Multimedia Tool and Application manuscript No.
(will be inserted by the editor)

Xueliang Liu · Benoit Huet

On the automatic online collection of training data for visual event modeling

Received: date / Accepted: date

Abstract The last decade has witnessed the development and uprising of social media web services. The use of these shared online media as a source of huge amount of data for research purposes is still a challenging problem. In this paper, a novel framework is proposed to collect training samples from online media data to model the visual appearance of social events automatically. The visual training samples are collected through the analysis of the spatial and temporal context of media data and events. While collecting positive samples can be achieved easily thanks to dedicated event machine-tags, finding the most representative negative samples from the vast amount of irrelevant multimedia documents is a more challenging task. Here, we argue and demonstrate that the most common negative samples, originating from the same location as the event to be modeled, are best suited for the task. A novel ranking approach is devised to automatically select a set of negative samples. Finally the automatically collected samples are used to learn visual event models using Support Vector Machine (SVM). The resulting event models are effective to filter out irrelevant photos and perform with a high accuracy as demonstrated on various social events originating for various categories of events.

Keywords Events, social media, multimedia semantics

1 Introduction

With the popularity of digital capture equipments and the easy use of media sharing web services, many photos and videos taken during public happenings

X Liu, B Huet
EURECOM, France
Tel.:+334-9300-8179
Fax: +334-9300-8200
E-mail: {xueliang.liu,benoit.huet}@eurecom.fr

are uploaded on the Internet and shared by participants. The resulting social media flooding gives rise to new challenge for these websites in terms of data query and management system. How to leverage these user contributed data in research is an open and challenging problem.

Due to the fact that much of the media data shared by users are taken during real world events, such as a wedding, a birthday party, a music concert, or even a holiday trip, it is reasonable to organize these media data into events.

Recently a new field of study concerning how to index the media data by social events has begun to emerge. Some researcher have proposed possible solutions for this problem [18,4]. These works aim at associating media data with events by exploring their rich contextual information, such as time, geographic coordinate, tags. However, it is well known that missing or inaccurate data is a frequent issue in user contributed data, which often limits the application of these methods.

Besides metadata, the main content in social media is the visual content, in the format of photos or videos (audio being only present in videos). State of the art visual concept modeling methods can be used for event based analysis. However, substantial amounts of labeled training data is required for learning the models and creating such a data collection (or ground truth) is a particularly expensive and tedious task.

In this paper, we propose a novel framework for automatically collecting high quality training data from social media and use them to model social events. The training samples are acquired based on the analysis of multimodal contextual information associated with social media and events. In particular, the positive samples are obtained based on specific tags which identify the events accurately, in the form of machine tags and abbreviation of events title. The negative samples are selected using a novel methodology, consisting in the retrieval of media from social media sharing platforms based on location and time, the analysis of media annotations for ranking purposes and the selection of the photos corresponding to the most common tags. Finally, both positive and negative samples are employed to train event models, which are verified against manually labeled ground-truth, exposing the accuracy of the approach and its effectiveness for associating visual media with existing social events. The contributions of this paper are twofold:

- We propose a framework to collect training samples automatically, based on the analysis of metadata associated with social media and events. Experiments performed on a number of events originating from various categories show that using the social media data as a basis for visual content based analysis can be effective.
- Compared with the latest work which builds visual filter on positive samples to prune irrelevant media [18], we take event visual modeling as the real classification problem. The visual properties of each event is learned using SVM whose high classification accuracy has been proved in many past applications and verified again in this paper.

The remainder of this paper is organized as follows: we review the related work in Section 2, and describe the whole framework in Section 3. Experimental results are presented in Section 4. Finally, we summarize the paper and discuss future work in Section 5.

2 Related Work

The study of events has been addressed in the computer vision community for many years [1]. In computer vision, the objective of event related research concerns the essentially the recognition and eventually the localization of special spatial-temporal patterns from a large collection of image sequence or video streams. This is a common yet challenging topic tackled by computer vision/video surveillance scenarios [3] which focus essentially on detecting abnormal or specific behaviors or activities. However, the concept of event addressed in this paper is drastically different compared with these works. Here, we define an event as a real life social happening, involving a group of person and occurring at a specific date (or time) and in a specific location. A live concert held in a club on a given night, an international scientific conference or a carnival (lively and animated street celebration) are among the types of events investigated in the work presented in this paper.

In the past few years, the study of new methods for organizing, searching and browsing media according to real-life events has drawn lots of attention in the multimedia research community. Much work has been done in very different areas. The methods found in the literature addressing this issue cover many multi-modal processing techniques. Therefore, we address the related work from a number of relevant research directions, including: event illustration by media documents; event detection from social media data; multimedia data tags analysis; as well as content based media analysis.

Illustrating events with media data studies the problem of how to leverage vivid visual content to represent events. In [10], the authors proposed a framework to generate photos collections of news to enhance user's experience while reading news articles. They computed the similarity between news text and image tags and obtained the relevant images using text retrieval techniques. In [18], an approach aimed at creating a vivid visual experience to users browsing public events, such as concerts or live shows, was proposed. They studied the user uploading behaviors on Flickr and YouTube, and matched events with medias based on different modalities, such as text/tags, time, and geo-location. The results is an enriched media set which better illustrates the event. In [13], the authors proposed a system to present the media content from live music events, assuming a series of concerts by the same artist such as a world tour. By synchronizing the music clips with audio fingerprint and other metadata, the system gave a novel interface to organize the user-contributed content.

The study of "how to detect events?" has also gained a lot of attention in the past years. The objective of event detection is to discover events out by sensing what is occurring at given location and time. To address the problem, the Social Event Detection Task in the MediaEval Benchmarking Initiative for Multimedia Evaluation focuses on discovering events and detecting media items that are related to either a specific social event or an event-class of interest [22]. A solution of this problems is proposed in [26] that studies how to exploit the social interaction and other similarity between media data to detect events. In addition, Quack *et al.* [23] presented methods to mine events and object from community photo collections. They clustered the photos with multi-modal features and then classified the results into events and objects. A similar problem was also studied in [11] where Firan *et al.* focused on building a Naive Bayes event models which classify photos as either relevant or irrelevant to given events. In [5,6], the authors followed a very similar approach, exploiting the rich "context" associated with social media content and applying clustering algorithms to identify social events.

Tagging is popular on media sharing web sites, and such additional information can be extremely valuable for identifying/representing the content the associated media. However, tags can be very diverse in nature. They might describe the visual content of media but can also refer to emotions, or be personalized for a user (or the media owner himself) with the sole aim of triggering his memory or to attract other users' attention. In [25], the authors took tags as a knowledge source and studied the problem of inferring semantic concepts from associated noisy tags of social images. Some other works are done to improve the tag quality. In [16], Liu *et al.* proposed a social image re-tagging approach that aims to assign better content descriptor to the social images and remove noise description. In [2], Arase *et al.* propose a method to detect people's trip based on their research of geo-tagged photos.

Much of the previous approaches aimed at mining the intrinsic connection between events and media are performed by metadata analysis (i.e. time, location, owner, tags, etc...). Only little work has been done on the analysis the visual content of medias in the context of event, and this is precisely the issue we address with this paper.

The usage of low-level visual features for improving content-based multimedia retrieval systems has made great progress [9]. To address the problem of web visual data analysis, some large scale datasets have been built using multimedia data crawler from shared portals [8]. Beside those web datasets, a number of learning techniques performed on these datasets have shown acceptable results [27,12]. Many works [14,24,15] have been done to study how to automatically or semi-automatically collect online data for training purposes. In [14], Li *et al.* proposed their work on how to train visual concept model by data collected from Internet automatically. The proposed OPTIMOL model employs a Hierarchical Dirichlet process to learn visual concepts and to make the decision rule on new images. An improved work is reported in [24], where the authors employed text, meta-data and visual information in order to achieve better performance. In [15], the authors tackle the problem of collecting negative training samples to model concepts automatically. This objective is somehow similar to the one addressed in this paper. However, their solution exploits the semantics between different visual concepts using related tags. In our work, the objective is to associate media with missing or inaccurate metadata to their corresponding social events. Clearly, the method proposed in [15] cannot be employed to solve our problem, since we cannot define the related and unrelated tags for each event as required by their approach. Our solution leverages the rich contextual information surrounding and defining events to automatically build the collection of on-

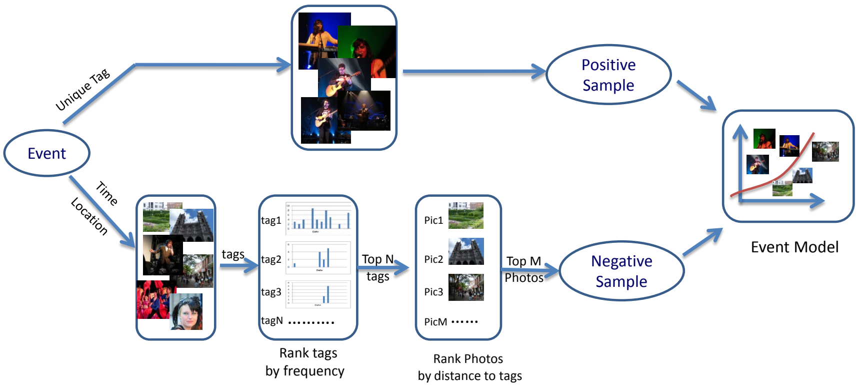


Fig. 1: Overview of the framework for modeling events semantic. The positive samples are collected based on event machine tags (or specific identifiers), and the negative samples are collected using a learning-to-rank approach, to sort the photos according to the common-ness of its tags with respect to the geographical location.

line media samples, using an approach inspired from ranking techniques, and training the classifiers individually for each specific event.

3 Training Samples Collections

We define a social event as the specific happening that takes place at a given location and time and involve several persons (i.e. concerts, conferences, exhibitions, etc...). This work investigates the feasibility of modeling event visually from automatically collected data. To build a visual event model, one needs a collection of images labeled as positive or negative with respect to the event. Unfortunately, labeling data is a labor intensive and time consuming task. In this paper, we propose an original scheme for collecting the training samples for modeling social events visual semantics without any human assistance. Figure 1 depicts the automated steps leading to the creation of the dataset to learn event models. The positive samples are collected directly from social media platforms using identification tag based query. The identification tags are the tags that refer to the event content accurately (i.e. event machine tag).

Collecting the representative negative samples is a more challenging task due to the vast amount of irrelevant data available. Here, negative samples are retrieved from online social media data using metadata analysis. We have observed while experimenting that when querying for photos originating from an event, based on its date and location, the negative samples (those photos which do not correspond to this particular event) are photos depicting general concepts for this location. Among such photos one typically finds, buildings, objects and portraits, etc... and some of the tags associated with these media are common for this location. For example, the city name is a popular tag in many situation yet it does not provide much discriminative information to accurately refer an event. In the work presented here, it is reasonable to assume that these photos captured at the same location as events and containing common tags as the most relevant negative samples for this specific event. Common tags, along with their corresponding photos, are identified based on a novel approach inspired from learning to rank [17], which will be detailed in section 3.2.

3.1 Positive Samples Collection

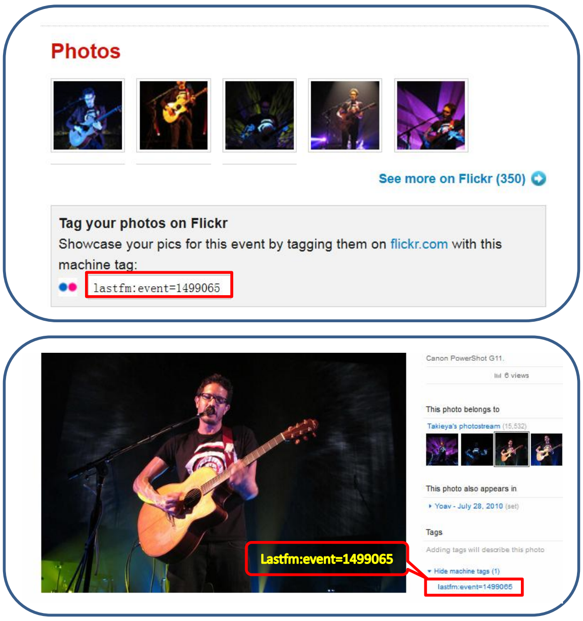
We collect social events visual positive samples by querying social media platforms with event identification tag. There are different kinds of tags to identify events in social media data. The machine tag is a overlap metadata that is available from some events repositories (such as LastFM[1], Upcoming[2] or Facebook[3]) and can be used to refer an event when users upload media data taken during the event. It is popularly used to connect events and photo/video in media sharing platforms, such as Flickr[4]. In these social event websites, machine tags are formatted as "$DOMAIN:event=$XXX", where "$DOMAIN" is the name of website, and "$XXX" is the unique event id provided by the event sites, for example, "lastfm:event=1842684" is an event registered in Last.FM whose id is 1842684, and "facebook:event=108938242471051" is a facebook event whose id is 108938242471051. When users take photos during the event, they can upload them to media sharing websites with such a tag in order to explicitly associate the photos with the event. The machine tags can be recognized by both kinds of web services and give explicit and accurate links between events and multimedia documents. Hence the media documents containing the appropriate machine tag are taken as positive samples for the corresponding event.

Although machine tags are becoming more frequently used, many events still do not feature such metadata. To overcome this issue, we also use the abbreviated events name to identify certain events. The events abbreviations are well known and popularly used among the attendees. For example,"ACMMM10" is short for the ACM International Conference on Multimedia which took place in 2010, without any ambiguity. All photos with such tag are assumed to be positive samples of this social event in the current work.

3.2 Negative Samples Collection

Since social events are characterized by a grouping of people at a given time and place, the most relevant negative samples are those images taken around the same period and location as the event but which do not originate from

[1] http:/www.last.fm
[2] http:/www.upcoming.org
[3] http://www.facebook.com/events/
[4] http:/www.flickr.com



Photos

See more on Flickr (350)

Tag your photos on Flickr

Showcase your pics for this event by tagging them on flickr.com with this machine tag:

lastfm:event=1499065

Lastfm:event=1499065

Canon PowerShot G11.

6 views

This photo belongs to

Takieya's photostream (15,532)

This photo also appears in

Yoav - July 28, 2010 (set)

Tags

Adding tags will describe this photo

Hide machine tags (1)

lastfm:event=1499065

Fig. 2: **Machine Tags Used in Last.fm(Top) and Flickr(Bottom), which provide explicit ground truth on events and media data.**

the event. Here is an example to motivate our assumption. Given an event held in a city near a famous landmark, it is likely that among the photos taken by attendees some will show the landmark. As a famous landmark, it is expected to be captured frequently by tourist. It is important that such photos are included in the negative samples in order to differentiate between the event and its surrounding. Based on this assumption, we collect negative samples with tags referring to the commonest concepts in that location. We measure the commonness of a tag by its frequency over a given period, and our approach to collect negative samples from localized data is composed of three steps.

The first step consists in gathering the photo candidates. For each event, online services such as Last.FM or Facebook/events are used to identify the location and date. These parameters are then employed to query the Flickr API for a photo set (P). The location is defined by a circle, whose center is determined by the GPS coordinates of the event venue and radius value (R). The time interval is the period of D days before and after the event's date. In order to obtain a large set of candidate photos, appropriate values should be set for both D (days) and R (kms). The influence of those two parameters will be studied in the experiment section 4.2.

The second step is to build the text ranking model to identify the "common tags". Here, we define "common tags" as tags that are commonly and frequently associated with a set of photos. In effect, a group of "common tags" represents the most general concepts associated with photos taken in a location. The commonness of a tag can be represented by the fraction of the number of days it appears within a given period. More formally, the commonness of tag t over a time period of D days can be calculated as:

$$Score(t) = \sum_{i=1}^{D} SD(t, i)/D$$

where the value of $SD(t, i)$ is 1 if tag t appears on day i, and 0 if not.

We rank the tags according to their $Score()$ decreasingly. The top N tags (with the largest $Score()$) are kept as the group of common tags $CTags$ for the given period at this location. These tags are prevalently used and highly relevant to the location but do not represent an event due to the fact that they cover a too large time-span. The effect of N, the number of common tags kept to represent the location, is also studied in the experiment section 4.2.

The last step is to select of the negative photo samples based on commonness ranking. For each photo p of P, we extract the title and tags as their text description $Text(p)$, and compute the similarity between these terms and the common tags obtained previously. The measure used here is the cosine distance [20].

$$Similarity(CTags, Text(p)) = \frac{CTags \cdot Text(p)}{\|CTags\| \|Text(p)\|}$$

All of the negative candidates are ranked by their textual similarity to the common tags set ($CTags$) and the top M photos are kept as negative samples for training the visual models.

Having collected both positive and negative visual examples of a particular event, machine learning approaches can be employed to learn the visual model. The methodology used to train the Support Vector Machines used in this work is detailed in 3.3.

3.3 Model Training

The collected data is adapted to training the event-specific models with different visual features and classifiers. Since SIFT feature is an effective feature to represent image content [19,28], we follow this opt for its use for representing the content of the photos. The classifier used in this work is Support Vector Machine, which has been popularly used in different domains [7] and is nowadays prevalently employed for modeling visual content in multimedia indexing and retrieval systems [21]. Each individual event model is obtained as follows; First, 128D Scale Invariant Feature Transform (SIFT) feature is computed over the local region detected by Difference of Gaussian (DoG) filter, then we cluster all the visual feature with K-means for each event, and the SIFT description is quantized to generate 400-dimensional Bag of Visual Words. The event model is learned by Support Vector Machine with Radial Basis Function kernel. Model parameters are optimized using cross-validation method.

4 Experiments

4.1 Data Set and Experiment Setting

Our proposed algorithm is evaluated on different types of events, including 10 concerts from LastFM, 3 scientific conferences and 1 popular street carnival. The photo source used here is Flickr, although other media and sources could be easily added to the framework. The details of each event in the dataset is presented in Table 1.

Table 1: The event dataset used in our experiments includes 10 concerts, 3 international conferences and 1 carnival.

EventID	Title	Date	Latitude	Longitude
lastfm:804783	Metallica	03/03/2009	54.964053	-1.622136
lastfm:1830095	Hole in the Sky Bergen Metal Festival XII	24/08/2011	60.389585	5.323773
lastfm:1858887	Duran Duran	23/04/2011	41.888098	-87.629431
lastfm:1499065	Osheaga en Ville	28/07/2010	45.509788	-73.563446
lastfm:1787326	The Asylum Tour: The Door	03/03/2011	34.062496	-118.348874
lastfm:1351984	Bospop 2010	10/07/2010	50.788893	5.708738
lastfm:1842684	Buskers Bern	11/08/2011	46.947232	7.452345
lastfm:2020655	Lacuna Coil - Darkness Rising Tour	18/11/2011	50.723090	-1.864967
lastfm:1301748	End Of The Road Festival	10/09/2010	50.951341	-2.082616
lastfm:1370837	Into The Great Wide Open	03/09/2010	52.033333	4.433333
ACMMM10	the ACM conference on Multimedia 2010	25/10/2010	43.777846	11.249613
SIGIR2010	ACM Special Interest Group on Information Retrieval,2010	19/07/2010	46.194713	6.140347
ACMMM07	the ACM conference on Multimedia 2007	24/09/2007	48.334790	10.897200
NICECarnival2011	the Carnival de Nice 2011	05/03/2011	43.701530	7.278240

For our experiments, three photo sets are created. The first set contains all the Flickr photos which match the identification tag (EventID) of the 14 selected events. We randomly split the positive photos originating from each event into two equal parts according to usage: 50% for training, 50% for verifying.

The second set contains the negative candidates. Photos that are taken within a given spatial distance (less than R Kms from) and a given temporal interval (less than D days away) of each selected events are retrieved from Flickr. The process of common tags generation and photos ranking is performed on each event photo set in order to retain only the 200 most common photos (which corresponds to the average number of positive training samples) for each event as negative samples for training the model.

The third set of media is called Real Online data (**RO**) and is used to evaluate our approach in a real life situation. The collection is obtained using Flickr queries combining text, location and time as presented in [18]. This collection process is somehow similar to the one anyone would use to gather photos about an event from any user contributed content platform. The irrelevant photos in this dataset can not be filtered just according their metadata. The ground truth on this collection is provided by manual human labeling.

The number of photos for each event of the three sets can be found in Table 2 while some photos samples can be seen in Table 5. Since the data is collected based on a realistic scenario, it is diverse in terms of size and content. Clearly the number of photos for each events ranges from very few to several hundreds, while the photos describe different concepts, such as performers, buildings, sky etc...

Table 2: The number of media collected for the 14 events. Positive samples are collected with unique tags, negative samples are the photos taken near the event location (pre-ranking and selection) and RO data is collected by the methods proposed in [18], and are manually labeled.

EventID	**Positive Samples**	**Negative Candidate**	**RO**	
			Pos	**Neg**
lastfm:804783	441	1063	466	64
lastfm:1830095	716	748	398	134
lastfm:1858887	408	745	431	266
lastfm:1499065	348	712	16	153
lastfm:1787326	446	913	0	313
lastfm:1351984	307	584	498	19
lastfm:1842684	602	1125	535	78
lastfm:2020655	538	745	750	6
lastfm:1301748	944	541	1157	80
lastfm:1370837	592	1025	592	115
ACMMM07	100	557	178	23
SIGIR2010	30	525	0	201
ACMMM10	118	64	15	44
NICECarnival2011	52	848	60	209
Total	5642	10195	5096	1705

We use half the positive samples and the negative samples to train the SVM model for each event, and optimize the parameters D, R and common vocabulary size N using the remaining part of the positive samples.

In our experiments, the results are measured in terms of accuracy, a criteria commonly used for evaluating classification tasks [20]. Accuracy is defined as the number of true predicted elements divided by the total number of elements in the dataset. To be more precise, four values, True Positive(**TP**), True Negative(**TN**), False Positive(**FP**) and False Negative(**FN**) can be used to measure the performance of a classification or recognition system. The terms Positive and Negative refer to the results that are predicted by a system, while True or False refer to whether the prediction is correct with respect to the ground truth. The accuracy measure is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

These measure will be used for comparing the performance of various approaches of this paper.

4.2 Location Distance, Time Interval and Tags Size

We investigate the impact of parameter R, and D, the location distance and time interval between photo taken and event held, to the final event model. We change the two parameters gradually and test the trained model accuracy on the verification dataset. Specifically, R is chosen from 4 to 20 kms with step of 4 kms, and D is set from 5 to 30 days with step 5 days. Cross-validation on the two parameters is performed in the process. Figure 3 shows 3 examples of resulting classification accuracy averaged over the different value of R, and D. Results for all selected events favor the use of rather large parameters for both time interval and location distance. This finding is supported by the fact that the larger the values of D and R, the more photos are retrieved from Flickr and this results in increased diversity within the selected negative samples. Based on the results obtained, the parameters of R and D are set as 20 kms and 30 days respectively.

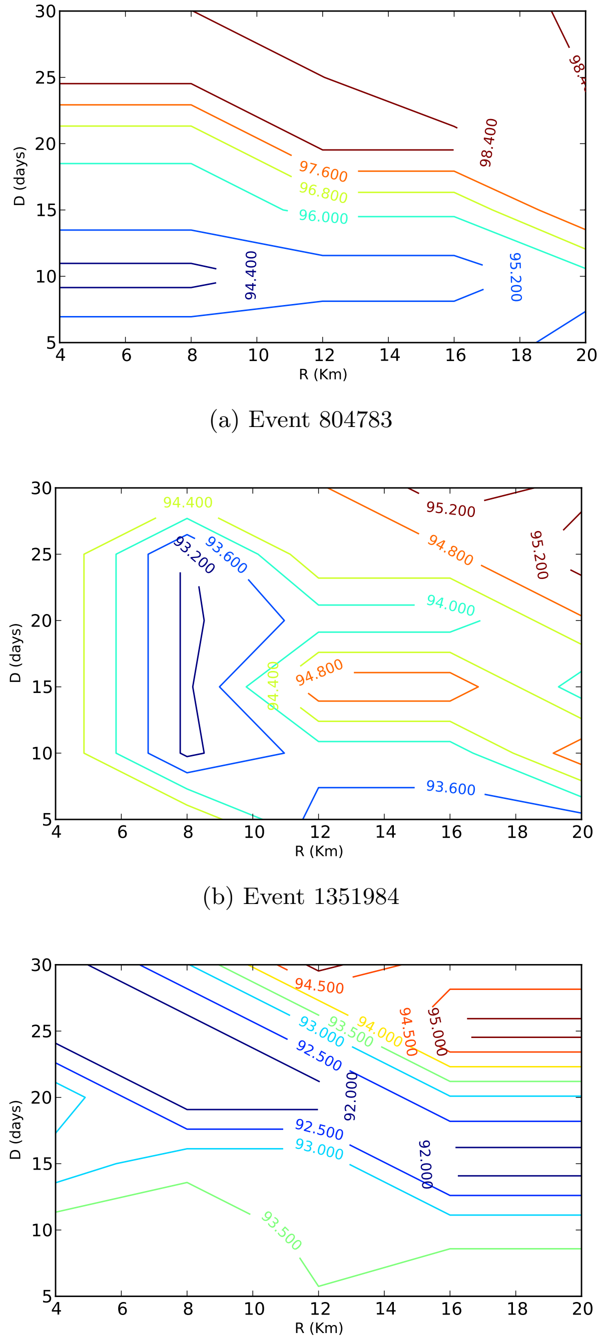
We also evaluate the influence of N, the number of common tags employed, with respect to the resulting event model accuracy. For each combination of parameters R and D, we optimize the model with vocabulary size varying from 5 to 50 tags. The results, presented in Figure 4, clearly indicate that the best performance is obtained when the negative vocabulary contains 10 tags.

4.3 Performance Evaluation

In our experiments, the automatically learned visual event models are compared with four other approaches at the task of mining online media illustrating events and collecting training sample effectively. The first and also the most basic approach, consist in simply running a Flickr query (the one used to create the real online data) and assuming all returned media are positive. In other words, the accuracy value reported in the column **Flickr Query**, indicates the precision in the **RO** test dataset. The second approach reported for comparison is similar to the K-NN visual filter proposed in [18]. In this approach, photos in the test dataset are assigned to the event if and only if their visual similarity with their nearest neighbor is above a high threshold (i.e. 95%). This approach is fast, since it does not require any training nor collection of negative samples. However, the pruning rule is based solely on the analysis of positive samples.

In addition, we compare our approach with two different negative sample collection methods. In the third approach (column **Localization Aware**), rather than ranking photos based on the commonest tags, we use the negative samples randomly selected from the localized negative candidates to train the SVM models. In order to evaluate the influence of “location”, a unique set of 200 negative samples is randomly selected from the entire set of (200 photos * 14 events) negative samples and used to train all SVM models. The results corresponding to this approach are reported in column **Localization Unaware**.

It should be noticed that the values in the column **Flickr Query** shouldn’t be compared with the values in the following 4 columns since it measures “Ac-



(a) Event 804783

(b) Event 1351984

(c) Event 1351984

Fig. 3: Cross Validation on R and D for 3 Events (Performance of classification, measured by accuracy)

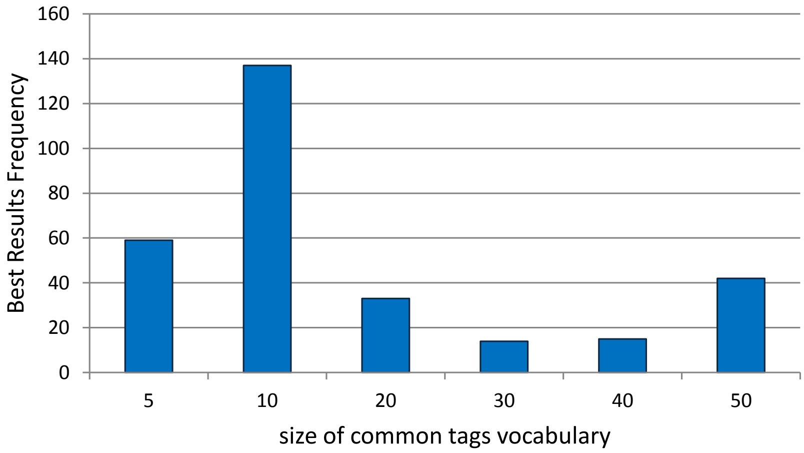


Fig. 4: Performance vs size of common tag vocabulary. The best results are achieved when the 10 most common tags are employed.

curacy" in different context. Nonetheless, it is interesting to bear in mind the ratio between the number of positive samples and negative samples in the RO dataset for each event in order to better interpret the results obtained using of the four classification alternatives.

Table 3: Performance (Accuracy) of alternative classification approaches for associating Media with their corresponding Event

EventID	Flickr Query	Our Algorithm	Pruning in [18]	Localization Aware	Localization Un-aware
lastfm:804783	87.92	88.68	46.98	50.00	75.85
lastfm:1830095	74.81	78.38	80.26	96.62	84.96
lastfm:1858887	61.84	63.41	63.56	76.47	73.89
lastfm:1499065	9.47	90.53	89.94	92.90	89.35
lastfm:1787326	0.00	98.40	92.65	97.12	42.49
lastfm:1351984	96.32	96.32	55.32	86.65	93.81
lastfm:1842684	87.28	87.93	67.86	79.28	87.11
lastfm:2020655	99.21	91.80	71.69	75.00	94.58
lastfm:1301748	93.53	93.53	73.73	64.83	93.21
lastfm:1370837	83.73	85.15	73.83	60.25	80.62
SIGIR2010	0.00	60.19	42.28	16.41	22.38
ACMMM07	25.01	57.62	46.61	28.81	27.18
ACMMM10	85.83	91.04	87.56	86.57	89.05
NICECarnival2011	22.30	76.58	59.10	55.39	56.51
Average	69.41	83.31	68.64	70.07	73.42

From the results presented in table 3, it is interesting to note that the approach proposed in [18] for analyzing visual content using K-NN filtering achieves, on average, almost the same performance as the **Flickr Query**. In other words, such a pruning approach is not very effective at identifying positive and negative illustrations of an event. When compared with the approach in [18], our learned visual model performs significantly and consistently better (83.3% vs 68.6% on average over all 14 events). This result shows the importance of exploiting negative samples to training the events visual content models where the margins between positive and negative samples can be maximized.

Out of the three modeling approaches, our method obtains the best performance with an overall accuracy of 83.31%. Compared with our proposed approach, the models trained using random negative samples expose degraded accuracy (from 83.3% to 70.1%), which shows the importance of carefully selecting the negative samples when building the training collection. The idea of employing the commonest tags to identify nonevent related media proved to be effective. Moreover, the performance of models trained with the uniform negative dataset is better than models trained with random negative event sample, but not as accurate as our approaches. Those results confirm our hypothesis, "location" information plays an important role in negative samples collection and our approach is effective in collecting such negative samples.

In addition, we detail the final statistical results from the four approaches in Table 4. In this table, the results are presented in terms of confusion matrix (**TP**,**TN**,**FP**,**FN**), Precision, Recall, and F1 measure. Clearly, although the Location Unaware method obtains the best **TP**=55.63 and Precision=94.74 ratio, however, it performs worst of all four approaches when dealing with negative sample (**TN**=17.82, **FN**=23.44). Both the K-NN pruning method from [18] and the Location-Aware method fail to correctly classify many positive samples (**FP** scores are 24.09 and 22.29 respectively) . While not achieving the best result in terms of **TP** alone, our proposed approach handles better than any others methods the negative samples, leading to the best performance overall (**F1**=85.58).

Table 4: The detailed classification performance of the four approaches, averaged over all 14 events, measured in terms of confusion matrix and Precision/Recall and F1

	TP	**TN**	**FP**	**FN**	**Precision**	**Recall**	**F1**
Our Algorithm	49.47	33.85	9.84	6.83	83.41	87.87	85.58
Pruning in [18]	35.56	33.07	24.09	7.28	59.61	83.01	69.39
Localization Aware	37.03	33.05	22.29	7.63	62.42	82.92	71.23
Localization Unaware	55.63	17.82	3.1	23.44	94.72	70.36	80.74

Overall, the experiments have clearly shown the value of using visual analysis to model social events content. Furthermore, we have demonstrated that the construction of the event model can be automated without compromising the resulting performance.

5 Conclusion and Future Work

We proposed a novel framework leveraging on the huge number of media documents available on social media website to gather the training data collection necessary to learn social event models. The positive samples are collected using photos with identification tags explicitly referring to the event. The negative samples correspond to those photos taken at the same period and in the vicinity of the event but for which the tags are identified as being common (repeatedly appearing over time). We evaluate the trained visual models on a manually labeled dataset, study the effect of the methodology related parameters and finally report accuracy results of 83% on a real world scenario.

As future work, we currently investigate approaches for collecting additional positive samples with extended coverage of the event while preserving accuracy, so that our system is able to handle more complex situation. The results reported in this paper reinforce the fact that models do not generalize well when the training data is not rich and diverse enough. We can address this problem further by making use of collective intelligence algorithm. For examples, our future work will utilize the "owner" metadata information associated with media originating from the events to enrich the training samples. In social network (social graph), it is possible to identify connected people who attended the same events and took photos at the time of the event, those photos are potential candidates to train the event models. This will further increase the training set size, augment its diversity and lead to even better media classification thanks to more robust event models.

Acknowledgments

The research leading to this paper was partially supported by the project AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) co-funded by the European Commission and the French Research Agency (ANR) in the Ambient Assisted Living (AAL) programme.

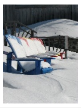
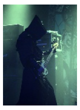
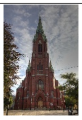
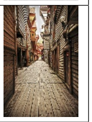
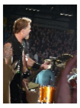
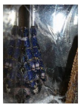
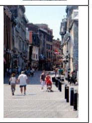
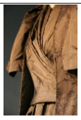
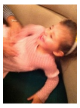
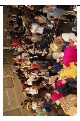
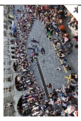
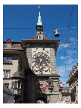
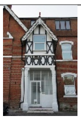
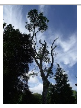
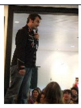
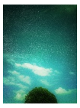
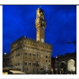
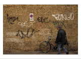
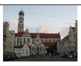
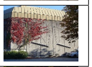
References

1. J. K. Aggarwal and Q. Cai. Human motion analysis: a review. In *IEEE Nonrigid and Articulated Motion Workshop*, pages 90–102, 1997.
2. Y. Arase, X. Xie, T. Hara, and S. Nishio. Mining People's Trips from Large Scale Geo-tagged Photos. In *18th ACM International Conference on Multimedia (ACM MM'10)*, pages 133–142, Firenze, Italy, 2010.
3. L. Ballan, M. Bertini, A. Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, Nov. 2010.
4. H. Becker, D. Iter, M. Naaman, and L. Gravano. Identifying content for planned events across social media sites. In *ACM conference on WSDM*, 2012.
5. H. Becker, M. Naaman, and L. Gravano. Event Identification in Social Media. In *12th International Workshop on the Web and Databases (WebDB'09)*, Providence, USA, 2009.

6. H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. In *3^{rd} ACM International Conference on Web Search and Data Mining (WSDM'10)*, pages 291–300, New York, USA, 2010.
7. C. M. Bishop. Springer, 2006.
8. T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece, 2009.
9. R. Datta, D. Joshi, J. Li, James, and Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40, 2008.
10. D. Delgado, J. Magalhaes, and N. Correia. Automated Illustration of News Stories. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 73–78. IEEE, Sept. 2010.
11. C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Bringing order to your photos: event-driven classification of flickr images based on social. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 189, New York, New York, USA, Oct. 2010. ACM Press.
12. R. Hong, G. Li, L. Nie, J. Tang, and T.-S. Chua. Explore Large Scale Data for Multimedia QA. In *ACM conference on CIVR*, Xi'an, China, 2010.
13. L. Kennedy and M. Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *18^{th} ACM International Conference on World Wide Web (WWW'09)*, pages 311–320, Madrid, Spain, 2009.
14. L.-J. Li and G. Wang. OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 88(2):1–8, 2007.
15. X. Li, C. G. Snoek, M. Worring, and A. W. Smeulders. Social negative bootstrapping for visual categorization. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.
16. D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *18^{th} ACM International Conference on Multimedia (ACM MM'10)*, pages 491–500, Firenze, Italy, 2010.
17. T.-Y. Liu. *Learning to Rank for Information Retrieval.* springer, 2011.
18. X. Liu, R. Troncy, and B. Huet. Finding Media Illustrating Events. In *ACM Conference on ICMR*, 2011.
19. D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004.
20. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 1 edition, July 2008.
21. P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, and A. F. Smeaton. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID.* NIST, USA, 2011.
22. S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *MediaEval'12*, pages –1–1, 2012.
23. T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *ACM conference on CIVR*, page 47, New York, USA, July 2008.
24. F. Schroff, A. Criminisi, and A. Zisserman. Harvesting Image Databases from the Web. In *IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.
25. J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *17^{th} ACM International Conference on Multimedia (ACM MM'09)*, pages 223–232, Beijing, China, 2009.
26. Y. Wang, H. Sundaram, and L. Xie. Social event detection with interaction graph modeling. In *ACM conference on Multimedia*, 2012.

27. Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. GRAPH-BASED SEMI-SUPERVISED LEARNING WITH MULTI-LABEL. *ACM Trans. Program. Lang. Syst.*, 20(5):97–103, 2009.
28. L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. *International Conference on Image processing*, 2(x):721–724, 2001.

Table 5: Event Training and Testing Samples for the 14 events. The negative training sets are collected by the proposed approach automatically. For LastFM event 1787326 and SIGIR10, the Flickr query returned only negative samples, hence no positive media are available for those events in the testing set.

EventID	Training				Testing			
	Positive		Negative		Positive		Negative	
804783								
1830095								
1858887								
1499065								
1787326					None Retrieved	None Retrieved		
1351984								
1842684								
2020655								
1301748								
1370837								
ACMMM10								
ACMMM07								
SIGIR10					None Retrieved	None Retrieved		
NICE carnival								