

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/254462624>

Discovering inherent event taxonomies from social media collections

Article · June 2012

DOI: 10.1145/2324796.2324852

CITATIONS

8

READS

91

3 authors, including:



Minh-Son Dao

Università degli Studi di Trento

23 PUBLICATIONS 162 CITATIONS

[SEE PROFILE](#)



G. Boato

Università degli Studi di Trento

138 PUBLICATIONS 2,126 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Face anti-spoofing [View project](#)



Device linking [View project](#)

Discovering Inherent Event Taxonomies from Social Media Collections

Minh-Son Dao
mmLAB - University of Trento
Via Sommarive, 5 I-38123
POVO(TN), Italy
dao@disi.unitn.it

Giulia Boato
mmLAB - University of Trento
Via Sommarive, 5 I-38123
POVO(TN), Italy
boato@disi.unitn.it

Francesco G.B. DeNatale
mmLAB - University of Trento
Via Sommarive, 5 I-38123
POVO(TN), Italy
denatale@ing.unitn.it

ABSTRACT

Events are becoming very popular as a tool to organize and access large media collections. An unsolved problem however, is how to define event models. Most part of the approaches so far proposed in the literature are based on a-priori knowledge, and translate into hierarchical data structures or taxonomies a more or less intuitive definition of what a given type of event is. The association of media and event models is then a consequent process, in which one tries to learn the distinctive characteristics of media associated to a certain event or sub-event. In this paper, we attempt to reverse this paradigm, inferring from a set of media collections belonging to the same event class the underlying taxonomy in an unconstrained way. As a result we obtain a hierarchy of natural clusters, largely shared by the different collections, which capture the essence of the event itself. Although it is not possible to compare the proposed approach with state-of-the-art method based on a-priori event structures, experimental results demonstrate that this approach may become an effective support for discovering and defining event models and managing event-related data collections.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models, Search process

General Terms

Algorithms

Keywords

Event Taxonomy, Social Media Collection, Natural Clustering, Gist of the Scene, Saliency Map

1. INTRODUCTION

The use of "events" as a tool to index and organize media collections is more and more popular, as events represent

the most natural way to remember and reproduce the facts that happen in our lives (see, for instance, [17][12][14] and references therein). Personal events, such as a wedding or a vacation, and large-scale social events, such as sports, music, politics, are the lens through which we observe and memorize our personal experiences. The number of events happening every day all over the world is countless, nevertheless, the categories to which such events belong is far more limited, thus making it possible to define models that fit large numbers of event instances. As an example, notwithstanding the diversity in cultural or social habits, "weddings" share a common cliché in that they typically include some kind of ceremony, a party or a gala dinner with friends and relatives, group pictures among spouses and other attendees, and so on.

Such similarities leave a kind of fingerprint in media that relate to a specific type of event, in the form of repetitive patterns that may include the presence of certain visual concepts, color combinations, backgrounds, etc. In this context, appropriate classification techniques fed with suitable descriptions of the media content as well as with additional contextual information (e.g., time, space, camera setting information) may learn from a set of examples how to discriminate among media depicting different kinds of events. Just to quote most recent contributions on this topic, in [9] authors propose a framework for vision-assisted tagging of personal image galleries using people, events, and location information, by exploiting a probabilistic model of context. In [16], images and videos are represented as vectors of responses given by various visual concept detectors, and analyzed with a subclass discriminant method to identify the most appropriate concept subspace and thus recognize the corresponding event. In [7], fast and robust event detection is achieved by extracting a unique signature from a set of photo collections related to a given event category, and by recognizing the event type of a new photo album computing and classifying its signature. In [19], authors propose to use geo-location information (informative features derived from traces of GPS coordinates) and bag of visual words for event recognition, by combining analysis of such features both on the entire collection and on individual photos. Both time information and GPS coordinates are used in the first step of the hierarchical method proposed in [6], where first event clustering is performed on personal photo collections and then visual feature are exploited for scene level analysis. Very recently, in [15] a new method is introduced which mines and ranks the discriminative object patterns for event classification and characterizes albums by the relative fre-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMR '12, June 5-8, Hong Kong, China

Copyright ©2012 ACM 978-1-4503-1329-2/12/06 ...\$10.00.

quencies of discriminative patterns.

All the above methods, however, imply the availability of some a-priori information about the event characteristics, in the form of a model, a description, or some type of ontological knowledge. This information is not always available and, more than this, there is no unique and shared understanding on how events should be described and how to code this information, although some researchers have recently suggested the possibility of standardizing the models for ontological knowledge representation (see, for instance, [3]).

In this paper, we propose a new approach that allows moving a step ahead along this direction but somehow reversing this paradigm, by providing an automatic tool for discovering the taxonomy of events in personal photo albums. The idea is to analyze a set of diverse photo collections referring to the same event type, with the objective of capturing the commonalities and the repetitive patterns shared among them. To this purpose, time information is first used to cluster subgroups of pictures that are densely concentrated in a given time frame (and therefore possibly connected to a salient moment in the event) [11]. Then, a rich visual description including visual saliency and gist information is extracted from the clusters to build a sort of cluster signature. Finally, clusters are matched across different collections based on the similarity of the descriptions and the consistency in time, to determine the common patterns. The emerging structures are used to extract the inherent event taxonomy contained in the original samples, which is then presented as a selection of representative shots. Although it is difficult to provide performance comparison with state-of-the-art methods relying on a-priori event models, experimental results demonstrate that the methodology described in the following sections may become an effective support for discovering and defining event models and managing event-related data collections.

2. METHODOLOGY

The main idea underlying the present work is that photo collections of similar events contain *common patterns* and *common semantics* which can be extracted and used to build event taxonomies, given a sufficient number of training samples. Although sub-events can obviously change from a specific event to another, the proposed method allows identifying representative moments within the single album and to define so-called *common factors* of different collections of the same event type. This allows the generation of specific event taxonomies.

We assume to work with personal photo albums, where several images are usually taken during interesting and important moments. This assumption implies the possibility of extracting and analyzing suitable patterns with common semantics, as described in the next paragraphs. In fact, the proposed method borrows some ideas from video processing techniques (especially *video shot* and *key frame* extraction) and is based on following considerations:

1. Since people use to take a high number of consecutive pictures around an interesting episode, images falling within the relevant time frame could be seen as *frames* of one *video shot* (i.e., images possibly sharing *common patterns*). Thus, we can extract a *key frame*, which can be used as a template of that group of photos.

As an alternative, the template can be calculated as a representative point in the space of image descriptors (e.g., the centroid of the cluster), called in the following *signature image*. Both representations are used in the proposed algorithm.

2. Since different instances of the same event type - thus different photo albums of similar events (e.g., weddings, graduations) - usually contain *common patterns* and *semantics*, *key frames* that appear frequently across different instances can be considered significant *factors* representing such event type. Clustering *key frames* to generate a dictionary of *factors* and analyzing their time relations could then lead to the definition of a sort of *inherent event taxonomy*.

Building upon these observations, we propose a method based on four processing steps (see also Figure 1):

1. *Shot* detection using time information (described in Section 3)
2. *Key frame* extraction (described in Section 4)
3. *Factor* generation (described in Section 5)
4. *Event taxonomy* construction (described in Section 6)

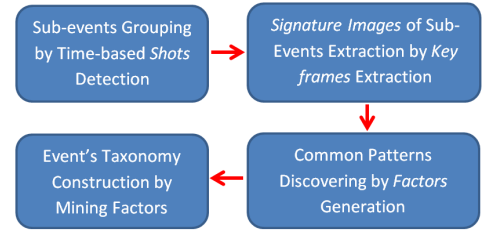


Figure 1: Overview of the proposed method

3. SHOT DETECTION USING TIME INFORMATION

Starting from the above observations, we assume that time differences between consecutive images associated to the same moment of an event, or *sub-event*, are typically much smaller than time differences between images belonging to different sub-events in the same photo album. Therefore, we propose to use the timestamp extracted from the EXIF of images to split the gallery into *shots*. Although there are many different methods to perform time-based clustering, we decided to adopt the algorithm described in [11], since it provides a good compromise between performance and computational cost. Details of the procedure are reported in Algorithm 1, where $\|time(I_i), time(I_{i+1})\|$ denotes the absolute distance between timestamps (in milliseconds) of two consecutive images I_i and I_{i+1} in photo album A , and $|A|$ is the number of images in A . An example of shot detection for three different albums of weddings is provided in Figure 2.

4. KEY FRAME EXTRACTION

The purpose of this second step is to identify each shot with a single key frame representing the most common pattern within the shot (see Figure 2, where key frames are

Algorithm 1 Time-based Shot Detection [11]

Input: Photo Album A **Output:** Set of shot $\{S_k\}$

1. Sort ascendingly all images I_i in photo album A by time (i.e., chronological order)
 2. FOR $i = 1$ TO $|A| - 1$ DO
 - (a) Calculate time difference $\Delta_i^t = \|time(I_i), time(I_{i+1})\|$
 - (b) Perform appropriate scaling for Δ_i^t of the time difference axis
 3. Compute histogram $H(\Delta_i^t)$ for each Δ_i^t
 4. Divide the histogram H in two parts using K-mean clustering algorithm ($K=2$). The cluster with higher values contains the time differences corresponding to separations between *shots*
 5. Identify $\{S_k\}$ based on these separations
 6. Return $\{S_k\}$
-

highlighted in red). This is done based on visual appearance, i.e., identifying the images within the shot that mostly resemble one each other. As a matter of fact, we assume that most significant photos in a shot, besides time contiguity (when), will share common information about actors (who) and place (where), thus showing a significant visual similarity.

There are several ways to choose visual features to measure visual similarity, including dominant color analysis, textures, interesting points, saliency maps, gist [10][14]. In this work, we exploit visual features able to capture *Gist*, *saliency*, and *structural information* from images and we use such information to compare images. Gist and saliency represent two different approaches to understand the meaning of images: scene-centered primitives, and object-centered primitives, respectively [13]. The former pays attention to general characteristics of the images, known as Gist (or Gist of the scene). For instance, when looking at a picture with dominant color "white", people tend to think about "snow" and events related to snow such as "skiing". The latter focuses on discovering regions in images that capture the attention of the observer, known as saliency areas. In fact, when taking photos people tend to place the objects or persons that attract their attention into specific places in the picture, depending on several factors. Finally, structural information is used to find similarity in visual objects. In that case, an important consideration is to ensure the independency of the similarity measure on the perspective view. In fact, people often takes shots zooming in/out or changing the position to better acquire interesting places or subjects. Figure 3 shows an example of two images which can be considered similar following to the above considerations.

From the algorithmic viewpoint, Gist descriptors are achieved through dominant colors analysis [8], saliency derives from the analysis of dominant colors within the salient area (computed as in [1]), while perspective-invariant correspondences are calculated based on SURF [4]. The following sections report the definition of the metric built on these descriptors and its use to extract the key frames.

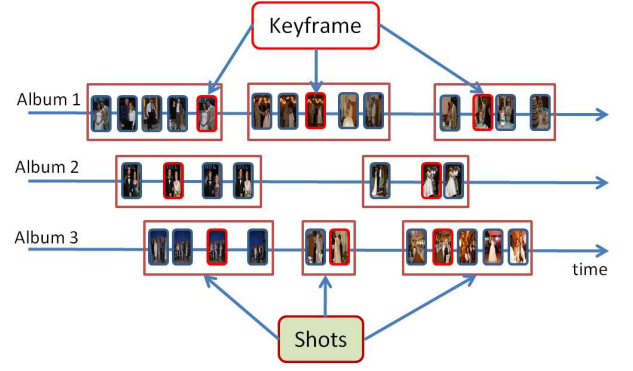


Figure 2: Time-based Shot Detection and Key Frame Extraction (key frames are emphasized by RED rectangles)

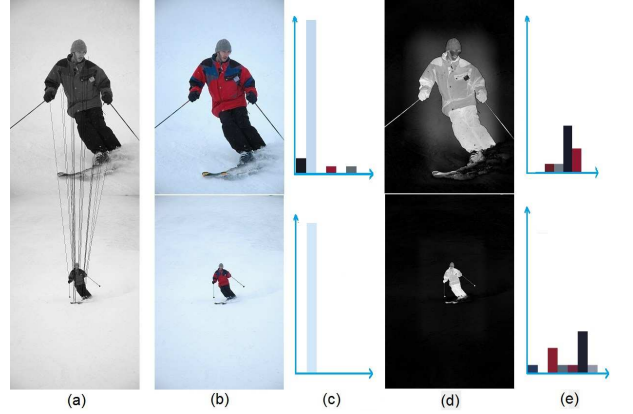


Figure 3: These two images (left-top and left-bottom) are considered similar: indeed Gist, saliency, and perspective allow to merge them into one cluster. (a) Perspective match; (b)(c) Dominant color match; (d)(e) Dominant color match within the salient region (dominant color histograms are arranged so that D_F is the best)

4.1 Visual Similarity Distance

To calculate the overall dissimilarity measure between image I_i and image I_j , we consider each of the three features mentioned above equally important and independent, thus defining the following formula:

$$VD(I_i, I_j) = VD_{ij} = VD_{ji} = \frac{D_{SURF}(SURF_i, SURF_j)}{3} + \frac{D_{DCD}(DCD_i, DCD_j)}{3} + \frac{D_{DCD^S}(DCD_i^S, DCD_j^S)}{3} \quad (1)$$

where $D_{SURF}(SURF_i, SURF_j)$, $D_{DCD}(DCD_i, DCD_j)$, and $D_{DCD^S}(DCD_i^S, DCD_j^S)$ are the similarity distances between SURFs, top 8 dominant colors extracted from a whole image, and top 8 dominant colors extracted from salient region of image I_i and I_j , respectively. Therefore, VD_{ij} is then proportional to the visual distance between I_i and I_j .

4.2 Key Frame Selection

Given a shot S_k that contains a set of images $\{I_i\}$ we aim at organizing them into several clusters so that all im-

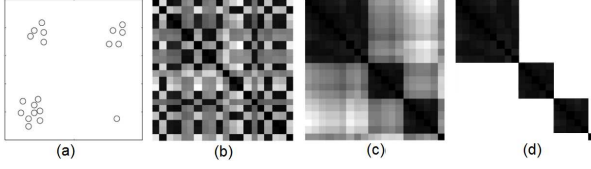


Figure 4: VAT cluster assessment for a set of 20 point in R^2 : (a) scatter plot of X , (b) unordered image, (c) VAT-ordered image, (d) number of clusters after thresholding ((a), (b), and (c) are from [5])

ages in each cluster share common patterns and differ from other clusters. Several state-of-the-art clustering methods could be used, such as K-means, X-means, Spectral clustering, SVM, to name a few [18]. Since our goal is to ensure the selection of a significant key frame, which is mostly connected to a perceptive analysis, we adopted a human-centric scheme to semi-automatically select the optimal number of clusters by exploiting the procedure described in [5]. In such work, Visual Assessment of clustering Tendency (VAT) is performed via a so-called VAT-image, introduced to visually manage and control the clustering tendency of a dissimilarity matrix representing the distances among every pair of items in a given dataset. In other words, this method rearranges the dissimilarity matrix so that a user can easily control the number of clusters to achieve the desired result. Figure 4 illustrates an example of VAT cluster assessment and user’s selection.

In our case, the VAT image is constructed exploiting the procedure described in [5] based on the previously calculated dissimilarity measure VD_{ij} . Then, the VAT image is used to control the clustering by an appropriate threshold selection. Finally, the largest cluster is selected as the most representative of the shot, and it is used for key frame extraction. Finally, the key frame is extracted with the following algorithm 2:

Algorithm 2 Key Frame Extraction

Input: Shot S_k

Output: Key Frame KF_k of S_k

1. Calculate a dissimilarity matrix VD_{ij} using Eq. (1) and imposing $VD_{ii} = 0$
 2. Run algorithm VAT [5] on VD_{ij} to generate the corresponding VAT image
 3. Record the threshold value T from users who assess VAT image
 4. Cluster shot S_k into set of clusters CS_k using T
 5. Calculate the centroid of the selected cluster and call it *signature image* SI_k
 6. Select the *real image* RI_k whose feature vector is closest to SI_k
 7. Return $KF_k = \{RI_k, SI_k\}$
-

As a result we achieve a real image (RI), representing the selected cluster (called template image), and a feature vector (SI), being the centroid of the cluster itself in the feature space (called signature image). The pair *template image-signature image* is called *key frame*.

5. FACTOR GENERATION

At this point, each photo album is represented by a set of key frames $\{KF_k\}$, where k is the index of shots identified within the collection. Figure 5 (a) shows a possible configuration of key frames generated on a set of training photo albums referring to the same event class, aligned along the time axis. In order to avoid the duplication of key frames, we further cluster them into sets, based on their visual similarity. To this purpose, algorithm 2 is applied again, using the key frames of all image collections of the same event type as input. After that, we have a set of distinct key frames for a given event type, consistent in time (see Figure 5 (b)). Such key frames contain most significant common patterns of an event.

Given a set of training photo albums $\{A_m^E\}$ of event E , $m = 1, \dots, M$, we aim at defining a set of *factors* containing both *common patterns* and *common structures* of E . Let $\{KF_k^E\}$ denote distinct key frames extracted from $\{A_m^E\}$, with $k = 1, \dots, K^E$. Let $T_m^E = (KF_{k_1}^E, KF_{k_2}^E, \dots, KF_{k_{n_m}}^E)$, $m = 1, \dots, M$ and $k_r \in \{1, \dots, K^E\}$, denote the *event structure* of album A_m^E . Then we define the set of factors representing E , which contain both common structures and common patterns of the analyzed photo collections, as follows:

$$Factor^E = \left(\left\{ T_m^E \right\}_{m=1, \dots, M}, \left\{ KF_k^E \right\}_{k=1, \dots, K^E} \right)$$

Table 1 reports the factors of event E created from albums A1, A2, and A3 of Figure 5.

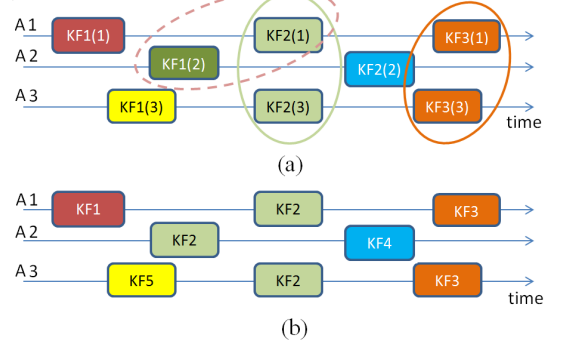


Figure 5: Factors Generation: (a) and (b) illustrates results before and after grouping similar key frames and assigning factor labels, respectively

Table 1: Factor $Factor^E$ of event E

Common structures	$(KF1, KF2, KF3) \rightarrow E$ $(KF2, KF4) \rightarrow E$ $(KF5, KF2, KF3) \rightarrow E$
Common patterns	$KF1, KF2, KF3$ $KF4, KF5$

6. EVENT TAXONOMY DISCOVERY

In previous steps we’ve been able to discover and record a set of common patterns and common structures that characterize a given event. In this last step, we discuss how to take into account the *common semantics* of events, thus defining an *inherent event taxonomy*.

It is well known that there are spoken and unspoken rules in events, and it is much easier construct taxonomies for spoken-rule events. For instance, it is well accepted that a *wedding* usually starts from *rehearsal*, continues with *ceremony*, and ends with *party*. Thus, we have a transaction representing the wedding as $\{rehearsal, ceremony, party \rightarrow wedding\}$ [2]. On the other hand, for unspoken-rule events is hard to build a compact taxonomy easily understandable under natural language form. For instance, a picnic could have diverse sub-events with no strict time order, such as $\{eating, playingongrass, eating, trekking, playingongrass\}$. In this case, it is not possible to identify the *picnic* event as a single transaction, but only with a set of possible transactions. In the following, we provide a inherent taxonomy definition which is based on previously defined factors and can deal with both types of structured and non-structured events.

Let us consider $Factor^E$ as a transaction database and use the associate rule mining [2] to discover all hidden transactions and to emphasize which transaction is the best for representing the taxonomy of event E (i.e., using "support" and "confidency" rule-strength measures). Using associate rule mining manner, the data in Table 1 could be analyzed and restructured as in Table 2. It is worth noticing that a *transaction database* is a set of *transactions*, where each transaction T_m^E is a set of distinct items, as defined in Section . For example, (KF1, KF2, KF3) are distinct items of transaction T_{A1}^E (see Figure 5 (b)).

Finally, the classical a-priori algorithm [2] is applied to mine all transactions in $Factor^E$. These transactions are then sorted in descending order, according to their confidence and support. We define the inherent event taxonomy using only the closed item sets (an item set is closed if none of its immediate supersets has the same support as the item set). The following equation describes the mathematic format of the inherent event taxonomy:

$$IET^E(ms, mc) = \left(\left\{ T_m^E \right\}_{m=1, \dots, M_{supp}}, \left\{ KF_k^E \right\}_{k=1, \dots, K^E} \right)$$

where $ms = min_support$, $mc = min_confidence$, $M_{supp} \leq M$ is the number of closed item sets; $\forall i \neq j: T_i^E \neq T_j^E$, and $KF_i^E \neq KF_j^E$.

The inherent event taxonomy is very flexible in term of understanding semantics of events and creating events' ontologies. Meanwhile the $\{KF_k^E\}$ component could give a hint as which sub-events/concepts should be used to represent an event; the $\{T_m^E\}$ component could answer how to link/order these sub-events/concepts temporally and logically towards to build event ontology. The parameters ms and mc could control the number of transactions by which the demand of large storage could be alleviated, and the speed of querying could be increased. Concretely, the higher value the ms , the higher appearant frequency the transaction is. In other words, if we want to discover the rare transaction, we should lower the value of ms , and vice versa.

For instance, the second row and the second column of Table 3 describe the inherent event taxonomy $IET^E(20\%, 80\%)$ of event E . If event E does not require the strict order relation among sub-events, we can say "event E should have sub-events KF2, KF3, KF5, KF1, and KF4". If event E requires the strict order relation among sub-events, we can say "event E should have sub-events KF1 and KF2 that happen before sub-events KF3". The second row and the third col-

umn shows that when the ms value is high, 50% in this case, some rare transactions cannot be discovered such as (KF1, KF2, KF3), (KF2, KF4) and the number of transactions decreases dramatically.

Table 2: Transaction database of event E

Transaction ID	KF1	KF2	KF3	KF4	KF5
A1	1	1	1	0	0
A2	0	1	0	1	0
A3	0	1	1	0	1

Table 3: Rules mined for event E

	Itemsets	
	ms=20%	ms=50%
Strong rules	KF5 KF5, KF2 KF5, KF2, KF3 KF5, KF3 KF1 KF1, KF2 KF1, KF2, KF3 KF1, KF3 KF2 KF2, KF3 KF2, KF4 KF3 KF4	KF2 KF2, KF3
Closed Itemsets	KF2 KF2, KF3 KF2, KF4 KF1, KF2, KF3 KF5, KF2, KF3	KF2 KF2, KF3

7. EXPERIMENTAL RESULTS

In this section, we describe the setup and the results of the performance evaluation carried out on the proposed method. In particular, we describe the dataset used, the assessment criteria, and a selection of experimental results.

To the best of our knowledge, this is the first attempt to extract the underlying structure of an event starting from the analysis of associated media, without any a-priori information or modeling. This makes rather difficult to discuss the results in quantitative terms (what we achieve is a kind of natural structuring of the information, which is hard to measure in terms of classical precision-recall parameters), and even more difficult to compare it with state of the art approaches (which are typically based on the use of event models, ontologies, or other knowledge representations). This said, we propose here a qualitative comparison, whose main objective is to demonstrate how the discovered inherent event taxonomy contains the majority of the characteristics that are commonly understood as the main features of a given event type.

7.1 Corpora

To test our tools on real data, we used a dataset collected and freely distributed in [12]. Such dataset contains multiple media collections for a certain number of event types (including personal events, social and sports events, etc.), thus making it possible to validate our discovery methodology over a large set of different examples with diverse characteristics in terms of time, visual content, environment, and so

on. Furthermore, it is very realistic as it is made up of images taken from the real world (mainly, social networks). In order to facilitate the evaluation phase, we focused on events that were already analyzed by researchers working on event analysis using personal photo album (see, for instance, [6]). This lead to the selection of 5 event types, with multiple instances per event and around 100 JPG images per album. All pictures have UTC time information (i.e., timestamp). Table 4 shows the details of the used dataset in term of size and number of instances.

Table 4: Dataset

Event	No. Albums	No. Images
SeaHoliday	15	2253
SkiHoliday	15	1878
Baseball	15	1635
Graduation	15	1815
Wedding	15	1776

7.2 Evaluation

As previously mentioned, it is difficult to compare our results with the state of the art, due to the very different target of our approach. Notwithstanding this premise, we attempt to validate our approach with respect to a ground-truth built on the basis of some related state-of-art methodologies. We consider a given a set of events for which a formal definition has been defined in the literature, and we try to see if our method is able to capture their structure and characteristics. In particular, we consider the following approaches:

1. *Method in [12]*: the authors define the semantic structure of different types of events (see Table 5): these structures are used in the following for evaluation of the proposed method, proving that the discovered inherent taxonomy captures the event characteristics (e.g., subevents) analyzed in [12].
2. *Method in [6]*: the authors propose the definition of each event using natural language (see Table 5), describing the scene and the subjects/objects usually present in the event: the analysis of the presence of this primitives will be exploited for evaluation of the discovered inherent taxonomies.

For the sake of simplicity, in Table 6 and 7, all *RI*s sharing the same semantic meaning are manually re-labeled and grouped in the same row. For example, (*RI*₅, *RI*₆, *RI*₇) of event *Wedding* could share the same semantic meaning, in our opinion, ‘couple picture’; or (*RI*₇, *RI*₈, *RI*₉) of event *Graduation* could represent for ‘group picture’.






Since two events *SeaHoliday* and *Baseball* do not contain sub-events according to results reported in [12], only the event definition of Cao et al [6] is used to see whether our taxonomies satisfy such a definition. As it can be noticed in Table 6, the *RI*s part of key frames totally satisfies the event definition. Moreover, the detected taxonomy also contains more information with respect to [6] about the event *SeaHoliday*, like sub-events or concepts such as ‘dinner’ (e.g. *RI*₁₁), ‘group on the beach’ (e.g. *RI*₄/*RI*₅), ‘sunset’ (e.g. *RI*₉/*RI*₁₀), ‘sea shore’ (e.g. *RI*₆, *RI*₇, *RI*₈).

Considering the three events *SkiHolidays*, *Graduation*, and *Wedding*, we can compare with both [12] and [6]. Results are reported in Table 7 for a photo album chosen at random and

Table 5: Event Structure and Definition

Event	Sub-Events [12]	Detailed Definition [6]
Graduation	group pictures celebration, party-eating, unknown	At least one subject in academic cap or gown
Wedding	group pictures, ceremony, party-eating, unknown	Bride must be present. Better with Groom
SkiHoliday	Skiing, walking in the city, eating at hotel, eating relax, party-eating	Containing both snow and skier; on the slope as opposed to a backyard. Not at night
SeaHoliday		Containing people playing on the beach
Ballgames		Containing players and the playing field with or without balls. The field can be baseball, soccer, or football

Table 6: Event Taxonomy Evaluation (1)

Detected Taxonomies: Key Frames (<i>RI</i> s part)		
SeaHoliday		Baseball
<i>RI</i> ₁ , <i>RI</i> ₂ , <i>RI</i> ₃		
<i>RI</i> ₄ , <i>RI</i> ₅		<i>RI</i> ₁ , <i>RI</i> ₂ , <i>RI</i> ₃
<i>RI</i> ₆ , <i>RI</i> ₇ , <i>RI</i> ₈		<i>RI</i> ₄ , <i>RI</i> ₅
<i>RI</i> ₉ , <i>RI</i> ₁₀		<i>RI</i> ₆
<i>RI</i> ₁₁ , <i>RI</i> ₁₂		<i>RI</i> ₇ , <i>RI</i> ₈

obtained i) with [12], plotting both resulting image clusters and recognized sub-events among the ones detailed in Table 5; ii) with the proposed method, reporting the representative real images *RI*s associated to the discovered inherent taxonomy. The right column of Table 7 shows the inherent taxonomy for these three event types, especially common patterns part, discovered using database of Table 4: it properly describe the event by capturing different important moments of it, various actors and various places, via a small representative set of images (*RI*s).

As mentioned above, we analyze results also by counting the number of sub-events in [12] depicted by selected *RI*s, and the number of *RI*s satisfying the definition in [6]. It can be noticed that the proposed method capture almost all characteristics defined in both [12] and [6]. As far as Sky-Holidays is concerned, we have almost all sub-events of [12] represented: skiing (1/5 *RI*s), walking in the city (1/5 *RI*s), eating at hotel (1/5 *RI*s), eating relax (1/5 *RI*s). Moreover, 2 over 5 *RI*s satisfy the definition in [6]. Similar results we have for graduation where all sub-events are depicted (group pictures (2/9 *RI*s), celebration (5/9 *RI*s), party-eating (2/9 *RI*s)) and all *RI*s satisfy definition in [6]. Finally, for event wedding we also have all sub-events represented (group pictures (2/15 *RI*s), ceremony (7 /15 *RI*s), party-eating (6/15

RI_s)) and 8 over 15 RI_s satisfy definition in [6]. Indeed, the proposed method can capture some rare sub-events such as walking in the city or difficult to cluster sub-events like group pictures, eating at hotel or relax during skiing.

It is also worth noticing that definition in [6] is based on visual object recognition and it is very strict. Since we are not using support of visual object detection it is almost impossible that all real images in the taxonomy satisfy their definition. On the other hand, we aim at discovery the taxonomy of a complex event, also capturing particular sub-events which cannot be considered using [6] (as already mentioned referring to Table 6).

As previously discussed in Section 6, the ms value can help understanding which sub-events (and their transactions) rarely appear in an event. This not only helps building suitable event ontology, but also understanding the habits of people when taking pictures of events. For instance, in the event *SkiHoliday* the RI_1 keyframe (in this case, could be interpreted as ‘walking in the city’ sub-events) only appears when ms is set smaller than 10%; meanwhile only RI_2 (‘skiing’) and RI_4 (‘relaxing’) are maintained when $ms = 80\%$, showing that most people prefer to take pictures when skiing than other actions. With event *Graduation*, when setting $ms = 20\%$, all keyframes appear in set of transactions but only $RI_5/RI_6/RI_7$ (‘celebration’) keyframe appears when $ms = 80\%$. When setting $ms = 20\%$, all key frames contribute to the set of transactions also for the event *Wedding* but RI_3 and RI_4 (‘ceremony’) keyframe is surprisingly not present when $ms = 80\%$.

8. CONCLUSIONS

In this paper, we introduce a novel methodology, which automatically infer from a set of personal photo albums belonging to the same event class the underlying event taxonomy. We obtain a hierarchy of natural clusters, capturing the commonalities and the repetitive patterns shared among the different galleries, thus representing the essence of the event itself. This is done by first exploiting time information to cluster subgroups of pictures that are densely concentrated in a given time frame, then extracting key frames from such clusters using rich visual information, and finally determine the common patterns among collections based on the similarity of the descriptions and the consistency in time. The emerging structures become an effective support for defining event taxonomies and managing event-related media collections, as demonstrated in the experimental section.


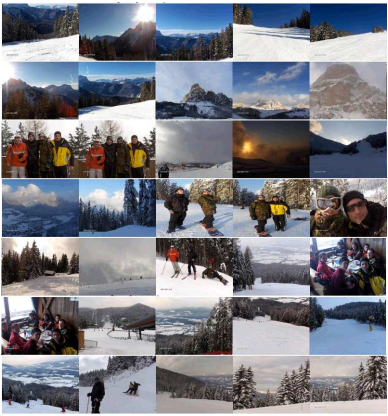
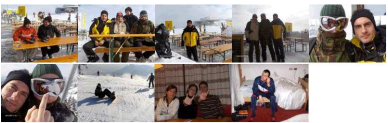
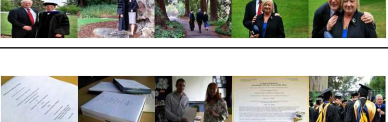




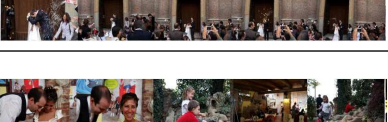



9. ACKNOWLEDGMENTS

This work has been partially supported by the EU Commission under the framework of the EU project Glocal, grant no. 248984.

10. REFERENCES

- [1] R. Achanta and S. Süsstrunk. Saliency detection using maximum symmetric surround. In *ICIP*, pages 2653–2656. IEEE, 2010.
- [2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, pages 207–216. ACM, 1993.
- [3] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura. Comm: Designing a well-founded multimedia ontology for the web. In *ISWC/ASWC*, pages 30–43. LNCS 4825, 2007.
- [4] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417. Springer Verlag, 2006.
- [5] J. C. Bezdek, R. J. Hathaway, and J. M. Huband. Visual assessment of clustering tendency for rectangular dissimilarity matrices. *IEEE TFS*, 15(5):890–903, October 2007.
- [6] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *IEEE TMM*, 11(2):208–219, Feb. 2009.
- [7] M.-S. Dao, D.-T. Dang-Nguyen, and F. G. B. DeNatale. Signature-image-based event analysis for personal photo albums. In *MM*, pages 1481–1484. ACM, Nov-Dec 2011.
- [8] A. Ibrahim, A. Al-Zou’bi, R. Sahawneh, and M. Makhadmeh. Fixed representative colors feature extraction algorithm for moving picture experts group-7 dominant color descriptor. *Journal of Computer Sciences*, 5(11):773–777, 2009.
- [9] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*, pages 243–256. Springer Verlag, September 2010.
- [10] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Journal of Pattern Recognition*, 40:262–282, 2007.
- [11] A. C. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE TMM*, 5(3):309–402, September 2003.
- [12] R. Mattivi, J. Uijlings, F. G. B. DeNatale, and N. Sebe. Exploitation of time constraints for (sub-)event recognition. In *EiMM*. ACM, December 2011.
- [13] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.
- [14] P. Sandhaus and S. Boll. Semantic analysis and retrieval in personal and social photo collections. *MTAP*, 51(1):5–33, 2011.
- [15] S.-F. Tsai, L. Cao, F. Tang, and T. S. Huang. Compositional object pattern: A new model for album event recognition. In *MM*. ACM, Nov-Dec 2011.
- [16] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *MIR*, pages 1–2. ACM, April 2011.
- [17] W. Wagenaar. My memory: A study of autobiographical memory over six years. *Cognitive Psychology*, 18(2):225–252, August 2004.
- [18] R. Xu and I. Wunsch, D. Survey of clustering algorithms. *IEEE TNN*, 16(3):645–678, May 2005.
- [19] J. Yuan, J. Luo, and Y. Wu. Mining compositional features from gps and visual cues for event recognition in photo collections. *IEEE TMM*, 12(7):705–716, November 2010.

Table 7: Event Taxonomy Evaluation (2)

Events	Sub-events detected with [12]		Detected Taxonomies	
	Sub-events	Grouped Images	Key Frames (<i>RI</i> s part)	
SkiHolidays	Eating			
				
	Skiing			
	Relaxing			
Graduation				
	Celebration			
				
	Celebration			
Wedding				
	Ceremony			
				
	Group			
	Party			
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				
Wedding				