplify these goals with user scenarios; **Section 1.4** describes the overall structure of the thesis.

1.1 The Motivation

The rapid growth in the number of digital photos and videos within the past decade has led to a rethinking of the traditional principles for media indexing and organizing. In fact, object-centred media analysis strategy has a limited ability to satisfy user needs in organizing personal media collections [2].

Recent developments in media industry show the interest towards an event-based multimedia indexing. Apple´s iPhoto[2] allows a user to organize manually his images by folders called events, Adobe´s Lightroom[3] lets the user build image collections. Both Google´s Picasa[4] and iPhoto let the user visualize their pictures according to detected faces and locations. Still, there is neither a combination of this information nor an extended knowledge extracted from these clusters. Opposed to these applications, the highly controversial timeline from Facebook[5] aims to represent the entire life of an user as a chain of events. However, the absence of automated or semi-automated approach for event-based media clustering of personal collection bothers a user with the tiresome and boring work of organizing it manually. Nevertheless, recent studies have demonstrated that at very least 70% of albums published at Google Picasaweb represent events related to a user´s life. Therefore, the new event-based paradigm of multimedia analysis has drawn a significant attention recently [52], [70].

Tremendous growth of personal media collection and the absence of appropriate tools for managing and organizing them have made the event-based media indexing techniques an emerging area of research. In this area the contextual

[2] `http://www.apple.com/ilife/iphoto/`

[3] `http://www.adobe.com/products/photoshoplightroom/`

[4] `http://picasa.google.com/`

[5] `http://www.facebook.com/`

features of media plays the crucial role. We contribute to this area developing our ecosystem that consist of framework for storing media and its metadata, event detection tools for minig events from digital photo collections and the exploitation of detected events.

1.2 Objectives

The main objective of this thesis is to investigate the concept of *events* for multimedia content indexing, organizing and management. In order to achieve this main goal, the thesis addresses a wide variety of design and technical challenges, which closely relate to the following detailed objectives:

- Provide a user with relevant model for event-based media content representation and management;
- Assist a user in media organization according to events and context to ease further collection management through fully automatic event detection approach;
- Improve the quality of contextual data of the personal media collection via the spatial context reconstruction;
- Derive knowledge for a user inferring social ties of a user via semantic layer of events;
- Enrich semantics of media collection with the most ”attractive” moments;
- Provide advanced indexing and retrieval techniques that expand complex event structure to the atomic sub-events with different level of granularity;
- Support a user within the media annotation process providing recommendations derived from the analysis and fusion of existing information;

1.3 User Scenarios

The presented below set of user scenarios shows how the proposed paradigm of event-based media analysis helps a user in different real life situations. In order to cover user's needs described in those scenarios possible services are proposed. Further in this thesis different sections have references to presented scenarios and services.

1.3.1 Trip to Portugal

Maxim and Mila have just returned from a well-organized by an agency touristic trip in Portugal. Mila being a professional photographer made a lot of nice shots by a professional camera. Maxim preferred to use his smartphone for taking photos and videos. They enjoyed the trip and want to share taken media with their friends via online service tools. For a better storytelling they are going to locate all the photos from the trip on a web mapping service application. However, the vast majority of taken photos and videos are not geo tagged. Annotating process is very boring and tedious for them so they decided to submit their photos and videos from the trip to the system. Based on the existing spatio-temporal information the system automatically assigns missing geo-stamps to some media within the collection.

Services:

· estimate geographical coordinates to photos and videos with missing geo information

· place photo on the map

· get location name of given geo-coordinates from external sources

1.3.2 Football match

Last night Yuri was on the football match of his favourite team BATE. He made a nice collection of photos during the match. A friend of Yuri missed it and asked him to show some of the photos with interesting moments of the match. Usually the frequency of taking photos by Yuri is higher within the attractive moments of an event. In order to find these moments Yuri uploads the photos to the system. It analyses the photos and propagates him a subset of photos related to the most attractive moments of the match.

Services:

· get an attractive moments of an event
· get an attractive moments of a sub-event
· search the location of the match

1.3.3 Protests in Syria

Serge is a professional blogger interested in protests in Syria. He is writing an article about it. He has photos of the protests created by a professional photographer. However, he would like to gather additional media taken by the participants of this event. So he uses the system for media propagation based on the existing set of photos.

Services:

· get related photos
· get related videos
· get related articles

1.3.4 Air show

George is an aviation fan and each year he visits some of the European air shows such as Farnborough air show, Paris Air Show. Last time he was on

MAKS in Moscow. Each time he takes a lot of media entities with automatically acquainted geo-spatial information. Being a communicative person he got to know a lot of people during these trips. Recently he has started to use the system. He would like to find the users in the system he may know in real life. To do this he uploads all his media taken within air show events to the system. The system retrieves a set of users and for each of them the number of co-attended with George events. Most likely George knows users with the high number of co-participated events and has similar interests with other propagated users.

Services:

· get a user with the given number of co-attended events

1.3.5 Concert

John was on a charity concert to aid the victims of a tsunami. He took a huge collection of photos and videos within the celebrity performances. He is eager to share it on-line but foremost he would like to properly annotate the media. He exploits the system to automatize this process. The system extracts media description of different users participated in the same concert and propagates it to John.

Services:

· get related textual annotation for photos
· get event description

1.3.6 Photo collection

The growing number of digital photos taken by Victoria becomes unmanageable. So she finally decides to sort them. In order to make this automatically she submits the collection to the system. It indexes the media collection according to the hierarchy of Victorias personal events.

Services:

· sort collection on space basis
· sort collection on event basis

1.4 Structure of the Thesis

The reminder of the thesis is compromised of three major parts. The first part describes a model for event-content representation with emphasis on image and video components of the model. The second part is dedicated to event mining from multimedia collections. The last part gives insights into the use of the detected events. More detail description of each chapter is given further down.

Chapter 2 contains the description of the framework that models the multimedia content and the events that may be depicted. The framework aims to describe the structure of events by answering to five W questions (i.e. Who, What, When, Where, Why). Moreover, this framework is able to represent possible relations (e.g. spatial, temporal) among different components of the proposed event structure such as actors, objects etc. We will briefly describe the architecture of the framework. This architecture allows preserving semantically meaningful to a user information that is codified in events and media content. This chapter also describes some of the features of the framework components and possible services for them. In particular, we concentrate on the components dedicated for multimedia content representation. Following publications are related to the chapter: [19].

Chapter 3 aims to describe our approach for mining events from personal and social multimedia archives. In particular, the first section of the chapter distinguishes personal events and social events. The second section describes the methodology for personal event detection from the multimedia collection of a user. Moreover, an approach for reconstruction of the hierarchy of events is presented. The underlying idea of this reconstruction is to discover possible sub-events within detected events. This, in turn, eases the navigation process for a user on different levels of granularity of events. The description of the collected data set and the experimental validation concludes the section. The last section of the chapter addresses an issue of personal information organization in social environment. We propose a fully automated approach for detection of social events from personal events. The section contains detailed description of the algorithm for social event detection and it experimental validation. Following publications are related to the chapter: [78], [58], [79], [59]

In **Chapter 4** we demonstrate how the detected events can be exploited. The main goal of the section is to demonstrate abilities of the event-centered approach to metadata reconstruction and information enrichment. The process of metadata reconstruction allows a user to recover missing metadata related to media. In contrast to it, information enrichment stands for providing a user with the new knowledge derived from the analysis of data fusion. Following publications are related to the chapter: [80], [81], [59]

Finally, **Chapter 5** concludes and describes possible future directions for research on event-based multimedia indexing.

Chapter 2

Event-based Content Representation

In this chapter we describe how a framework for event-based media content management is built. The framework is based on the entity-oriented model. The model represents a complex structure of entities of different types, their relationships and set of services for them. The underlying idea of this model is to describe the world in terms of entities and relationships between them. An entity type describes instances of a particular class by defining the attributes that can be used to describe common properties of these instances. A type of an entity depends on the scope of application and can be defined by a user. In the course of our research we have developed media entity types (Etypes) and defined the interaction process with other Etypes such as event, location, person, organization.

This chapter is organized as follows. In **Section 2.1** state of the art models and standards for event-content representation is considered. **Section 2.2** gives an essential insight into the framework for event-based media content management. **Section 2.3** and **Section 2.4** describes accordingly how the Image and Video are represented in the framework. Finally the chapter concludes with **Section 2.5**.

2.1 State of the art on media content modelling

Recent studies have shown growing interest towards event-centred multimedia modelling. Well-structured information about an event provided by an ontology can facilitate organization and indexing of multimedia resources. The authors in [32] propose to use ontology (top down approach) with standard content-based image retrieval technique (bottom up approach) as a way to bridge the semantic gap. In [85] the authors use Wikipedia-based ontology and low-level visual feature ontology for improving performance of web-based image retrieval. An event-centred approach for conceptualizing, storing, and querying multimedia information is presented in [57]. The main aim of this work is to use events as a unifying concept to model various types of multimedia data. Design of a formal model of events (F model) based on the foundational ontology DOLCE+DnS Ultralite is considered in [71]. This formal model is used to facilitate the interchange of event related information between different event-based components and systems. The work of Jain et al. [65] mainly focuses on the issues of event composition using the sub-event-of relationship between events. In order to represent the possible semantics of a composite event they propose to compute event attributes as a function of its sub-event attributes. A set of requirements for a base model of events is presented in [87] that categorises all the properties and relations of an event into six aspects: temporal, spatial, informational, experiential, structural and causal. This work is based upon the event model E for e-chronicle applications [86]. In [72], this model is extended and specialized in order to support description of events in multimedia data. The model in [65] is based on the generic models E and F, as well as novel sets of event composition operators are presented. LODE [75] also addresses the temporal, spatial and informational aspects by integrating different event ontological models. In [90], the Event Ontology (EO) [66] is implemented in OWL and is used to describe music events in several granularity levels. The Simple Event Model is proposed in [83] to represent not only who did what, when and where, but also to model the roles of each actor involved, when and for how long this is valid and according to whom. MediAssist [55] organises digital photo collections using time and location information combining it with content-based analysis (face-detection and other feature detectors. In [26], a semantic-syntactic video model (SsVM) is proposed, which provides mechanisms for video decomposition and semantic description of video segments at different granularity levels. In [27], the video event representation language (VERL) and the video event markup language (VEML) are presented for the description and annotation of events in videos respectively. In both models, events are not treated as first class entities, i.e., the instantiation of an event presumes the existence of a video segment depicting the event. Wide variety of approaches for event-content modelling can be found in different standards developed for practical purposes. A good example of this kind of approaches is IPTC G2 group of news exchange standards. EventML is one of these standards exploited for describing public events

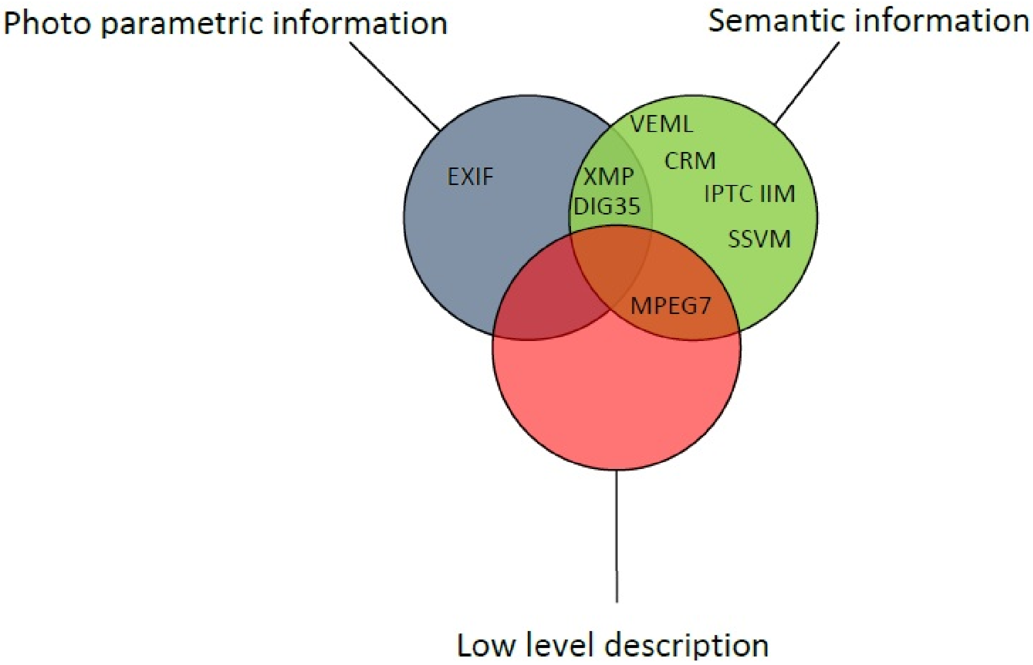


Figure 2.1: Covering different informational aspects by current standard.

with deep semantic annotation (see green circle on Figure 2.1) in a journalistic fashion. Since this model is follow the idea that images are only used to complement a story the support for media is limited. In [76], the conceptual reference model (CRM) ISO standard of the International Committee for Documentation (CIDOC) is described, aiming to provide formal semantics for the integration of multimedia data in cultural heritage domain applications MPEG-7 is a powerful standard for describing multimedia semantic but does not face the reality of state of the art in low-level features for image content description (red circle on Figure 2.1). Lack of a predefined set of attributes for describing technical information of a media item is one of the disadvantages of the standard. EXIF is used to standardize such essential technical information (blue circle on Figure 2.1) like camera setting, timestamp and geographical coordinates, that automatically acquired by the device at the moment of creation of digital media. XMP and DIG35 standards are dedicated for describing technical and semantic information of digital media. The considered digital media standards show their appliance for different tasks. However, the existing standards are limited to some extent in different aspects for applying them for event-centred media description. Most of the standards do not describe a digital media from the both content and context sides (see Figure 2.1).

2.2 Event Content Media Modelling

The crucial role of the entities and their linking relationships has been widely recognized in different research fields such as Semantic Web (see [3] and [6]), Information Extraction (see [14]), Digital Libraries (see [69]). In [3] entities presented as single units which are used for reasoning and linking Semantic Web applications and represented with a conceptual model. Moreover the schema with different entity classes and types is presented also. In [14], researches facilitate the task of named entity recognition and relation extraction by categorizing entities. In [31], employ linked data in order to support a user fact-finding and question answering tasks. Globally unique identifiers proposed in [6]. The authors have developed a framework where those identifiers are used for data integration and the development of entity-centric applications.

In [69], possible infrastructure for contextualized search in the Digital library domain is described. It is used to link different types of informational sources among the Internet (e.g., calendar directories, location gazetteer, biographical dictionaries). ”Linked Data” principle and practice is described in [34]. More specifically, the mechanism to access data sources in uniform way from the application layer is proposed.

The initial version of proposed model for our framework is described in [29] and uses similar to Linked Data and Resource-oriented mechanism for unique entity identification. Unique identifier is used to enable interaction with the entity representations over a framework. The detail description of the mechanism can be found in [69] and [34]. During the framework development I was involved in the development of the event-media related part of the model. However, in order to have a deeper understanding of the framework we will consider the essence of the model (see Figure 2.2).

The entity (en) is defined in the framework as a pair: $en =< uri, Etype >$ Where:

· uri, is a unique identifier of an entity;

· $Etype$, is a non-empty set of entity types. It represents the class to which entity belongs e.g. the entity “Air show” is of type Event.

An entity type $Etype$ defined as follows:
$EType =< Attr, S, Rel >$
Where:

· $Attr$, is a non-empty set of attributes, where each attribute $attr$ consists of attribute name $attr\_name$ and its value $attr\_value$.An attribute aims to describe the properties (e.g., "duration time") of the entity;

- S,stands for a set of services that can be executed on the entity level for example, a service "*get causal events*" can be enabled on the Event Etype.
- Rel, is used to describe the entity relation, where relation rel consists of relation name rel_name and relation value rel_value, for example "George" is a $participant\_of$ "Air show";

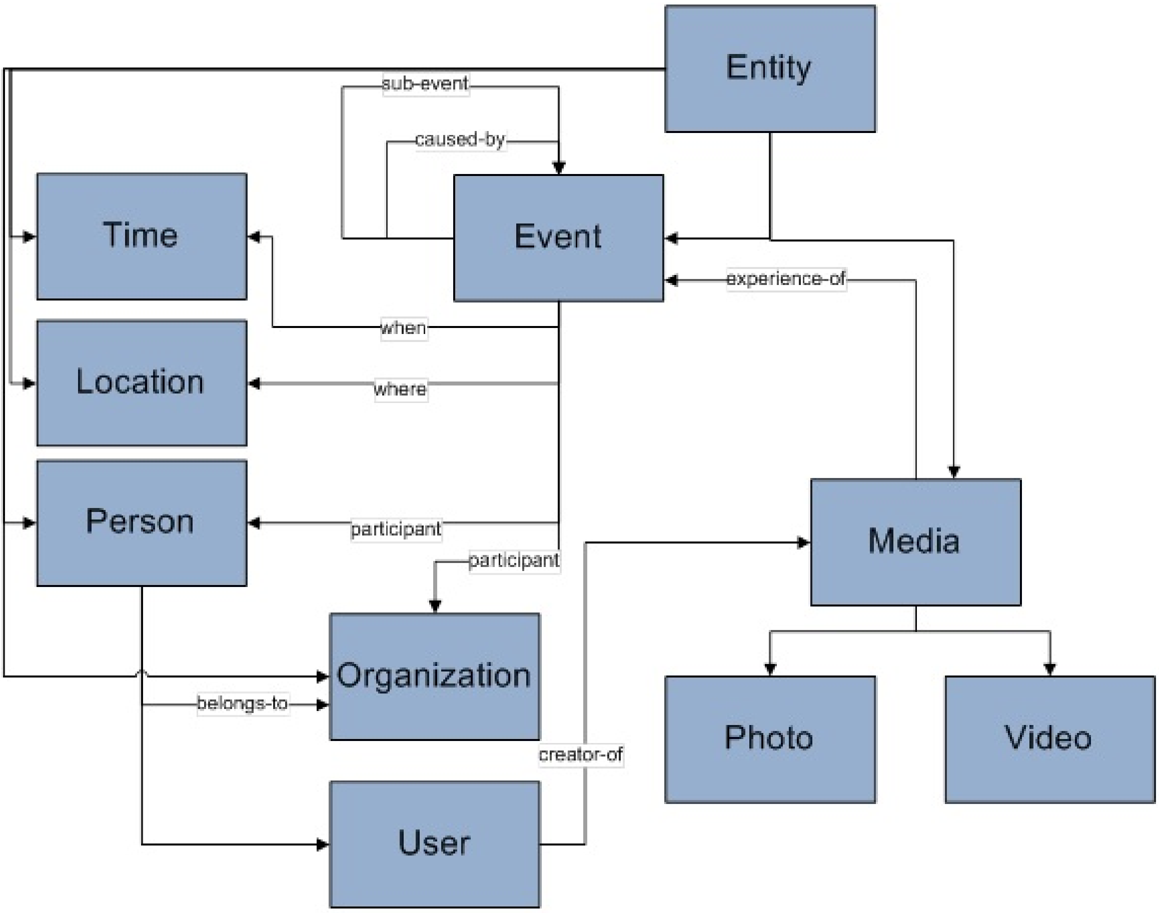


Figure 2.2: Example of simplified entity-oriented model for event and media management.

We would like to emphasize that the framework contains a list of Entity Types, so called "Etype Lattice" (see Figure 2.2). The generic entity type has the following mandatory attributes:

- uri, a unique identifier assigned to the entity at creation time

· $name$, the label given to the entity (e.g., “Trento”, “trip to Italy”, etc.)
· $description$, a verbose description of the entity
· $start$, the time when the entity is started to exist
· end, the time when the entity ceased to exist
· $partof$, specifies the relation parent-child relationship with other entities.

Among the services associated to the generic entity type, we consider Create-Read-Update-Delete (CRUD) operations:

· *create entity*, creates a new Entity of the given type and with the given set of attributes
· *create attribute*, creates an instance of an attribute with the given value(s), to later be assigned to an entity
· *add role*, assigns the given role with the given set of attribute instances to the entity
· *read entitiy*, returns the set of entities that instantiate this Etype
· *delete entity*, deletes the given Entity.

Assigning the role to an entity is arbitrary and depends on the nature of the entity.

2.2.1 Structural components of the framework

In this section, we focus on the most related entity types to the event-media content modeling. Among these entities, event entity type plays one of the key roles in the framework. Therefore the description of this Etype is given in more detailed manner.

Event

Events are regarded as domain-level entities, i.e., things that happen and have positions in space-time. Event is a structured entity that can be characterized by

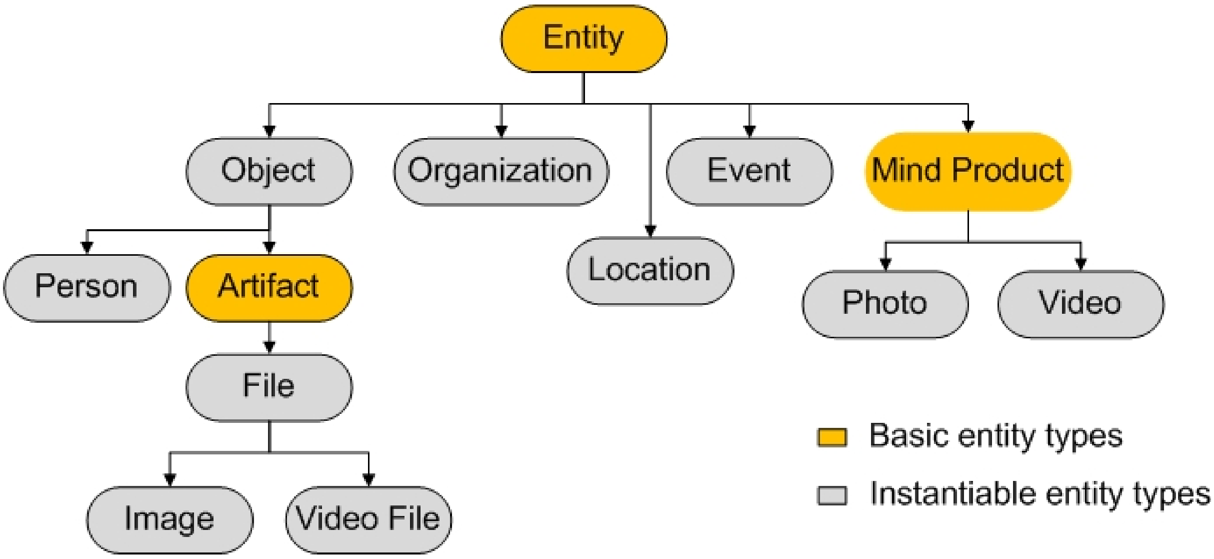


Figure 2.3: Simplified Lattice of Entity Types.

the hierarchical nature. Moreover, there are sometimes different views of the same event; for this reason, in our model an event is considered as an entity that is subjective to the user who describes it. We group attributes and relational attributes on five aspects that are similar to some extent with the six event aspects described in [87]. These attributes separated by the following aspects: informational, temporal, spatial, associated and experiential. Below is a more detailed description of each aspect.

Informational aspect provides general description of events along with the inherited from parental Etype attributes (e.g. name, description). It contains the following attributes:

- *category*, the category to which the event belongs (e.g., concert, wedding, soccer match, trip)
- *status*, specifies the state of the event (e.g., postponed, expected)
- *participant*, a set of relational attributes specifying event participants and links to agentive entities, i.e. entities of type Person and Organization.

Each event participant could have a role. The role is assigned on the entity layer and inherited by Event Etype. For example George “plays” a spectator role on an Air show event and this attribute can be valid only within the event scope. In the context of our model, we define the role attribute similarly to the non-substantial predicate specified by Guarino et al. in [30].

Temporal aspect describes time-related attributes of the event entity type. On this stage of work we just consider absolute-time as the temporal location of an event. Such attribute defines a temporal interval whose endpoints and length might be specified with varying degrees of accuracy. Note that, in general, the information on the date can be relative (e.g., "in two week after the Wedding"). For a basic event-model, we assume the temporal location as always being a single time interval.

Spatial aspect is a relational attribute which links to an entity of type Location (see the detail at Section 2.2.1).

Associated events aspect stands for description of relationship between events. There are two types of relations between events: structural and compositional (in [87], they are separated categories). Structured relation is used to represent the decomposition of an event into two or more sub-events. Each sub-event is a self-standing event. Causal relation links a given event to events caused by it and to the events that generated it. The kind of relationship between these events are indicated by the bidirectional links: sub-event of/parent of and generated by/caused by for structured and causal aspects of event accordingly. The ability to represent these two categories by a unified way allows us to merge them in our framework.

Experiential aspect is represented by relational attributes that links an event to its associated media. The links either point to media entities as wholes or to part of them: for example, a link might identify a piece of video, the one describing a given event.

The following services are offered for entities of Event Etype:

· show events on map, displays the given entity on a map, where map provider is a free on-line service (e.g., Google, Yahoo!)

- show events in timeline, displays all the event entities of a given type (if specified) in a timeline
- show associated events, shows related events in maps and/or timeline according to the relation types
- get latest events, gets all event entities added to the system since the time specified
- sort sub-events, sorts the sub-events belonging to a given macro-event according to their temporal aspect
- detect events, detects event by analysing spatio-temporal data extracted from media collection
- guess event location, guesses event location from other metadata (e.g. by analysing data from media related to event)
- guess time, guesses event time from other metadata (e.g. by analysing data from media related to event).

Person

The Person type is identified by a set of attributes describing different types of information about a person in the real world and a set of related services. For example the hockey player W. Gretzky can be described with an instance of Person Etype with the full name, date of birth, nationality, etc. specified. Attributes are taken from existing vocabularies such as FOAF [8] for personal information (e.g., name, gender, nationality, etc.), Freebase [5] for carrier information (degree, major, etc.), XFN for social relations (e.g., friend, colleague, etc.).

The following services are provided for this entity type:

- find related people: finds entities representing other people that are socially related to the current entity. Several conditions could be defined as parameter, e.g., which social relation to use, if the relation should be transitive and with which depth

· recommend people by event: finds and recomends people who co-attended the given number of events with the given person

· search people: searhes people with the given set of attributes.

Organization

The Organization type comprises group of persons organized for some end or work, or social institutions such as companies, societies etc. As stated in FOAF specifications, an organization type identifies ad-hoc collections of individuals. A basic relational attribute of an Organization type is member: it links to an entity of type Person. Among the other attributes of such type are: date-of-foundation, statute, acronym, main-seat, homepage, domain-field (e.g., educational, science, sport, etc.), area-of-activity. An example of the services is:

· show on map: display the given entity on a map.

Location

The Location type includes all those entities for which the spatial dimension is significant. It identifies spatial objects, real entities occupying regions of space (e.g., regions, cities, natural-bodies, buildings, etc.). The Location Etype inherits all attributes from the generic entity etype and includes the following ones: latitude, longitude, and altitude (optional) in accordance with the WGS84 standard. Besides the just mentioned attributes, new metadata can be added depending on the location category: for example, if the location identifies a building an address will be specified. For instances of Location Etype we consider following services:

· show on map, display the given entity on a map.

· get related media, retrieves a list of media related to this entity.

2.3 Image and Photo

Photo and Image types contain a set of attributes related to a static two dimensional entity that represent something or somebody. Our model distinguishes between Image that stands for container and Photo which represents content. An Image is the digital representation of a Photo. Therefore Image type identified by a set of attributes describing pixel-level parameters (e.g. Byte per pixel) and the way of displaying the image on the screen (e.g. Width, Height, Color representation). Photo Etype aims to describe the content of the medium and contains links to other Etypes (e.g. Person, Location, Event). The relationship between Photo and Image Etypes is one to many.

The list of services proposed for these Etypes is following:

- show the entity, displays the given photo
- read entity metadata, extracts the embedded metadata from the given image
- sort collection, sorts the collection of photos by space, time, events
- estimate geo coordinate, estimates geographic coordinates of a photo based on its position within the collection
- annotate the entity, annotates photo with location name and related event
- find the momentum of attraction, find the attractive moment of a given photo collection.

2.3.1 Image Etype

An image entity is a digital form of a photo entity in the framework. Image Etype contains all the attributes related to the creation and representation of an image entity. Image Etype being a descendant of File Etype 2.2 inherits all the attributes from it such as source of creation, device model, etc. The set of basic attributes of Image Etype includes but are not limited by the image size, color model, etc.

Since any kind of low level processing is performed on the level of Image Etype an attribute for low level image representation is required. To this end we propose to use a pair of attributes for low level image description:

· *feature name*, is a type of low level image feature

· *File Etype*, is the file that contains the low level feature of the image

2.3.2 Photo Etype

A photo entity inherits the set of attributes from Mind Product Etype along with the Entity Etype. These attributes deals with authorship and copyrighting issues.

At the first glance it seems that all the contextual information captured by camera (i.e. EXIF, see details in [1]) in the moment of image creation is related to Image Etype. However, the major part of this contextual information corresponds to Photo Etype. Further we will consider example that helps us distinguish them. The set of Photo Etype attributes is separated on three categories: camera setting, geographical attributes and content attributes.

Camera settings attributes stand for representation of the camera settings at the moment of photo creation e.g. focal length, exposure time, flash. They are used by photographers during the creation of the mind product. For example the exposure and light source plays a crucial role in creation of art photography. Therefore we refer this category of attributes to Photo Etype.

Geographical attributes describe information related to the place where the photo is taken. The set of attributes contains GPS coordinates and the direction of the camera at the moment of photo creation. Geographical coordinates can be translated to a human readable form such as name of location. This is done via reverse geocoding tools.

Content attributes aims to describe the content of a photo. In order to provide a detail description of photo content we have developed a system for referring spot of interest within the photo. The region of interest (ROI) on the photo is specified as a bounding box by a user or system. In case of absence of a bounding box we assume that the ROI relates to the whole image. The image is described in a Cartesian system. Starting point of the system [0,0] is a pixel on the top left corner where the X-axis horizontal and directed to the right and the Y-axis vertical and directed downward. The rectangle of interest on an image is described as coordinates of starting point of a rectangle (top left corner) and coordinate of end-point (down right corner). However, there is an issue in this system: each time when the image is cropped the position of ROI may become disoriented.

The system of referring ROI within an image allows a user to tag any kind of object or entity on it. This flexibility facilitate for a user or a system the process of linking content data with entities e.g. Person, Location, etc.

2.4 Video and Video File

Video being a sibling Etype to Photo shares some attributes with it. These attributes were inherited from parental Etypes and include time of creation, creator and others attributes. Similarly to the previous section we define Video and Video File as content and container accordingly. Video Etype describes the content in the semantically meaningful to a human way referring it to a Persons, Location, Events, etc. Video file describes the digital representation of a video with attributes like frame rate, decoding algorithm, etc. Video File Etype also contains link to low level features of the video entity. The following list of services is proposed for the these Etypes:

- show the entity, displays the given photo
- sort collection, sorts the collection of video by time and events.

2.4.1 Video File

After a deep analysis of state of the art standards we have chosen the list of attributes which from the one hand are essential to describe video file entity and from another hand fit the needs of our entity model. Three categories of attributes is defined for Video File Etype: basic information, audio information and low level representation information.

Basic information contains the information required for video file decoding, the length of video, frame rate, frame size, etc.

Audio information provides information related to the soundtrack of a video file. The examples of these attributes are bit rate and number of channels.

Video File follows the similar to Image Etype (see section 2.3.1) approach of addressing *low level representation information.*

2.4.2 Video

Video being a complex media entity requires appropriate tools for content description and its semantic representation in the framework. Thus, we treat it as the composition of a set of images and accompanying sound. This approach allows us to employ existing mechanism of Photo content description for Video content description. The additional attribute to address frame ID is required. This attribute allow us to reefer a specific frame of the video. The coordinates of a bounding box describe the region on the framework.

The sound track of a video treated separately. In order to refer an interesting part of the sound track we specify an attribute that contains the value of Δt i.e. the beginning and the end of this part (see Figure 2.4).

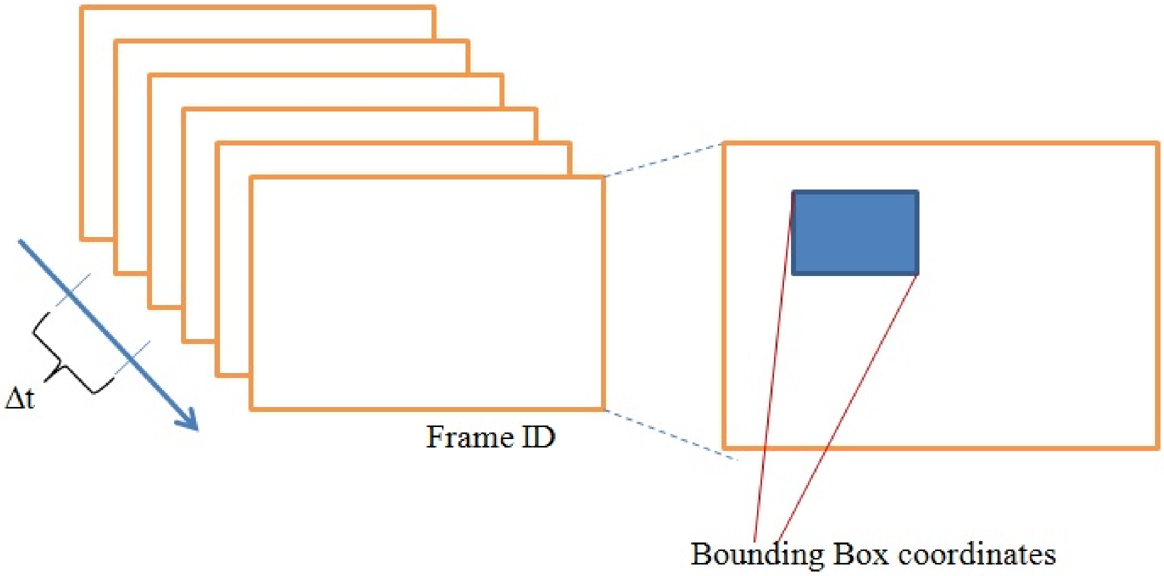


Figure 2.4: Example of Video Content description.

2.5 Conclusion

In this chapter we have described the model for media content and event representation. The state of the art approaches and standards for event modelling are considered and discussed. The essential insight given by the considered state of the art allows us to define possible improvements that can be done. The analysis of metadata shows that events can be described by the common set of attributes like space, time, participants, etc.

The developed model is able to refer specific pieces of media and link them to another Etypes such as Entity, Person, Location, Organization. Moreover, the flexibility of this model eases the process of event media analysis and allows a system to preserve the semantics of media content.

So far we have focused on videos and photos because those media are used further in the thesis. Event mining and exploitation use the spatio-temporal context extracted from images and videos. Further in the thesis we will demonstrate how Location, Person and Event Etypes are involved in the process of data and knowledge enrichment. However, the model integrates other media entities such as music or Sound Entity and text.

Chapter 3

Multimedia Event Mining

3.1 Introduction

In this chapter we describe an approach for multimedia event mining separated on two parts: (i) personal event detection using individual, unsorted photo collections, in which we make use of the spatio-temporal context embedded in digital photos to detect event boundaries within the collection; (ii) social event detection for which we use a tailored similarity measurement between personal events of different users. We introduce the notion of personal events and distinguish them from social.

The chapter is organized as follows: in Section 3.2 we distinguish personal and social events and how they correlate with personal and collective experience. Section 3.3 contains a methodology and experimental validation of an approach for personal event detection. In Section 3.4 a novel approach for detection of social event from personal events is given, while Section 3.5 concludes.

3.2 Personal events vs Social events

Life can be seen as a chain of events that chronologically pace our everyday activities and index our memories. Different events such as a birthday, a business trip or a winter vacation are the lens through which we see and collect our own personal experiences. An **autobiographical memory** records a connected set of personal events and can be characterized by a high personal belief of accurateness, certain perceptual qualities, detailed accounts of the personal circumstances and highly vivid details. Smith et al. [50] hypothesize that memory improves with higher levels of experienced emotion. Their study examined the memories of Canadian students for the (here: social) events of September 11, 2001 in New York. They found that the memory for the 9/11 events was less consistent than for the personal event memory. With higher arousal at the time the news broke, the quality of memory of the 9/11 events did not increase, but the accuracy of the related personal events did ("Where have *you* been at 9/11?"). Therefore the autobiographical memory is positively correlated with the level of emotional arousal. The richness of details shows significant decline by the time passed since the event occurred. In the scope of this work, we therefore argue that a higher frequency in the photo-taking activity correlates with higher personal relevance.

Brown [9] showed that the *retrieval time* in our autobiographical memory is hardly affected from the time that passed since a personal event occurred. Therefore we argue that this is the best way to index media over a long period of time. Autobiographical memory is the structure that can be retrieved in the most convenient way for the user. Details about a certain personal event may be vanished already, but in our memory the personal event itself is still very vivid and at hand. Connections and the **hierarchy of personal events** are predominantly invariant to time. Rarely, personal events "stand alone" after a certain passage of time in the auto-biographical memory, not being connected to other events any more. Interestingly, the younger the test persons are, the less connections are established between personal memories. With age, people seem to draw more and more conclusions based on their personal experiences, but do not change their conclusions drawn at a younger age. A personally meaningful hierarchy of personal events does not become meaningless after time.

Personal events are put into context and are stored in our autobiographical memory. In this sense, we refer to our personal experiences as noumena in the sense of Kant [41]: a personal, possibly inexpressible, reality, that is the sum of personal reasoning in an attempt to understand the world in which we exist. In the course of producing and collecting personal media, people try to represent the memorable events. Unfortunately, the representations are not able to provide the full context of these memories.

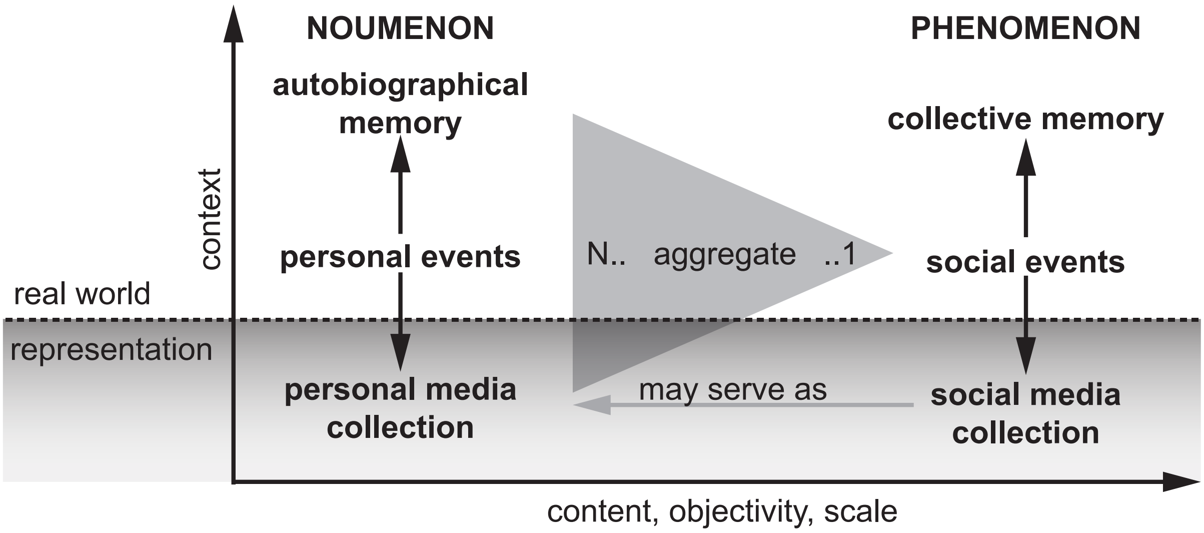


Figure 3.1: Distinction between personal events and social events.

Personal events form a hierarchical narrative-like structure that is connected by causal, temporal, spatial and thematical relations [9]. Figure 3.1 gives a classification in relation to social events.

In contrast, **social events** are phenomena aggregated by multiple personal events. Social events build the basis for the **collective memory**, our common denominator for social communication. Since personal media is shared between people who are either participating or interested in specific social events, social media collections are built. When these collections serve as personal media collections, a loss in context is observed: A picture collection taken by multiple people is not necessarily a representation for one's personal memory, and therefore an automated approach to reason about such an index is to the greatest possible extent unfeasible.

Events are more than simple aggregators for our memories, being essential for collective experiences to take place. These collective experiences arise when two people exchange their individual views of the world and discover common interests and goals based on the roles they play in the events they co-participate in. Thus, we can say that events have a social impact in the shaping of interpersonal ties between individuals.

In contrast, a social event is a convergence of multiple personal events. A social event, being one of the nodes in the collective memory, builds the basis for the collective experience, our common denominator for social communication. Since personal media is shared between people who are either participating or interested in specific social events, social media collections are built. When these collections serve as personal media collections, a loss in context is observed: A media collection taken by multiple people is not necessarily a representation for ones personal memory, and therefore an automated approach to reason about such an index is to the greatest possible extent unfeasible. Events provide the common framework inside which the local experience-driven contextual information can be not only codified but also shared and reduced to a common denominator. Consider a simple example of a photo taken within an party that can be contextualized to a specific place, time or set of participants. We can say that the context accumulated by an event entity contains values for the five Ws (Who, What, When, Where and Why) to fill the complete story.

Naturally, one could argue that there is a strong causal connection between hierarchically close personal events. This would constitute an unfeasible problem for automated approaches aiming at indexing media: a profound understanding of the content of the media in the sense of *strong* artificial intelligence [73] would be necessary for such a task. Luckily for automated approaches, we do not remember causatively: In Brown's experiments, only 2% of the 904 judged personal event connections were causally related. It seems that real "media understanding" is not necessary for a personally meaningful indexing. People tend to connect personal events by (i) a non-narrative relation (17%), which is predominantly a closeness in time, and (ii) participants (13%), and (iii) a common location (7%), and others that are not so significant. These findings give a good explanation why the state-of-the-art in visual analysis of media largely fails in this problem (e.g. [22, 23]).

We build context from time, location and visual appearance by putting them into a higher level of abstraction: Personal photos are visual representations of our personal memory; events are mental pivots which are used by humans to facilitate the extraction of their memories. Therefore, we aim at organizing photo collections following this new paradigm and advocate the use of a personal focus for media indexing using personal events.

3.3 Personal Event Detection

In the last decade we have been observing a tremendous growth in the number of digital images in personal photo collections. Standard photo software lets the user assign photos to *albums*, the metaphor originating from the traditional way of storing developed photos, or solely clusters the media by certain tags. The lack of a proper automated indexing of personal photos leaves the users with the tedious task of manually organizing their photos.

Therefore, we propose a fully automated approach for managing photo collections in the most intuitive, flexible and semantically meaningful way. We aim to aggregate the entities of the autobiographical memory as personal events. This is done via the convergence of multi-modal information in time, space and image content.

We build a meaningful hierarchy of personal photo events. To this end, we learn iteratively from the data-set and discover the routine and non-routine patterns of taking photos. Based on this fundamental distinction, we interpret the location, the frequency and the color layout of the images to extract a personal distinction between events.

Emulating the autobiographic memory in a personal photo collection is the most natural and convenient way for the user. Personal events as memory episodes are milestones in our autobiographical memory.

3.3.1 State of the Art

Many popular photo applications were developed around the album metaphor. This approach comes from how people have been organizing their printed photographs before the advent of digital photography. However, there is a new trend that aims at studying new metaphors of organization to leverage more complex indexing and search.

Although Flickr is mainly designed using the album metaphor, it has introduced a very loose organization system focusing on tags to group photos. Andrews et al. [2] study tagging in Flickr and other services, pointing out that these services do not take into account the semantics that users want to encode in each tag. To avoid ambiguous annotations the researchers provide an alternative model where tags are tied to concepts instead of being free strings. These services can also make use of GPS-enabled devices that "geotag" media automatically. MyLifeBits [28] is an authoring tool for storytelling that use location information to organize media in a map, thus enabling a richer way to present stories. Moreover, recent face recognition capabilities can be used to provide search and navigation based on acquaintances depicted in the photo.

Most popular media management services continue to model tasks around the "album" metaphor as people need to group media together in ways that are more meaningful to them than just a location or time grouping. To model this higher-level intent of the users when they group their media, researchers now focus on the *event* metaphor to combine metadata. This interest in events as media aggregators predates digital photography [11]; however, Chalfen argues that people do not share photos *per se* but use them to tell a story. Miller et al. [53] also concluded that users take photos mainly to archive important events and share within their community.

Time- and visual-based clustering are the most popular approaches for photo organization. [15] and [16] show a fusion for clustering photo collections based on temporal information and visual content of images. [10] presents an innovative approach to event recognition of collections of images: the essence of the paper is to build a collection of photos with time stamp and geographical coordinates, and to define a compact ontology of events and scenes. Their model takes into account two types of correlations: the first is a correlation by time and GPS tags, the latter is a correlation between scenes represented in images and corresponding events. Temporal intervals between photos are used in [62] in order to group these photos using an adaptive threshold. The same feature has been used in [47], reporting very robust results. Low level features of photo content and creation time of the photo are exploited in [44] for the task of automatic *summarization*. Event-level Bag-of-Features (BOF) representation is considered in [37] to model typical events. There, sets of concepts are linked to an event to define an event-related pattern. A similar approach of Bag-of-Features and time constraints have been used for (sub-)event recognition in [51]. Spatio-temporal information and visual appearance have been proposed to build multilayer hierarchical models of scene and event in [10] for image annotation within the context personal photo collection. Detecting the significant subset of events within a personal photo collection using a technique based on time series is studied in [21].

3.3.2 Methodology

This section shows how the contextual information of a photo can be used to build a hierarchy of personal events. We suggest a multi-modal clustering (MMC) approach described in the following Section 3.3.2. The proposed framework around MMC is presented in Section 3.3.2.

Multi Modal Clustering (MMC)

We use a method based on 2-means clustering as shown in [47] and extended in [78] which is independent to the density and robust to extreme variation in the distribution of the data. The data can consist of a high number of (also highly unbalanced) dimensions, but must be sequentially sortable. It allows for a robust clustering without knowing the final number of clusters.

Differently from [47], we apply the 2-means to a combined time-space-routine-visual clustering. The method locates significant gaps in assorted data, where for each sample n_i in the set $n_{i-1} \leq n_i \leq n_{i+1}$ holds, where n_i can be a timestamp, a color vector or GPS coordinates. The method uses the distances Δn between the samples n and clusters them using k-means with $k = 2$. It divides efficiently one cluster with many small distances in the data and another cluster that defines fewer, but significantly more distant gaps between given samples. Every distance is then assigned to be a member of either the first or the latter cluster. For each distance that is a member of the cluster of larger distances c_i, a boundary in the data is marked. For the final cluster estimation, we iterate over all $n_{1..N}$ once. Every n is merged with the current cluster until a boundary is encountered. Then, a new cluster is built. For a data-set D with the samples $n_{1..N}$, the pseudo code is given in Algorithm 1.

For temporal clustering in photo collections, the time intervals between photos are of a too large variation to provide satisfying results. We face problems when a large time difference among events is interpreted as the only point in

Algorithm 1 multi modal clustering

```
MMC(D)
for all $n_{1..N-1}$ do
    $\Delta n_i \leftarrow n_i - n_{i+1}$ // estimation of distances
end for
c = kmeans($\Delta n$,2) // cluster $\Delta n$ with $k = 2$
// boolean c gives true for significant gaps in D
cluster $\leftarrow 1$
for all $n$ do
    $n_i \leftarrow$ cluster // assign cluster number to sample
    if $c_i$ then
        cluster++
    end if
end for
```

the cluster corresponding to event separations (compare Figure 3.2). Therefore, following [47], a scaling function ψ to scale Δn has to be introduced. Using semantically meaningful intervals as there are $y =$ [quarter days, days, weeks, months and years] we apply

$$\psi(\Delta n) = \sqrt[i]{\Delta n}, \tag{3.1}$$

where i is the index of the according element in y. To this end, we are able to adjust this function based on statistical data obtained from a given data-set (i.e., mean time between events, mean duration of an event, etc.) or from semantic terms. Therefore, the only parameter y of the approach is providing flexibility regarding the granularity of the resulting clusters.

Event Detection

Figure 3.3 shows a flowchart of the proposed approach. The numbers (1) to (4) link to the corresponding paragraphs in this section. People tend to think about events in terms of spatial entities leveraged by personal context like *home*, *work*, etc. Moreover, moving away from routine locations typically establishes last-

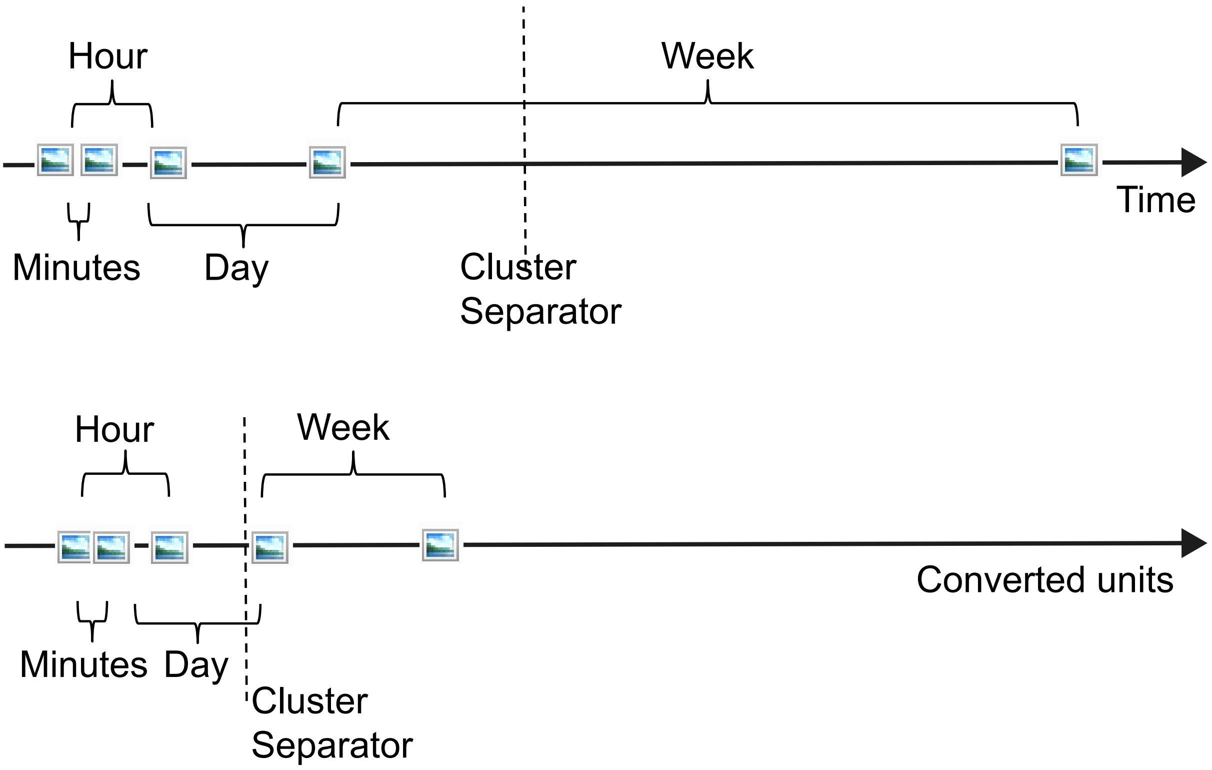


Figure 3.2: Scaling by $\psi(\Delta n)$ and MMC on temporal scale

ing memories. Therefore we advocate discriminating events in two categories: *home* and *away-from-home*.

One fundamental presumption of the approach is that breaking a routine – and returning to it – frames one semantically connected memory. This entity provides the borders of a root event being *away from home*. Since a trip typically starts on routine places, e.g. taking photos at the home airport, photos taken at routine places may belong to an event away from a routine place. A movement out of the routine locations starts a new root event, for which we hierarchically detect sub events until a routine location is entered again. This coincides with a memory of ”my trip to the States”. Typically, in such a memory, there are many very different sub events. Still, the events are connected by the one predominant property of being away from one’s routine locations. This does not apply for *home* events: spatial-temporal clusters can be either root events or sub events.

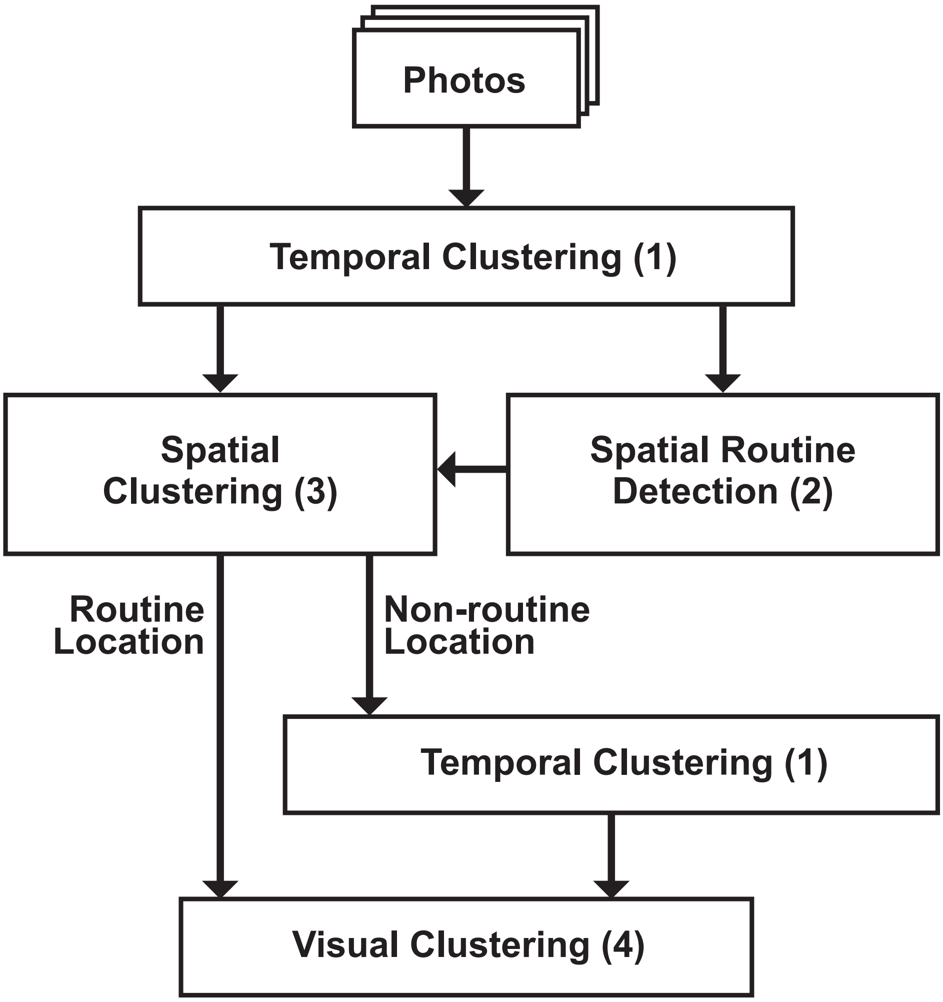


Figure 3.3: Schematic overview over the proposed framework.

No additional context is given a priori. The *home* events span sparsely a long temporal period but occur on a very narrow space scale.

Temporal Clustering (1) Temporal information is the most essential and reliable type of information for detecting events within the personal photo collection. The main reason is that the time stamp is unique, whereas for events the spatial location is not. Time stamps can be easily extracted from the EXIF metadata embedded in digital images. Photos captured within an event are typically characterized by relatively small temporal gaps between them. Therefore the time intervals between chronologically neighboring images are fed to the

MMC algorithm. The clustering results in an assignment of every photo in the collection to the clusters $c_1, c_2, .., c_n$.

Spatial Routine Detection (2) First we map the GPS information of the photos to meaningful GeoLocation[1] by reverse geocoding. Reverse geocoding is the process of converting geographic coordinates into a readable address or place name. We use a granularity of city level, or, if not available, province. This gives us already a meaningful clustering of our photo locations. If several images are missing GPS information, the spatial movement is linearly interpolated between the existing data points. A density function Φ of GeoLocations is built based on the number of days with one or more photos taken – accumulated by location.

The function Φ is given in Fig. 3.7 (b). The presented function is invariant to the absolute number of images taken in one location (compare Fig. 3.7 (a)). In simple terms, we want to define home as where you take the most photos on different days.

The routine locations p_h are determined by the MMC described before: Δn_i are the sample points of ψ. Since the spatial distance of the GeoLocation does not play a role (e.g. moving from New York to Paris is similar as moving from Boston to Washington in the change of routine), we disregard the actual location of the GeoLocation in ψ. Leaving the city or province is already a change of routine.

Spatial Clustering (3) Based on the temporal clusters C_i we discriminate between photos taken at routine locations p_h and photos that are taken outside p_h. As soon as one photo of an event is taken outside of p_h, the event is regarded as being a non- routine event. We spatially cluster all non-routine photos by their GPS coordinates. At this stage, geocoding does not give precise enough results any more. Therefore similarly to temporal clustering we cluster the locations by the spatial distances between chronologically neighboring images into clusters C_{ha} $\{C_{l1}, C_{l2}, .., C_{\mathrm{ln}}\}$. The resulting spatial clusters give us clusters of locations

[1] http://code.google.com/apis/maps/documentation/geocoding/

that may disseminate over various locations. Therefore we are able to overcome mistakes being made when city borders are crossed often. There is no relation between taking two photos in the same location and them being part of the same event. For example, c_{l1} is a cluster of photos taken in Milan. This cluster contains photos related to a Christmas time and Italy Republic Day which should be separated into two clusters. For each cluster C_{ha}, we perform temporal clustering as described above. This results in the final root event clusters and the corresponding sub-events.

Visual MMC (4)

A semantically meaningful color similarity [82] is chosen to make up for varying lighting conditions and camera settings. Following a user study on images downloaded from the Internet the mapping of 11 English color names to RGB coordinates is learnt, thereby creating a look-up table of each RGB coordinate to one of the eleven color names[2].

Without any presumption about the nature of the content of the image, we do not rely on color models when matching visual content, but we merely ask figuratively "*Can this object still be regarded as brown?*". Following the look-up table of colors, as long as people would agree that the color stays the same, we can successfully match two images. A typical example is given in Fig.3.4, where the color of the castle changes significantly, still maintaining similar information from the perceptual point of view. The images are matched successfully to belong to one scene.

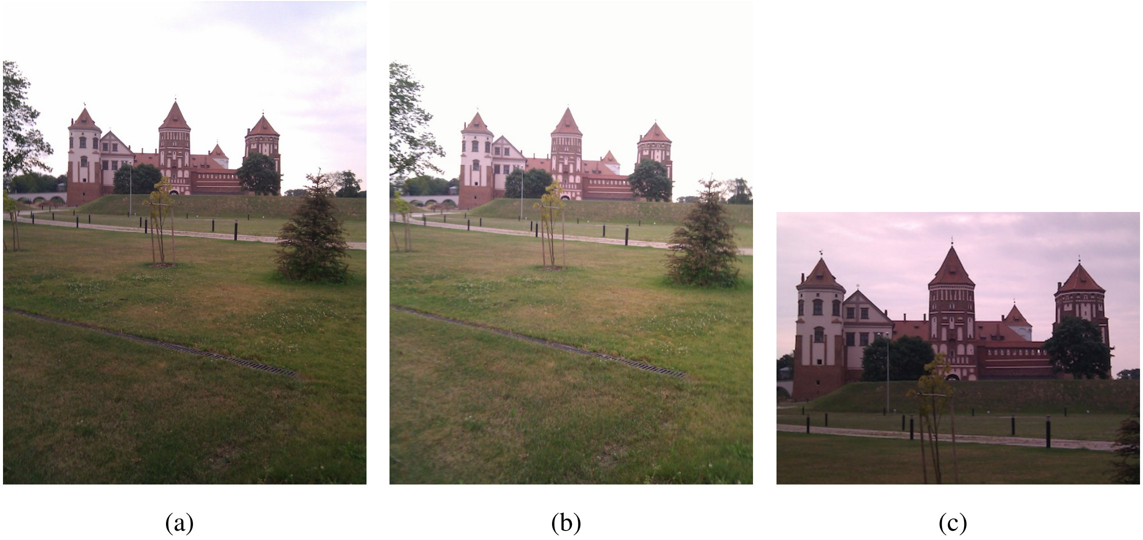
For histogram creation the 11-dimensional feature vector is extracted for each image. Euclidian distances between feature vectors of neighboring images are computed on the next step. MMC clustering is performed to separate event boundaries from similar images related to one event. As shown in Section 3.3.5, the results highly coincide with the event boundaries given by the previous steps of the framework. This leads to a finer subdivision of the events

[2] `http://lear.inrialpes.fr/people/vandeweijer/color_names.html`

User	U1 ♂	U2 ♂	U3 ♂	U4 ♀	U5 ♂	U6 ♂
home	San Francisco, CA	Miramar, FL	San Jose, CA	Tel Aviv, Israel	Newport, GB	Trento, Italy
# pics	2186	5803	16041	1366	16527	1008
# events	36	219	180	22	190	79
# pics/events	61	26	89	62	87	13
sub-events	no	65 (1509 pics)	no	no	51 (4522 pics)	9 (212 pics)
# locations	34	47	71	9	58	15
∅ pics/loc	66.24±63.5	127.91±215.2	522.5±737.51	60.22±78.77	268.55±839.37	67.2
time period	3 years	6 years	11.6 years	4.4 years	12.4 years	0.75 years
movement	193542 km	157799 km	260327 km	9767 km	68823 km	7214 km
km/pic	88.54	26.25	16.54	10.54	4.23	7.16
∅ pics/day	2.07±7.91	2.77±16.30	3.70±22.41	0.05±0.78	3.95±22.67	3.76±12.32

Table 3.1: Overview of the data-set used in the experiments crawled from Picasaweb.

in event *scenes*.



(a) (b) (c)

Figure 3.4: Photo of Castle Mir using different camera settings and view points.

3.3.3 Experimental Data-set

Based on the 9548 albums that are selected as probably referring to events, the album publishers (the Picasaweb users) were crawled. These users were sorted by the number of public albums they published. In this order, the pictures were analyzed to determine if there was complete EXIF information available. Some pictures missed even the time stamp, so these users were disregarded. Spatial information was also mandatory for our approach, therefore users without GPS information or an automatically detected location in the titles were disregarded, too. Moreover we have included one user (user **U6** in 3.4.3) who was asked to create event hierarchy from his personal photo collection. This collection is characterized by highly accurate manually sorted event hierarchy according to the user perception.

We ended up with 6 users. Their albums were refined and 18 event-irrelevant albums such as “favorites”, “nature”, etc. were disregarded. The final data-set consists of a total number of 42,931 photos. Images without GPS coordinates were automatically tagged by geo-coordinates provided by the user annotation. Personal photo collections are heterogeneous in terms of time periods and geographical sparseness. The users classified their images in 726 event-related albums, leading to 59.13 images per album on average. An overview of the data-set is given in Table 3.1. 104 people are recognized by the Picasaweb API with Name and Surname, 23,976 faces are detected in the whole data-set.

The collection of users is very diverse. **U1** is a *nomad* user. The number of events almost corresponds to the number of locations. He shares images when he is traveling, and does not return to prior locations. Spatial clustering is essential for this user. **U2**, **U5**, **U6** try to build the event/sub-event hierarchy by naming their albums following an ad-hoc systems like “Wedding_Ceremony”, “Wedding_Reception”, ... **U4** is characterized by sharing a comparably small number of distinguished images over the years. All the users like traveling, being the most common type of personal event. The travelling distance is diverse over the continents. **U5**, being the only European user, travels a lot, but within a smaller range. Finally the dataset of **U6** consist of the photo that he made by his smartphone but not shared publicly.

3.3.4 Experimental Set-up

The given data-set exemplifies six personal photo collections. As ground-truth, it provides a manual and subjective hierarchy of events done by the very user. Five users are unaware of the experiments. The ground-truth is only justified by their personal experience. We regard the data-set as a temporally serial stream of photos, where we detect event borders. Therefore, we evaluate a 2-class classification problem. For each experiment we compute standard information retrieval measures and refer to them in percent.

$$\begin{aligned}
\text{Precision} &= \frac{100\,tp}{tp+fp}, \\
\text{Recall} &= \frac{100\,tp}{tp+fn}, \\
\text{Sensitivity} &= \frac{100\,tp}{tp+fn}, \\
\text{F1-measure} &= \frac{2\,\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}},
\end{aligned} \tag{3.2}$$

where tp denotes a true positive detection of event borders, tn a correct non-detection of an event border, fp an event being parted wrong and fn an actual event border not being detected. The hierarchical interpretation of the data-set is seen as a 3-class classification problem. Event borders, sub-event borders and no borders. The measures are then estimated analogically.

Every event boundary not detected gives a higher *false negative*, every boundary separating one event gives a *false positive*. In this sense, false positives are probably less problematical, as it is still easy for the user to retrieve a desired photo. On the other hand, a false negative detection ”hides” an event from the user making it harder to find in the final hierarchy. From this perspective, the recall rate is probably more meaningful when organizing one’s photos.

3.3.5 Event-based Album Recreation

We consider the state of the art reference approach proposed by [47] as a baseline. It is based on the temporal information of images only. Throughout the experiments, we refer to it as the **time** approach. With an F-measure of about 60, it already shows promising results. This clearly illustrates that the unique key of the recording time is the most essential information for detecting events. On the other hand, it gives no contextual information, making it impossible to derive a meaningful hierarchy of events. Detailed numerical results are shown in Table 4.2.

User		**U1**	**U2**	**U3**	**U4**	**U5**	**U6**	∅
Time [47]	Prec	27.87	71.01	32.67	73.68	52.11	67.31	54.11
	Recall	94.44	77.17	89.67	71.79	71.96	90.91	82.65
	F1	43.04	**73.96**	47.90	72.73	60.44	77.35	62.57
Spatial	Prec	81.58	96.97	56.10	94.74	98.48	44.00	78.65
	Recall	86.11	14.61	25.14	46.15	34.95	14.10	36.84
	F1	**83.78**	25.40	34.72	62.07	51.59	21.36	46.48
Visual	Prec	2.74	7.42	2.30	6.95	2.81	17.86	6.68
	Recall	58.33	68.49	71.74	76.92	61.38	76.92	68.96
	F1	5.23	13.39	4.38	12.74	5.37	28.99	11.68
Combined approach	Prec	72.34	95.76	66.11	80.00	87.60	80.26	80.35
	Recall	94.44	51.60	92.40	75.68	58.89	67.12	73.36
	F1	81.93	67.06	77.07	77.78	70.43	73.10	74.56
Proposed approach	Prec	75.00	84.76	94.71	76.74	78.75	76.09	81.01
	Recall	94.29	65.29	80.50	89.19	70.00	90.91	81.70
	F1	83.54	73.74	**87.03**	**82.50**	**74.12**	**82.84**	**81.35**

Table 3.2: Experimental results for automatic event detection in relation to the user's album organization.

For the second set of experiments we use **spatial** information only. Spatial clustering gives worse results when the temporal information is disregarded since many events are merged. On the other hand, it provides a precision of almost 86. In this sense, location information improves precision, while time information improves the recall rate.

Visual matching is performed to show the shortcomings of the contextual interpretation of visual appearance. It provides a high recall rate, but separates the data-set in too many different scenes, more than a person would want to organize his pictures. Compared to the given ground-truth, we receive 4 of precision, 70 of recall and 8 of F1-measure. We do not claim to use the best visual event matching system available ([77, 45]). However, we use an efficient approach for matching the color layout. More sophisticated approaches will probably increase the precision, but still would not overcome the semantic gap in the way contextual information does.

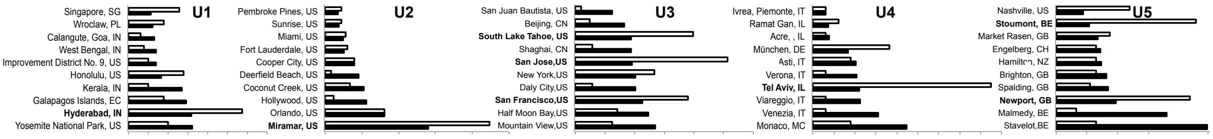


Figure 3.5: Relative number of pictures of the 10 most active locations (black bars) per user. Picture density Φ per location (white bars). Routine locations are marked bold, from left to right: U1 to U5.

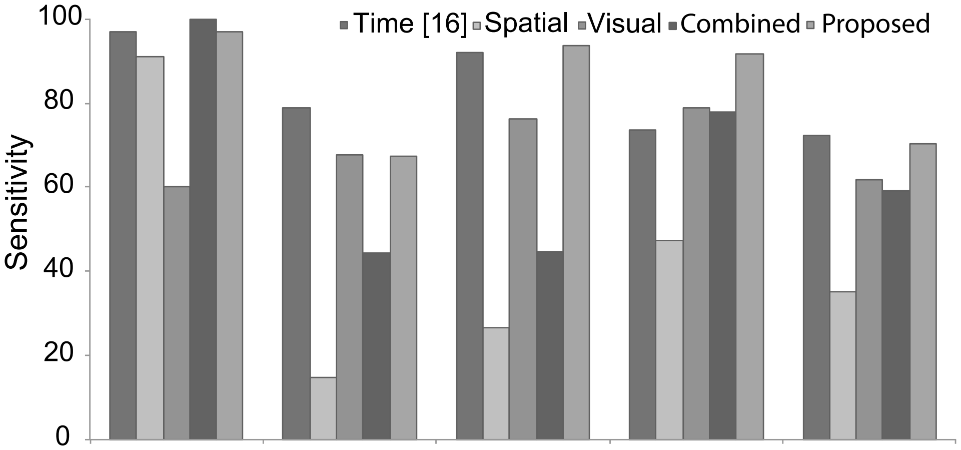
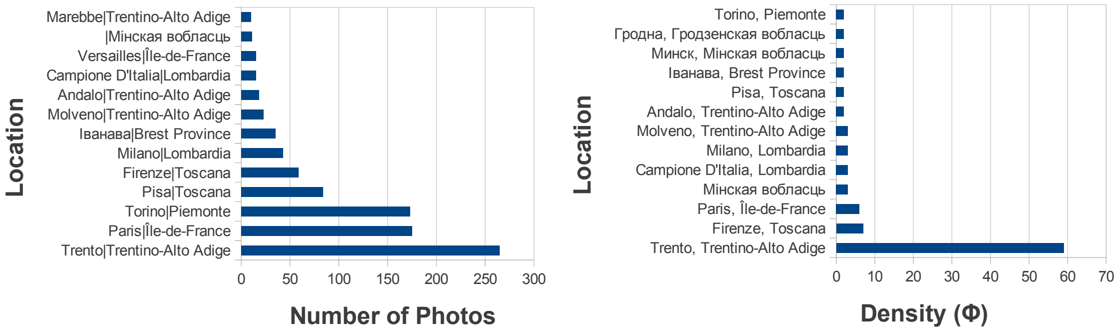


Figure 3.6: Sensitivity per User. Bars from left to right: Time [47], spatial, visual, combined approach and proposed approach.

The combination of the three approaches (**combined approach**) overcomes the weaknesses of the previous approaches and makes the results more balanced and robust, providing a mean F1-measure of 75: we use the method described in Section 3.3.2, but do not perform the spatial routine detection. Adding this provides the full **proposed approach**. Leveraged with this essential context of being in a routine place or not, we are able to further improve the performance to an F1-measure of 80.

The results of the spatial routine detection are given in Figure 3.5. It is best seen for U4 that the actual number of images do not give an indication for routine places. The estimation of the "home" location is correct for U2 - U6. Only for the nomad U1 it is completely wrong. However, there are no photos of his home location in the data-set. For U3, two other "routine" places are found: The place were he was born and raised, and the holiday location were he goes skiing on a regular basis. For U5 the second "routine" location is the location where his motor club meets frequently.



(a) (b)

Figure 3.7: (a)Distribution of photos per location; (b) density (Φ) of days with photos taken per location

The c function Φ for the data-set is given in Fig. 3.7 (b). It clearly shows the significantly higher density of Φ for the home cluster, which is Trento, Italy for U6. The MMC algorithm gives this location as the only routine place, correctly determining it for this experiment. Note that this approach is invariant to the number of photos or photo density, taken per location, as seen in Fig. 3.7 (a) per location and per day.

Sensitivity provides a measure for the ability to identify positive detections. Therefore we evaluate how reliably the event borders are found. The results are given in Fig. 3.6. It is shown for the nomad U1, time, spatial or the combination provides satisfactory results. Since there is no routine in the data-set, the proposed approach is slightly outperformed by the combined approach. Only for U2, time clustering outperforms the proposed approach significantly. U2 builds a very fine granularity of events with only 26 pictures per event. Additionally, he builds a hierarchy of sub-events, which are regarded as a sequence of root events in these experiments. U3 moves a lot between recurrent locations. Therefore sole spatial clustering is of not much help. The combination of the features does not perform that well either. Only by adding the context of routine detection, time clustering is outperformed. U4 shows a very different habit of taking pictures at routine locations or abroad. Therefore, the proposed approach outperforms all others. U5, member of a motor club, travels a lot in a comparably small area. Here time clustering is slightly more reliable than the proposed approach. For U6, the proposed approach works better than other.

To overcome the decreased performance for users that use a notion of sub-events for their album names, we interpret related album names as root events in a hierarchy of sub-events. The hierarchical subset of the data provides 6031 photos, 37 events and 116 sub-events. The results are given in table 3.3. It is shown that we can recreate this mental hierarchy of events with a F1-measure of 45.

User		**U2**	**U5**	**U6**	∅
Proposed approach	Prec.	39.68	36.51	72.22	49.47
	Recall	64.10	46.94	68.42	59.82
	F1	49.01	41.07	70.27	53.45

Table 3.3: Experimental results for automatic sub-event detection in relation to the user's album organisation.

3.3.6 Momentum of attraction

In Section 1.3.2 we have described scenario where a user need to find the most interesting moments. To this end, we propose a measure of saliency or relevance for single scenes in the event hierarchy. The underlying idea is that every photo is taken intentionally. Therefore, the more photos that are taken in a short time, the more interesting or exciting one event should be 3.8.

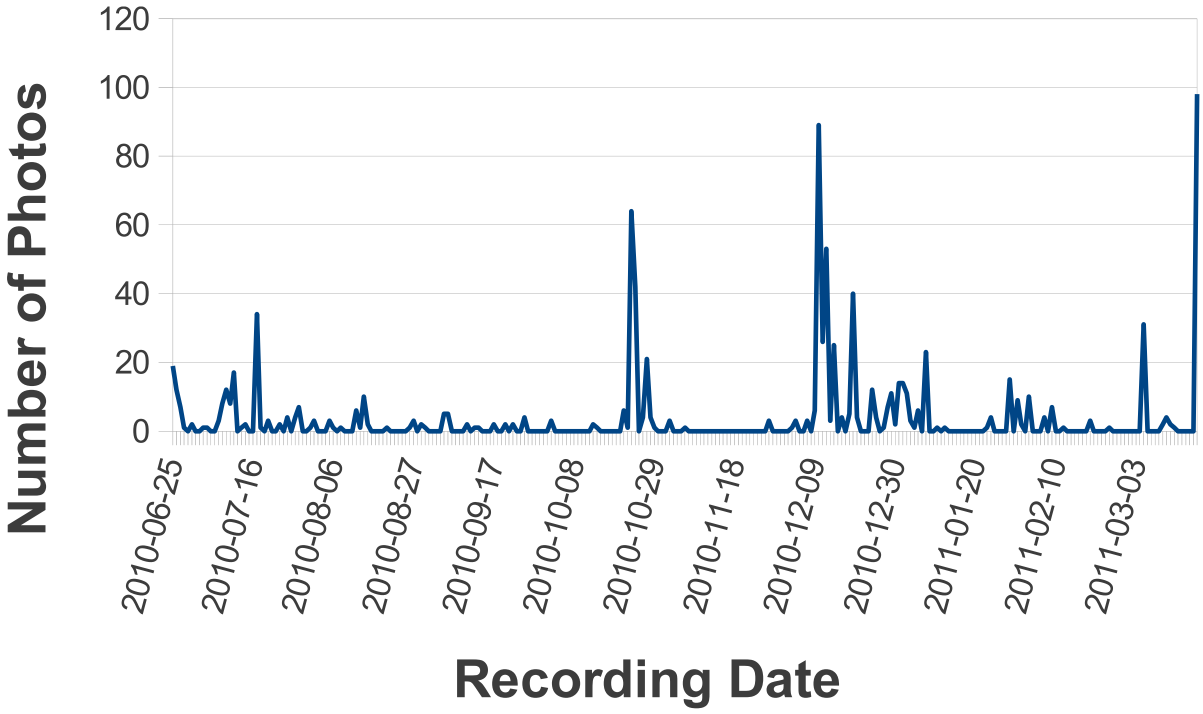


Figure 3.8: Number of photos per day of U6.

We observe the relative change in the recording frequency of photos by measuring the acceleration of the time differences 3.9. This measure is defined as the *momentum of attraction* (MoA). With this measure, we provide a straightforward cue on how scene changes related to personal behavior. The assumption is that things which change our behaviour rapidly are important to us. This allows us to retrieve the most interesting shots conveniently 3.10.

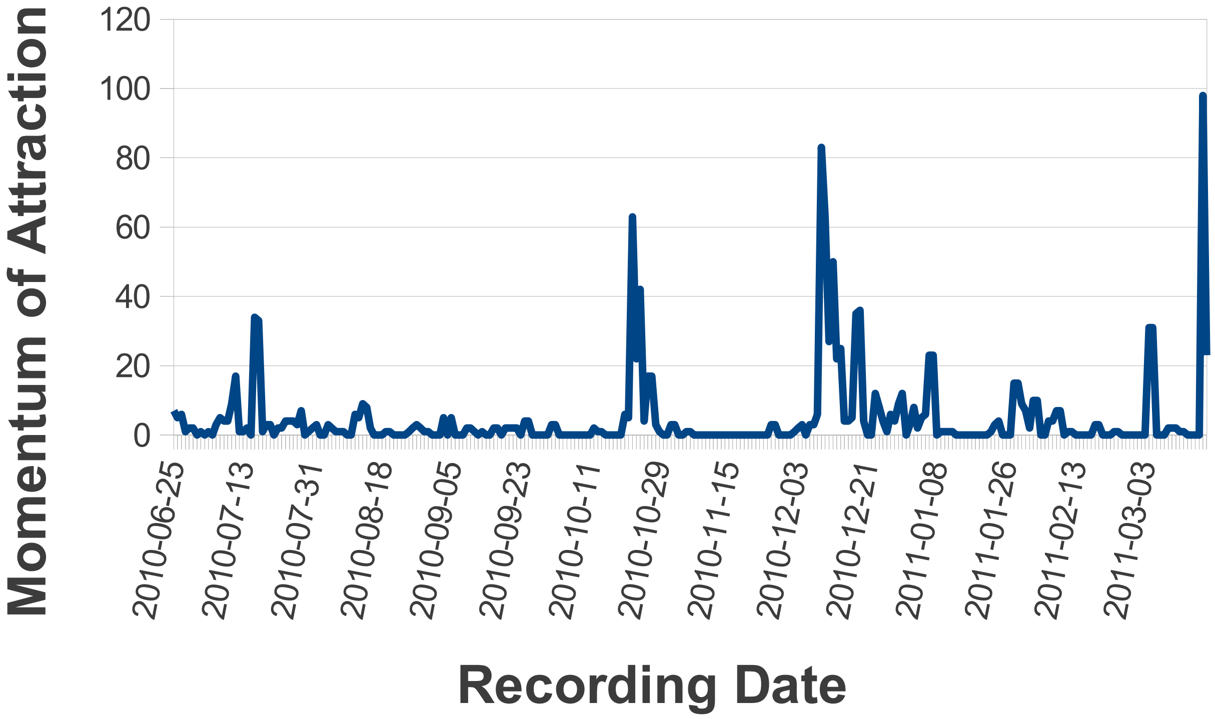


Figure 3.9: Momentum of Attraction of U6: Acceleration of change of behavior on temporal scale.

3.3.7 Growing Data-set

The typical use case for a personal photo collection is the continuous addition of new photos. We simulate this behavior in letting the data-set grow over the years. We start with the first 6 months of each user, and evaluate the proposed approach, the time-based approach and a random detector for the small subset. Then we continuously extend the data-set to 1, 2, 3, 5, 7 years and the full data-set. The collection of U6 was not included in the experiment because of the small time period of his data set. Figure 3.11 shows the average F1-measure and its standard deviation in the error bars over the experiment. The random detector gives the detection prior.

We observe a relation between the performance and the amount of data. The approaches increase their performance, whereas the detection prior decreases. The time-based approach increases its performance steadily, converging after 5

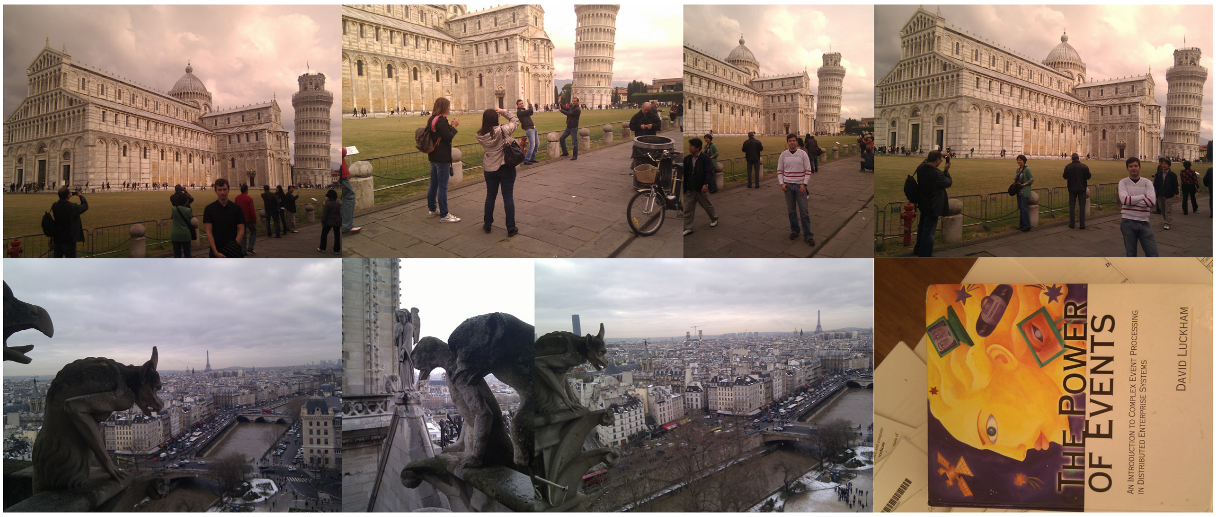

THE POWER OF EVENTS
AN INTRODUCTION TO COMPLEX EVENT PROCESSING IN DISTRIBUTED ENTERPRISE SYSTEMS
DAVID LUCKHAM

Figure 3.10: MoA of U6: Most attractive scene in the upper row, the tower of Pisa. Second attractive scene: on the roof of Notre- Dame, second row. Most unattractive event: a photo of a new book (lower right).

years or 25% of the photos in the data-set. The proposed approach increases its performance with a growing data-set as well, but converges faster: For the first 12 months, the performance is stable. After the second year, the performance is almost final already. After 3 years, some of the users show outstanding performance with a F1-measure of about 90. The reason for this faster convergence lies in the routine detection. People repeat their patterns mainly in the period of one year. Christmas, Thanksgiving, the summer vacation: We implicitly learn this re-occurrence to define personal events more reliably. Already after the third year we are fully aware of a user's routine patterns. The only reason for lower performance is when the user changes his routine frequently. This is the case for U3, where routine places are introduced every other year: The first years he used to live at his parent's, then he moved to San Francisco. Eventually he made enough money to drive around and to go skiing on a regular basis. In times of such fundamental change of the routine, the approach needs some time to re-adjust again, but does not forget about old routines. One's parent's house will always be a well-known place.

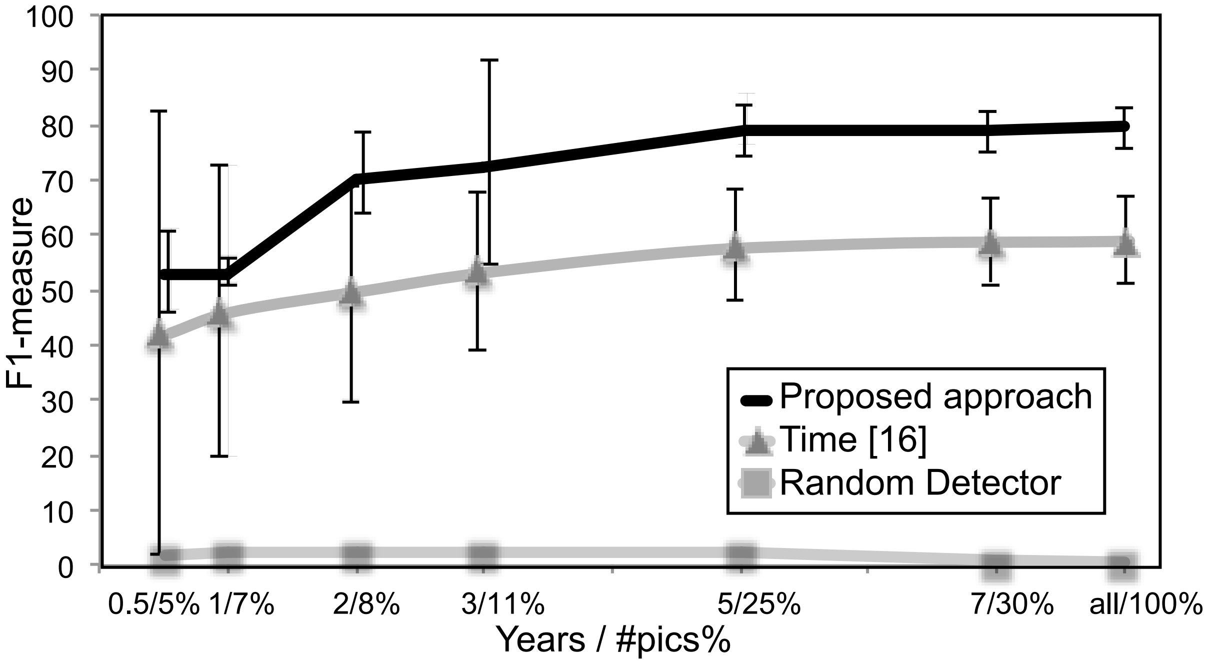


Figure 3.11: Mean F1-measure on the growing data-set.

3.3.8 Discussion

Going back to the scenario described in Section 1.3.6, there is a important attention of how a user sorts personal photo collection. In this section we described an approach for the event-based media indexing. Large scale experiments on Picasaweb show that more than 70% of the publicly shared albums are using this event to describe an album. Based on these findings, we build a data-set of about 42,000 photos with 647 event-related albums. Our approach allows us to improve the current state of the art in event-based media analysis significantly. Through the context of personal routine and personal events we are able to automatically build the index of user images in coherence to the structure of one's individual autobiographical memory. The notion of this phenomenon can be found in the studies of cognitive science. Using very simple contextual cues given by time stamp, spatial location and perceptual color distribution, we are able to mine in one's personal life and behavior. This minimizes the effort for the user to organize his personal photo collection and makes the retrieval of desired photos more convenient. The proposed method uses a flexible MMC approach for all data-sources.

Experiments show that we are able to outperform the state-of-the-art from an F1-measure of about 62. Combining time, spatial and visual information, we reach an F1-measure of about 74, adding the context of the automated spatial routine detection we reach 81.35. Additionally, the performance of the proposed approach converges faster than previous methods in scenarios with growing data. Providing the experiments on the different time frames we conclude that the growing amount of data allows us to increase the performance of the algorithm.

3.4 Social Event Detection

So far, we have considered event detection task as the problem of mining personal events from individual's media collection. However, research on the task of detection of social events have shown a considerable attention recently [52], [70].

Some of the on-line social media platforms (e.g. Flickr[3], Facebook) allow users to organize their media items according to social events. Web services such as Flickr offer the use of machine tags for this purpose. Still, even if social events can be represented in this way, the absence of automated solutions hampers the use of this feature, and is the reason why contextual information which can be used to facilitate media indexing and information enrichment is still relatively unexploited.

In this section we approach the problem of **social events** detection through the layer of **personal events**. Once personal events are detected, the next step consists on finding the intersections between these personal events coming from

[3] www.flickr.com

different users. These intersections are found by calculating the proximity measurement of personal events in the spatial and temporal dimensions. At this stage we make the following assumption: if two or more personal events from users intersect, then they are assumed to be *different personal accounts of the same social event*. This property of an event unveils an enormous potential for annotating personal photo collection based on the description of social events given by other users. Thus, we can employ this event feature to cover issues presented in scenarios of Section 1.3.5 and Section 1.3.3.

3.4.1 State of the art

Recently, numerous studies have been devoted to the social event detection problem. Authors in [35] and [7] use on-line sources of publicly available information (e.g., LinkedData[4], Freebase[5]) to find additional information related to social events. This information is used as a hint to detect the relevance of photos to the given set of social events in the MediaEval benchmark [70]. For the same task, authors in [46] utilize mainly three kinds of features: temporal, spatial and textual. Moreover, they involve visual features for removing noisy pictures. The final step of their approach enriches the detected set with photos that lack textual features using metadata about the owner of the collection.

A multimodal clustering algorithm is proposed for social event detection in [60]. The essence of the approach lies in the classification of distance matrices. These matrices are created computing pairwise distances for each modality. Modalities are based on the following features: temporal, spatial, visual and textual. The authors in [68] exploit machine-learning techniques to identify events in social media streams. They apply support vector machines to classify the dataset of Flickr photos related to events and annotated by machine tags

[4] http://linkeddata.org/

[5] http://www.freebase.com/

from LastFM[6]. Becker et al. define the problem of event identification in social media [4]. They present an incremental clustering approach for assigning documents to event-related clusters. The similarity metric learning approach is utilized by the authors in order to increase the performance of their technique.

The methodology of [40] initially classifies the dataset according to the location of the media, and then relies on a set of topics that match the description of the event detection challenge and are likely to be found within textual metadata in the media analyzed. [24] employs "same class" model for organization of photos into a graph. A community detection algorithm is applied for the analysis of the graph to derive candidate clusters. The last processing step classifies on the relevance to the given challenges. Zeppelzauer et al. [49] approach the task using a spatio-temporal clustering technique to come up with the event candidates, and in further steps these candidates are filtered taking into account textual and visual information.

Some approaches rely on metadata coming from external sources. In [48] the authors detect events related to football matches relying on DBpedia and WordNet to establish if events are relevant to the challenge. Dao et al. [20] employ a watershed-based image segmentation technique for social event detection. In this approach "markers" are produced by using keywords and spatial features with involvement of external data sources.

In contrast with approaches that employ textual and visual cues, our approach is purely based on spatial and temporal features for the social event detection task.

3.4.2 Methodology

In our approach, social events are detected by clustering personal events that are significantly similar. We compute a similarity score in order to match two events using a combination of metadata features. According to [43], of all these

[6] `http://www.last.fm`

metadata features, time and space are the most important to segment our perception of reality in events.

Consider our driving example from section 1.3.4. For this step we assume George's photo collection of air show visits is already organized into personal events (this step is automatized using the approach described in previous section). We can use these "personal" air show events to find other personal events belonging to other users. If we can match his personal events with personal events from other users we will end up with a richer set of photos that will better represent a given air show event, and not just the scenes that were subjectively interesting for George or for each of the other users.

We can group personal events into richer social events working with temporal and spatial proximity. Personal events happening at roughly the same time in the same places are likely different personal accounts of the same social event.

Accordingly, the similarity function used in our approach combines two separate functions that work with time and space. Consequently, we represent an event E as a pair of temporal and spatial information like this:

$$E =< t, \mathbf{P} > \tag{3.3}$$

where t is the time period at which the event occurs and $\mathbf{P}$ is a set of geographic points (such as the ones provided by embedded GPS metadata in photos).

The algorithm first determines if there is a time overlap between events. If two events E_A and E_B are defined as:

$$E_A =< t_A, \mathbf{P_A} > \quad \text{and} \quad E_B =< t_B, \mathbf{P_B} >$$

then time similarity is computed by taking into account how much of t_A overlaps with t_B. An example can be seen in Figure 3.12a.

We take the inverse of the time overlap function, called t_{sim}, to compare how similar two event instances are:

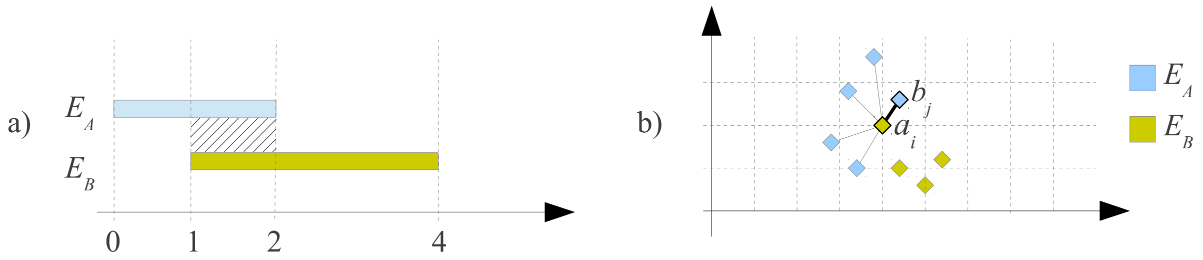


Figure 3.12: a) Time overlap of t_A and t_B and b) selecting the minimum spatial distance of all points p_{bj} of P_B to a particular point p_{ai} or P_A

$$t_{sim}(t_A, t_B) = 1 - \begin{cases} \frac{t_A \cap t_B}{t_A \cup t_B}, & \text{if } t_A \cup t_B \neq \varnothing. \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Events that are completely temporally-similar will have $t_{sim} = 0$ while events that are not similar at all will have $t_{sim} = 1$.

For spatial analysis we use the Haversine distance (denoted $d(p, q)$ for two geographic points p and q) to build another function that selects the minimum distances from a spatial point p to the closest spatial point of a set $\mathbf{P}$:

$$d_{min}(p, \mathbf{P}) = \min_{i=0}^{n} d(p, p_i) \tag{3.5}$$

where $\mathbf{P} = \{p_0, p_1, ..., p_n\}$.

The spatial similarity s_{sim} between two sets of geographic points $\mathbf{P_A}$ and $\mathbf{P_B}$ is then given by:

$$s_{sim}(\mathbf{P_A}, \mathbf{P_B}) = \frac{\sum_{i=0}^{n} d_{min}(p_{ai}, \mathbf{P_B})}{n} \tag{3.6}$$

where $\mathbf{P_A} = \{p_{a0}, p_{a1}, ..., p_{an}\}$ and $\mathbf{P_B} = \{p_{b0}, p_{b1}, ..., p_{bm}\}$.

Both t_{sim} and s_{sim} are *directed*, that is, $t_{sim}(t_A, t_B)$ is not the same as $t_{sim}(t_B, t_A)$. Non-directed versions of these function are:

$$
\begin{aligned}
& nt_{sim}(t_A, t_B) = min(t_{sim}(t_A, t_B), t_{sim}(t_B, t_A)) \\
\text{and} \quad & ns_{sim}(P_A, P_B) = min(s_{sim}(P_A, P_B), s_{sim}(P_B, P_A))
\end{aligned}
$$

Two events E_A and E_B are similar when:

$$nt_{sim}(t_A, t_B) < c_T \quad \text{and} \quad ns_{sim}(P_A, P_B) < c_S$$

where $c_T = 0.5$ and $c_S = 1.0$. These thresholds are learned experimentally as described in detailed in Section 3.4.3.

3.4.3 Experiments

Data Set

To build our dataset we collected events from Yahoo! Upcoming[7], photo albums associated to them from Flickr and their corresponding contact lists.

Our ground truth provides validation for social event detection based on metadata from photos that were annotated with *machine tags*[8] of the form 'upcoming:event=ev_{id}' where ev_{id} refers to an event instance in the Upcoming service. In this way an event from Upcoming can be linked to photos in Flickr that constitute the depiction of such event. In a first instance, photos of the same user that are grouped by the same machine tag are ground truth for that user's personal events. Moreover, personal events from different users that have the same machine tag indicate that these personal events are different accounts of an underlying social event.

Social ties are validated using information about Flickr contacts for each user. Two users that took photos and tagged them using the same machine

[7] http://upcoming.yahoo.com/

[8] http://tagaholic.me/2009/03/26/what-are-machine-tags.html

tag provide evidence that they were co-participants in the same event identified by that tag. We exploit this fact to build participant lists from machine tags contained in photos from different users. To build our ground truth we extracted Flickr contact lists for all the users involved. By testing if two users have each other in their contact lists we can establish whether users that co-participate in a particular event are socially connected or not.

The dataset consists of more than 11 180 events with photos contributed by more than 4100 different users. Of these events, 1291 have photos owned by more than 1 user, a condition necessary if we are to analyze co-participation and its relation to acquaintance. The dataset includes metadata of photos uploaded in the years 2007 - 2012. For this study we only required the Flickr user ID of the owner of each photo, tags for Upcoming events and the list of Flickr contacts for each user. Nevertheless the crawler also retrieved timestamps and, when available, geo-tagging information. The average number of events per user in the part of the collection that we analyzed is 2.05. In the same manner, we computed the average number of participants per event which is 3.71, and the average number of contacts per user which is 213.52. This dataset is available on request.

Experimental results

To derive social events from personal photo streams we proceeded in two phases. First, we found personal events within the user's photo stream using the approach described in previous section. After all individual photo streams were processed and personal event instances detected, we analyzed each personal event against all others, using the spatial-temporal similarity function described in Section 3.4.2.

This similarity function works with two thresholds c_T and c_S above which we do not consider that two personal events are accounts of the same social event. To set these two thresholds we divided our dataset in two groups, using 50% of the detected personal event instances to tune these thresholds and the remaining 50% to test. Pairs of personal events for which the similarity function produced results below the thresholds tested were considered to be part of a bigger social event. In this way we were able to combine several personal events into social event candidates.

To validate a detected social event, each one of them was automatically scanned looking for upcoming tags. If the detected social event had only one upcoming tag throughout its whole photo set then we considered this case as a true positive. It was not possible to do automatic evaluation on detected social events without Upcoming tags or with multiple Upcoming tags. Manual inspection was required in these cases, in order to see if photos coming from different users were still about the same event.

Table 1 shows the parameter learning process. We have selected $c_T = 0.50$ and $c_S = 1.00$ where the percentage (column "TP%") of true positives (column "True Pos.") is higher. Incorrect detection happens due to under-joining (column "U-joint") and over-joining (column "O-joint") of social events. We can also see that there are small variations for temporal threshold c_T of 0.75 and 0.50, with $c_T = 0.50$ resulting in less over-clustering, while a value of 0.25 results in an increased number of partial social events. Based on the set of experiments run, cT = 0.50 and cS = 1.00 produce the best results.

After submitting the testing subset to evaluation we found that the event similarity function produced correct results ("True Pos.") in 78.76% of the cases, partial social events ("U-joint") in 9.56% of the cases and social events with multiple tags ("O-joint") in 11.68% of the cases.

3.4.4 Discussion

This section presents the novel and robust approach for social event detection task. The essence of the approach lies into the exploitation of personal events for detection of social events. We use 50% of our dataset to train our algorithm and

c_T	c_S	U-joint	True Pos.	O-joint	TP%
0.75	0.50	28	265	32	81.54
0.75	1.00	26	265	32	82.05
0.75	5.00	23	264	34	82.24
0.75	10.00	23	263	34	82.19
0.50	0.50	28	267	31	81.90
0.50	**1.00**	**26**	**267**	**31**	**82.40**
0.50	5.00	23	264	34	82.24
0.50	10.00	23	263	34	82.19
0.25	0.50	45	254	14	81.15
0.25	1.00	45	254	14	81.15
0.25	5.00	43	254	15	81.41
0.25	10.00	43	253	15	81.35

Table 3.4: Results for the parameter learning phase, comparing true positives against under-joint (U-Joint) and over-joint (O-joint) social events. Best results where achieved using $c_T = 0.50$ and $c_S = 1.00$.

select the correct parameters for temporal and spatial similarity. Afterwards we use the remaining 50% for evaluation. We validate our assumptions in the wild using 1.8 million public images of more than 4100 users, reaching a precision of 78.76% for full events. If we consider partial events as valid, our precision measurement rises to 88.32

3.5 Conclusion

In this chapter we introduced the notion of personal events and distinguish them from social events. Moreover, we show the correlation of this theory and psychological studies. Following this theory we develop the methodology for mining personal and social events from multimedia data.

We detect and arrange events in private photo archives by putting these photos into context. The problem is seen as a fully automated mining in ones personal life and behavior without actually recognizing the content of the photos. To this end, we build a contextual meaningful hierarchy of events. With the analysis of very simple cues of time, space and perceptual visual appearance we are refining and validating the event borders and their relation in an iterative way. Beginning with discriminating between routine and unusual events, we are able to robustly recognize the basic nature of an event. Further combination of the given cues efficiently gives a hierarchy of events that coincides with the given ground-truth at an F-measure of 0.81 for event detection and 0.53 for its hierarchical representation. We process the given task in a fully unsupervised and computationally inexpensive manner. Using standard clustering and machine learning techniques, sparse events in the collection would tend to be neglected by automated approaches. Opposed to these methods, the proposed approach is invariant to the distribution of the photo collection regarding the sparsity and denseness in time, space and visual appearance. This is improved by introducing a momentum of attraction measure for a meaningful representation of personal events.

We approach social event detection task by analysing the spatio-temporal characteristics of personal events. The essence of the approach lies in the proximity measurement of personal events characteristics. The analysis of 1.8 million of photos demonstrates that approach is able to detect about 79% of social events.

Chapter 4

Exploitation of Detected Events

4.1 Introduction

Events can be seen as aggregator of semantically meaningful information. Therefore events play significant role in the process of information enrichment and metadata reconstruction. This section describes how events can be used for the restoration of missing data and enrichment with the new one. User scenarios given in Section 1.3.1 and in Section 1.3.4 inspired us to devise methods that exploit the semantic power of events.

The chapter is organized as follows. Section 4.1 stands for reconstruction of missing data in particular it describes the problem of digital media with missing geo-information; while Section 4.3 provides us with the hints for the process of data enrichment and demonstrates how social networks can benefit from event exploitation. Section 4.4 concludes the chapter.

4.2 Event-based Geoannotation

The widespread of GPS-enabled digital cameras and camera phones leads to the increasing number of geo-annotated photos. The wide use of spatial information in multimedia is supported by photo management software and on-line sharing tools. Recent studies have shown the importance of geographical information to a user for organizing personal photo collection [58]. This unveils for a user the possibility of sorting and organizing one's digital media collection in geospatial modality. Moreover additional services can be provided based on spatial information extracted from personal media collection [13].

However the vast majority of photos and videos uploaded to on-line sharing services are not geotagged. If they are, the GPS information is not available for all images, or manual annotation is only done for a few images. Therefore automatic techniques for assigning geographical coordinates to the digital media are required [64]. Current state of the art techniques approach this problem using textual and visual analysis. Both techniques require prior training of classifier and availability of a training set for this task. All this leads to a decrease in efficiency. In contrast to the current state-of-the-art methods our approach analyses the context. By the context we mean the spatio-temporal information related to the image provenance. We claim that in the scope of the entire collection of an individual user, the spatio-temporal context information is at least as important for analysis as it is visual content.

We aim to leverage personal events for the task of geo annotation. The importance of event-based indexing for personal photo collection have been recently studied in [58]. Events can be seen as useful entities that provide a way to encode contextual information, and aggregate media that constitute the experience of such event. Events being context aggregators bring semantically meaningful information for a user. Due to the nature of an event space and time information is the most important data to identify an event. However, time information is the primary attribute for detection events in personal media collection. Therefore once we detect temporal boundaries of an event it became easier to estimate missing spatial information for media entities within the detected event. That makes the event metaphor important for the reconstruction of spatial information for media with missing geographical coordinates. Moreover, the analysis of spatio-temporal information is computationally cheaper in comparison with the analysis of visual features, since time stamps and GPS coordinatenes can efficently be extracted from the EXIF[1] metadata embedded in digital images.

We propose Event-based Semantic Interpolation (EBSI) approach that includes two steps:

1. Detection of events and their temporal boundaries within an unsorted and not tagged personal media collection.
2. Assigning missing GPS information for each sample within the temporal boundaries of each event. This is performed by interpolation or extrapolation techniques based on temporal distances between samples. For this purpose we use free on-line navigation services.

Interpolation and extrapolation methods require the presence of geotagged photos within the collection. So we assume that some of the samples in the media collection either were captured by GPS-equipped device (e.g smart phone, camera) or annotated by the owner of the collection.

4.2.1 State of the Art

Current state of the art techniques for automatic geotagging can be separated on the following categories: visual analysis, text analysis and their combination.

Visual analysis

Placing an image based only on visual content on global scale is a challenging task. It is difficult to assign location for an image without any context not only for computers but also for humans. At first glance classification of famous landmarks seems solvable to some extend. But considering more generic scenes like sky, forest or indoor images the appropriate geo-annotation become more

[1] http://www.exif.org/

complex. It happens because of an ambiguity of the image content especially for photos captured indoor. Moreover, *visual analysis* is a significant more time consuming approach than just read the GPS coordinates.

One of the first attempt to place images automatically within the world map is presented in [33]. The proposed approach automatically assigns geo-coordinates for 16% of test images within 200km accuracy. The approach is based on combination of low level features extracted from the training set of geotagged images collected from Flickr[2]. Authors in [91] tackle the problem of placing an image within the urban environment. The work on scene recognition [56] and [67] is related to the image localization task. The work of Hoare et al. [36] presents the approach to triangulate the location of historical images. Their system also able to reconstruct the 3D-model using the old archive photographs.

Annotation analysis

Any kind of textual description assigned to an image is analyzed in order to estimate its location. In contrast with previously discussed approaches placing images and videos on the map requires user involvement in form of textual description. The process of assigning geographical coordinates to an image based on location name provided by a user is called geocoding. Due to the ambiguity of location names (e.g. Paris, France and Paris, Denmark and Paris Hilton) the problem of distinguishing between them may arise. The problem becomes more complex when a user does not mention any location in the textual description. Authors in [74] approaches the problem of geoannotating by creating language model from user's tags. They place a grid over the world map where each cell on this grid defined by geo-coordinates. The approach is similar to bag-of-word technique. The main idea is to assign set of tags and their scores for each cell in the grid. Laere et al. [84] presents two-step approach where on the first stage they use classifier in order to propose the most likely area where a given photo

[2] `http://www.flickr.com`

was captured, and on the next step similarity search is needed to propagate the location with the highest likelihood within the area estimated on the previous step.

Fusion of textual and visual analysis

The combination of visual and textual modalities recently demonstrated promising results [70]. The framework presented in [17] trains classifier based on combination of textual, visual and temporal features. The authors of the framework point out that photos taken at nearby places and nearly in the same time are probably to be related. It is worth to mention that they limit their task to choose one landmark in the city from a given set of ten examples. [42] presents an hierarchical approach for the task. There, textual and visual modalities are used to determine the region where a video was taken and then - based on visual features - propagated towards geographical coordinates. A similar approach is used in [39]

4.2.2 Methodology

We present an Event-based Semantic Interpolation (EBSI) approach for estimating missing coordinates for images with absent geo information. At the first step the system separates a photo collection on a set of event-related clusters ($e_1 - e_4$) based on temporal information (Δt) only. The detail description of the method for event-based clustering of media is presented in [58]. The example is visualized in Figure 4.1, markers with letters indicate photos with GPS data, dark ones are photos without GPS data. Considering the position of the image in accordance to temporal boundaries there are two possible cases for assigning missing data points:

1. Extrapolation (Figure 4.1 (2)) is the task of extending a known sequence of values A_{e_1} or C_{e_4} .

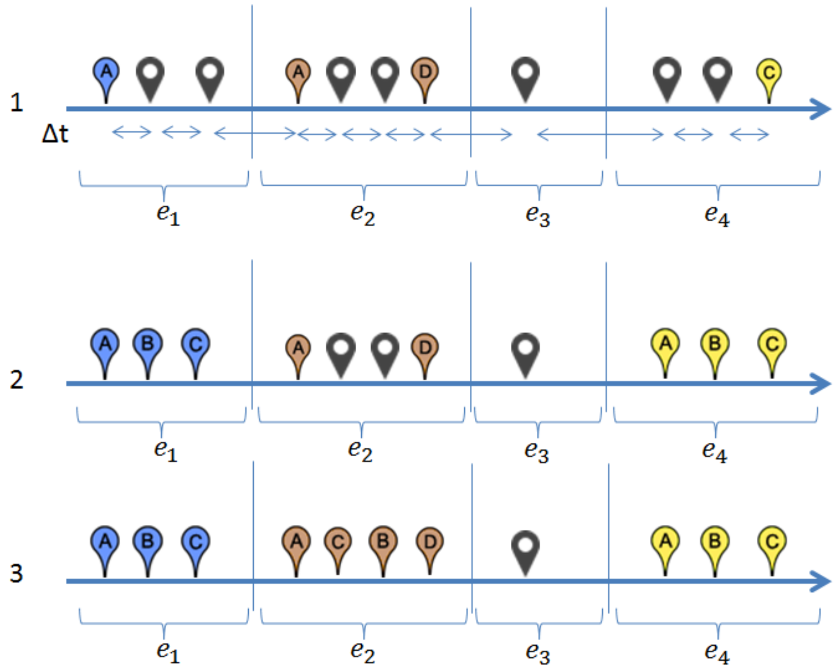


Figure 4.1: Examples of extrapolation (2) and interpolation (3).

2. Interpolation (Figure 4.1 (3)) is the task of estimation of a unknown sequence of samples within two known data points A_{e_2} and D_{e_2}. The linear interpolation can be described by the formula 4.1, where the interpolant y can be computed between two point (x_a,y_a) and (x_b,y_b) on a given x.

$$y = y_a + (y_b - y_a)\frac{x - x_a}{x_b - x_a} \tag{4.1}$$

In the case of extrapolation we extract from the first A_{e_1} or last C_{e_4} geotagged image within an event e_1,e_4 and assign it coordinates to all images without GPS-stamp B_{e_1}, C_{e_1},A_{e_4},B_{e_4} towards the event boundary.

In case of interpolation we do the following steps. Knowing the coordinates of two points where user made photos (A_{e_2} and D_{e_2}) during the event e_2 EBSI requests on-line navigator in order to understand how user moves between those two points. The are three different variants of travel mode: walking, bicycling and driving. As soon as the travel mode is identified the system queries navigator again. This time it quires the coordinates of a point with the given coordinates of initial point, travel mode and temporal distance to the next sample without coordinates. As the result the semantic analysis is done based on suggestions of travel routes using the Google Maps API[3]. If no route is provided, the locations are linearly interpolated based on temporal distances. In case of absence of geotagged samples within an event e_3 interpolation can be done with help of samples from previous or next event (D_{e_2} and A_{e_4}).

4.2.3 Experiments

The data-set consists of 1615 images taken within 1 year and 9,5 month-period. The data-set was produced unintentionally, meaning the owner was not aware that it would be used for this research. All images have time stamps and 901 (55.79%) images have GPS stamps. The images have been captured in five countries and 29 cities and towns. The photos are taken by a Google Nexus One[4] smartphone with a 5MP resolution of 2592 × 1944, sRGB IEC-61966-2 color profile and a fixed focal length of 4,31. For scientific purposes, the data-set is available on request. The given data-set exemplifies a typical private photo

collection. The ground-truth provided by the owner of the collection. The user reconstructed missing spatial information manually with the help of Google Street View[5]. He reported at least 200 m accuracy of placing for each sample. We compared his manual annotation with GPS coordinates automatically assigned to photos by the camera. The results can be seen on the Figure 4.2. The device is able to place only 71% of images within 200 meters error. The results clearly indicate that GPS reception of the device is not always correct. For evaluation of our approach (**EBSI**) we propose to use **linear interpolation** (**LI**) as a baseline.To this end we implemented standard linear interpolation approach.

[3] https://developers.google.com/maps/documentation/geocoding/

[4] http://www.google.com/phone/detail/nexus-one

[5] http://maps.google.com/help/maps/streetview/

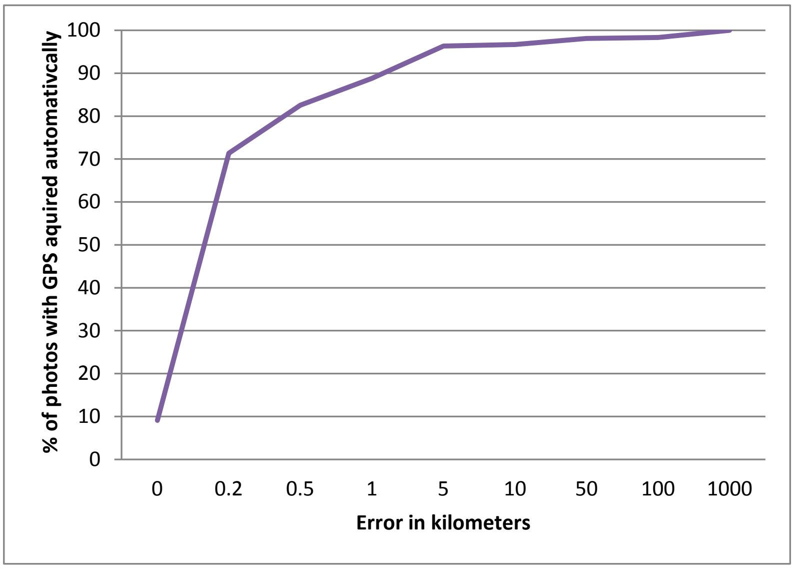


Figure 4.2: Comparison of images with manual geoannotation and assigned by GPS-enabled device.

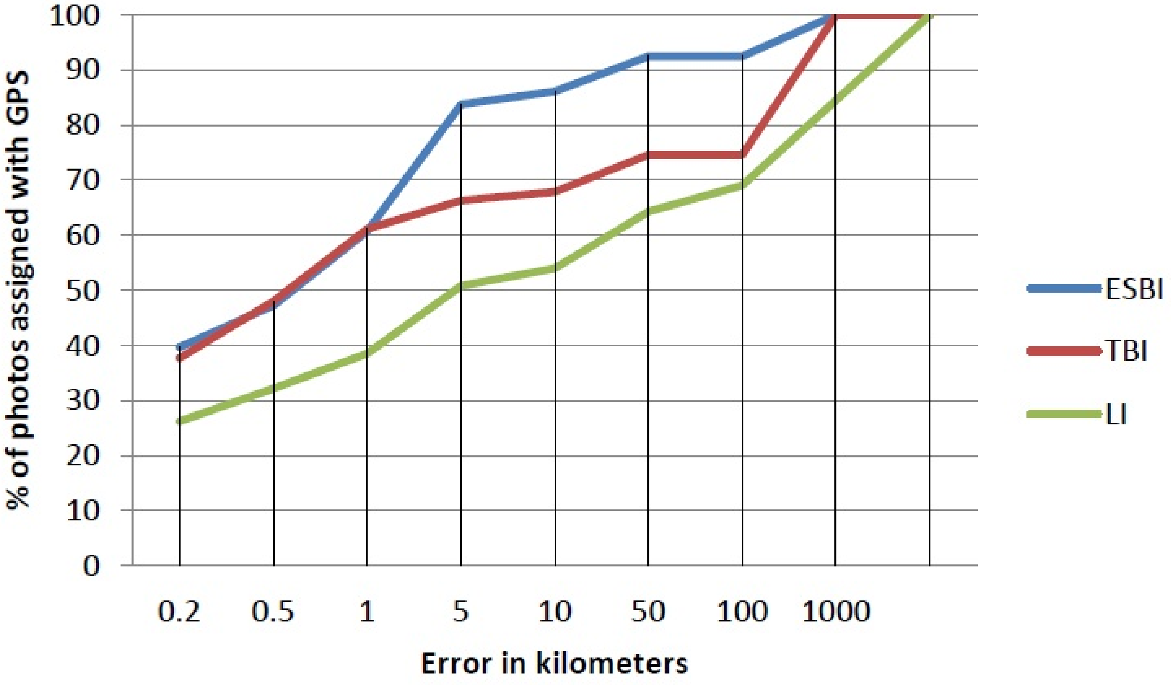


Figure 4.3: Comparison results for different approaches .

Error in km	0.2	0.5	1	5	10	50	100	1000	>1000
EBSI % of images	39.28	47.22	60.71	83.73	86.11	92.46	92.46	100	100
TBI % of images	37.70	48.02	61.11	66.27	67.86	74.60	74.60	100	100
LI % of images	26.19	32.14	38.49	50.79	53.97	64.29	69.05	84.52	100

Table 4.1: Experimental results for *Event-Based Semantic Interpolation* (*EBSI*), *Time-Based Interpolation* (*TBI*) and *Linear Interpolation* (*LI*)

We also tested **temporal based interpolation** (**TBI**) in order to estimate the influence of temporal information for interpolation process. For **TBI** we compute time distances between samples and on their basis perform interpolation. Achieved results presented on the Figure 4.3 and Table 4.1. It is worth to mention that EBSI was able to assign geographical coordinates only for 35.5% from the total number of images with missing geo information. This clearly indicates that vast majority of event-related clusters does not even contain a single sample with geo information. For such a case TBI can be used or the user should be involved. TBI and EBSI shows the similar accuracy till 1 km precision and both significantly outperform LI. However from the next threshold EBSI performance increases noticeably. This leap in performance allows to the system automatically place on the global map more than 83% of test images within the 5km error (Figure 4.3).

4.2.4 Discussion

In this section we describe the novel method for automatic geotagging based on the context of personal media collection. Event-based interpolation of images with missing geographical information demonstrates promising results. The approach unveil the significant role of events which they play in reconstruction of missing geo-spatial information. The experiments show that we are able to assign geographical coordinates for 83% of images within an error of 5 km. This is done without looking at the content of the image. In some photos (Figure 4.4) content information does not provide any cues to distinguish the location where

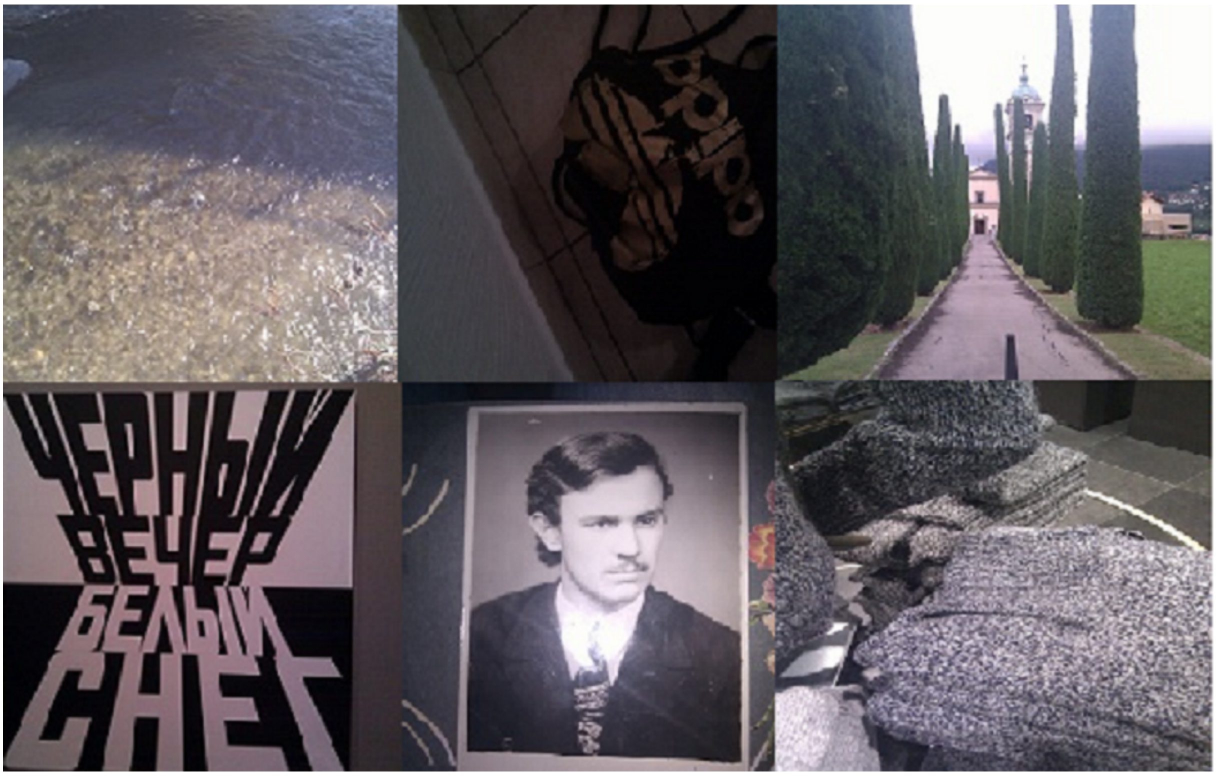


Figure 4.4: Images geoannotated by EBSI with minimal error.

it was captured. The approach does not require any kind of prior training. However the accuracy of the proposed method highly depends on the number of images with assigned GPS coordinates within the collection. We believe that the combination of contextual, visual and textual information can significantly increase the robustness of the automatic geotagging.

4.3 Event-based Social Connections Discovery

Further analysis of social events and their impact on information enrichment leads to the novel and efficient approach of discovery social ties through event co-participation. The key idea of this approach is that the link between the participation in events and the creation of interpersonal ties can be exploited to recommend new user contacts or friends in cases where the number of events that a pair of users co-participate in is above significant threshold. The proposed approach has the following advantages: (i) it suggests new interpersonal connections independently from existing social links and (ii) it can establish new friendship links to users with similar interests (e.g., concerts, sport competitions). Additionally, we found a correlation between the degree of co-participation in events and social closeness. User scenario presented in Section 1.3.4 allows us to understand the approach from the user perspective.

4.3.1 State of the Art

There are previous studies that infer social ties using knowledge about temporal and spatial proximity, i.e. co-occurrences in time and space. [25] predicts social ties based on mobile phone data. Temporal and spatial metadata from photos, as analyzed in [18], can be seen as another source to predict social ties. We argue that metadata from photos are an indirect evidence of *intersections* between users' lives and instead it is co-participation in the same events the underlying phenomena that points towards the existence of social ties.

Some of the potential sources of bias identified in [18] are related to users having potentially inaccurate GPS-enabled equipment or being overly active photographers. We can infer social ties avoiding these issues since our approach depends exclusively on the co-ocurrence of event-related tags without regard to their frequency and independently of the availability of GPS information or even timestamps.

A wide variety of social ties recommendation techniques have been presented recently. They differ from each other mainly in the nature of the information being processed.

Friend-of-Friend (FoF) is one of the most popular and simple techniques. Here we consider the concept of "friend" as a single or bi-directional social link. FoF exploits an existing social network to find the "social proximity" number [12] of shared contacts between two users. The probability of friendship between the users depends on this number. Facebook[6] and LinkedIn[7] have social recommendation services (e.g., "People You May Know") that are based on this approach.

Authors in [12] recommend social connections based on similar interests between users. The approach does an analysis of the content generated by users. A similar approach is used in [89]. The proposed friendship-interest propagation framework combines random walk in social network analysis and a coupled latent factor model. A fusion of FoF and interest-based friendship propagation is presented in [12].

Mobile phone short-range technologies (e.g. Bluetooth) are used by Quercia et al. [63]. They exploit *duration* and *frequency* in order to infer social closeness between two persons. Duration is computed from the total time those persons spend co-located while the number of meetings per time unit between them indicates frequency.

4.3.2 Methodology

The last decade has shown an enormous growth in interest towards on-line social networking services[8]. The effectiveness of these services strongly depends on the ability to reconstruct existing social connections between their members acquainted off-line and on the usefulness for building new connections on-line.

The social function of events plays a significant role in the development of interpersonal ties. Events can be seen as compositions of semantically meaningful entities of which participants, their relationship and roles represent the social dimension of events. In this way we can work with events as a leverage to extract the social connections of an individual.

It is a very boring and tedious task to indicate those connections manually,

[6] `http://www.facebook.com`

[7] `http://www.linkedin.com`

[8] `http://www.searchenginejournal.com/the-growth-of-social-media-an-infographic/32788/`

especially for newcomers. Therefore, automatic discovery of interpersonal links is a vital tool for social networking services. We found that events are useful for reconstructing already existing social ties and for aiding the user in enlarging his social network. To some extent, they encode the interests (e.g., sport events, concerts) of the users that participated in them. Moreover, it is through co-participation in events that one can find other people to share interests and, consequently, build and strengthen interpersonal ties.

Since the presented approach only relies on event co-participation to recommend interpersonal ties, it constitutes an alternative to other approaches based on the analysis of shared contacts.

In our example, George could use our approach to find other users that attended the same air shows (from user scenario in Section 1.3.4). The higher the number of events that George shares with one particular user, the higher the probability that both can have shared interests. This makes them suitable candidates to establish a social tie.

According to our results in Section 4.3.3, people that attend 2 or more events establish a social tie in more than 76% of the cases. That is, when George uploads his photos of air shows, the system would only need spatial and temporal features to suggest present or future acquaintances, even if George did not have any contacts already registered on the social network.

To come up with the results in Section 4.3.3, we have to test if event co-participation is a good predictor of social ties. In order to achieve this we calculate the probability of one user having another user as a contact based on the number of events that they share. If there is at least 1 photo of user u tagged with the ID of an event E_i we say that $u\, participated - in\ E_i$ holds $(u \in E_i)$. Co-participation of two users u_A and u_B in an event E_i is verified if both $u_A \in E_i$ and $u_B \in E_i$ hold.

The degree of co-participation represent the number of events that are shared between two users. That is, let $\mathbf{E_{part}(u_0)}$ be the set of events in which user u_0

was a participant, and $\mathbf{E_{part}}(\mathbf{u_1})$ the corresponding set for user u_1. We define:

$$N_c(u_0, u_1) = |\mathbf{E_{part}}(\mathbf{u_0}) \cap \mathbf{E_{part}}(\mathbf{u_1})| \tag{4.2}$$

as the *degree of co-participation* between users u_0 and u_1.

P_{st} of existence of a social tie between two users u_0 and u_1 follows the natural logarithm of the degree of co-participation N_c:

$$P_{st}(u_0, u_1)) = k \ln(N_c(u_0, u_1)) \tag{4.3}$$

where *k* indicates proportionality.

4.3.3 Experimental results

A first rapid analysis reveals that of the 1291 events with more than 1 participant, 1039 (80.48%) have at least 1 pair of participants that have each other as contacts. This is a good hint that further analysis can give an answer to our hypothesis.

Table 4.2 shows the complete analysis for all the degrees of co-participation that are present in the dataset. We compare occurrences for users that are contacts against occurrences for users that are not contacts. In Figure 4.5 we can see that for a co-participation degree of 1 only 41.39% of all analyzed pairs are of users that have each other as contacts. For a co-attendance of 2 the ratio increases to more than 76% and for degrees of 3 and greater the ratio consistently stays above 90%.

The coarse analysis of the derived data shows that the probability confirms our equation given in 4.3

Additionally we can see a trend in Figure 4.6 showing that for increasing degrees of event co-participation there is a corresponding increase in the average of shared contacts. This supports the idea that events are suitable containers of social information. An increasing number of shared events indicates a growing

Co-participation degree	Contacts	Non-contacts	Contacts (%)
1	3649	5167	41.39
2	534	166	76.29
3	166	18	90.22
4	49	5	90.74
5	27	3	90.0
6	21	1	95.45
7	7	0	100.0
8	7	0	100.0
9	2	0	100.0
11	3	0	100.0
12	2	0	100.0
13	1	0	100.0

Table 4.2: Experimental results for the degree of co-participation in events. The comparison is made counting occurrences of pairs of users participating in the same event. We establish two groups to clasify user pairs depending on wheter these user pairs have each other as contacts (the “Contacts row”) or not (the “Non-contacts” row). Additionally, we provide the results of the “Contacts” row as a percentage relative to the total number of user pairs (the “Contacts (%)” row)

degree of social closeness. This is sustained by the results we obtained taking the alternate approach of measuring the social closeness which is done by calculating the number of shared contacts.

4.3.4 Discussion

In this section we show how to automatically discover social connections via a semantically meaningful analysis of the layer of events. Presented approach is simple and robust. The experimental results show that two users know each other in more than 76% of the cases if they co-participated in at least 2 events.

We also show that the degree of event co-participation is an indicator of social closeness. This is sustained by the analysis performed on the dataset using the well-known FoF approach (see Figure 4.6).

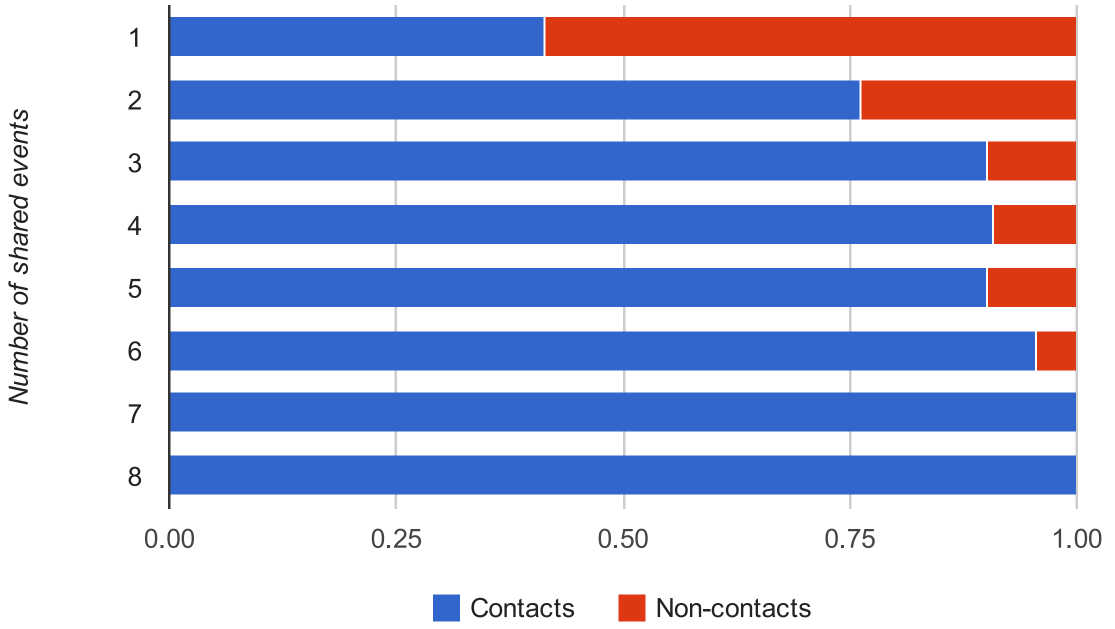


Figure 4.5: Ratio of users with social ties vs. users without social ties per degree of co-participation in the same events.

People that co-participate in an event are likely to have similar interests, especially when we are talking about events such as sport competitions, concerts, festivals and others. Therefore, events can be used not only for reconstructing existing social ties but also for creating new connections based on similar interests. The approach can be applied easily by social networking services in order to enrich a users interpersonal ties.

4.4 Conclusion

In this chapter possible approaches for event exploitation is shown. Being the contextual aggregators events are able to reconstruct data of the described media. Moreover events can play intermediary role for knowledge enrichment.

We employ events for the purpose of geographical annotation of digital pho-

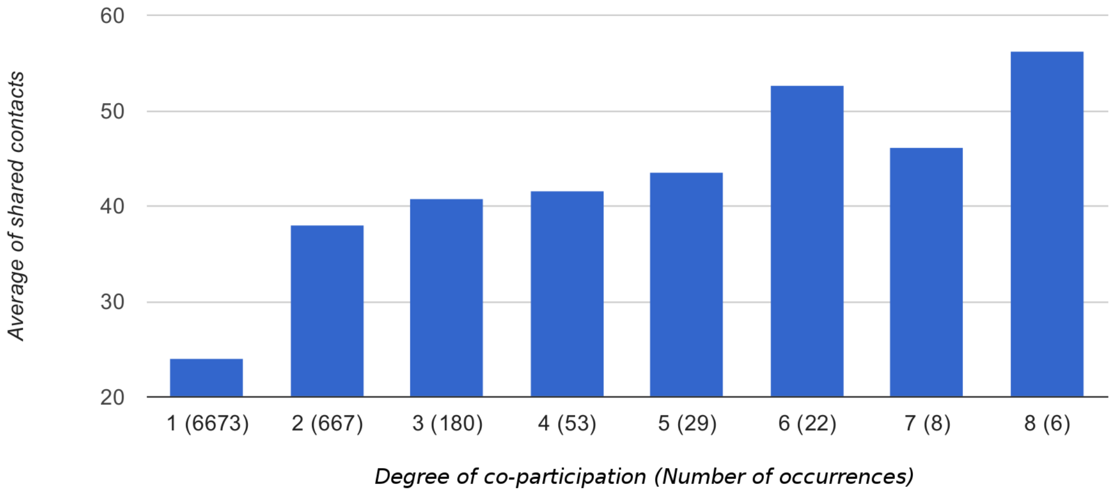


Figure 4.6: Shared contacts per degree of event co-participation. Included in parentheses are the number of occurrences for pairs of users with such characteristics.

tos. The approach for event-based geoannotation of photo with missing geographical information is fully rely on the spatio-temporal context. Based on this context images are combined on the event basis. Existing geo- information about an event is exploited for location estimation of images with missing geostamp. The approach shows promissing results estimating more than 80% of images within an error of 5km.

The approach for social ties discovery proposed another interesting use of events. This simple but robust approach unveils a social ties between two people in case of co-attendance of one or more events. Experiments shown that two people are socially connected in 76% of cases if they co-attended at least two events.

Chapter 5

Conclusion

In this thesis we investigate a new event-based paradigm for multimedia integration, indexing and management. The results of the investigation indicate that event-centered approach for media management has a promising potential in assisting users with multimedia organization and management. However, the review of the popular online tools for sharing media demonstrates the absence of such ability. Considered media sharing sites do not provide the infrastructure even for publishing personal media in terms of events. This gap between user needs and proposed to him facilities is defined during the analysis of the user sharing behaviour and annotation process. Furthermore, to the best of our knowledge, existing state of the art standards are not able to cover all the aspects of event media content representation. To bridge this gap we provide an ecosystem that allows us not only preserve semantic of the multimedia data but also enrich it. The ecosystem consists of the framework for storing and management multimedia content and metadata as well as the set of pluggable services. The proposed framework is able to represent events and media content in a unified entity-centric way. The proposed entity-centric model is the core of our framework. The model treats events, information about them and related media as a set of entities. These entities and relationship among them are codified in the model. This design unveils the ability to operate on the layer of entities. This, in turn, facilitates the data processing by the proposed services e.g. search, event detection, etc.

Within the scope of the thesis we introduce the distinction between social and personal events. Following this strategy we provide separate techniques for mining events from multimedia streams. These techniques solely exploit spatio-temporal metadata for detection personal and social events. Personal event detection algorithm demonstrates promising results outperforming the state of the art approaches. Social event detection algorithm uses detected personal events as input. Both techniques are implemented and validated on the large scale dataset.

The implemented services for personal event detection can assists users in organizing his personal media collection on event basis. Social event detection service also detects photos of different users from a social event and. This service is able to provide access to related media of co-participants of the same social event. Thus the service supports a user within the process of media sharing and enriching the personal media collection. Further analysis of social events allows us to provide a user with the novel and robust approach for social ties recommendation. The analysis of existing data set demonstrates significant potential of the approach. This approach demonstrates how the fusion of the semantic information about events is able to provide a user with the new knowledge. Meanwhile detected personal events can be exploited for the reconstruction of missing geographical metadata of a media item. The service for event based automatic geoannotating of photos and videos enriches the personal collection with valuable geographical information increasing the efficiency of further indexing process.

This thesis shows the value of events for management, indexing, organizing and aggregation of multimedia data. It is convenient to use events while sharing the media with the purpose to tell a story. Therefore from the users perspective events play significant role in media organization. Moreover the power of events to preserve semantic and aggregate context of multimedia makes them the first-class citizens in multimedia analysis approaches. Future work on context analysis includes event type recognition via the analysis of spatio-temporal patterns, investigation of causality aspect of events for event prediction. From the multimedia content perspective we are especially interested in face detection and recognition for event participants identification from images and video.

Bibliography

[1] Paul Alvarez. Using extended file information (exif) file headers in digital evidence analysis. *International Journal of Digital Evidence*, 2(3):1–5, 2004.

[2] P. Andrews, J. Pane, and I. Zaihrayeu. Semantic disambiguation in folksonomy: a case study. *Advanced Language Technologies for Digital Libraries*, pages 114–134, 2011.

[3] Barbara Bazzanella, Themis Palpanas, and Heiko Stoermer. Towards a general entity representation model. In *Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on*, pages 431–432. IEEE, 2009.

[4] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 291–300. ACM, 2010.

[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[6] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and A Maa. Entity name system: The back-bone of an open and scalable web of data. In

Semantic Computing, 2008 IEEE International Conference on, pages 554–561. IEEE, 2008.

[7] Markus Brenner and Ebroul Izquierdo. Mediaeval benchmark: Social event detection in collaborative photo collections. In *CEUR Proceedings of the MediaEval'11 Workshop*, 2011.

[8] Dan Brickley and Libby Miller. Foaf vocabulary specification 0.98. *Namespace Document*, 9, 2010.

[9] Norman R. Brown. On the prevalence of event clusters in autobiographical memory. *Social Cognition*, 23(1):35 –69, oct.-dec. 2005.

[10] Liangliang Cao, Jiebo Luo, H. Kautz, and T.S. Huang. Annotating collections of photos using hierarchical event and scene models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[11] Richard Chalfen. *Snapshot Versions of Life*. Bowling Green State University Popular Press, 1987.

[12] Jilin Chen, Werner eyer, Casey Dugan, Michael Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 201–210, New York, NY, USA, 2009. ACM.

[13] Maarten Clements, Pavel Serdyukov, Arjen P. de Vries, and Marcel J. T. Reinders. Personalised travel recommendation based on location co-occurrence. *IEEE Transactions on Knowledge and Data Engineering*, 1106.5213, 2011.

[14] Linguistic Data Consortium et al. Ace (automatic content extraction) english annotation guidelines for entities, version 5.6. 1 2005.05. 23, 2004.

[15] M. Cooper, J. Foote, and A. Girgensohn. Automatically organizing digital photographs using time and content. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III – 749–52 vol.2, sept. 2003.

[16] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(3):269–288, August 2005.

[17] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World Wide Web*, WWW '09, pages 761–770, 2009.

[18] D.J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[19] Minh-Son Dao, Fausto Giunchiglia, Javier Paniagua, Ivan Tankoyeu, Gaia Trecarichi, Nikolaos Gkalelis, Anastasios Dimou, Vasileios Mezaris, and Alexis Joly. Initial set of multimedia content and event models, 2010. Glocal: Project Deliverable 1.2.

[20] M.S. Dao, G. Boato, F.G.B. De Natale, and T.T. Nguyen. A watershed-based social events detection method with support of external data sources. In *Proceedings of the MediaEval'12 Workshop, Pisa, Italy*, 2012.

[21] M. Das and A.C. Loui. Detecting significant events in personal image collections. In *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, pages 116 –123, sept. 2009.

[22] Aiden R. Doherty, Ciarán Ó Conaire, Michael Blighe, Alan F. Smeaton, and Noel E. O'Connor. Combining image descriptors to effectively re-

trieve events from visual lifelogs. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 10–17, 2008.

[23] Lixin Duan, Dong Xu, I.W.-H. Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1667 –1680, sept. 2012.

[24] S. Papadopoulos Y. Kompatsiaris E. Schinas, G. Petkos. Certh at mediaeval 2012 social event detection task. In *Proceedings of the MediaEval'12 Workshop, Pisa, Italy*, 2012.

[25] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36), 2009.

[26] A. Ekin, A. M. Tekalp, and R. Mehrotra. Integrated semantic-syntactic video modeling for search and browsing. *Transactions on Multimedia*, 6(6):839–851, December 2004.

[27] 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.

[28] J. Gemmell, A. Aris, and R. Lueder. Telling stories with mylifebits. In *IEEE International Conference on Multimedia and Expo*, pages 1536 – 1539, july 2005.

[29] Fausto Giunchiglia, Pierre Andrews, Gaia Trecarichi, and Ronald Chenu-Abente. Media aggregation via events. In *Workshop on Recognising and Tracking Events on the Web and in Real Life*, 2010.

[30] Nicola Guarino, Massimiliano Carrara, and Pierdaniele Giaretta. An ontology of meta-level categories. In *Principles of Knowledge Representa-*

tion and Reasoning: Proceedings of the Fourth International Conference (KR94). Morgan Kaufmann, San Mateo, CA, pages 270–280, 1994.

[31] Lynda Hardman, Jacco Van Ossenbruggen, Raphaël Troncy, Alia Amin, and Michiel Hildebrand. Interactive information access on the web of data. 2009.

[32] J.S. Hare, P.A.S. Sinclair, P.H. Lewis, K. Martinez, P.G.B. Enser, and C.J. Sandom. Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. 2006.

[33] J. Hays and A.A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[34] Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.

[35] T. Hintsa, S. Vainikainen, and M. Melin. Leveraging linked data in social event detection. In *CEUR Proceedings of the MediaEval'11 Workshop*, 2011.

[36] Cathal Hoare and Humphrey Sorensen. On automatically geotagging archived images. *Libraries in the Digital Age Proceedings*, 12, 2012.

[37] Naveed Imran, Jingen Liu, Jiebo Luo, and Mubarak Shah. Event recognition from photo collections via pagerank. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 621–624, New York, NY, USA, 2009. ACM.

[38] R. Jain. Eventweb: Developing a human-centered computing system. *Computer*, 41(2):42 –50, feb. 2008.

[39] Dhiraj Joshi, Andrew Gallagher, Jie Yu, and Jiebo Luo. Inferring photographic location using geotagged web images. *Multimedia Tools Appl.*, 56(1):131–153, January 2012.

[40] P. A. Mitkas K. N. Vavliakis, F. A. Tzima. Event detection via lda for the mediaeval2012 sed task. In *Proceedings of the MediaEval‘12 Workshop, Pisa, Italy*, 2012.

[41] Immanuel Kant. *Critique of Pure Reason*. 1787.

[42] Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*, SBNMA ’11, pages 15–20, New York, NY, USA, 2011. ACM.

[43] Christopher A. Kurby and Jeffrey M. Zacks. Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2):72–79, 2006.

[44] Jun Li, Joo Hwee Lim, and Qi Tian. Automatic summarization for personal digital photos. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1536 – 1540 vol.3, dec. 2003.

[45] Li-Jia Li and Li Fei-fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, oct. 2007.

[46] Benoit Huet Liu Xueliang and Raphaël Troncy. Eurecom at mediaeval2011 social event detection task. In *CEUR Proceedings of the MediaEval‘11 Workshop*, 2011.

[47] A.C. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *Multimedia, IEEE Transactions on*, 5(3):390 – 402, sept. 2003.

[48] E. Izquierdo M. Brener. Qmul at mediaeval 2012: Social event detection in collaborative photo collections. In *Proceedings of the MediaEval'12 Workshop, Pisa, Italy*, 2012.

[49] C. Breiteneder M. Zeppelzauer, M. Zaharieva. A generic approach for social event detection in large photo collections. In *Proceedings of the MediaEval'12 Workshop, Pisa, Italy*, 2012.

[50] Uri Bibi Marilyn C. Smith and D.Erin Sheard. Evidence for the differential impact of time and emotion on personal and event memories for september 11, 2001. *Applied Cognitive Psychology*, 17(9):1047–1055, oct.-dec. 2003.

[51] Riccardo Mattivi, Jasper Uijlings, Francesco G.B. De Natale, and Nicu Sebe. Exploitation of time constraints for (sub-)event recognition. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, J-MRE '11, pages 7–12, New York, NY, USA, 2011. ACM.

[52] Vasileios Mezaris, Ansgar Scherp, Ramesh Jain, Mohan Kankanhalli, Huiyu Zhou, Jianguo Zhang, Liang Wang, and Zhengyou Zhang. Modeling and representing events in multimedia. In *Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, Arizona, USA*, 2011.

[53] A.D. Miller and W.K. Edwards. Give and take: a study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 347–356. ACM, 2007.

[54] L. Nyberg, A.R. McIntosh, R. Cabeza, R. Habib, S. Houle, and E. Tulving. General and specific brain regions involved in encoding and retrieval of events: what, where, and when. *Proceedings of the National Academy of Sciences*, 93(20):11280–11285, 1996.

[55] N. Ohare, H. Lee, S. Cooray, C. Gurrin, G. Jones, J. Malobabic, N. OConnor, A. Smeaton, and B. Uscilowski. Mediassist: Using content-based analysis and context to manage personal photo collections. *Image and video retrieval*, pages 529–532, 2006.

[56] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[57] D. Pack, R. Singh, S. Brennan, and R. Jain. An event model and its implementation for multimedia information representation and retrieval. In *IEEE International Conference on Multimedia and Expo*, volume 3, pages 1611 – 1614, june 2004.

[58] Javier Paniagua, Ivan Tankoyeu, Julian Stöttinger, and Fausto Giunchiglia. Indexing media by personal events. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 41:1–41:8, New York, NY, USA, 2012. ACM.

[59] Javier Paniagua, Ivan Tankoyeu, Julian Stöttinger, and Fausto Giunchiglia. Social events and social ties. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, ICMR '13, pages 143–150, New York, NY, USA, 2013. ACM.

[60] Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on*

Multimedia Retrieval, ICMR '12, pages 23:1–23:8, New York, NY, USA, 2012. ACM.

[61] David B. Pillemer. Momentous events and the life story. *Review of General Psychology*, 5(2):123, 2001.

[62] John C. Platt, Mary Czerwinski, and Brent A. Field. Phototoc: Automatic clustering for browsing personal photographs. Number MSR-TR-2002-17, 2002.

[63] Daniele Quercia and Licia Capra. Friendsensing: recommending friends using mobile phones. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 273–276, New York, NY, USA, 2009. ACM.

[64] Adam Rae, Vannesa Murdock, Pavel Serdyukov, and Pascal Kelm. Working notes for the placing task at mediaeval 2011. Proceedings at MediaEval'11 Workshop, 2011.

[65] S. Rafatirad, A. Gupta, and R. Jain. Event composition operators: Eco. In *Proceedings of the 1st ACM international workshop on Events in multimedia*, pages 65–72. ACM, 2009.

[66] Y. Raimond and S. Abdallah. The event ontology. 2007.

[67] Laura Walker Renninger and Jitendra Malik. When is scene identification just texture recognition? *Vision Research*, 44(19), 2004.

[68] Timo Reuter and Philipp Cimiano. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 22:1–22:8, 2012.

[69] Leonard Richardson and Sam Ruby. *RESTful web services*. O'Reilly Media, Incorporated, 2007.

[70] V. Mezaris R. Troncy I. Kompatsiaris S. Papadopoulos, E. Schinas. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *Proceedings at MediaEval'12 Workshop, Pisa, Italy*, 2012.

[71] A. Scherp, T. Franz, C. Saathoff, and S. Staab. F–a model of events based on the foundational ontology dolce+dns ultralight. In *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, pages 137–144, New York, NY, USA, 2009. ACM.

[72] Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. F–a model of events based on the foundational ontology dolce+ dns ultralight. In *Proceedings of the fifth international conference on Knowledge capture*, pages 137–144. ACM, 2009.

[73] John R. Searle. Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417–457, oct.-dec. 1980.

[74] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 484–491, 2009.

[75] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. Lode: Linking open descriptions of events. In *Proceedings of the 4th Asian Conference on The Semantic Web*, ASWC '09, pages 153–167, 2009.

[76] P. Sinclair, M. Addis, F. Choi, M. Doerr, P. Lewis, and K. Martinez. The use of crm core in multimedia annotation. In *Proceedings of First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, 2006.

[77] J. Stottinger, J.R.R. Uijlings, A.K. Pandey, N. Sebe, and F. Giunchiglia. (unseen) event recognition via semantic compositionality. In *Computer*

Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3061 –3068, june 2012.

[78] Ivan Tankoyeu, Javier Paniagua, Julian Stöttinger, and Fausto Giunchiglia. Event detection and scene attraction by very simple contextual cues. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, J-MRE '11, pages 1–6, New York, NY, USA, 2011. ACM.

[79] Ivan Tankoyeu, Javier Paniagua, Julian Stöttinger, and Fausto Giunchiglia. Sounding out semantic event detection in electronic health records. In *Proceedings of the 7th Russian Bavarian Conference (Erlangen, October 10-14, 2011)*, RBC'11, 2011.

[80] Ivan Tankoyeu, Julian Stöttinger, and Fausto Giunchiglia. Context-based media geotagging of personal photos. Technical report, University of Trento, 2012.

[81] Ivan Tankoyeu, Julian Stöttinger, Javier Paniagua, and Fausto Giunchiglia. Personal photo indexing. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1341–1342. ACM, 2012.

[82] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.

[83] Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011.

[84] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Finding locations of flickr resources using language models and similarity search. In

Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11, pages 48:1–48:8, New York, NY, USA, 2011. ACM.

[85] Huan Wang, Xing Jiang, Liang-Tien Chia, and Ah-Hwee Tan. Ontology enhanced web image retrieval: aided by wikipedia & spreading activation theory. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 195–201, New York, NY, USA, 2008. ACM.

[86] U. Westermann and R. Jain. E - a generic event model for event-centric multimedia data management in echronicle applications. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, ICDEW '06, pages 106–, Washington, DC, USA, 2006. IEEE Computer Society.

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

[88] Helen L Williams, Martin A Conway, and Gillian Cohen. Autobiographical memory. *Memory in the real world*, page 21, 2008.

[89] Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 537–546, New York, NY, USA, 2011. ACM.

[90] Mark Sandler Frederick Giasson Yves Raimond, Samer Abdallah. The music ontology. In *Proceedings of the International Conference on Music Information Retrieval*, pages 417–422, 2007.

[91] W. Zhang and J. Kosecka. Image based localization in urban environments. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pages 33–40. IEEE, 2006.