Event-based media organization and indexing

(In Invited Paper)

Riccardo Mattivi
University of Trento/Department of Information Engineering and Computer Science, Trento, Italy
Email: rmattivi@disi.unitn.it

Giulia Boato and Francesco G. B. De Natale
University of Trento/Department of Information Engineering and Computer Science, Trento, Italy
Email: {boato, denatale}@disi.unitn.it

Abstract—As highlighted in the recent literature, events may provide important information to associate content and context for media indexing and retrieval. Nevertheless, a thorough implementation of such concepts to viable applications is still lacking, due to the difficulty in defining a suitable representation of events but also in producing effective automatic tools to create and instantiate event structures. In this paper we give an overview of the existing work in the field and we present a showcase in which event structures are exploited to better organize and access personal media contents, allowing to automatically detect events and sub-events in photo collections and annotate media accordingly.

Index Terms—multimedia retrieval, events and sub-events recognition, photo collection organization.

I. INTRODUCTION

In the last years, the ever-increasing use of digital compact cameras allowed producing an enormous amount of multimedia contents. Nowadays, people use to capture the most important moments of their lives with pictures and videos and to store them on their pc, on mobile devices, or on social networks. This last instrument is also becoming the most popular solution to share a wide variety of multimedia information among individuals and groups. For a single user there is a clear link between such contents and their own experiences (e.g., their last holiday, their kids playing softball, the wedding of their best friend) and this is the most natural way of indexing their contents (e.g., “Winter 2010 Italy” or “Bob’s Wedding”). Indeed, recent studies in neuroscience have demonstrated that humans remember their life using past experience structured into events [1].

Our life is a constellation of events that, one after the other, pace our everyday activities and index our memories. A birthday, a marriage, or a summer vacation, are the lens through which we see and memorize our own personal experiences. Global events, such as world sport championships or global natural disasters or, on a smaller scale, a local festival or a soccer match, build collective experiences that allow us to share personal experiences as part of larger phenomena that we could call collective events. Events can therefore be used as the primary means for organizing and indexing media (e.g., photos, videos, journal articles), but also to share them (e.g., through social networks).

The typical approach to content-based media retrieval is to start from media and trying to extract information (descriptors) that helps understanding their content. Event-based analysis provides a way to reverse this approach, making events become the main key for organizing our memories and media an attribute of such memories, which represent our experience of that specific event. Events have a local dimension, which allows for a local mapping between tags and the personal experience, as represented in the data, thus making easier and doable understanding media contents. At the same time events have also a global dimension, in a twofold sense: first, the semantic of an event has a common ground for the peers involved in it, thus providing a way for sharing a specific event among involved people; second, events of the same type (e.g., weddings) share similar structures, in the sense that it is usually possible to identify a more or less standard sequence of episodes that characterize that specific event type (e.g., ceremony, group pictures, banquet), thus providing a way of sharing events of the same type within user communities.

From the user’s viewpoint, organizing and indexing photo collections through events may be beneficial, as it allows accessing their data in a natural and intuitive way. Furthermore, on this basis it is possible to develop powerful facilities to support users and communities in media management. As an example, let imagine a family coming back home from a winter holiday on the snow. At present, they would most probably download their photos from the camera to the pc and store them in a folder, perhaps with some manual annotation. An event-based indexing tool, exploiting a suitable model for that type of event, may automatically provide a structuring and labelling of the media collection with appropriate tags, sub-event recognition (e.g., skiing, taking a break, shopping, etc.), and possibly adding information on time, geographical location, people involved, and so on. This will be useful at a later time to retrieve their data based on multiple keys, or to match similar events (e.g., the previous year holyday), or to share it with friends that participated to the same event.
Event-based media organization and indexing

In this paper we overview the state of the art on event definition and recognition, by first describing basic image analysis tools exploited for this purpose (Section II), and then detailing the structure of events and various application scenarios where such a concept can be exploited (Section III). Section IV presents some results of automatic organization and classification of images based on three types of structured events (graduation, marriage, ski holiday), while Section V draws some concluding remarks.

II. IMAGE ANALYSIS TOOLS FOR EVENTS

In this section, we will review the basic existing tool for image analysis, which are extensively used in the techniques described in the following sections for event detection and recognition.

The question is: what kind of information a picture can provide? First of all, nowadays cameras or smartphones store a lot of useful information in the form of standard EXIF data. According to the device capabilities, such data may include date and time, camera settings - such as camera model, shutter speed, aperture, focal length, ISO settings - thumbnail information (for image preview in file managers or on the camera), but also GPS-based geolocation information, if available. In addition to this, the picture itself carries a lot of information in its pixel data, which can be analysed and exploited. In the last decades several global features were studied: colour histograms, which provide an estimation of colour distribution in images; invariant features, which are features invariant with respect to certain transformations; texture features and high-order moments, which give information about the relationships among neighbouring pixels. Among texture features, Gabor features computed as the result of a set of Gabor filters at different scale and direction have been widely used in this context. Tamura features have been used as well, for their ability to mimic the human visual perception in terms of coarseness, contrast and directionality.

In 2001 the MPEG7 group defined a standard set of visual descriptors that describe the visual content in terms of colour (five tools to compute colour distribution in a single image and colour relations into sequences of images), texture (three tools that describes the region homogeneity and edge histograms), shape (three tools to describe contours, regions or shapes after a segmentation phase), and motion (several algorithms that describe the motion in video sequences).

Colour and Edge Directivity Descriptor (CEDD) features represent a compact set of colour and texture statistics, which can be exploited in large image databases or in video segmentation. GIST descriptor aims at describing real world scenes bypassing the segmentation and the processing of individual objects or regions and it is based on a set of Gabor filters evaluated at different scales and at different orientations.

Although the typical approach is to calculate such descriptors on the entire image, thus providing a unique representation of the whole image, it is also possible to compute them in particular regions of the image, corresponding to some visual object or simply to a subblock. In the following we detail local descriptors, which are widely used for media content characterization.

Scale-Invariant Feature Transform (SIFT) algorithm collects features in interesting parts of the image. In its original version, SIFT detects some interest points in an intensity image, extracts a region around them and describes the relevant content. The features computed with this technique are invariant to translation, rotation and scale and have been widely used for different purposes such as image stitching, object recognition, wide line matching, and robotic mapping. In the last years several modifications have been proposed, such as the extension to several colour spaces, different method for computing the description of local patches, or fast approximations for the computation of both localization and description of interest points in the so-called SURF algorithm. Recently, a fast and dense computation of local features has been proposed in [16], which outperforms previous methods for object and concepts recognition.

Once visual features are obtained from raw pixel values, further steps are required to obtain more information for retrieval, recognition, or other goals. An adage says, "A picture is worth a thousand words". What about a group of pictures? Several clustering algorithms exist in literature whose aim is to gather photos into clusters exploiting time, geo-location, similarity of visual content, or tag information. The k-means algorithm finds k clusters based on the distribution of data in the feature space. Several modifications have been proposed, such as approximate k-means or hierarchical clustering. Grouping of pictures can be done also by analyzing distances inside a matrix that represents the distances between each feature vector, by applying probability latent semantic analysis to feature vectors, or by applying dominant sets.

In the following section, these basic tools will be used for image clustering and event recognition purposes.

III. EVENTS AND EVENT RECOGNITION

Event recognition is a term widely used in computer vision and image analysis, which encompasses several phases in the process of data interpretation, from the recognition of actions in video sequences, to the characterization of behaviours, the detection of specific situations, and the recognition of activities from single images. An event, with all the attributes that will be described later in this section, can be considered as a point in a multi-dimensional space, where each dimension brings a different facet of such a complex entity. This representation highlights the diversity that is inherent the concept of event, and includes the notions of time, space, activities, actors, emotions, etc. Photos and videos (among other media) depicting an event can be regarded as additional sources of information, or the "experiential" dimension of the...
event. The concept of diversity in itself is gaining special
attention in information retrieval, as witnessed by the
recent developments of search engines such as Yahoo!
and Google, and the big investments in research (see,
e.g., the EU FET IP Project LivingKnowledge and
related references [29]).

If one thinks to the typical way an event in one’s life is
recalled, typically the above dimensions are explored in
our memories answering simple questions such as: where
(place), when (time), who (people involved), what (type
of event), how (various attributes that characterize the
event). Once localized the event, people typically search
for the relevant photo albums, if available, to refresh the
visual memory of the event. Accordingly, the pioneering
work by Westermann and Ramesh [30] defines an event
model as characterized by the following aspects/axis:

Temporal: events are strictly related to the concept of
time; this information is easily stored in the time stamp
of multimedia content and it can be absolute or relative to a
previous event.

Spatial: an event occurs in a geographic area, usually
captured by a GPS receiver, or even in a region of interest
of a picture.

Informational: the event model should provide
additional information about the involved actors and the
involved entities: this information can be taxonomically
organized.

Experiential: since the event model is applied to
multimedia applications, the model should support
connections to different media contents and should
provide the users with an easy way to explore them.

Furthermore, events can be linked each other by
several types of relationships. For instance, an event may
be the cause of another one, or may simply precede or
follow it. An event can be also part of another larger
event (thus leading to the concept of event granularity).
On this basis one can define the following relational
attributes to an event:

Structure: event models may apply at different levels
of abstraction and granularity (e.g., sub-events, similar
events, etc.).

Causality: an event model should support the
description of cause-effect relationships.

This structure allows a flexible collection and
management of all information, data, and multimedia
contents connected with a certain event. In the following
subsection we describe how such different aspects can be
analysed and exploited for event-based media
organization and indexing.

III.A Event detection and event-based organization

The problem of recognizing events from visual data
inherits researches from recent years in the field of scene,
object and landmark recognition. In the following, a brief
survey on these methods is given, relying on the basic
tools described in the previous section, with references
and focus on the event recognition task.
Event-based media organization and indexing

classification. Such a model has been used for object recognition [34], scene recognition [34], and human action recognition [35][23].

More specifically for event recognition, Li et al. in [36] classify sport events in single images by fusing object and scene recognition. For instance, a rowing competition is recognized by the presence of a given scene (e.g., a lake) and some typical objects (e.g., boats, athletes, etc.). Thus, event recognition is achieved taking into consideration as much semantics as possible in the image interpretation. In [37], Jiang et al. exploit local visual features by building a 21-concept space instead of dealing with a higher dimensional low-level feature space. Rasiwasia et al. in [38] propose a similar approach to classify events in an intermediate space, by exploiting a low dimensional semantic theme for image representation. Indeed, each theme induces a probability density on the space of low-level features, and images are represented as vectors of posterior theme probabilities, thus reducing the computational time both for training and testing phase. In [39], Imran et al. mine a set of pictures belonging to specific events and describe it as a histogram of word occurrence, using the BoW approach. They try to discover the most informative features of different events using the PageRank technique, achieving a small increase in the performance. Das et al. in [40] investigate the use of high-level visual and temporal features and determine a subset of features that provide good correlation with the event class.

More recently, Lin et al. in [41] build a framework for vision-assisted tagging of personal photo collections using people, events, and location information. They exploit a probabilistic model of context that couples the multiple domains through a set of cross-domain relations, where each relation represents the probability of co-occurrence in two domains. In [28], Tsampoulidis et al. represent images and videos as vectors of responses given by a high number of trained visual concept detectors. The high-dimensional vector in the concept-space is then analyzed with a subclass discriminant method for identifying the most appropriate concept subspace to detect and recognize the target event. A novel method for fast and robust event detection is introduced by Dao et al. in [42]. The key idea is to define and extract a unique signature from a set of photo albums related to a given event category. The event type of a new photo collection is then given by the classification of its signature.

Usually, image annotation systems consider a single photo at a time and label the photos individually. However, as highlighted in [42], collections of personal photos contain information that could be exploited at a global level. In recent years, automatic clustering of digital photo sets has received increasing attention and picture timestamp information is one of the most exploited features to achieve this task (see Graham et al. [43]). Together with time information, content-based features have also been used to build systems able to summarize photos into events: Cooper et al. [19] present a multi-scale temporal and content similarity clustering to define salient moments in a digital photo library, while Lim et al. [44] summarize collections combining content and time information and use predefined event taxonomy to annotate new photos.

Space information, ever increasingly available in terms of GPS coordinates or geographic landmarks, is also relevant data that can be exploited to browse and organize picture archives. For instance in [45], Jaffe et al. present different frameworks for generating summaries of geo-referenced photographs with map visualization. In [46], Cao et al. make use of the contextual information naturally provided by the associated GPS position and time metadata in a collection of pictures. They first employ a constrained clustering algorithm to partition a photo collection into event-based sub-collections, taking into consideration both time and place. Then, a conditional random field is used to model the correlation among photos, based on time-location constraints and on the relationship between collection-level annotation (i.e., events) and image-level annotation (i.e., scenes). Another approach for exploiting the geo-location information is introduced by Yuan et al. in [47]. They mine informative features derived from traces of GPS coordinates and from the bag of visual words. Such features are based both on the entire collection and on individual photos. In [48], Luo et al. exploit the photo location in a different way: they obtain satellite images corresponding to the picture position and investigate a novel use of these data by recognizing the picture-taking environment, as if the picture taking activity has been seen by a third eye above the user. Moreover, they combine this inference with classical vision-based event detection methods and they study the synergistic fusion of the two approaches.

In the next Section we propose a novel methodology for event and sub-event detection and recognition, which exploits content-based features and EXIF data to support the user in photo album management and annotation.

IV. EVENT-BASED PHOTO COLLECTION MANAGEMENT

As a proof of concept, we propose here a completely automatic recognition system able to discover events from a collection of pictures and able to suggest a structure according to the relevant sub-events. A preliminary version of this approach was proposed in [49].

The system works as a cascade of classifiers, as follows:

Event level: the system first classifies a given photo album into “social” or “sport” event (binary classifier) and successively labels it more specifically according to given macro-category (see events description in Table 1).

Sub-event level: on the basis of the detected event type, a further understanding of users’ activities is conducted and a deeper analysis is performed in order to recognize all sub-events present in the image collection (see examples of sub-events recognized for social events in Table 2).

The cascade of classifiers supports events and sub-events discovery from general information to a more specific one. As an example, a photo album depicting a
wedding is first classified as belonging to social class, successively to wedding class and further on its specific moments are classified as ceremony, group-pictures or other sub-event types.

In the following we provide details of the dataset used, the experimental setup and the obtained results.

**IV.A Dataset**

In the literature, different datasets have been collected and made available to the research community for testing and comparing different algorithms, such as action recognition [50][51], object recognition [34], scene recognition [52]. Unfortunately, this is not the case for event detection from media where, to the best of our knowledge, no extensive corpora have been made available so far. Although many social networks are used to share event-related media among private users, no publicly available datasets contain consistent event-related tagging. In [40] and [47], authors use databases obtained from several users spanning different events, but not available to the community. Recently, MediaEval benchmarking initiative for multimedia evaluation [53] has released a dataset for social event detection (limited to testing purposes), where images are crawled from Flickr with associated tags and time data.

Due to the increasing interest in this field of research, we initiated the collection of an event-based image database, which consists of two different event families: sports and social events. Within them, we selected several event types, as reported in Table 1. The pictures of our database were downloaded as entire album collections from PicasaWeb Album service and manually labelled at the image level. For some specific social events that show a quite well defined structure (Graduation, Wedding, and SkiHoliday), we further manually labelled each image with sub-event classes as specified in Table 2 and some of them are shown in Figure 1. We define as event-albums a collection of pictures spanning one event and collected from one user.

| Event classes for social and sports events |
|-------------------------------|-----------------------------------|
| **Main event**                | **Event-albums classes**          |
| Social events                 | Concert, Graduation, Wedding,     |
|                               | Meeting, Mountain Trip, Pic-nic,  |
|                               | Sea Holiday, Ski Holiday          |
| Sport events                  | Baseball, Basketball, Bike,       |
|                               | Cycling, F1, Golf, Hockey, Rowing |
|                               | Figure Skating, Swimming          |

<table>
<thead>
<tr>
<th>Table 2 Sub-event classes for some social events</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social events</strong></td>
</tr>
<tr>
<td><strong>Wedding</strong></td>
</tr>
<tr>
<td>Group pictures, ceremony, party-eating, unknown</td>
</tr>
<tr>
<td><strong>Graduation</strong></td>
</tr>
<tr>
<td>Group pictures, celebration, party-eating, unknown</td>
</tr>
<tr>
<td><strong>Ski-Holiday</strong></td>
</tr>
<tr>
<td>Skiing, walking in the town, eating at hotel, eating, relax during skiing, partying, unknown</td>
</tr>
</tbody>
</table>

**IV.B Experimental Setup**

We make use of time information from EXIF data and visual information following the pipeline of Uijlings et al. [16]. We densely extract RGB-SIFT features, we project the descriptors with a Random Forest and finally we use Support Vector Machines (SVM) with histogram intersection kernel for classification. The codebook has been built on the Pascal 2007 database and has 4096 visual words. Such configuration results as the best tradeoff between accuracy and computational time. We proved in [49] this visual pipeline to be state-of-the-art in a standard sport event database.

We designed the system in two different ways:

- **Single-step classifier**: the system is trained with all social and sports event classes at once for the purpose of event recognition, and with all sub-events classes for the purpose of sub-event recognition. In this approach there is no previous discrimination between sports and social classes and no a-priori information about the event type being analyzed for the sub-event recognition phase.

- **Multi-step classifier**: the system is built as a cascade of classifiers working from coarse recognition to a more specific sub-class categorization. For instance, an event-album depicting a graduation ceremony is first classified as belonging to the social-event class, then to graduation, and finally to its specific sub-event types.

At the event level we followed two approaches for the recognition task:

- **Single image**: this approach provides the classification of single images, discarding event-related information, and represents the baseline for our comparisons.

- **Post processing - majority info**: since each event-album belongs to a single event, the event class label of each image is given by the majority of the single image event labels in the analyzed test instance.

Regarding sub-event analysis, several activities can be identified in a photo collection. Typically such activities are bounded in time. Therefore, the first step is to group the pictures into homogeneous clusters based on time information [19]. We exploit such information in two different ways: by applying post-processing algorithms on the single image classification output or by dealing with the visual features of each cluster as a whole.

- **Single image**: this represents the classification of single images taking in consideration only the visual information, regardless the belonging to a cluster.

- **Cluster BoW**: in this approach we exploit the information of the cluster in the classification phase. The signature of the cluster is obtained by averaging the BoW signatures of each image belonging to the cluster itself. The system is trained with the signature of single images, but tested with the signatures of the clusters. The label of the cluster is then assigned to each single image inside the cluster.

- **Post processing, median filter**: the result of the classification is filtered with a median filter with a fixed window size. The assumption is that successive
pictures belong to the same sub-event class. This method represents a baseline for the post-processing algorithm.

- **Post processing, cluster information:** the classification of images belonging to a cluster is given by the majority vote inside the cluster. The assumption is that pictures grouped into clusters belong to the same class. This method is similar with the approach used at the event level.

As our results are influenced by the reliability of the clustering algorithm to give homogeneous clusters, in the analysis of results we report the upper-bound as the classification achieved with the “post-processing, cluster information” method having the ground truth as test set (post processing, upperbound).

We conducted the experiments using the Leave-One-Out methodology: having n event-albums, we trained the system with n-1 albums and we performed the tests on the remaining one. Such task is repeated for n times and the results are averaged over all tests. As metric of performance we use the f-measure, defined as \( f = \frac{2 \cdot P \cdot R}{P + R} \), where the precision P is defined as \( P = \frac{t_p}{t_p + f_n} \) and the recall R as \( R = \frac{t_p}{t_p + f_p} \). In the previous formulas, \( t_p \) stands for true-positive, \( f_p \) for false positive, and \( f_n \) for false negative. The results refer always to the classification of single images.

**IV.C Results**

1) **Single-step vs Multi-step**

We are first interested in comparing the single-step vs. the multi-step classification strategies for both events and sub-events. As can be seen in Figure 1, the SVM classifier trained with both sport and social events (18 classes in total) gives an average f-measure of 0.66. The multi-step classifier performs better, with an average of 0.73, at the cost of a slightly higher complexity (due to two step classification). More specifically, in the second case the first step of binary classification of the event-album into sports or social event classes gives an average f-measure of 0.91 for sports and 0.88 for social events. The application of a post-processing algorithm (as mentioned in the experimental setup and detailed in the following) allows recovering also the erroneous cases, thus achieving a perfect categorization of sport or social event class. The classification within each macro-class is achieved through a separate set of classifiers dealing with simpler problem (8 classes for social or 10 classes for sports. This results in an average f-measure of 0.81 for sport events classification and of 0.62 for social events.

The same effect is noticed for sub-events recognition within structured events (Graduation, Wedding and SkiHoliday): single-step classifier trained with all sub-events from the three social event classes provides an average f-measure of 0.58, while multi-step classifier yields an improvement of 0.14. Also in this case, the duty of the classifier is simplified, having a reduced number of sub-events to recognize. Table 2 shows the results obtained for sub-event recognition (single-step and multi-step labels).

We conclude that the exploitation of events at different granularity level helps improving the overall recognition results.

2) **Exploiting time at the event level**

We now use the time information to improve the results both at event and sub-event level. At the event level, we apply the post-processing - majority info method. We focus on the results of Figure 1 to show the f-measure for the sport and social classes for single image classification (in the figure named also as multi-step classifier) and for our post-processing algorithm as majority voting. First of all, we can see that the post-processing significantly outperforms the baseline for sports events with an improvement of 0.17 in the f-measure (thus achieving an almost perfect classification). As a matter of fact, the post-processing is able to correct the confusion between visually similar classes (see for instance Bike vs. Cycling vs. F1, Hockey vs. Figure Skating, Baseball vs. Golf) mainly due to similar environments (speed circuit, ice-hockey stadium, or grass) where only minor parts of the scene change.

Similar results are shown for social event classification. We score an f-measure of 0.62 as baseline (in Figure 1 named as multi-step classifier), while f-measure increases to 0.89 with post-processing majority voting. Also in this case we observe a larger confusion between images with similar environment or sharing similar activities (e.g., outdoor eating in pic-nic or mountain trip, or people standing in a large meeting room in graduations and meetings). However, the representative images for each class are mostly correctly classified. Thus, the post-processing helps in disambiguating the Mountain trip, Sea Holiday and Ski Holiday classes and a significative improvement is obtained for the Graduation class.

We can conclude that considerable improvements can be achieved over single-image classification by defining them as events and by using majority voting over the event-albums.

3) **Exploit time at sub-event level**

We now exploit time information at sub-event level, since an event-album may contain multiple sub-events. We compare the classification of single images (the baseline) with an approach that uses the signature of the entire cluster and two post-processing strategies. We also provide an upper bound measure about the quality of the clustering algorithm. All the results are summarized in Table 3.

We first evaluate the ability of the clustering algorithm to give uniform clusters. Results, named as post processing, upperbound, are shown in Figure 2. The upperbound for Graduation and Wedding is very high with an f-measure of 0.98 and 0.96 respectively. However, performance for the Ski-Holiday class is only 0.87, due to the confusion between bar-relax and skiing sub-events. This is explained by the fact that many pictures are taken before or after a break instead of directly in action, thus inducing errors in the time-based clustering.
Table 3 Values of f-measure for sub-events recognition obtained with different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Grad.</th>
<th>Wedd.</th>
<th>SkiH.</th>
<th>Av.ge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.7406</td>
<td>0.7284</td>
<td>0.7017</td>
<td>0.7236</td>
</tr>
<tr>
<td>Cluster BoW</td>
<td>0.7564</td>
<td>0.7541</td>
<td>0.6726</td>
<td>0.7277</td>
</tr>
<tr>
<td>Post-median filter</td>
<td>0.7671</td>
<td>0.7807</td>
<td>0.6228</td>
<td>0.7235</td>
</tr>
<tr>
<td>Post-cluster info</td>
<td>0.7598</td>
<td>0.7631</td>
<td>0.7232</td>
<td>0.7487</td>
</tr>
<tr>
<td>Post-upperbound</td>
<td>0.9788</td>
<td>0.9568</td>
<td>0.8663</td>
<td>0.9340</td>
</tr>
</tbody>
</table>

The post processing - median filter helps by leveraging the sparse classification errors, thus performing better than the baseline approach for all classes. This method does not include information on the cluster borders, and therefore is prone to errors at the border of the sub-events. The Graduation gives an f-measure of 0.74 for the baseline and an increment of 0.02 is achieved with this method.

The post processing - cluster info method allows achieving better performances. The Graduation and Wedding classes reach the best performance of 0.77 and 0.78, respectively, while Ski-Holiday shows a slight decrease in accuracy, due to the aforementioned confusion in the clustering. This method is able not only to handle and correct sparse classification errors inside the cluster, but also at its borders, provided that the clustering is of sufficient quality.

Finally, we exploit cluster information with the cluster BoW method, which gives better results than the baseline but slightly worse than the post-processing cluster info method. Hence by using the cluster signature one can save computational effort at the expense of a small reduction in classification quality.

We conclude that the post-processing - cluster info is the best strategy to boost the classification accuracy of sub-events, if the quality of the cluster algorithm is sufficient.

V. CONCLUSIONS

In this paper we proposed an overview of the existing research about event recognition and event-based media indexing, showing how “events” can boost the performance of conventional content-based retrieval systems, making it possible to conveniently associate content with context. Moreover, we demonstrated a possible application of this concept in the field of personal photo collection organization. In particular, we showed that event structures may be exploited to better organize and access personal media, by introducing events and sub-events recognition in photo collections and automatically providing album organization.

ACKNOWLEDGMENT

This work is developed within the GLOCAL Project, co-funded by the EU Research Framework Programme 7 under DG INFSO Networked Media Systems.
Figure 1. Events: Single classifier, Multi-step classifier and Multi-step classifier with post-processing technique.

Figure 2 Sub-events: single classifier, multi-step classifier and post-processing techniques. Multi-step classifier label is the baseline for the comparison of the post-process techniques.
REFERENCES


Riccardo Mattivi was born in Italy in 1984. He received BA and MSc in Telecommunication Engineering from the University of Trento, Italy. He is currently a PhD candidate in Information and Communication Technologies at DISI department at the University of Trento.

His researches focus on detection and recognition of events in multimedia data within the EU project GLOCAL. He is interested in pattern recognition, image analysis and understanding, image retrieval and multidimensional data management and analysis.

Giulia Boato was born in Italy in 1979. She received the M.Sc. in Mathematics in 2002, and the Ph.D. in Information and Communication Technologies in 2005, both from the University of Trento, Italy. In 2006 she was visiting researcher at the Signal Theory and Communications Department of the University of Vigo (Spain). Currently, she is Assistant Professor at the Department of Information Engineering and Computer Science (DISI) at the University of Trento, working in the Media Laboratory.

Her research interests are focused on image and signal processing, with particular attention to multimedia data protection, data hiding and image forensics, but also intelligent multidimensional data management and analysis. On these topics she is active in the European Projects coordinated by DISI: in LIVINGKNOWLEDGE for the development of image forensics methods supporting opinion mining on the web, in GLOCAL for the design of multimedia annotation and retrieval systems based on the concept of event.

Giulia Boato is a Member of IEEE.

Francesco G.B. De Natale (M.Sc. in Electronic Engineering, 1990, Ph.D. in Telecommunications, 1994) is Professor of Telecommunications Engineering at University of Trento, Italy.

His research interests are focused on multimedia communications, with particular attention to multidimensional signal processing, analysis, archiving and transmission. He was General Chair of the Packet Video Workshop (PV-2000), TPC 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 member of the IEEE Signal Proc. Society Technical Committee on Multimedia Signal Processing (MMSP). In 1998 he was co-recipient of the IEEE Chester-Sall Best Paper Award. He is the Coordinator of the EU IP GLOCAL project.

Prof. De Natale is a Senior Member of IEEE.