Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

EVENT DETECTION AND SCENE ATTRACTION BY VERY SIMPLE CONTEXTUAL CUES

Ivan Tankoyeu, Javier Paniagua, Julian Stöttinger, Fausto Giunchiglia

August 2011

Technical Report # DISI-11-470

Event Detection and Scene Attraction by Very Simple Contextual Cues

Ivan Tankoyeu, Javier Paniagua, Julian Stöttinger, Fausto Giunchiglia
DISI, University of Trento
via Sommarive 14
38123 Povo,Trento, Italy
[tankoyeu | javier.paniagua | julian | fausto]@disi.unitn.it

ABSTRACT

We detect and arrange events in private photo archives by putting these photos into context. The problem is seen as a fully automated mining in one's personal life and behavior. To this end, we build a contextual meaningful hierarchy of events based on personal photos. With the analysis of very simple cues of time, space and perceptual visual appearance we are refining and validating the event borders and their relation in an iterative way. Beginning with discriminating between routine and unusual events, we are able to robustly recognize the basic nature of an event. Further combination of the given cues efficiently gives a hierarchy of events that coincides with the given ground-truth at an F-measure of 0.83 for event detection and 0.70 for its hierarchical representation. We process the given task in a fully unsupervised and computationally inexpensive manner. Using standard clustering and machine learning techniques, sparse events in the collection would tend to be neglected by automated approaches. Opposed to these methods, the proposed approach is invariant to the distribution of the photo collection regarding the sparsity and denseness in time, space and visual appearance. This is improved by introducing a *momentum of attraction* measure for a meaningful representation of personal events.

Categories and Subject Descriptors

I.4.8 [**Computing Methodologies**]: Image Processing and Computer Vision—*Scene Analysis*; I.2.10 [**Computing Methodologies**]: Vision and Scene Understanding

General Terms

Algorithms, Theory, Human Factors

Keywords

Event Detection, Context, Personal Photo Collection

1. INTRODUCTION

In the last years, a vastly growing number of digital cameras and a continuous reduction of the cost of mass storage are observed. Modern devices such as GPS equipped cameras and smart phones provide extensive additional information, including spatial and temporal data of the captured image. This omnipresent availability of taking and storing photos typically results in a vast personal collection. These photos tend to be organized in an unsound manner and cannot be retrieved in a way that is convenient to the user. The main reason for this is that managing and annotating personal collections of digital photographs is difficult and tiresome. Additionally, one's personal organization might not pass the test of time for the owner, e.g. the favorite organization changes over the years, making the previous photos unclassified. Therefore, there is a strong interest for an automated, reliable and flexible image classification and annotation solution for personal photo collections.

There is strong evidence [1] that people classify intentionally taken photos according to events in their lives [22]. These events are seen on different *scales* or *granularity*, as there are unrelated root events, which are increasingly refined into semantically connected sub events. When photos are seen as frozen moments of our memories, the same can be applied for photo collections [16]. To this end, this work aims to organize photos in the most natural and convenient way for the user: By emulating the semantic hierarchy the memory is built of. This is done by combining very simple cues of time, space and color in an efficient and robust way, building a semantic hierarchy of the events of life. Semantic event recognition in images is a challenging and complex problem. It becomes even more difficult when a data-set of images is of a heterogeneous structure. State of the art content-based computer vision systems are not powerful enough to solve the aforementioned problem. In the best case, they are able to give professional users intelligent tools at hand to retrieve the desired photo in a semi-automated way [18]. Recent works in the computer vision community show good results in event recognition and classification, but only for specific domains, on very limited data-sets, or number of classes [3]. Recent studies show the growing interest in processing contextual information of automatic image classifications [19].

In this work it is shown that the fusion of visual content and context information is crucial for improving the performance of automatic image clustering. While the aim of content analysis is to process the visual part of photos on a low-level information only, the context information is used for describing environmental conditions of an image and the basic property of the event it belongs to. The context information is accompanying the information of a photo that can be captured by a camera, or extracted from related images in the same photo collection, or it might be even a textual description provided by user.

The paper is organized as follows. In the following Section 2 the state of the art is given, Section 3 describes the proposed framework in detail. Experimental validation is given in Section 4, while Section 5 concludes.

2. STATE OF THE ART

The importance of sorting images around events is discussed in [16]. According to this study, people tend to group their personal photo collections around events. [12] and [13] show an approach for clustering photo collections based on temporal information and visual content of images. [1] presents an innovative approach to event recognition of collections of images: The main idea of the paper is to build a collection of photos with time stamp and geographical coordinates, and to define a compact ontology of events and scenes. Their model takes into account two types of correlations; the first is a correlation by time and GPS tags, the latter is a correlation between scenes represented in images and corresponding events. Temporal intervals between photos are used in [9] in order to group these photos using an adaptive threshold. The same feature has been used in [11]. Low level features of photo content and creation time of the photo are exploited in [10] for the task of automatic *summarization*. The semantic gap [17] is a challenging problem that is still unsolved in the multimedia community. Different approaches have been introduced in order to solve this problem [5]. Recently studies have shown a growing interest in context processing as one of the step towards bridging the semantic gap [8, 2]. For some tasks such as event-based indexing of contextual information it is shown that time and space are more important than visual content.

2.1 Events

Our life is a constellation of events which, one after the other, pace our everyday activities and index our memories. Events such as a birthday, a marriage, a summer vacation, or a car accident are the lens through which we see and memorize our own personal experiences. In turn, global events, such as world sport championships or global natural disasters (e.g., the 2004 tsunami, climate change, or the world recession) 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 a more social phenomenon that we could call *collective events*. When describing events, we ground in our experience, our common and abstract understanding of the world and the language that we use to describe it. The generic notion of "beach" is then associated to a specific time and place, which is probably frozen in the photo we have taken back then.

Events provide the common framework inside which the local experience-driven contextual information can be not only codified but also shared and reduced to a common denominator. Thus, for instance, the photo of a person on a beach taken on vacation can be contextualized to the specific time (night or day? which season?), to the specific location (which part of the world?) and to the specific event (summer vacation or some more specific sub-event).

The aim of matching content and concept must be done while taking context into account [4].

2.2 Events in Context

From a psychophysical point of view the importance of context for human perception has been discussed in [15]. Events can be seen as useful entities that provide a way to encode some contextual information, and aggregate media that constitute the experience of such event. Context can be defined as the totality of environmental conditions. We define several types of image context:

Photo-parametrical context comprises camera parameters which accompany the moment of image capturing. Mostly it is metadata from the EXIF[1] standard. It includes a set of attributes related to temporal information (timestamp), spatial information (GPS coordinates), device information (camera maker, model, etc.), lens information (focal length, exposure time), flash information (flash, light return).

Image environmental context covers knowledge about environmental conditions during image acquisition. It includes knowledge about the season of the year (summer, winter), weather information (raining, sunny) and place information (indoor, outdoor). This kind of context could be broadly used for improving the performance of image analysis.

User-generated context is any information generated by the user or his actions related to a photo. These data play a valuable role in enriching the semantics of a photo. Textual description, tags, comments in social networks, paths to the folder of the corresponding photo collection in the file system, file name, surrounding images, all are parts of the user-generated context.

The compositional aspect of events have been presented recently in [21]. An example of events representation with different granularities of abstraction can be seen in weddings, which typically include the stag party, the main ceremony, the first dance, cutting the cake and several other parts. The importance of building event hierarchies is shown in recent studies like [14], where the authors mainly focus on the issues of event composition using the sub-event-of relationship between events. In order to represent the possible semantics of a composite event, the event attributes should be computed as a function of its sub event attributes.

3. SEMANTICS FROM CONTEXT

This section shows how the contextual information of a photo can be used to build semantics. The processing of the number of photos, the temporal and the spatial information generate semantically meaningful semantics to the user. We suggest a multi-modal clustering (MMC) approach described in the following Section 3.1, the proposed framework is presented in Section 3.2.

3.1 Multi Modal Clustering

A straightforward approach for hierarchical analysis of multidimensional data would be the hierarchical clustering, as there are agglomerative ("bottom-up") approaches and divisive ("top-down") approaches. Due to the excessive memory and run-time requirements of hierarchical clustering [7], partitional clustering, such as k-means, is the method of choice for large scale applications.

We use a method of [11] which is independent of the density and robust to extreme variation in the distribution of the data. It allows for a robust clustering without knowing the final number of clusters. Opposed to their work, we apply the 2-means to a combined time-space-visual clustering. The method locates significant gaps in assorted data, where for each sample n_i in the set $n_{i-1} \leq n_i \leq n_{i+1}$ holds. Where n_i can be timestamp or color vector or gps coordinates. The method uses the distances Δn between the samples n and clusters them using k-means with $k = 2$. It divides efficiently one cluster with many small distances in the data and another cluster that defines fewer, but significantly larger gaps in the data. Every distance is now assigned to be a member of either the first or the latter cluster. For each distance being a member of the cluster c_i with the larger distances, a boundary in the data is marked. For the final cluster estimation, we iterate over all $n_{1..N}$ once. Every n is merged with the current cluster until a boundary

[1] www.exif.org

is encountered. Then, a new cluster is built. For a data-set D with the samples $n_{1..N}$, the pseudo code is given in Alg. 1.

Algorithm 1 multi modal clustering

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

For temporal clustering in photo collections, the time intervals between photos are of a too large variation to provide satisfying results. We face problems when a large time difference among events is interpreted as the only point in the cluster corresponding to event separations. Therefore, a scaling function ψ (compare Fig. 1) to scale Δn has to be introduced. We start with the exemplary time intervals of 1 day (x=1440 min), and proceed further to 1 week (x=10080 min) and 1 month (x=43200 min). For one day periods, we scale with the square root function. Thus, f_1 is given by

$$f_1(x) = \sqrt{(360(x-b))} + c. \tag{1}$$

In order to find constants b and c the continuity of the function and its derivative is used

$$\begin{aligned} f_1(360) &= f_0(360) = 360, \\ f_1'(360) &= f_0'(360) = 1. \end{aligned} \tag{2}$$

For general time periods, the function is therefore estimated as follows

$$f_i(x) = \sqrt[y^i]{(A_i(x-b))} + c_i. \tag{3}$$

To optimize the scaling of ψ, the variation over A_i ends up with $A_1 = 360$ (1/4 day) , $A_2 = 1440$ (1 day), $A_3 = 10080$ (1 week), $A_4 = 43200$ (1 month). Thus,

$$\begin{aligned} f_2(x) &= \sqrt[\frac{12}{5}]{1440(x-1078)} + 588(\bigvee x \in [1440, 10080]) \\ f_3(x) &= \sqrt[3]{10080(x-7880)} + 1197(\bigvee x \in [10080, 43200]) \\ f_4(x) &= \sqrt[4]{43200(x-38815)} + 1788(\bigvee x \geq 43201) \end{aligned} \tag{4}$$

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

3.2 Event Detection

Fig. 2 shows a flowchart of the proposed approach. The numbers (1) to (7) link to the corresponding paragraphs in this section.

People tend to think about events in terms of spatial entities leveraged by personal context like *home*, *work*, etc. Moreover, moving away from routine locations typically establishes lasting memories. Therefore we advocate discriminating events in two categories: *home* and *away-from-home*.

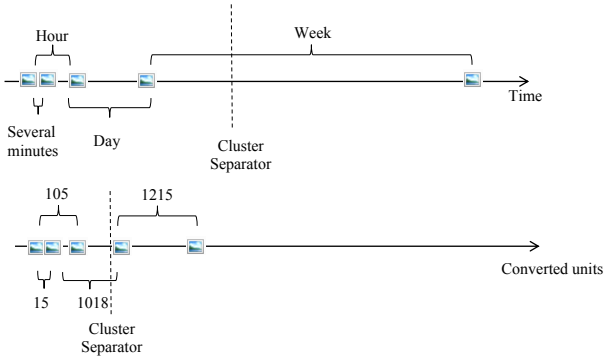

Figure 1: Scaling by $\psi(\Delta n)$ and MMC on temporal scale

One fundamental presumption of the approach is that breaking a routine – and returning to it – frames one semantically connected memory. This entity provides the borders of a root event being *away from home*. Since a trip typically starts on routine places, e.g. taking photos at the home airport or photos of saying good-bye to the family, photos taken at routine places may belong to an event away from routine place. A movement out of the routine locations starts a new root event, which is hierarchically detecting sub events until the routine location is entered again. This does not apply for *home* events: spatial-temporal cluster can be either root events or sub events. A priori no more context is given. The *home* events span sparsely a long temporal period but occur on a very narrow space scale.

Routine Detection (1) First we map the GPS information of the photos to meaningful GeoLocation[2] by reverse geocoding. Reverse geocoding is the process of converting geographic coordinates into a readable address or place name. We use a granularity of city level, or, if not available, province. This gives us already a meaningful clustering of our photo locations. A density function Φ of GeoLocations is built based on the number of days when photos have been taken – accumulated by location. The function Φ is given in Fig. 4. Note that the function is invariant to the absolute number of images taken in one location (compare Fig. 3). In simple terms, we want to define home as where you take the most photos on different days.

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

Temporal MMC (2) Temporal information is the most essential and reliable information for detection of events within the personal photo collection. The main reason is that the time stamp is unique, whereas the spatial location is not. Photos captured within an event are typically characterized by relatively small temporal gaps between them. Therefore the time intervals between chronologically neighboring images are fed to the MMC algorithm. All extracted information from image time stamps are converted to minutes, which are set to be the basic unit of the algorithm. Therefore, images of one event have often no time difference. The clustering results in an assignment of every photo in the collection to the clusters $c_1, c_2, , c_n$.

Spatial DenClue (3) Based on the temporal clusters C_i we discriminate between photos taken at routine locations p_h and photos that

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

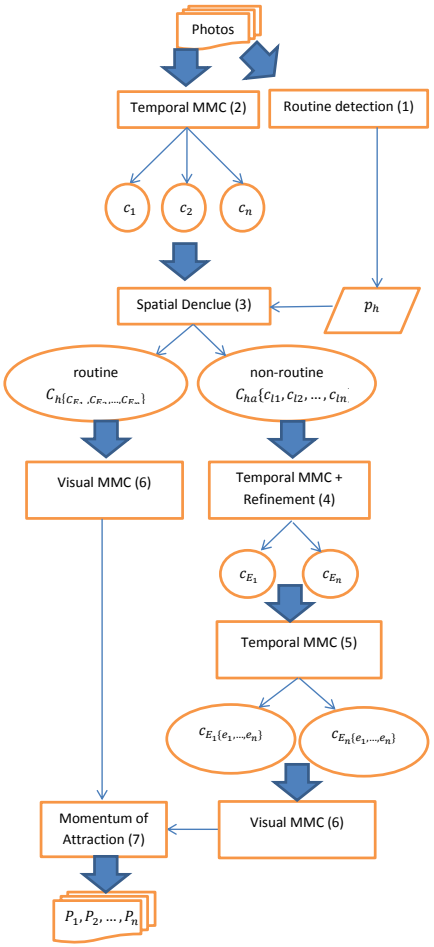

Figure 2: Schematic overview over the proposed framework.

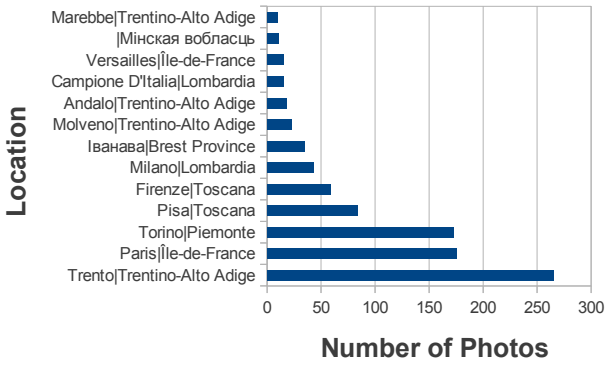

Figure 3: Distribution of photos per location.

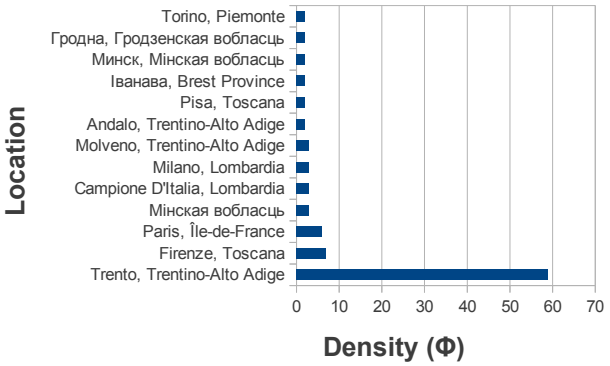

Figure 4: Density (Φ) of days with photos taken per location

are taken outside p_h. As soon as one photo of an event is taken outside of p_h, the event is regarded as being a non-routine event. We spatially cluster all non-routine photos by their GPS coordinates. In this step we use a temporally unsorted set of absolute GPS coordinates, we cluster the locations by the density-based clustering algorithm (DenClue) [6] with σ=10.0 km. The resulting spatial clusters give us density clusters of locations that may disseminate over various locations. There is no relation between taking two photos in the same location and them being part of the same event. For example, c_{l1} is a cluster of photos taken in Milan. This cluster contains photos related to a Christmas time and Italy Republic Day which should be separated into two clusters. Therefore, for each cluster C_{ha} $\{C_{l1}, C_{l2}, .., C_{\ln}\}$, we perform **temporal MMC + refinement (4)** with a higher level of granularity by scaling the spatial distances by f as defined in Section 3.1. This results in the final root event clusters.

For every C_{En} we perform **temporal MMC (5)**. These clusters are sub events. C_{ha} $\{C_{l1}, C_{l2}, .., C_{\ln}\}$ is a set of "away-from-home" clusters where each cluster corresponds to some location.

Visual MMC (6) A semantically meaningful color similarity [20] is chosen to make up for varying lighting conditions and camera settings. Following a user study on images downloaded from the Internet the mapping of 11 English color names to RGB coordinates is learnt, thereby creating a look-up table of each RGB coordinate to one of the eleven color names[3].

Without any presumption about the nature of the content of the image, we do not rely on color models when matching visual content, but we merely ask figuratively "*Can this object still be regarded as brown?*". Following the look-up table of colors, as long as people would agree that the color stays the same, we can successfully match two images.

For histogram creation the 11-dimensional feature vector is extracted for each image. Euclidian distances between feature vectors of neighboring images are computed on the next step. MMC clustering is performed to separate event boundaries from similar images related to one event. As shown in Section 4.3, the results highly coincide with the event boundaries given by the previous steps of the framework. This leads to a finer subdivision of the events in event *scenes*.

Momentum of Attraction (7) We propose a measure of saliency for single scenes in the event hierarchy. The underlying idea is that every photo is taken intentionally. Therefore, the more photos that are taken in a short time, the more interesting or exciting one event should be. We observe the relative change in the recording frequency of photos by measuring the acceleration of the time differences. This measure is defined as the *momentum of attraction* (MoA). With this measure, we provide a straightforward cue on how scene changes related to personal behavior. The assumption is that things which change our behavior rapidly are important to us. This allows us to retrieve the most interesting shots conveniently.

4. EXPERIMENTAL VALIDATION

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

This section gives the experimental validation of the proposed approach. The approach tries to emulate these results fully automatically. In the following Section 4.1 the data-set is described, Section 4.2 gives the experimental set-up. Section 4.3 provides the numerical results of the evaluation.

4.1 Data-set

The data-set consists of 1008 images taken from 25.06.2010 to 20.03.2011 by one person. Therefore, the data-set consists of the photos of 268 days, or almost 9 months. For the first 6 months, the data-set was produced unintentionally, meaning the owner was not aware that it would used for this research. All images have time stamps and 646 images have GPS stamps (in real life, GPS reception is not always available). 39 people are depicted on the collection (for future work on face recognition and participant reasoning). The images have been captured in four countries and 25 cities and towns. 79 events were defined manually, 9 events have at least 1 sub event, summing up to 36 sub events. The photos are taken by a Google Nexus One[4] smartphone with a 5MP resolution of 2592 × 1944, sRGB IEC-61966-2 color profile and a fixed focal length of 4,31. For scientific purposes, the data-set is available on request.

Figure 5: All locations of photos in the data-set.

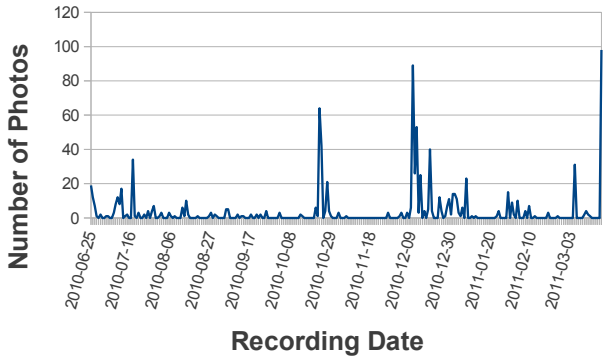

Figure 6: Number of photos per day.

4.2 Experimental Set-up

The given data-set exemplifies a typical private photo collection. As ground-truth, it provides a manual and subjective hierarchy of events that are semantically connected. The ground-truth is only justified by the personal experience of the user. Therefore, a perfect result is very unlikely to be achieved. A perfect experiment would result in all *true positive* detections having exactly the same event boundaries as what the ground truth provides. Every event boundary not detected gives a higher *false negative*, every boundary separating one event gives a *false positive*. In this sense, false positives are probably less problematical, as it is still easy for the user to retrieve a desired photo. On the other hand, a false negative detection "hides" an event from the user making it harder to find. From this perspective, the recall rate is probably more meaningful when organizing one's photos.

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

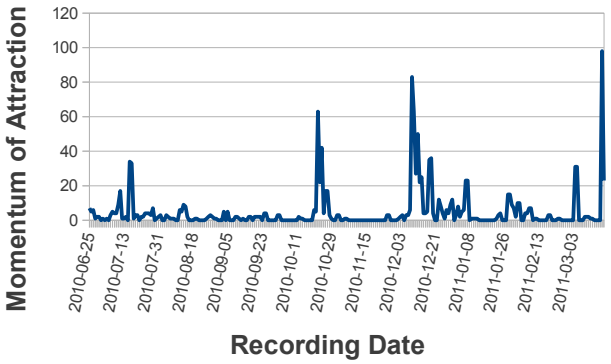

Figure 7: Momentum of Attraction: Acceleration of change of behavior on temporal scale.

4.3 Results

As a baseline and state of the art reference, we use the approach of [11]. In the numerical overview in Tbl.1, it is denoted as *temporal MMC*. It is based on the time stamps of images only. With an F-measure of 77,35%, it already provides promising results for the given task of event detection. This is a clear sign that the unique key of the recording time is the crucial information for detecting events but unfortunately it gives no contextual information, making it impossible to derive a hierarchy.

Using other cues, we can improve these results and enhance the final classification with a hierarchy of events, a scene detection and a saliency measure per scene (compare Fig.8), adding additional semantics to the detection. Spatial clustering gives the worst results as the temporal information is lost and too many events are merged. It provides a precision of 44%, recall of 14,10% and F-measure of 21,36%. Visual matching provides a high recall rate, but separates the data-set in too many different scenes, more than a person would want to organize his pictures. Compared to the given ground-truth, we receive 17,86% of precision, 76,92% of recall and 28,99 of F-measure. But more importantly, by using this efficient color similarity, subsequent events are merged only 4 times in the data-set. In this sense, semantically-right visual classifications of sub-scenes are done with an accuracy of 94,5%. The most attractive and the most unattractive scenes are shown in Fig. 8.

The density function Φ for the data-set is given in Fig. 4. It clearly shows the significantly higher density of Φ for the home cluster, which is Trento, Italy. The MMC algorithm gives this location as the only routine place, correctly determining it for this experiment. Note that this approach is invariant to the number of photos or photo density, taken per location, as seen in Fig. 3 per location and per day in Fig. 7. Carrying out the proposed approach we achieve 76,09% of precision, 90,91% of recall, results and 82,84 of F-measure in event detection. This outperforms the state of the art.

	precision	recall	F-measure
temporal MMC [11]	67,31	90,91	77,35
spatial DenClue	44,00	14,10	21,36
visual MMC	17,86	76,92	28,99
proposed approach	**76,09**	**90,91**	**82,84**

Table 1: Numerical results event detection.

	precision	recall	F-measure
temporal-spatial MMC	44,44	66,67	53,33
proposed approach	**72,22**	**68,42**	**70,27**

Table 2: Numerical results of hierarchical event representation.

For the hierarchical representation (compare Tbl. 4.3), we evaluate the temporal-spatial MMC without knowledge of daily routine with the proposed approach. This lack in context decreases the results in precision to 44.44%, but still gives a recall of 66.67%. The proposed approach gives a significantly higher precision of 72.22% which leads to an improved F-measure of 70.27%.

Figure 8: MoA: Most attractive scene in the upper row, the tower of Pisa. Second attractive scene: on the roof of Notre-Dame, second row. Most unattractive event: a photo of a new book (lower right).

5. CONCLUSIONS

People are often overwhelmed by the number of photos they produce or get shared by their friends. Using very simple cues given by time stamp, spatial location and perceptual color distribution, we are able to mine in one's personal life and behavior. It is shown that we are able to build a semantically meaningful hierarchy of events in a fully automated way. Of course, the actual implementation of such a hierarchy is always very subjective. Future work will include very sparse user interaction and user correction. This will give additional context, enabling for better classification throughout the data-set. The important point is that the user has to do as little as possible to be able to browse his visual memories. The experiments validate this outlook: We are able to show that the hierarchy of events independently defined by the user coincides with the proposed automated solution. This minimizes the effort of the user to organize his photo gallery and makes the retrieval of desired photos more convenient. The proposed method of using a flexible MMC approach for all data-sources makes the approach efficient in computational time. The visual scene classification arranges the photos in a meaningful way. Combined with the moment of attraction, we provide the ability to retrieve the personally most important images to the user.

6. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP7-248984 GLOCAL.

7. REFERENCES

[1] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Annotating collections of photos using hierarchical event and scene models. In *CVPR*, 2008.

[2] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, pages 762–775, 2010.

[4] F. Giunchiglia. Contextual reasoning. *Epistemologia, edizione speciale su "I Linguaggi e le Macchine"*, 16:345 – 364, 1993.

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

[6] A. Hinneburg and H.-H. Gabriel. Denclue 2.0: Fast clustering based on kernel density estimation. In *ISIDA*, pages 70–80, 2007.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[8] R. Jain and P. Sinha. Content without context is meaningless. In *ACM MM*, pages 1259–1268, 2010.

[9] M. C. J.C. Platt and B. Field. Phototoc: automatic clustering forbrowsing personal photographs. In *ICM*, 2003.

[10] J. Li, J. H. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *PCM*, pages 1536–1540, 2003.

[11] A. C. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *Multimedia*, 5(3), 2003.

[12] A. G. Matthew Cooper, Jonathan Foote. Automatically organizing digital photographs using time and content. In *ICIP*, pages 749–752, 2003.

[13] A. G. Matthew Cooper, Jonathan Foote and L. Wilcox. Temporal event clustering for digital photo collections. *TOMCCAP*, 1(3), 2009.

[14] S. Rafatirad, A. Gupta, and R. Jain. Event composition operators: Eco. In *EiMM*, pages 65–72, 2009.

[15] I. Rock. *The logic of perception*. MIT Press, 1980.

[16] K. Rodden and K. Wood. How do people manage their digital photographs? In *CHI*, pages 409–416, 2003.

[17] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *PAMI*, 22:1349–1380, 2000.

[18] J. Stottinger, J. Banova, T. Ponitz, N. Sebe, and A. Hanbury. Translating journalists' requirements into features for image search. In *VSMM*, pages 149–153, 2009.

[19] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32:1582 – 1596, 2009.

[20] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *CVPR*, 2007.

[21] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *Multimedia*, 14:19–29, 2007.

[22] J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L. Buckner, and M. E. Raichle. Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, 4(6):651–5, 2001.