Multimed Tools Appl
DOI 10.1007/s11042-012-1153-6

Robust event discovery from photo collections using Signature Image Bases (SIBs)

Minh-Son Dao · Duc-Tien Dang-Nguyen · Francesco G. B. De Natale

Abstract Analyzing personal photo albums for understanding the related events is an emerging trend. A reliable event recognition tool could suggest appropriate annotation of pictures, provide the context for single image classification and tagging, achieve automatic selection and summarization, ease organization and sharing of media among users. In this paper, a novel method for fast and reliable event-type classification of personal photo albums is presented. Differently from previous approaches, the proposed method does not process photos individually but as a whole, exploiting three main features, namely *Saliency*, *Gist*, and *Time*, to extract an *event signature*, which is characteristic for a specific event type. A highly challenging database containing more than 40.000 photos belonging to 19 diverse event-types was crawled from photo-sharing websites for the purpose of modeling and performance evaluation. Experimental results showed that the proposed approach meets superior classification accuracy with limited computational complexity.

Keywords Gist of the scene · Saliency map · Approximate string matching · Human vision system · Personal photo albums · Holistic approach

M.-S. Dao (✉) · D.-T. Dang-Nguyen · F. G. B. De Natale
MultiMedia Signal Processing and Understanding LAB (mmLAB),
Department of Information Engineering and Computer Science,
University of Trento, Via Sommarive, 5, 38123 Povo (TN), Italy
e-mail: dao.minhson@gmail.com

D.-T. Dang-Nguyen
e-mail: dangnguyen@disi.unitn.it

F. G. B. De Natale
e-mail: denatale@disi.unitn.it

Published online: 16 June 2012

1 Introduction

People use photos to preserve memories of major events of their lives. Studies in cognitive science show that, when looking at pictures, the first thing people recall is the event itself, then, who was involved, and finally, where and when that event happened [32]. Therefore, personal photo albums can be used to pinpoint highlights of our life experiences.

In early days of photography, photo albums contained just a few, very significant images, due to the price and complexity of taking and collecting photos. This made it easier to organize and annotate photo albums. Currently, low cost digital photography encourages people to take photos of every significant, interesting or even curious thing that happens around them. This increases the amount and variety of data, and makes it more complex to organize, annotate and access photo albums. Photo album management becomes even more complicated when people upload, share and tag their photos on the Web and in social networks. As a matter of fact, a huge amount of data on one side and the subjectivity of media annotation and organization on the other side make it extremely difficult to search and retrieve what users really want.

Content-based retrieval would be a valuable tool in this context. Unfortunately, although content-based media retrieval has been significantly improved recently, successful automatic tools to organize and manage image databases are still lacking [21, 28]. The possibility of assigning a "unique" label to events of the same type will solve many ambiguous problems in data organization, and will also provide a clear context for more specific annotation, e.g. adopting suitable event taxonomies or ontologies. Accordingly, the need of powerful tools for event-based photo album organization has become an emerging trend [28]. In this context, the availability of tools able to provide event-based classificationof media at reduced computational load can be considered a basic requirement for further steps, such as automatic or semi-automatic annotation, topic-based organization, event-based indexing and retrieval, and automatic summarization.

So far, most methods try to detect the event type from the observation of individual images, in a bottom-up manner. The main drawback of such strategy is that it requires the capability of recognizing the content of an image without any a priori knowledge about the context or domain, which is still an unsolved problem in computer science. Consequently, the information used for event classification is largely unreliable, and such uncertainty reflects of the final result. The opposite approach would be to consider a media collection as a whole, try to capture the essence of the collection itself, and then refine the classification, in a top-down fashion.

As discussed in [5], the point when comparing between top-down and bottom-up approaches for event detection is how to generate more reliable annotations by analyzing a set of media instead of individual photos. It could be observed that, when analyzing a single picture, one cannot take into account important sources of information such as relationships in time, space or semantics: in a single word, the context. Besides, most images in an album have no cues to predict which event they belong to, when analyzing them individually. For instance, an image of "flowers" may perfectly fit a wedding, a mountain hike, or a visit to a botanic garden. Methods dealing with individual photos are usually able to detect some degree of accuracy visual concepts, and sometimes could derive from the combination of such concepts with the understanding of scenes or activities [21, 28]. Instead events are characterized by complex combinations of scenes, concepts, and activities happening in a

given combination or order. For example, the description of the event "Having fun at the beach" should contain the following concepts: PEOPLE (entity), PLAYING (activity), and BEACH (place) [5].

In this paper, we introduce a novel method that addresses this problem from a different perspective, namely, analyzing the photo collection of an event as a whole, and trying to characterize it accordingly. The motivation of the proposed method has been inspired by the following criteria:

- **Gist, Saliency, and Time:** Gist and Saliency represent two different approaches to understand the meaning of images: scene-centered primitives, and object-centered primitives, respectively [25, 26]. The former pays attention to the general characteristics of images, known as Gist (or Gist of the scene); the latter focuses on discovering prominent regions in images, known as Visual Saliency. Psychophysics experiments show that humans can get at first glance not only a holistic perception of the scene as a whole, but also the location of one or more key spots, which allow interpretating the message conveyed by the scene [11]. In this respect, the fusion of Gist and Saliency is very powerful, as Gist captures a structural representation of the scene without segmentation or grouping operations [25, 26], whereas Saliency uses local features to improve the ability of discovering key objects in the scene [3, 12, 16]. Although individual features have been widely investigated in scene classification, event recognition, and target detection [9, 10, 13, 16, 25, 27, 29], their integration deserves further consideration [17, 19, 31]. As far as temporal information is concerned, the availability of multiple shots of the same event taken at different times may significantly improve the detection, given that time evolution is an important dimension of the event. For instance, in [8] the authors observed that there is a strict correlation between the frequency of shots taken and the importance of an event, and they used such parameter to detect significant events in Personal Image Collections. Other researchers used time as a key to cluster images into semantically coherent groups [5].
- **Common patterns:** events of daily life are not infinite in nature, and share many common characteristics, even when they are possibly influenced by personal, cultural or geographical aspects. Also their representation through photo collections inherits some commonalities. For instance, people usually like taking images of known landmarks where important events happened, or they associate different colors to different event types, e.g., preferring bright colors in association to happy events. Also, they tend to put the most interesting elements of their photos in particular positions. Taking into account these considerations, some methods focus on composing panorama images around special places. For instance, in [14], similar regions in a huge collection of images were treated as a key to reconstruct a composite scene. In [30], the authors exploited the habit of taking multiple photos of the same scene to achieve a panorama by viewpoint clustering. By analogy, common patterns inside an event could be used to achieve a mosaic representation which is characteristic of that event.
- **Common semantics:** every real-life event contains a lot of implicit semantics by which humans can understand its general content without explicit explanations. Most social events are structured along sets of unspoken rules that people follow by tradition, cultural background or instinct. For example, in [5], authors used this argument to build a dataset of events. Each event in the dataset is associated

with a detailed definition that represents the distinguishing characteristics of that event (e.g., images of 'beach fun' event must contain 'people playing on the beach'; likewise, images of 'Christmas' must include 'Christmas decorations'). Therefore, events of a similar type usually contain some common semantics that need to be discovered.

Accordingly, we propose a method that combines Saliency, Gist and Time to create a composite event signature, used for event-type identification and classification. Gist and Saliency are first combined into a 2D histogram, called Gist-Saliency Signature Image Base (GS-SIB), which takes into account dominant colors and saliency maps of all the images belonging to an event-related photo collection. In other words, each photo is projected into a point of the GS-SIB space, according to its dominant color and saliency map pattern. Then, time information is used to build a sequence of symbols called Temporal SIB (T-SIB), which captures the temporal evolution of images associated with the event. The differences and similarities among photo collections referring to various events are then measured in terms of weighted distances between the corresponding GS-SIBs and T-SIBs.

The rest of the paper is organized as follows: Section 2 provides a short overview of related work in the field. Section 3 describes in detail the proposed methodology for event type classification. Sections 4 and 5 present the results of thorough experimental validation, assessing the accuracy of the proposed method under different setups and experimental conditions. Computational complexity is also investigated. Section 6 draws the conclusions and discusses future developments as well as possible extensions of this method to other application domains.

2 Related work

In this section, some of the most significant related works are briefly reviewed, with special attention to papers dealing with the analysis of personal photo collections.

In [22], the authors proposed a two-level method for event-based clustering of photo collections by using (i) time information, and (ii) block-based histogram correlations. The application to low quality images extracted from the Kodak database reached a precision of 79% and a recall of 80%. They also showed some significant differences between time and date information of photos of the same event and photos of different events. Their method used color information of individual images, instead of global information of the whole event.

In [20] an event-detection method for personal photo albums was proposed based on a conceptual graph, in which concepts were detected from the images. They estimated the event type by matching visual concepts associated with event models using a weighted metric based on histogram intersections. Experimental results on 2,400 photos with an event taxonomy of five layers and 20 categories showed an average precision of around 60%. A main drawback of this method is that time information is not used, thus making it difficult to classify the whole event, which is typically well characterized in terms of temporal evolution. Furthermore, their proposed approach can only be applied to selected photos, because images that are not characteristic of the specific event type, do not allow discriminating the event unless additional information is provided.

In [34], Yu et al. designed an approach for photo grouping based on similarity of pictures taken by different users on the network. The authors built a network of relationships among groups and tags, based on the SimRank. The final result was achieved by spectral clustering. To compare images, they used Gist and CEDD as visual descriptors, while for topic analysis they used LDA and pLSI techniques. The results on real users' collections showed that the accuracy of 63%.

Another interesting method was introduced in [5]. In this paper, the authors used conditional random field models to calculate the correlation among photos in a collection based on time, location, event, and scene. Results showed that this approach made it possible to efficiently annotate images at both scene and event levels. The major problem of this method was the limited generalization capability. For instance, we must be able to infer that images containing 'table and dishes with more than two people' represent a semantic key of events such as 'dinner', or analogously that 'cake, balloon, and/or birthday hat' recall a 'birthday' event. This implies a rich a priori knowledge on the semantic keys that characterize each event type.

In [18], a method for event detection from a single photo was presented, with application to sports events. Compared to previous works, they achieved significant improvement by using graphical models for labeling and SIFT features for matching. Since only one photo was used, their approach was suitable for specific domains and very representative pictures. In the case of photo collection analysis, their approach could obtain additional information on some particular images once the global information was available.

In [8] the authors translated user's picture-taking behaviors into a time series, to detect special events. The underlying assumption was that a significant photo corresponded to a burst in the time series. Based on the authors' observation, when an outlier appeared in the time-series, there was a high probability that a significant event was happening at that time.

In general, it can be observed that most of the existing methods treat images as individual objects and try to extract as much information as possible from each of them, instead of considering them as a part of the whole. Besides being suboptimal from a philosophic viewpoint (already Aristotle in his **Metaphysics** suggested that 'the whole is different from the sum of its parts'), such approach has strong limitations in terms of computational complexity, effectiveness of the description and capability to render the complexity of an event. In fact, one needs to analyze the content of each image before being able to extrapolate the meaning of the whole collection, thus requiring huge computation. Furthermore, not all the images contain significant event-related information, in terms of contents or tags, thus causing erroneous interpretations of the relevant semantics, which may cumulate and impact the analysis of the event interpretation.

3 GST-SIB: Gist-Saliency-Time Signature Image Base

The main idea of the proposed method is to observe the photo collection associated with an event in a holistic way, and to extract a representative signature based on the most important characteristics perceived by human visual system: saliency, global appearance, and temporal information. Those features are captured by saliency maps, Gist and timestamps, respectively. In this sense, the proposed method synthesizes the

pictorial dimension of an event by mimicking some of the basic characteristics of the human visual system.

Since the only information we use for the 'event' is the 'photo collection' or 'album' associated with it, from this point forward we will use those terms interchangeably. Moreover, we assume that an event is associated with an album captured by a single device, thus allows us to associate UTC time with each picture. In the following sections, we will describe how we build a low-dimensional descriptor from a photo collection, and how we use it to detect the relevant event-type. Figure 1 shows a general overview of our method.

3.1 GS-SIB: Gist-Saliency Signature Image Base

In our approach, the two most important visual descriptors are *Saliency* and *Gist*. As mentioned in previous sections, such features could help create a kind of fingerprint that captures both conceptual and perceptual dimensions of an event. To build a compact descriptor, the two features are integrated into a unique 2D histogram, called GS-SIB, which synthesizes information of the salient areas and dominant colors of the whole event. To achieve this goal, each photo is projected into one point of the GS-SIB depending on its dominant color and saliency pattern.

Color is usually considered as a rich source of information as well as an easy feature to be extracted from images. For this reason it has been widely employed in content-based image retrieval, and it plays a major role in image description standards [21]. In Gist perception, color plays also an important cue to capture the semantics of images [24, 26, 29, 31]. In [24, 26, 31], the authors pointed out that the organization of color blobs in an image could provide important information about the semantics conveyed by the image. For example, images with green as dominant color are typically associated with natural landscapes, whereas images with blue or white as dominant colors are most probably depicting the ocean or the sky. In our method, dominant colors are used as a low-level feature to build the scene Gist. The

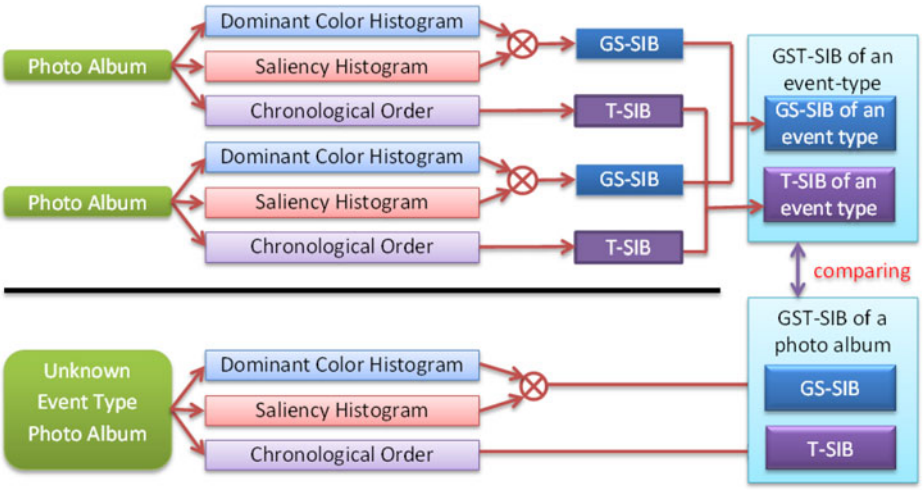


Fig. 1 Block diagram of the proposed method

fixed representative color features extraction algorithm [15] is used to capture the dominant color of each image in the collection, using **38 perceptual colors** in **RGB color space**. A photo album is then represented by **1D histogram** with **38 bins**, called **1D-38-HIS**.

As far as Visual Saliency is concerned, human visual attention has been studied in several disciplines, including neuroscience and psychology. It has also been widely used in computer science to recognize the most important contents of an image in a single glance [1, 12, 21, 29, 33], as well as to study the human reactions to a given subject or his/her preferences. Applications of saliency span from image coding to personalization, focus of attention, watermarking, and so on. In our method, saliency is used to detect regions of interest in images, with the aim of capturing common patterns possibly connected to the semantics of an event. As a matter of fact, the usual way of taking pictures is strongly connected to the subject and to the context, and follows some composition rules that are implicitly or voluntarily used by the photographer.

In fact, when taking pictures users do not take shots without an aim: they always focus on the most significant points that could convey the meaning of an event. Furthermore, even low-end cameras nowadays provide software tools that automatically focus on visual attention areas, such as faces or sharp details. Therefore, in our method saliency maps help discover common patterns and common semantics by exploiting user's picture-taking habits.

The saliency map is then used as a computational visual attention predictor to extract saliency features. Next, a histogram of the saliency map of the whole album is created using Algorithm 1.

Algorithm 1 Calculating Histogram of Saliency Map of a whole album

Require: 1. photo Album P
2. binary threshold β

return Histogram of Saliency Map of P

for each image I_i in P **do**

1. The visual saliency of I_i is calculated by applying the method in [1]
2. The visual saliency image is binarized by applying the binary threshold β, and divided into 3x3 equal regions
3. Regions having a majority of '1' values are marked as salient, thus form a 3x3 binary saliency pattern (see Fig. 3). Possible patterns are therefore 2^9, and each image can be mapped into a **1D histogram** with 2^9 bins, called **1D-512-HIS**.

end for

Finally, the GS-SIB of the event is obtained by combining the two above histograms, **1D-38-HIS** and **1D-512-HIS**, into a **2D histogram 38x512** GS-SIB (see Fig. 2).

3.2 T-SIB: Temporal Signature Image Base

As discussed above, the temporal relationship among episodes/images within an event is a very important cue to discover implicit semantics of events/event-types

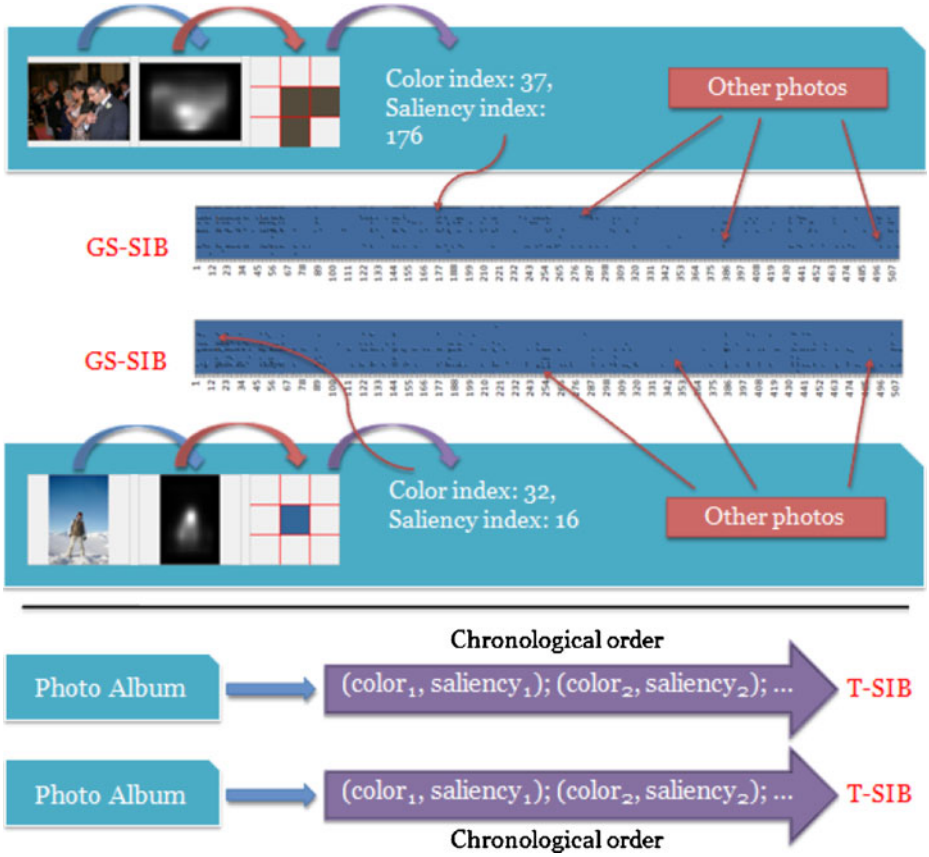


Fig. 2 GS-SIB and T-SIB generation

as well as to measure the similarity among them. For example, a traditional western wedding typically starts with a rehearsal party which is then followed by a religious ceremony, and concludes with a celebration party; likewise a penalty in a soccer match should be connected to a yellow card or red card. Besides, time-series also play an important role to group images that share the same semantics, instead of focusing on every single image. This relates to the habit of taking series of shots within short intervals around an interesting place or moment. Furthermore, temporal information may allow inferring new or hidden patterns for unknown or rare events without stating explicit rules. It is therefore important to take into account such information, and to integrate it with Gist and Saliency features.

After building the GS-SIB, timestamps associated with the event pictures are used to create an ordered sequence of $(color\_idx_i, saliency\_idx_i)_{i=1..N}$, where N is the number of images, while $color\_idx_i$ and $saliency\_idx_i$ are the indices of dominant color and saliency map of image i in GS-SIB, respectively. Such sequence is called T-SIB. Each $(color\_idx_i, saliency\_idx_i)$ pair and its position in T-SIB jointly represent the photo coordinates in GS-SIB and time order (see Fig. 2).

3.3 Event model and similarity measures

GS-SIB and T-SIB represent the signature of a specific event. Since our goal is to capture the similarity of events through those signatures, we need to build a template for each event of interest, and define an appropriate metric to measure the similarity between templates and actual event signatures (Fig. 3).

The GS-SIB model ($GS-SIB^*$) of a given event-type is built by simply accumulating and normalizing the 2D histograms associated with a set of training photo albums referred to the same event type. Figure 4 illustrates a 2D visual representation of the GS-SIB model for three event-types: Basketball, Ski Holiday, and Wedding. The first mostly contains indoor scenes and people; the second outdoor scenes and natural landscapes; the last both indoor and outdoor scenes, with people and natural landscapes. One can easily recognize the big difference among these relevant signature images.

In order to measure the similarity between a generic event i and a template in terms of GS-SIB, we define the distance:

$$d_{GS-SIB}(GS-SIB_i, GS-SIB^*) = |GS-SIB_i, GS-SIB^*| \quad (1)$$

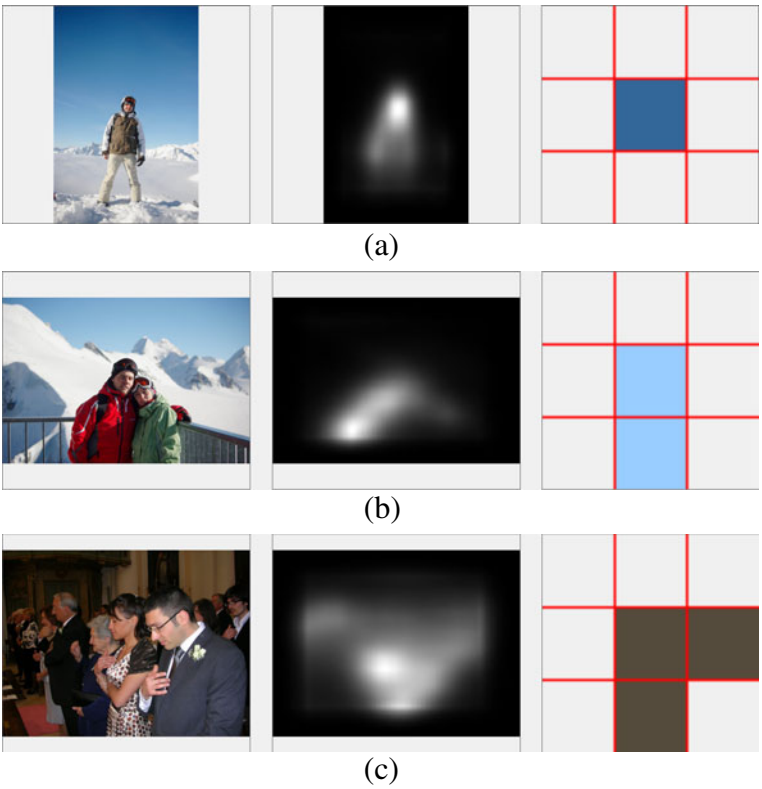
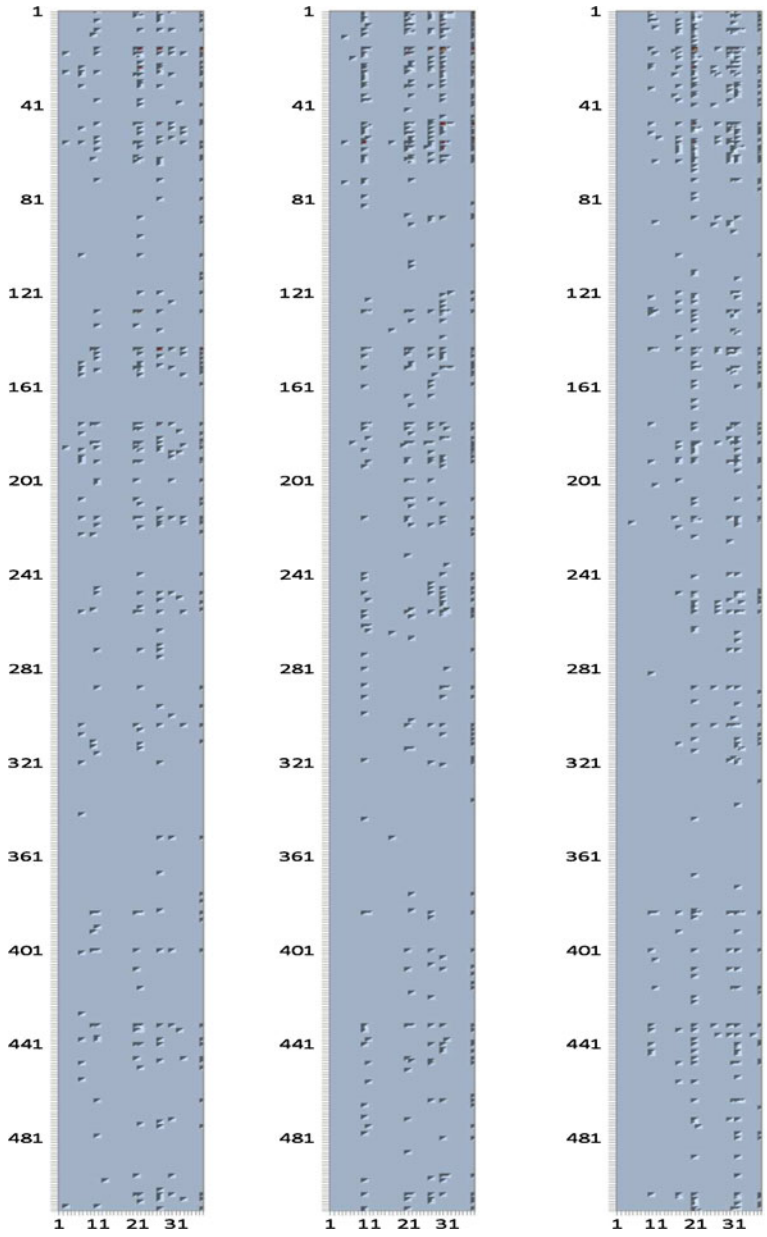


Fig. 3 Saliency Map Patterns: from left to right: original image, its saliency map, its pattern



(a) Baseketball's GS-SIB (b) Marriage's GS-SIB (c) SkiHoliday's GS-SIB

Fig. 4 2D visual representation of $GS - SIB^*$ associated with event types (512×38 2D-histogram)

where $\|.\|$ is the selected metric. In our experiments, the Bhattacharyya distance is chosen due to the superior performance achieved in this context with respect to other metrics (see Fig. 8).

As far as T-SIB model ($T-SIB^*$) is concerned, the T-SIBs of similar events are collected to create a *document*. In this case the information is treated as symbolic, then, it should not be averaged as in the previous case, but simply concatenated. Consequently, comparing an event to an event model in terms of temporal consistency can be seen as matching a string within a text document, to check whether it can be associated with a sentence in the document.

The temporal similarity between a generic event $T-SIB_i$ and a model $T-SIB^*$, $d_{T-SIB}(T-SIB_i, T-SIB^*)$, is then measured by using a classical *approximate string matching technique* [2, 23]. In our case, *dynamic programming* is applied to calculate the similarity between two time models.

Finally, we integrate the GS-SIB and T-SIB into a unique event signature called $GST-SIB(i) = \{GS-SIB, T-SIB\}$. Correspondingly, the integrated event model $GST-SIB^*$ will be obtained from $GS-SIB^*$ and $T-SIB^*$. The similarity between an event and a model will be then defined as follows:

$$\begin{aligned} & d_{GST-SIB}(GST-SIB_i, GST-SIB^*) \\ & = \alpha d_{GS-SIB}(GS-SIB_i, GS-SIB^*) \\ & \quad + (1-\alpha) d_{T-SIB}(T-SIB_i, T-SIB^*) \end{aligned} \tag{2}$$

where the weighting parameter α is used to balance the visual and temporal components, i.e., to give priority to visual similarity or time consistency. This parameter can be adjusted based on the application domain. For events that have a very structured nature (e.g., a wedding) time consistency is more important, while in events where the sequence of sub-events is unpredictable (e.g., a football match) visual similarity is dominant.

The Algorithm 2 is used to detect event types using SIB descriptors, where the term SIB could be replaced with GST-SIB, GS-SIB, or T-SIB depending on the descriptor adopted. Figure 5 illustrates how the algorithm works.

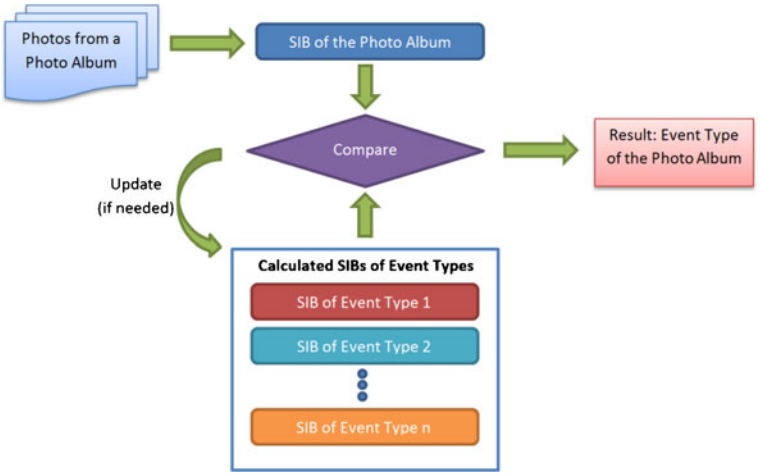


Fig. 5 Detect event-type using SIB family

Algorithm 2 *(pseudo-code)* Detecting event-type using SIB family

Require: 1. photo album P representing an instance of an unknown event-type
2. weighting parameter α
3. set of event-type models $\{SIB^k\}$, k being the number of available event-types
4. updating parameter γ

return event-type name

Calculate SIB of photo album P using algorithms mentioned in Sections 3.1 and 3.2;

Initialize the set of similarity distance $D = \{0\}$;

for all event-type model SIB^k **do**

1. Calculate similarity distance between SIB and SIB^k: $d_k = d(SIB, SIB^k)$ (using (2));
2. $D = D \cup \{d_k\}$;

end for

Calculate $d_{min} = min(d_k)$, and $idx = argmin_k(d_k)$, where $d_k \in D$;

if $d_{min} \leq \gamma$ **then**

1. Update SIB^{idx} with SIB using Algorithm 3;
2. Return idx (i.e. the given photo album belongs to idx^{th} event type);

else

the given photo album is assigned as **unknown event type**;

end if

Algorithm 3 *(pseudo-code)* Updating event-type model SIB^k with current SIB

Require: 1. event-type model SIB^k
2. weighting parameter α
3. updating parameter λ

return updated event-type model SIB^k

if $\exists GS - SIB$ **then**
  $GS - SIB^{temp} = GS - SIB^k \cup GS - SIB$;
  if $d(GS - SIB^{temp}, GS - SIB^k) > \lambda$ **then**
    $GS - SIB^k = GS - SIB^{temp}$;
  end if
end if

if $\exists T - SIB$ **then**
  if $T - SIB$ is NOT a subsequence in $T - SIB^k$ **then**
    $T - SIB^k = T - SIB^{temp}$;
  end if
end if

if $\exists GST - SIB$ **then**
  Update $GST - SIB^k$ using Equation 2;
end if

3.4 Computational complexity

Given M predefined event models $GST - SIB^*$ and a new photo album A containing N images, we want to calculate the computational complexity for detecting the

event-type of A, namely $CC(A/GST - SIB^*)$. The detection process is made of two steps:

1. calculating GST-SIB for album A
2. comparing the GST-SIB of A to a set of event-type models $GST - SIB^*$ and choosing the minimum distance one

It should be observed that we normalize all images of photo album A to the same size (n pixels). Then, we have:

$CC(A/GST - SIB^*)$ = Computational Complexity of Step 1 + Computational Complexity of Step 2, where

- *Computational Complexity of Step 1* = N(Computational Complexity of computing saliency map for each image [1] + Computational Complexity of calculating dominant color for each image [15])

  $= N(O(n) + O(nk + logk))$

  $= O(nN + nkN + Nlogk)$, where k is the number of perceptual colors.

- *Computational Complexity of Step 2* = M(Computational Complexity of calculating Bhattacharyya distance on two 2D histograms with p bins + Computational Complexity of calculating approximate string distance with dynamic programming)

  $= M(O(p) + O(Nq_j)$, where N and q_j are the length of $T - SIB_i$ and $T - SIB^*$

  $\approx O(Mp + MNQ)$, where Q is the longest string of q_j

  Therefore

  $$CC(A/GST - SIB^*) \approx O(N(n + nk + logk) + M(p + NQ))$$

  Since k (=38 bins), n (=64 × 64 pixels) and p (=38 × 512 bins) are fixed, the computational complexity of the proposed method is linearly influenced by the number of images in album N, the number of pre-defined event-type models M, and the longest length of $T - SIB^*$ Q.

4 Dataset and testing strategies

In this section, we introduce the dataset and the strategy used in the experimental validation of the proposed method.

4.1 Corpora

Since most of the existing methods focus on the extraction of concepts from single images [18] or small groups of images [5], it is difficult to find extensive corpora providing large event-based photo collections suitable for testing and comparing the proposed method. The few available datasets that contain event-related information, still have strong limitations. The one used in [5], does not suit the general case for it

is built upon some restrictive conditions, e.g., 'wedding' events must contain images of 'bride', or 'Graduation' events must contain images with at least one subject in academic cap or gown, etc., as shown in Table 1 of their paper. In [18], the authors collected a dataset for sports events on the Internet. Unfortunately, the scope of the selected events is too narrow for our purposes.

Therefore, we decided to create our own dataset which covera diverse topics including social, sports, and personal events, indoor and outdoor, different environments, and mostly taken from real-life user's collections. The approach we used was to crawl images from volunteers or social networks. In particular, images were crawled from Picasa, collecting available photo albums tagged with specific event types by users. For each event type (19 in total), a variable number of instances were collected due to the availability on the network. On an average, each album was made of approximately 110 JPEG images. The smallest collection contains 53 images (*meeting* event-type), while the largest 210 (*wedding* event-type).

One thing to point out is that the selection of the event-types did not follow any domain-related restriction or preference, while being driven by other practical considerations, such as: (i) to include a sufficient number and variety of common "real-life" events; (ii) to include categories that show some inter-class similarities (e.g., different sports events) as well as intra-class diversity; (iii) to select event types for which it is possible to crawl a sufficient number and variety of photo collections from the network. Table 2 provides a further insight on this. The left column shows event types with a high potential of misclassification due to inter-class similarity, whereas the right column describes the relevant misclassification causes.

Finally, we divide the dataset into two categories, D1 and D2, where D1 contains event types for which less than 30 instances are found, while D2 contains larger collections. Table 1 shows the detail of the corpora in terms of event types, instances, and total pictures crawled. Together with images, a complete ground-truth

Table 1 Datasets

Category	Event type	No. albums	No. photos
D1	Baseball	10	1,358
D1	Bike	10	1,005
D1	Concert	15	1,085
D1	Cycling	10	1,424
D1	F1	10	1,351
D1	Golf	11	1,481
D1	Hockey	13	1,383
D1	Meetings	15	795
D1	MountainTrip	15	2,051
D1	Picnic	18	2,105
D1	Rowing	10	2,100
D1	SeaHoliday	15	2,253
D1	Skating	10	1,877
D1	Swimming	15	2,293
D2	Basketball	30	2,498
D2	Cricket	35	3,525
D2	Graduation	34	3,634
D2	Marriage	75	6,279
D2	SkiHoliday	33	3,165
Total		384	41,662

Table 2 Events with a high potential of misclassification

High potential misclassification event types	Some common patterns and semantics (that could be easily recognized by human beings)
Bike, F1, cycling	One showy object on a race track, crowd with gaudy color, similar dominant color over a whole album
SeaHolidays, MountainTrip, SkiHolidays, golf, picnic, meetings	Similar dominant color over a whole album (e.g. too much white (clouds, snow), blue (sky, ocean), green (grass, trees)), the ratio of outdoor scenes and natural scenes is large
Rowing, swimming	Objects are almost focused on a center of image with different perspective angles, similar dominant color over a whole album (color of water)
Skating, hockey	Similar stadium, dominant color white (ice stadium)
Baseball, cricket	Similar playing rules and dominant color (stadium and uniform of players)
Marriage, graduation	Have several distinct sub-events (ceremony, indoor, outdoor, group people, couple people), a lot of people with uniforms
BasketBall	Special color of stadium, focus on players, total indoor scenes
Concert	Various illumination condition, indoor or outdoor scene with a lot of artificial architectures (e.g. stages)

information is stored at both album and picture level, including event type, available metadata, and EXIF.

4.2 Testing strategies and comparisons

In our tests we evaluated two essential features of the proposed method:

- **Discrimination capability:** different event-types should have different $GST - SIB^*$ models, and
- **Representation capability:** events of the same type should have close GST-SIBs.

In order to measure the representation and discrimination capabilities, we estimated the distances among events belonging to different categories in a large training set: the smaller the distance is between two event-types, the more similar they appear. The relevant SIB distances were calculated and mapped into a distance matrix. In order to avoid misclassification, inter-event SIB distances have to be large (*discrimination*), while intra-event distances have to be as small as possible (*representation*). Successively, a set of events has been used to test the accuracy of the classifiers through the usual precision and recall measures. Here, two different testing strategies were applied: first, a leave-one-out classification is applied to the category D1 of 14 event-types; second, five different test cases are randomly created by selecting one-third and two-thirds of the images of each category to form training and test datasets, respectively.

As far as comparisons are concerned, to the best of our knowledge there is no other method that shares the idea of determining event types from personal photo albums without the need to understand the meaning conveyed by single images, or the distinguishing characteristics of events. This makes it rather difficult to perform fair comparisons. Therefore, we limit our comparative analysis to two methods: the first [7] (*Color-SIB*), builds upon the importance of color features in event recognition, and creates an holistic description of an image collection by selecting

the dominant color of each image and mosaicking all event-related images as time-ordered color blobs. The second [5] uses time, location, and object/scene features to detect events. Instead of using the album as a whole, this method deals with single images: first it analyzes each picture to detect some 'semantic keys' representative of some predefined scenes, then it performs the event detection on the basis of such semantic keys.

5 Experimental results

In this section, we report the *Discrimination Capability* and *Representation Capability* of the proposed method, and compare it to the method presented in [7]. Moreover, merits of the proposed descriptors and overall advantages of the proposed method are discussed thoroughly.

5.1 Discrimination capability

In this section, we use the research of Bezdek et al. [4] to evaluate the discriminative capability of the proposed method. In other words, we would like to evaluate the degree of misclassification among the proposed event models SIB^k.

In [4], the authors introduced a new visual approach (VAT) that first represents a dissimilarity matrix of objects as a *dissimilarity image*, then produces a new image, namely the *VAT-ordered image* to visualize the existence and number of potential clusters. According to Bezdek et al [4], the dissimilarity image of the given dissimilarity matrix is defined as follow:

Dissimilarity image definition Given a set of objects $O = \{o_i\}_{i=1..n}$, and let $DM = \{dm_{ij}\}_{i=1..n;\, j=1..n}$ be a dissimilarity matrix computed from O, where $\forall i, j$: $dm_{ij} \geq 0$, $dm_{ij} = dm_{ji}$, and $dm_{ii} = 0$. The dissimilarity image of DM is an intensity image (i.e. gray-scale image) $I = \{g_{ij}\}_{i=1..n,\, j=1..n}$ so that the value $dm_{ii} = 0$ corresponds to $g_{ii} = 0$ (pure black); the largest dissimilarity value in DM is denoted by $g_{ii} = 255$ (pure white). Therefore, the darker color the g_{ij}, the more similar objects o_i and o_j is.

By applying this method to our event models SIB^k, we could see which event models have a higher probability of misclassification. One thing to emphasize is that each 'Dissimilarity Matrix' represents the distance of event models SIB^k among each other. Therefore, all the values on the diagonal are null, and the matrices are clearly symmetric.

Tables 3 and 4 report the upper triangular dissimilarity matrices achieved with the Color-SIB method in [7] and the proposed GST-SIB method on D1 category, respectively.

Figure 6a and b illustrate dissimilarity images corresponding to dissimilarity matrix in Tables 3 and 4, respectively; Fig. 6c and d denote VAT-ordered images of the dissimilarity matrix showed in Tables 3 and 4, respectively: the brighter the intensity of the blob/cell is, the higher the dissimilarity of event models. As shown in Fig. 6, we could see that the Color-SIB method has higher misclassification potential than GST-SIB method.

The comparison shows that the GST-SIB outperforms Color-SIB with much larger normalized distances, and consequently the capability of discrimination among

Table 3 Dissimilarity matrix—color-SIB [7]

Baseball	0.00	0.40	0.50	0.50	0.50	0.48	0.48	0.42	0.55	0.49	0.51	0.46	0.89	0.40
Bike	–	0.00	0.57	0.26	0.16	0.39	0.59	0.26	0.42	0.32	0.62	0.40	0.85	0.47
Concert	–	–	0.00	0.62	0.52	0.64	0.69	0.40	0.67	0.41	0.70	0.60	0.88	0.61
Cycling	–	–	–	0.00	0.35	0.30	0.52	0.37	0.35	0.31	0.47	0.32	0.91	0.47
F1	–	–	–	–	0.00	0.48	0.66	0.25	0.51	0.25	0.55	0.46	0.85	0.49
Golf	–	–	–	–	–	0.00	0.55	0.41	0.45	0.31	0.49	0.48	1.00	0.57
Hockey	–	–	–	–	–	–	0.00	0.59	0.75	0.57	0.63	0.50	0.61	0.79
Meeting	–	–	–	–	–	–	–	0.00	0.43	0.27	0.59	0.42	0.83	0.44
MountainTrip	–	–	–	–	–	–	–	–	0.00	0.45	0.56	0.41	1.00	0.40
Picnic	–	–	–	–	–	–	–	–	–	0.00	0.44	0.47	0.87	0.54
Rowing	–	–	–	–	–	–	–	–	–	–	0.00	0.51	0.82	0.53
SeaHoliday	–	–	–	–	–	–	–	–	–	–	–	0.00	0.74	0.45
Skating	–	–	–	–	–	–	–	–	–	–	–	–	0.00	0.91
Swimming	–	–	–	–	–	–	–	–	–	–	–	–	–	0.00

different events is much higher, even when they share some common patterns. For example, let us consider the two events 'Bike' and 'F1' (see Fig. 7). If we look at the two image groups as a whole, we could perceive several common patterns that would lead to misclassification. For instance, visual concepts such as 'one showy object on a race-track', 'crowd with flashy colors' in two sides of race-track or as a background of athletes, 'race-track' with motley advertising board or roadside, are present in both collections. Figure 6c–d shows how these two event models present some risk of misclassification.

5.2 Evaluation of the importance of single features in GST-SIB, and the representation capability of SIB model

In this section, we provide some details on the importance of single features in GS-SIB, in order to demonstrate that the combination of Saliency and Gist is much more discriminated than each single feature. Figure 8 clearly shows how the integration of color and saliency into GS-SIB significantly outperforms each single feature for all

Table 4 Dissimilarity matrix—GST-SIB

Baseball	0.00	0.77	0.87	0.80	0.78	0.78	0.91	0.82	0.83	0.79	0.87	0.82	0.94	0.83
Bike	–	0.00	0.90	0.69	0.64	0.72	0.77	0.80	0.76	0.79	0.73	0.76	0.87	0.79
Concert	–	–	0.00	0.88	0.85	0.92	0.91	0.79	0.82	0.84	0.88	0.80	0.85	0.88
Cycling	–	–	–	0.00	0.68	0.77	0.78	0.80	0.75	0.76	0.78	0.75	0.85	0.79
F1	–	–	–	–	0.00	0.80	0.78	0.77	0.73	0.75	0.71	0.74	0.86	0.77
Golf	–	–	–	–	–	0.00	0.87	0.83	0.81	0.78	0.86	0.83	0.91	0.83
Hockey	–	–	–	–	–	–	0.00	0.88	0.80	0.86	0.75	0.78	0.64	0.82
Meeting	–	–	–	–	–	–	–	0.00	0.79	0.76	0.84	0.78	0.91	0.83
MountainTrip	–	–	–	–	–	–	–	–	0.00	0.74	0.72	0.64	0.87	0.74
Picnic	–	–	–	–	–	–	–	–	–	0.00	0.83	0.78	0.92	0.82
Rowing	–	–	–	–	–	–	–	–	–	–	0.00	0.69	0.83	0.76
SeaHoliday	–	–	–	–	–	–	–	–	–	–	–	0.00	0.83	0.74
Skating	–	–	–	–	–	–	–	–	–	–	–	–	0.00	0.87
Swimming	–	–	–	–	–	–	–	–	–	–	–	–	–	0.00

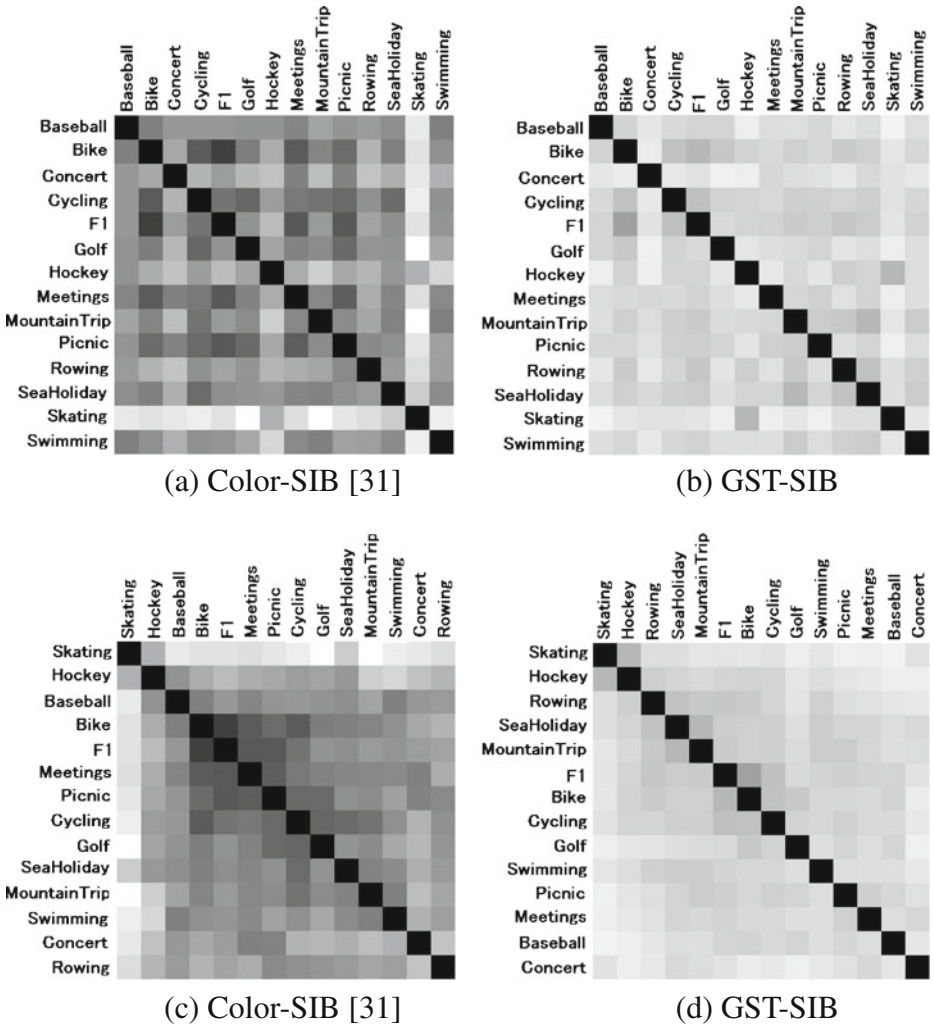


Fig. 6 **a**, **b** Dissimilarity images; **c**, **d** VAT-ordered dissimilarity images (corresponding to dissimilarity matrices in Tables 3 and 4)

the considered distance metrics. The chart also shows the fact that Bhattachryya and Chi-square distances provide best result, thus leading to the choice of the first in our experiments.

Furthermore, Fig. 9 compares the result of GST-SIB with the method in [7] to prove the higher representation capabilities of GST-SIB. The former integrates both Saliency and GIST features, while the latter uses only GIST information. It is clear that GST-SIB provides better performance for any performance measure (precision, recall, f-score) and for almost every type of event, with rare exceptions where the performances are rather equivalent. On average, the F-Score obtained by GST-SIB is 0.85, against 0.78 of the method in [7]. These results confirm that the integration of Saliency and Gist improves the accuracy of event detection.

(a) 'Bike' event

(b) 'F1' event

Fig. 7 Common patterns of two events 'Bike' and 'F1' (*one showy object on a race-track*, *crowd with gaudy colors* (in two sides of race-track or as a background of athletes), *race-track* (with motley advertising board or roadside))

Next, we assessed the method with dataset D2, on five different test cases. The results of this second test are illustrated in Fig. 10, while Fig. 11 shows the accuracy on each event type of the category D2. This test aims at checking possible 'overfitting' phenomena when training with larger datasets, where common patterns appearing with high frequency could bias the GST-SIB model. Results confirm that no overfitting is present.

The method of Cao et al. [5] is then compared with the proposed one. The comparison is performed on four common event-types: *Wedding, Graduation, Skiing, SeaHoliday*, using our dataset. Figure 12 illustrates how the proposed method outperforms the other, even if the latter uses additional information such as GPS, and introduces some constraints in the scene contents.

Finally, in order to evaluate the influence of the weighting parameter α on the overall performance of GST-SIB, we assign different values to α, ranging from 0 to 1, and repeat the experiments over the whole database, averaging the results. Figure 13

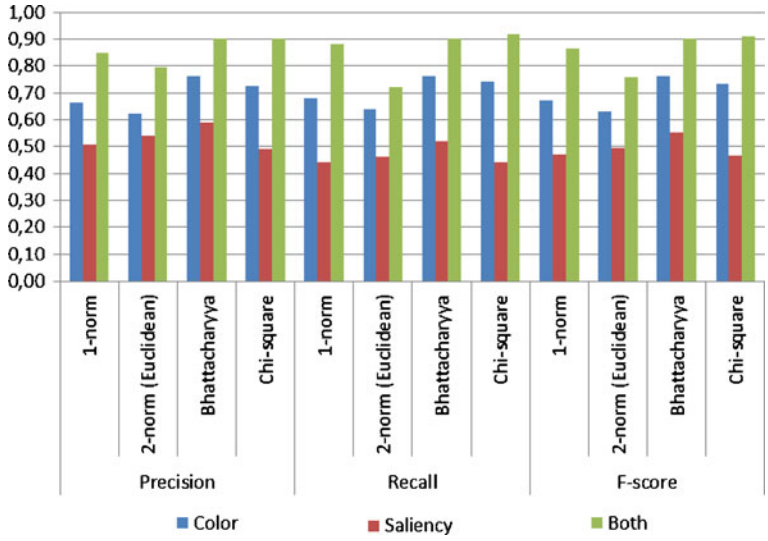


Fig. 8 Comparing between 'gist' and 'saliency' in GS-SIB, average precision, recall, and F-score calculated on all events in the dataset with different metrics

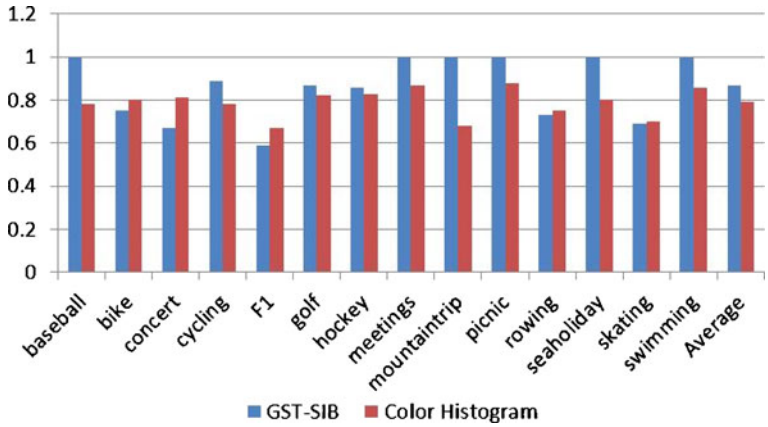
shows the result of such experiment. It is easy to argue that best values for α are in the range 0.5 to 0.6, showing that visual and temporal information have an almost equivalent weight in the representation of events.

5.3 Convergence

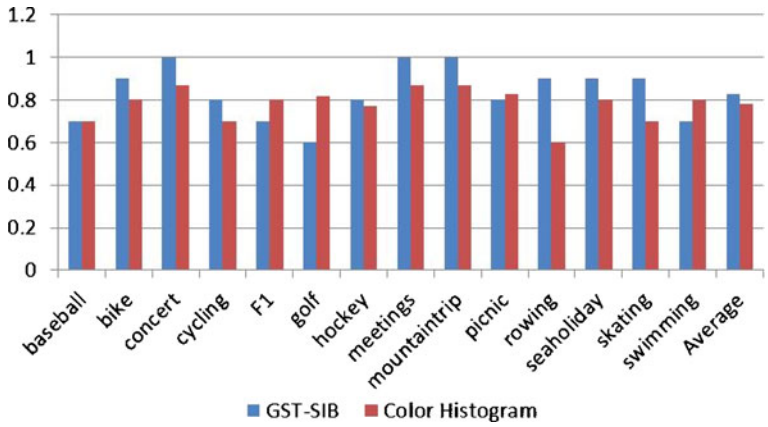
The last test concerns the convergence of GST-SIB. Basically, GST-SIB of an event-type could be considered as a continuous learning progress: the more samples the system is fed, the more knowledge the system gains. As a matter of fact, GST-SIB is able to capture common patterns that are present in similar real-life events, as soon as they appear in the training samples. To prove this fact, we measure the precision and recall obtained by progressively feeding additional training patterns. The result is illustrated in Fig. 14. It is possible to observe how convergence typically happens after about ten patterns of continuous learning, as expected.

This convergence pattern ensures that GST-SIB is a stable similarity measure, and that it has the ability to capture most of common patterns of an event. This is also an interesting feature of the proposed method: while in other approaches there is a need to understand or define manually and subjectively the conceptual information of events to set up a model, the proposed method can automatically discover common patterns shared by photo albums which belong to the same event type, without any manual interaction.

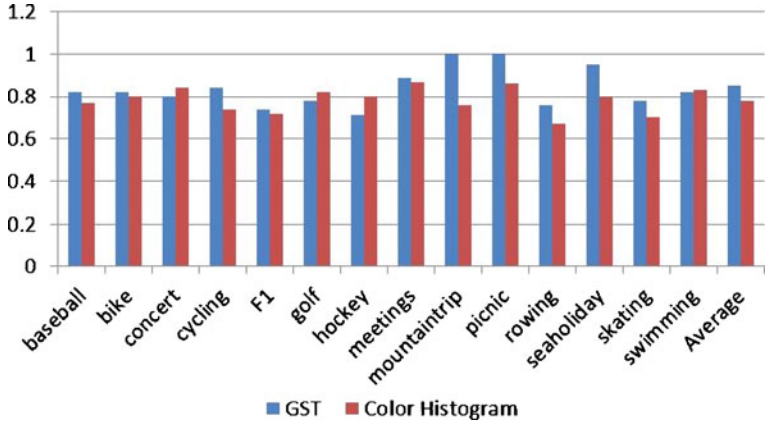
In fact, after reaching the convergence, the GST-SIB of an event-type contains enough information (i.e., common patterns) to classify a unlabeled collection independently of the dimension. The learning, of course, can be made incremental, including the new examples in the model. Clearly, in that case the support of the user is needed, in order to verify the classification before using the new information.



(a) Precision



(b) Recall



(c) F-scores

Fig. 9 Performance of GST-SIB vs. color histogram (color-SIB) [7]

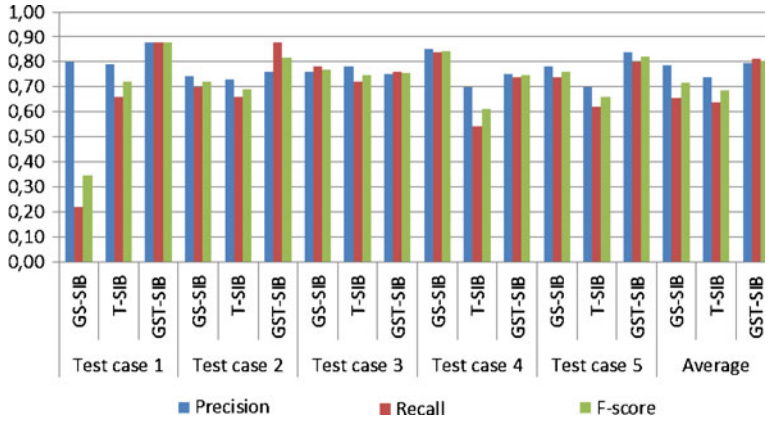


Fig. 10 Results on large dataset D2

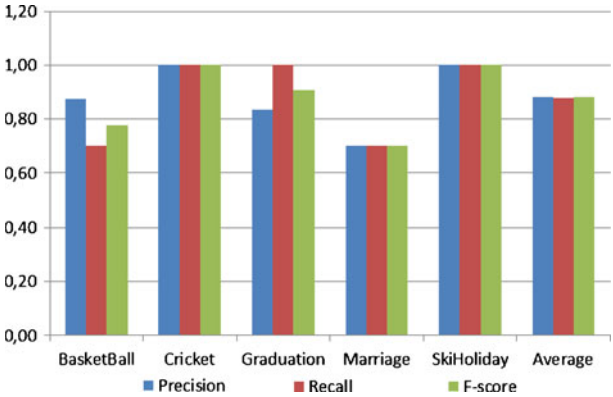
5.4 Computational performance

In order to evaluate the computational performance of the proposed method, we tested it on a PC with Intel(R) core i7 CPP 920 @2.67 GHz and 8GB RAM. As discussed in Section 3.4, the total cost for detecting an event from a given album is made up of two parts: calculating the GST-SIB of the current album, and comparing it with a set of models.

Figure 15 illustrates the average time needed to calculate the GST-SIB out of an album. Since there is a dependency on the size and number of images, the chart shows the processing time for various combinations of image resolutions and album cardinalities. It can be observed that for an average size album (100 photos rescaled to a standard size 256×256) this task requires less than 30 s.

As to the second step, it implies two distance computations; GS-SIB and temporal. Since the size of GS-SIB is fixed, the time requested to compare it does not change and is almost negligible (0.1 s) with respect to the time requested to calculate the temporal consistency with dynamic programming.

Fig. 11 Results on large dataset D2



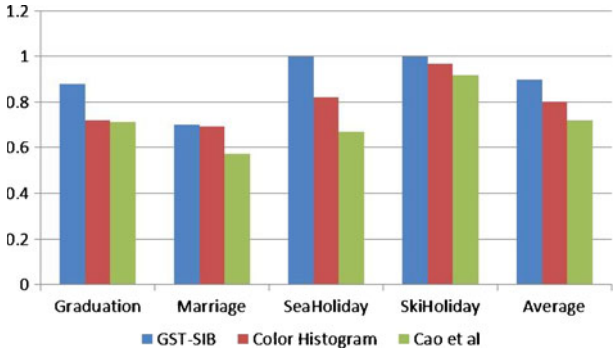


Fig. 12 Comparing perfomance of GST-SIB, color histogram (color-SIB) [7], Cao et al [5]

Figure 16 shows the total time requested to compare the GST-SIB of an album with our whole dataset. We could observe that the size of the album does not affect significantly the computation, whereas albums associated to events with more complex structures (e.g., wedding, graduation) require more computation than simpler ones (e.g., bike, F1). Also the average computation for step 2 anyway is below 30 s.

On the whole, the average time requested to detect an event starting from a typical photo collection is lower than 1 min, which makes our algorithm still usable also for on-line recommendation systems.

5.5 Discussion

The most important feature of GST-SIB is that it is able to capture a set of representative and discriminating information about events, without the need of explicit modeling of semantics, concepts, or taxonomies of the events themselves. Besides, whenever a new event type needs to be defined, the system is able to learn the relevant GST-SIB signature from a set of representative photo collections. In this sense, the proposed approach can be considered intrinsically scalable.

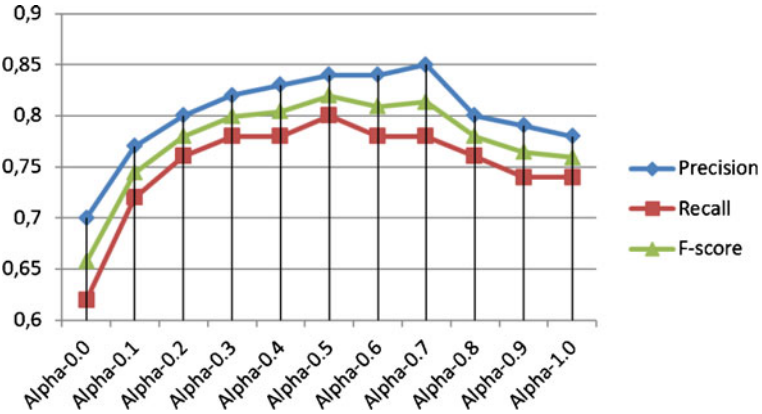


Fig. 13 The trade-off between GS-SIB and T-SIB

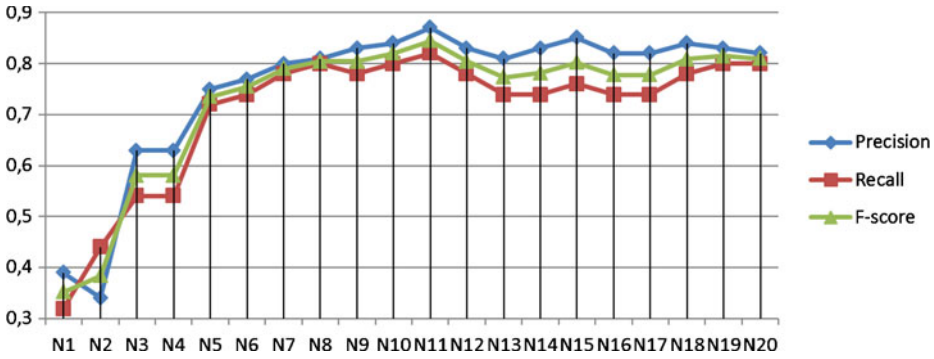
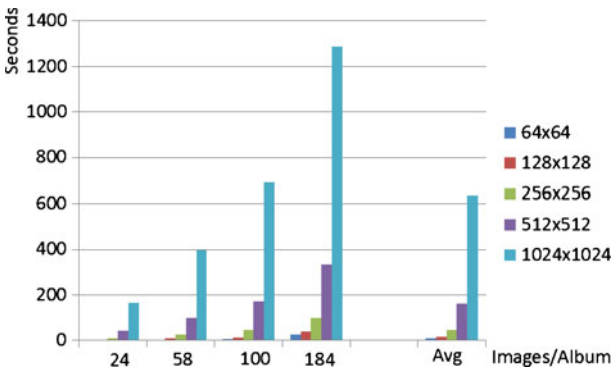


Fig. 14 The convergence of GST-SIB model

The possible dependency of the event detection accuracy on the characteristics of the photo album have been also investigated. The number of photos taken at an event and their layout partially depends on the individual photographer's habits and style. Some photographers take series of shots, whereas others select few significant instants. Amateur photographers typically take photos to preserve the memory of an event, whereas professional photographers consider their photos as art products. Those attitudes also influence the style of photos: a simple point-and-shot with full camera automation for most users, more sophisticated choice of subject and perspective with manual selection of parameters for more experienced photographers. Such phenomena may potentially affect the GST-SIB signature, and especially the saliency-related component. In our experiments however, we did not observe any significant dependency of the classification accuracy on the photographer. This is mostly related to the fact that large part of the material used to build our datasets comes from social networks, where the great majority of users are non professional, and this is the usual situation for most personal events. Furthermore, the holistic nature of our approach makes it more robust to the presence of single pictures with special or unusual characteristics that may appear in a collection. Nevertheless, it would be interesting to further investigate the presence of "photographer signatures"

Fig. 15 Average time for calculating GS-SIB for one album



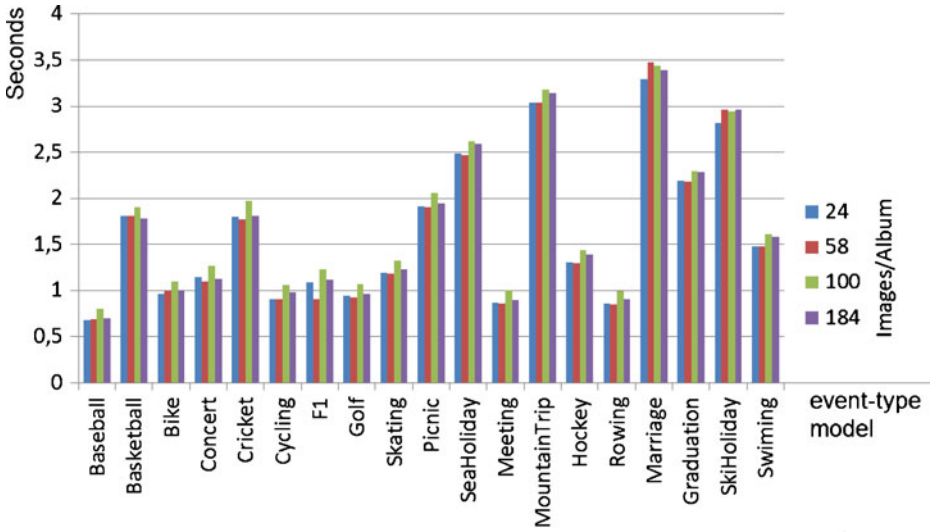


Fig. 16 Time for comparing GST-SIB of given album with GST-SIB models

left on the images, as a way to detect the author of a set of photos, which is a completely new problem.

As far as the album size is concerned, we observed that the classification accuracy is influenced more by the significance of contents than the number of photos. If the photo collection contains enough common and discriminatingve patterns, the resulting classification accuracy will be high. This is quite natural also for human beings: we cannot deduce that a photo album refers to a 'wedding' if we just observe a few pictures with elegantly dressed people, while we have evidence of it if we have the spouses in traditional dressing or the photos of the ceremony. Consequently, in our tests we were able to correctly detect an event even from as low as 25 photos, provided that they are representative and diverse. When the number grows to 50 or more, the accuracy reaches its maximum values.

The possible applications of the proposed approach are manifold. On one side, users may use it to organize and tag their own pictures, even if just for a few hundreds at a time, which is the typical average size when participating to a significant event. On a larger scale, it is worth having a structuring/annotation system being able to provide a uniform and homogeneous tagging across different users, to ease the management of shared collections. This may be the case of Flickr or Picasa users, where content producers may be supported in organizing and/or tagging their uploaded data in a rational way, while consumers may benefit of a more structured and coherent way of accessing data.

6 Conclusion and future work

In this paper, we introduced a novel method to detect the event-type associated with a photo collection with high accuracy and low computational cost. While

other related methods try to extract semantic information from the analysis of single images, the proposed approach considers all images of an event as a whole without the need of understanding the semantics of each image. By using this holistic approach, *Gist*, *Saliency*, and *Temporal* information of events are captured and simplified into a unique signature called *GST-SIB*. Thorough experimental analysis showed that such signature is capable of representing events and discriminating with high accuracy different event types. Besides, it does not require any a-priori or structured knowledge and it has an ability to increase its knowledge whenever new examples are provided, as well as to discover hidden common patterns/semantics without human intervention.

The main underlying assumption is that each photo album is associated with a single event. Evidently, an event may be composed of a sequence of sub-events at finer granularity. Consequently, the proposed method can provide two main functionalities:

1. Assign a label to the whole photo album: the proposed method will return the label of the detected event, which is then associated with the whole collection and/or to each image for future event-based search and retrieval.
2. Provide a context to the album, for further analysis, organization and tagging: in this case, the conditional knowledge of the main event can be used to drive more specific classification, for instance trying to discover possible sub-events with arbitrary granularity.

In the second case, appropriate extensions of the proposed method can be designed to deal with sub-events recognition.

An interesting point of the proposed GS-SIB is that it is apparently able to capture significant hidden elements of the event type. A thorough interpretation of these facts will be carried out in future research, where we will investigate the 'inverse' problem of building a dictionary of hidden or unknown common patterns/semantics discovered from GST-SIB model. This dictionary will be helpful for further tasks in event-based automatic tagging of photo album such as automatically building taxonomy of events. An interesting evolution of the proposed method concerns the possible analysis of fingerprints left by users in photos for further use in personalization or forensics. In this case, one should try to find user traces in media (see, for instance [6]), by detecting possible dependencies of the GST-SIB on the photographer, e.g., a particular habit or style in taking photos or arranging the subjects.

Acknowledgement This work has been partially supported by the EU Commission under the framework of the EU project grant no. 248984 "GLocal".

References

1. Achanta R, Susstrunk S (2010) Saliency detection using maximum symmetric surround. In: ICIP. IEEE CS Press, pp 2653–2656
2. Baeza-Yates R, Navarro G (1998) Approximate string matching in a dictionary. In: SPIRE. IEEE CS Press, pp 14–22
3. Begum M, Karray F (2011) Visual attention for robotic cognition: a survey. IEEE TAMD 3(1):92–105
4. Bezdek JC, Hathaway RJ, Huband JM (2007) Visual assesment of clustering tendency for rectangular dissimilarity matrices. IEEE Trans Fuzzy Syst 15(5):890–903

5. Cao L, Luo J, Kautz H, Huang T (2009) Image annotation within the context of personal photo collections using hierarchical event and scene models. IEEE Trans Multimedia 11(2):208–219
6. Conotter V, Boato G (2011) Analysis of sensor fingerprint for source camera identification. Electron Lett v. 47(25):1366–1367
7. Dao MS, Dang-Nguyen DT, De Natale F (2011) Signature-image-based event analysis for personal photo albums. In: ACM multimedia
8. Das M, Loui AC (2009) Detecting significant events in personal image collections. In: IEEE int. conf. on semantic computing. IEEE press, pp 116–123
9. Doran MM, Hoffman JE (2009) The role of eye fixations in concentration and amplification effects during object tracking. Taylor & Francis JVC 17(4):574–597
10. Douze M, Jégou H, Sandhawalia H (2009) Evaluation of GIST descriptors for web-scale image search. In: ACM CIVR. ACM press, pp 1–8
11. Fei LF, Lyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene? J Vis 7(1):1–29
12. Frintrop S, Rome E, Christensen HI (2010) Computational visual systems and their cognitive foundations: a survey. ACM TAP 7(1, 6):1–39
13. Han Y, Liu G (2010) A Hierarchical GIST model embedding multiple biological feasibilities for scene classification. In: ICPR. IEEE press, pp 3109–3112
14. Hays J, Efros AA (2007) Scene completion using millions of photographs. In: ACM SIGGRAPH. ACM Press, pp 1–7
15. Ibrahim A, AlZoubi A, Sahawneh R, Makhadmeh M (2009) Fixed representative colors feature extraction algorithm for moving picture experts group-7 dominant color descriptor. Journal of Computer Sciences 5(11):773–777
16. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
17. Jiang A, Wang C, Xiao B (2008) Scene modeling in global-local view for scene classification. In: ACM CIVR. ACM press, pp 179–184
18. Li L, Fei-Fei L (2007) What, where and who? Classifying events by scene and object recognition. In: ICCV, pp 1–8
19. Li Z, Itti L (2010) Saliency and gist features for target detection in satellite images. IEEE Trans Image Process 86(10):1–27
20. Lim JH, Tian Q, Mulhem P (2003) Home photo content modeling for personalized event-based retrieval. IEEE Multimed 10:28–37
21. Liu Y, Zhang D, Lu G, Ma W-Y (2007) A survey of content-based image retrieval with high-level semantics. J Pattern Recogn 40:262–282
22. Loui C, Savakis A (2003) Automated event clustering and quality screening of consumer pictures for digital albumin. IEEE Trans Multimedia 5(3):390–402
23. Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv 33(1):31–88
24. Oliva A, Schyns PG (2000) Diagnostic colors mediate scene recognition. Cogn Psychol 41:176–210
25. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelop. Int J Comput Vis 42(3):145–175
26. Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155:23–36
27. Renniger LW, Malik J (2004) When is scene identification just texture recognition? Elsevier JVR 44:2301–2311
28. Sandhaus P, Boll S (2011) Semantic analysis and retrieval in personal and social photo collections. Multimed Tools Appl 51:5–33
29. Siagian C, Itti L (2007) Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans Pattern Anal Mach Intell 29(2):300–312
30. Sibiryakov A (2008) Photo-collection representation based on viewpoint clustering. In: SPIE, vol 68(1). SPIE press, pp 6833-02
31. Vogel J, Schwaninger A, Wallraven C, Bulthoff HH (2007) Catergorization of natural scenes: local vs glocal information and the role of color. ACM TAP 4(3, 19):1–20
32. Wagenaar W (2004) My memory: a study of autobiographical memory over six years. Cogn Psychol 18(2):225–252
33. Wang W, Wang Y, Huang Q, Gao W (2010) Measuring visual saliency by site entropy rate. In: CVPR. IEEE press, pp 2368–2375
34. Yu J, Jin X, Han J, Luo J (2009) Mining personal image collection for social group suggestion. In: IEEE ICDMW. IEEE press, pp 202–207

Minh-Son Dao (M'04) received the B.Sc. degree in computer science in 1995 and the M.Sc. degree in computer science in 2000, both from the University of Ho Chi Minh City, Ho Chi Minh, Vietnam. He received the Ph.D. degree from the Dipartimento Informatica e Telecomunicazioni (DIT), Università degli studi di Trento, Trento, Italy, in February 2005. He worked as scientist at GraphiTech, Italy from 2005 to 2007. He was as a JSPS Postdoc at Media Integrated Communication Lab. (MICL), Graduate School of Engineering, Osaka University from 2007 to 2010. Currently, he serves as a researcher in MultiMedia Signal Processing and Understanding LAB (mmLAB)—University of Trento, Itsaly. His main interests include multimedia retrieval, event detection, data mining, image processing, computer vision, and pattern recognition.

Duc-Tien Dang-Nguyen finished his Master's thesis when he was at Toyota Technological Institute in Japan in 2009. He is currently working in MultiMedia Signal Processing and Understanding LAB (mmLab) in Trento University, where he's studying for his PhD. His interesting topics are: Event Analysis, Object Recognition and Digital Image Forensics.

Francesco G. B. De Natale graduated in Electronic Engineering (M.Sc. level) in 1990 at the University of Genova (Italy) and got a Ph.D. in Telecommunications in 1994 at the same University. In 1996 he got a position of Assistant Professor at the University of Cagliari and successively moved to the University of Trento, Italy, where he is Full Professor of Telecommunications Engineering (from 2003). He has been the Head of the Department of Information Engineering and Computer Science (DISI) from 2006 to 2009, and is the current Dean of the Bachelor and Master degrees in Telecommunication Engineering. He also leads the Research Lab on Multimedia Communications at DISI (mmlab.disi.unitn.it) and the MMSPI (Multidimensional Multimodal Signal Processing and Interpretation Lab) of the Italian branch of the European Institute of Technology (EIT-ICTLabs@Italy). His research interests are focused on multimedia communications, with particular attention to multidimensional signal processing, analysis, and archiving. His results are witnessed by the publication record, with more than 200 works published on major international peer reviewed scientific journals and conferences. He was General Co-Chair of the Packet Video Workshop (PV-2000), Program Co-Chair of the IEEE Intl. Conf. on Image Processing (ICIP-2005), and General Chair of the ACM Intl. Conf. on Multimedia Retrieval (ICMR-2011). He is Associate Editor of the IEEE Trans on Multimedia and of the IEEE Trans. on Circuits and Systems for Video Technologies. He is also member of the IEEE Signal Proc. Society Technical Committee on Multimedia Signal Processing (MMSP), chairing the Technical Directions Subcommittee. From 2009 he coordinates the EU-IP GLOCAL project, one of the key EU initiatives in media retrieval. Prof. De Natale was appointed evaluator for several international bodies, including the European Commission, and the NSFs of US and Ireland. Prof. De Natale is a Senior Member of IEEE and a member of ACM.