Exploitation of Time Constraints for (Sub-)Event Recognition

Riccardo Mattivi, Jasper Uijlings, Francesco De Natale, Nicu Sebe
DISI department
University of Trento, Italy
{rmattivi, jrr, denatale, sebe}@disi.unitn.it

ABSTRACT

The aim of this paper is threefold: (a) to introduce a dataset for the recognition of events and sub-events in photographs taken by common users; (b) to propose event-based classification to achieve a more accurate labeling of event-related photo collections; (c) to use time clustering information to improve the sub-event recognition in an efficient Bag of Features classification approach. The dataset is organized according to event models and provides a collection of sample instances that allow the comparison of different recognition systems. On this basis, we will demonstrate how the use of time clustering together with multiple image visual features can outperform single image classification.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]; I.4 [**Image Processing and Computer Vision**]; I.5 [**Pattern Recognition**]

General Terms

Algorithms, Performance, Experimentation.

Keywords

Events, Sub-Events, BoW, Clustering

1. INTRODUCTION

In the last years, the ever-increasing spread of digital compact cameras allowed producing an enormous amount of multimedia contents. People take a lot of pictures and videos of their moments of life, and store all these digital data on their own pc or on web albums. Moreover, the increasing spread of social networks allows sharing lots of multimedia information among individuals and groups. For each user there is a clear link between such contents and their own life experiences (e.g., their last holiday, the wedding of their best friend) and this is the most natural way of indexing their contents (e.g., "Summer 2010 Spain" or "Carol's Wedding"). Also, most of these experiences have some kind of common structure, in the sense that it is usually possible to identify a more or less standard sequence of episodes that characterize a specific event type. Recent studies in neuroscience have shown that humans remember real life using past experience structured in events [15]. Organizing images through events is therefore beneficial from a user point of view and would enable to develop powerful facilities to support users and communities in managing their media.

For example, imagine a user back home from the graduation ceremony of his nephew or a family back from a winter holiday on snowy mountains. Nowadays, these users would download their pictures from the camera to the pc and store them in a specific folder for future memories. However, it would be better to store each picture labeled as 'graduation' and to know which picture describe the ceremony or which one is a group picture.

Earlier work on automatic categorization addressed the classification of concrete concepts from single images, such as mountains, grass or beach [13], or concrete objects such as cats, cars, and persons [11]. In this paper we aim for a more semantic categorization into events and sub-events, which is very meaningful from a user point of view [12]. We want to leverage the organization of images in applications such as Picasa, iPhoto or Windows Media Center, from tags into an event-based structure.

An event imposes logical constraints on the automatic organization and classification of images which we intent to exploit. A case study is presented as a proof-of-concept and several types of social and sport events have been defined, together with three types of structured social events further labeled into sub-events.

We use time as a logical event constraint to (a) classify images within an event as a whole. Afterwards the images are classified as one of its event or sub-events. This leads to a multi-step classification process. (b) Within an event we cluster images into sequences and classify these sequences into sub-events. We show that in both cases considering collections of images using time yields considerable improvements over the use of single images for event classification.

The main contributions in this paper: (a) we introduce a photographic dataset designed for events and sub-events recognition; (b) we show the exploitation of event model to be useful for a more accurate recognition of structured events; (c) we use the time clustering information for a more accurate sub-events recognition in an efficient Bag of Words classification approach.

J-MRE'11, November 30, 2011, Scottsdale, Arizona, USA.

2. RELATED WORK

Event recognition is a term widely used in image analysis that includes the recognition of activities in video sequences, the detection of anomalous situations happening in a scene, the characterization of a given behavior. More in general, people tend to classify as an event every significant moment occurring in their lives. Such events can be described in words, usually answering to a number of basic questions such as: where (the place), when (the time), who (the people involved), what (the type of event), how (the attributes that characterize the event) [12]. In this work we propose a database for categorization of events as defined in [12]. In recent years, the computer vision community has adopted and adapted the Bag of Words model, which is commonly used in text retrieval. In the image domain, an image is represented as an unordered collection of "visual words" and the model is usually named as Bag of Words (BoW) model. This model has been used in, e.g., object and scene recognition [11], visual concept classification [10] and human action recognition [4]. We adopt BoW for event classification.

Regarding event recognition, Li et al. [8] classify sport events in single images by fusing object and scene recognition. Event recognition is achieved considering into consideration as much semantic-level image interpretation as possible. Jiang et al. [7] exploited local visual features by building a 21-concept space instead of dealing with a higher dimensional low-level feature space. In [6], Imran et al. mined a set of pictures belonging to specific events and describe them following the BoF methodology. They try to discover the most informative features of different events using the PageRank technique and the results show a small increase in the performances. We use an efficient BoF approach exploiting time constrains for event and also for sub-event recognition.

Using other sources of information, Cao et al. [1] annotate a collection of photos by exploring the GPS and time information associated with them. The authors employ a conditional random field model to exploit the correlation between pictures based on time and location, on level annotation and on scene types. Das et al. [3] focuses on investigating the use of a variety of high-level visual and temporal features, and determining a set of features that show good correlation with the event class. They proposed a Bayesian belief network for event classification that computes the a-posteriori probability of event classes given the input features. We address a different problem, namely sub-event categorization.

Automatic clustering of digital photo sets has also received increasing attention in recent years. Picture timestamp is one of the most exploited features to achieve this task, e.g. in the work of Graham et al. [5]. Content-based features have also been used to build systems able to summarize photos into events: Lim et al. [9] built an event-based retrieval system combining content and a pre-defined event taxonomy and tested on four categories. Cooper et al. [2] present a multi-scale temporal and content similarity clustering to define salient moments in a digital photo library. We employ the time-based clustering algorithm of [2] to find image sequences that correspond to sub-events.

3. THE DATASET

In this work we are interested in the recognition of events and their structure. Although many social networks are used to share event-related media among private users, to the best of our knowledge no dataset for event and sub-event recognition is publicly available. We therefore initiated an event-based image corpora, which consist of two different kind of event families: sports and social events. We chose 10 event types for the sports activities and 8 event types for the social class and we define *event-album* a collection of pictures spanning one event and belonging to one user. The pictures of our database were downloaded as entire album collections from Picasa Web Album service. We labelled each image according to the event classes of Table 1, examples are shown in Figure 1a. The social event Graduation, Wedding and SkiHoliday were further divided into sub-event classes according to Table 2. Examples of sub-events for Graduation are shown in Figure 1b. The selected event types provided a good variety of situations in terms of time span, location

(a) Sports and social events, example of images. From the 1st row to bottom: Baseball, Bike, Cycling, Meeting, Mountain trip, Ski-Holiday (skiing and bar-relax sub-events)

(b) Graduation class, example of images for different sub-events. From the 1st row to bottom: ceremony, group-pictures and party-eating class

Figure 1: Example of images for sports and social events.

Table 1: Event-classes for social and sports events. In brackets the number of images per each class. Each class has 15 instances

Main event	Events description
Social Events (13219)	*Concert (1085), Graduation (1815), Wedding (1776), Meeting (795), Mountain Trip (2051), Pic-nic (1627), Sea Holiday (2253), Ski-Holiday (1817).*
Sport Events (19754)	*Baseball (1635), Basketball (1196), Bike (1366), Cycling (1735), F1 (1606), Golf (2138), Hockey (2487), Rowing (2512), Figure Skating (2786), Swimming (2293).*

Table 2: Event and sub-event labels for 3 classes of Social Events. In brackets the number of images per each event/sub-event class. Each class has 20 instances.

Social ev.	Sub-event classes
Graduation (2180)	group pictures (1099), celebration (748), party-eating (135), unknown (198).
Wedding (2120)	group pictures (601), ceremony (514), party-eating (910), unknown (95).
Ski-Holiday (2141)	Skiing (1286), walking in the city (115), eating at hotel (156), eating-relax during skiing (233), partying (267), unknown (84).

(indoor/outdoor, environmental conditions, point of view, etc.) and visual content.

Grouping of pictures into sub-events was performed on a heuristic basis, asking several users to group the pictures according to their personal taste. Results were uniform, except for some differences in the level of granularity: some users gave more sub-event labels, such as cake cutting and party-dancing within the Wedding event. By ensuring that we have enough images per sub-event, we obtained a three-level event structure (no further split of sub-events).

Compared to Flickr albums, we found out Picasa albums to be more realistic, as they were less edited or manipulated. We found that the picture collections uploaded at once belong to a single event and the collection is seen as a whole for the classification task. This assumption holds for most real-world data. However, when it does not (e.g. pictures taken in several days and across multiple events), a clustering algorithm can be applied, similar to what we do in this paper for sub-event recognition.

We collected images for 10 sports events and 8 social events. On average, one event-album of sports event contains 132 images while one event-album of a social event contains 110 images. The total number of pictures is 32973. Images of the social events Graduation, Wedding and Ski-Holiday, may not fall into the subcategories defined in Table 2. These are labelled unknown. Note that different events show different properties, e.g., Graduation event spans only one day and usually has a precise order of its sub-event: first celebration, then group-pictures and finally party-eating. In contrast, SkiHoliday could span several days with no fixed order of the sub-events.

The dataset is available to the community at the web address http://mmlab.disi.unitn.it/.

4. EXPLOITING TIME CONSTRAINTS FOR (SUB-)EVENT RECOGNITION

In this paper we use, together with the analysis of visual features, time information to achieve a completely automatic recognition system for events and sub-events in common user's picture collections. We reach this goal at two levels:

Event level: We are interested in the categorization into the social or sports event and the subsequent categorization into concert, graduation, meetings, etc.. In this dataset each set of images covers one event only.

Sub-event level: In this part we want to achieve a further understanding of users' activities and a deeper analysis is performed in order to recognize sub-events. Unlike with events, there are often multiple sub-events in an image set.

4.1 Exploiting time at the event level

We designed a cascade of classifiers to recognize the different user's event activities. An initial classifier is used to distinguish between social and sport event classes. Successively another classifier, trained on the 10 sports classes or the 8 social classes, is used to categorize the test instance as one of the sports or event type of Table 1. We followed two approaches for the recognition:

Single image: this approach represent the classification of single images discarding the information of event; it represents a baseline.

Post processing - majority info: since each test instance belongs to only one event, a simple post-processing algorithm is performed to exploit time information: the event class label for each image is the majority of the single image event labels in the analyzed test instance.

Regarding the recognition of sub-events we followed two different methods.

Single-step classifier: we do not take into consideration the event class a specific test instance belongs to. We trained a classifier with all the different sub-events classes of table 2. This approach represents a baseline.

Multi-step classifier: we first recognize the event classes of the test instance, as explained before, and only afterwards we exploite a classifier trained with only the sub-events of that specific event. This approach follow the initial concept of classifier cascade and the sub-event categorization represent the third level.

4.2 Exploiting time at the sub-event level

Once the event has been recognised, we want to divide it into sub-events. In contrast with events, one set of images spans multiple sub-events, where we do not know when one sub-event ends and the other begins. But as a sub-event often consist of multiple, consecutive images taken shortly after one another, this can be exploited. We do this both using a time-based clustering algorithm and by using a simple median filter. We exploit the time information given by the aggregation of pictures into clusters in two different methods: by applying post-processing algorithms on the single image classification output or by dealing with the visual features of each cluster as a whole. Specifically, we compare the following methodologies for sub-event recognition:

Single image: this is the baseline of our comparisons and is given by the classification of single images regardless the belonging to a cluster or to any time information.

Cluster BoW: this method exploits the information of the clustering technique in the classification phase. The signature of the cluster is given by averaging the 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 (whose number is about a order of magnitude less then the number of single images). 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 equal to 5. 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 should belong to the same class.

Post processing, upperbound: we use the majority vote of the ground truth labels inside a cluster to assign the labels to each image. The result of this classification gives an upperbound measure about the quality of the clustering algorithm.

4.3 Implementation details

Following the pipeline of Uijlings et al. [10], we densely extract RGB-SIFT features, project the descriptors with a Random Forest and we use Support Vector Machines (SVM) with histogram intersection kernel for classification. The codebook has 4096 visual words and it has been built on the Pascal 2007 database. This configuration gives us the best tradeoff between accuracy and computational time. Our experiments have been conducted as a Leave-One-Out methodology. The results shown in the following section are therefore the average over n simulations, where n is the number of instances as specified in Table 1 and 2. In each simulation we use n-1 event instances for training and the remaining one for testing. The photos labeled as unknown are excluded from the training and from the testing phases. The clustering we use in our test is our implementation of [2] considering only the time information.

5. EXPERIMENTS

In this section we first verify that our BoW classification pipeline is state-of-the-art in a standard dataset. Then we tested the exploitation of time at both event and sub-event levels as described in the previous section. As measure of our performances, we used the F-measure, defined as $F = (2 \cdot P \cdot R)/(P + R)$, where precision P is defined as $P = (t_p/(t_p + f_p))$ and recall R as $R = t_p/(t_p + f_n)$. t_p stands for true-positive, f_p for false positive and f_n for false negative. The accuracy Ac used for the comparison of our pipeline is computed as $Ac = (t_p + t_n)/(t_p + f_p + t_n + f_n)$. The results of our tests refer to the recognition of single images.

5.1 Sports Event Dataset

We first verified if our BoW classification pipeline is state-of-the-art in event recognition. Since our approach gives an initial classification result per each single image, we used the *Sport Event Dataset* of Li et al [8] for comparison. Following their testing methodology we used 70 random images per class during training and 60 random images during testing. We repeated the experiment several times to compute an averaged accuracy.

Table 3: Comparison on a standard image sport dataset [8]

Method	Accuracy
Li et al. [8]	73.4 %
Wu et al.* [14]	78.50 %
Our pipeline	87.97 %

Table 3 summarizes the performances achieved by different methods. Li et al. [8] reached an accuracy of 73.4% by using a full integrative model. Wu et al [14] achieved an overall accuracy of 78.5%, but the result is achieved for scene recognition and has to be compared with the 60% obtained in [8]. The dense color SIFT extraction together with the other optimization done in [10] permits us to reach a higher accuracy equal to 87.97%. We conclude that our BoW pipeline is state-of-the-art for event categorization.

5.2 Exploit time at event level

5.2.1 *Experimental Setup*

This experiment investigates how we can increase recognition performance by exploiting time at the event level. We perform two experiments. First of all we investigate how time at the event level can improve categorizations into the main events. As a baseline we use single image classification. We compare this with using all images within a collection to classify the event, where we use majority voting. Second, we investigate how time at the event level can improve categorization into the sub-events. Again we use the single image classification as a baseline. We compare with our multi-step classifier as proposed in section 4.1. All results refer to the recognition of single images.

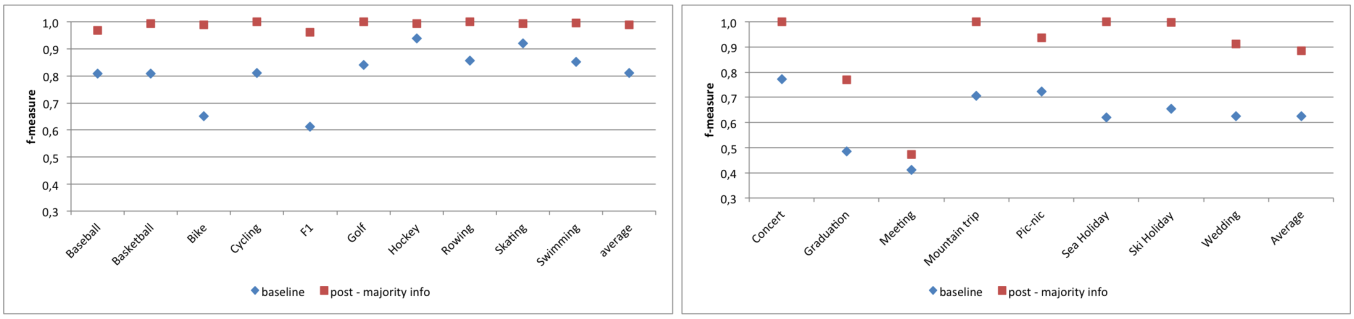
5.2.2 *Results: sport and social events*

The first step of our framework yields an F-measure for Sports and Social events classes of 0.91 and 0.86, respectively. Taking the majority vote yields an F-measure of 1, i.e. we get perfect sport vs social event categorization result.

We now use this perfect categorization result and create a separate set of classifiers for events within the social category and for events within the sports category.

Figure 2a shows the F-measure for the sport classes for the single image classification and for our post-processing as majority voting. First of all, we can see that the post-process greatly outperforms the baseline: the baseline achieves an average F-measure of 0.81 while the post-processing yields 0.98, an improvement of 0.17 and an almost perfect classification. For the baseline, most confusion is between Bike, Cycling and F1 classes, between Hockey and figure Skating classes and between Baseball and Golf, since they share the same environment (e.g. speed circuit, ice-hockey stadium or a field) and mainly only the objects in the scene change (bike vs car or ice-hockey player vs. figure skater). The post-processing method is able to correct these confusions for most of the cases and representative images inside an event-album are correctly classified. For example, while individual images of ice-hockey and figure-skating may be confused, a whole event album contains many characteristic images, such as a close-up of a figure skater during a spin or a goal during a ice-hockey match, both which are almost always recognised. The characteristic images are numerous enough to make majority voting an effective strategy.

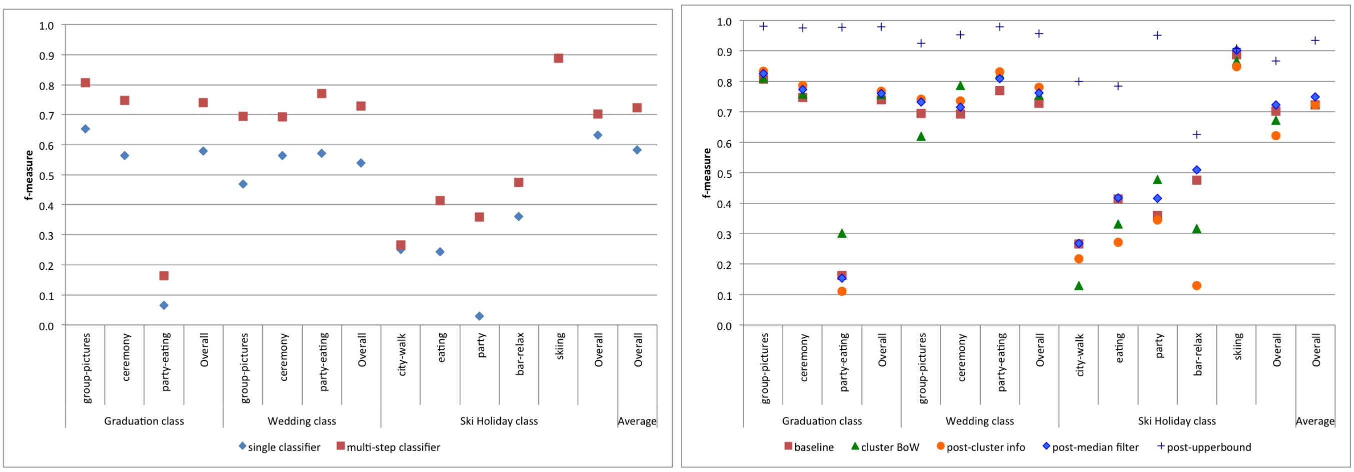
Regarding social event classification, we score an F-measure



(a) Classification of sports event classes.

(b) Classification of social event classes.

Figure 2: Classification of sports and social events.



(a) Single classifier vs multi-step classifier

(b) Different test methodologies.

Figure 3: Classification of sub-events: exploitation of time at event and sub-event level.

of 0.62 as baseline, while 0.89 is achieved with the post-processing majority voting, achieving, also in this category, a considerable improvement of classification performances over the baseline. For the baseline, we observe the confusion of classification to be between the Sea Holiday, Ski-Holiday, Mountain trip and Pic-nic classes and between Graduation and Meeting classes. In the first confusion example, the events shares common activities, as eating in an outdoor environment during a pic-nic or in a mountain trip; in the second example, some activities are shared within the classes, as people standing in a large meeting room. However, the representative images for each class are mainly correctly classified, as the group pictures during a Graduation or the outdoor walking in a Mountain trip. For the post-processing method, the disambiguation between Mountain trip, Sea Holiday and Ski Holiday are correctly solved and a considerable improvement is obtained for Graduation class.

We conclude that by defining events and sub-events and by using majority voting over event-albums, we can obtain considerable improvements over single-image classification.

5.2.3 Results: sub-events

Given an accurate event categorization result, we now use this categorization result for recognizing sub-events in our multi-step classifier. We compare post processing - majority info method the with the direct classification of the single images into the respective sub-event categories. Results are shown in Figure 3a. As can be seen, the use of the post-processing permits to significantly increase the average sub-event classification performances from F=0.58 to F=0.72. We conclude that majority voting is a powerful strategy for boosting the classification results of a single sport or social event-album.

5.3 Exploit time at sub-event level

5.3.1 Experimental Setup

In this experiment we will compare our system for the sub-events classification. Unlike with events, a set of image may span multiple sub-events. We compare the baseline, which is the classification of single images without any notion of event and sub-events, with using a cluster signature and two post-processing strategies. We also provide an upper bound given as output of our clustering algorithm.

5.3.2 Results

We first evaluated the ability of the clustering algorithm to give uniform clusters. Results, named as *post - upperbound*, are shown in Figure 3b and in Table 4. The upperbound for Graduation and Wedding is very high with an F-measure of 0.98 and 0.96 respectively. Performance for the Ski-Holiday class is lower, 0.87. This is because there is high confusion between the bar-relax and skiing sub-events. This is logical as many skiing pictures are taken just before or after a break, when it is convenient to take out one's camera. This means that using only the time-line for clustering is sufficient for

Table 4: F-measure for the different test methodologies

Method	Grad.	Wedd.	SkiH.	Av.ge
baseline	0.7406	0.7284	0.7017	0.7236
cluster BoW	0.7564	0.7541	0.6726	0.7277
post - median filter	0.7598	0.7631	0.7232	0.7487
post - cluster info	0.7671	0.7807	0.6228	0.7235
post - upperbound	0.9788	0.9568	0.8663	0.9340

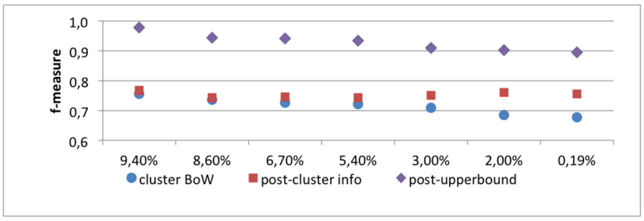


Figure 4: Graduation class: F-measure for different granularity of clusters. the x-axis represents the ratio between the number of cluster and the number of total images.

Graduation and Wedding. For Ski-Holiday, using also the appearance information could be beneficial, which we leave for future work.

The *post processing - median filter* performs better than the *baseline* approach for all classes and it helps by leveraging the sparse classification errors. This method does not have any information of the cluster borders and therefore is prone to make mistakes at the borders of the sub-events. The Graduation has an F-measure of 0.74 for the baseline, while 0.76 is achieved with this method.

The *post processing - cluster info* permits to achieve higher performances than the median filter. The Graduation and Wedding class reach the best performance of 0.77 and 0.78, respectively. Ski-Holiday shows a decrease of accuracy, from 0.70 as baseline to 0.62. This is due to the confusion in the clustering we observed earlier. This method is able not only to handle and correct the sparse classification errors inside the cluster, but also at its borders; provided the clustering is of sufficient quality.

The *cluster BoW* method is better 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 the *post-processing - cluster info* is the best strategy if the quality of the cluster algorithm is sufficient.

5.3.3 Effect of clustering on accuracy

The clustering algorithm we used can give as output different levels of granularity. In Figure 4 different levels of granularity are tested. The x-axis represents the ratio between the number of cluster and the number of total images. The performances increase having a higher number of clusters, reaching the maximum at the level of the finer granularity given by the clustering algorithm. This level of granularity has been used in all our experiments.

6. CONCLUSIONS

In this paper we made the following contributions: a) We introduce a new dataset for event and sub-event recognition. The dataset contains realistic photo-collections collected from Picasa. Each photo-collection spans a single event which is either a sport or social event. Additionally, three social events - Graduation, Wedding, and Ski-Holiday - have additional annotation into sub-events. b) We have shown that categorization into events and sub-events, coupled with time information, can significantly improve image categorization of structured events. Using post-processing on photo-collections results in a perfect sport/social categorization, and improves the F-measure from 0.81 to 0.98 for sport events and from 0.62 to 0.89 for social events. A multi-step classifier improves categorization into sub-events from 0.58 to 0.72. c) We also exploited time information to increase the performance of sub-event recognition. Using a clustering technique coupled with majority voting yielded improvements of 0.03 and 0.05 for the sub-event Graduation and Wedding, respectively. For Ski-Holiday time information alone did not yield adequate clusters. Instead, post-processing using a median filter improved results by 0.02. We conclude that the specification of events and sub-events together with exploiting time constraints can significantly improve event-recognition.

Acknowledgement

This work is developed within the GLOCAL Project, co-funded by the EU Research Framework Programme 7 under DG INFSO Networked Media Systems.

7. REFERENCES

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

[2] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on MCCA*, 2005.

[3] M. Das and A. C. Loui. Event classification in personal image collections. *ICME*, 2009.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. *IEEE International Workshop VSPETS*.

[5] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *ACMIEEECS*, JCDL, 2002.

[6] N. Imran, J. Liu, J. Luo, and M. Shah. Event recognition from photo collections via pagerank. In *Proceedings of the 17th ACM MM*, 2009.

[7] W. Jiang and A. C. Loui. Semantic event detection for consumer photo and video collections. *ICMR*, 2008.

[8] L.-J. Li. What, where and who? Classifying events by scene and object recognition. *ICCV*, 2007.

[9] J.-H. Lim, Q. Tian, and P. Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE MultiMedia*, 10:28–37, October 2003.

[10] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-Time Visual Concept Classification. *IEEE Transactions on Multimedia*, 12(7):665–681, Nov. 2010.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–96, 2010.

[12] U. Westermann and R. Jain. Toward a Common Event Model for Multimedia Applications. *IEEE Multimedia*, 2007.

[13] R. C. F. Wong and C. H. C. Leung. Automatic semantic annotation of real-world web images. *PAMI*, 2008.

[14] J. Wu and J. M. Rehg. Where am I: Place instance and category recognition using spatial PACT. *CVPR*, 2008.

[15] 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*, 2001.