J. Vis. Commun. Image R. 40 (2016) 546–558

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Event recognition in photo albums using probabilistic graphical models and feature relevance ☆

Siham Bacha [a], Mohand Saïd Allili [b,*], Nadjia Benblidia [a]

[a] LRDSI Laboratory, Saad Dahlab University - Blida1, Blida, Algeria
[b] Department of Computer Science and Engineering, University of Quebec in Outaouais, Gatineau J8X 3X7, QC, Canada

ARTICLE INFO

Article history:
Received 9 March 2016
Revised 25 June 2016
Accepted 24 July 2016
Available online 29 July 2016

Keywords:
Photo albums
Event recognition
Object/scene relevance
Probabilistic graphical models (PGM)

ABSTRACT

This paper proposes a method for event recognition in photo albums which aims at predicting the event categories of groups of photos. We propose a probabilistic graphical model (PGM) for event prediction based on high-level visual features consisting of objects and scenes, which are extracted directly from images. For better discrimination between different event categories, we develop a scheme to integrate feature relevance in our model which yields a more powerful inference when album images exhibit a large number of objects and scenes. It allows also to mitigate the influence of non-informative images usually contained in the albums. The performance of the proposed method is validated using extensive experiments on the recently-proposed PEC dataset containing over 61 000 images. Our method obtained the highest accuracy which outperforms previous work.

1. Introduction

The proliferation of digital cameras has contributed to produce an increasing amount of personal photos with an exponential rate. Therefore, the need for efficient and advanced methodologies regarding personal photo collections management emerges as a challenging and imperative issue. In the last decades, a number of research works have focused on the development of techniques for effective organization of personal photo collections [3]. These works process image visual content to infer high-level semantics as perceived by humans [41]. Researchers have incorporated semantic cues, such as faces [5,32] and person identification [33] to help photo collections management. Moreover, contextual cues such as time-stamps and GPS information have been also used for the same objective [15,21,31,46,49].

In real-world scenarios, people usually take photos that are related to particular events (e.g., *birthdays*, *sport events*, etc.), and the photos are arranged later on into albums. Events can also be considered as an important semantic clue for recalling photos content [60]. Therefore, automatic event recognition in personal photo collections plays an important role for intelligent photo management and advanced retrieval. It is also important for applications such as semantic image indexing and summarization [15,18] and security enforcement [33]. Several methods have been proposed to deal with event recognition on single images or group of photos. For example, methods based on bag-of-features models have been used to predict event categories [10,17]. Recently, features based on deep Convolutional Neural Networks (CNNs) [22]) have been successfully used with classifiers such as neural networks [52] and nearest neighbor [38] for event prediction. In addition to visual features, contextual information (e.g., time-stamps, GPS, etc.) has been also used to enhance event recognition [8,23,39,45,55]. The major limitation of the above methods, however, is that they heavily rely on classifiers based on low-level features. Since these features have no explicit semantic meaning and can be shared by several events, event recognition becomes less efficient and interpretable.

To address this issue, several works propose to use high-level semantically meaningful features for event recognition. To recognize events, [6] use correlation between scene categories (e.g., mountains, urban areas, etc.) and events. Although scene information can provide some good clues about events, it is insufficient to discriminate events sharing the same scene categories. For example, *Wedding* and a *Birthday* events can be associated to the same scene types, but can contain different objects. Therefore, foreground objects are important for event recognition. For instance, a *Hiking* event can be intuitively derived from a 'snowy mountain' scene, whereas a *Wedding* event is usually characterized by the presence of white-dressed 'bride'. To integrate object information, another line of works has been proposed recently for event

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: mohandsaid.allili@uqo.ca (M.S. Allili).

http://dx.doi.org/10.1016/j.jvcir.2016.07.021
1047-3203/

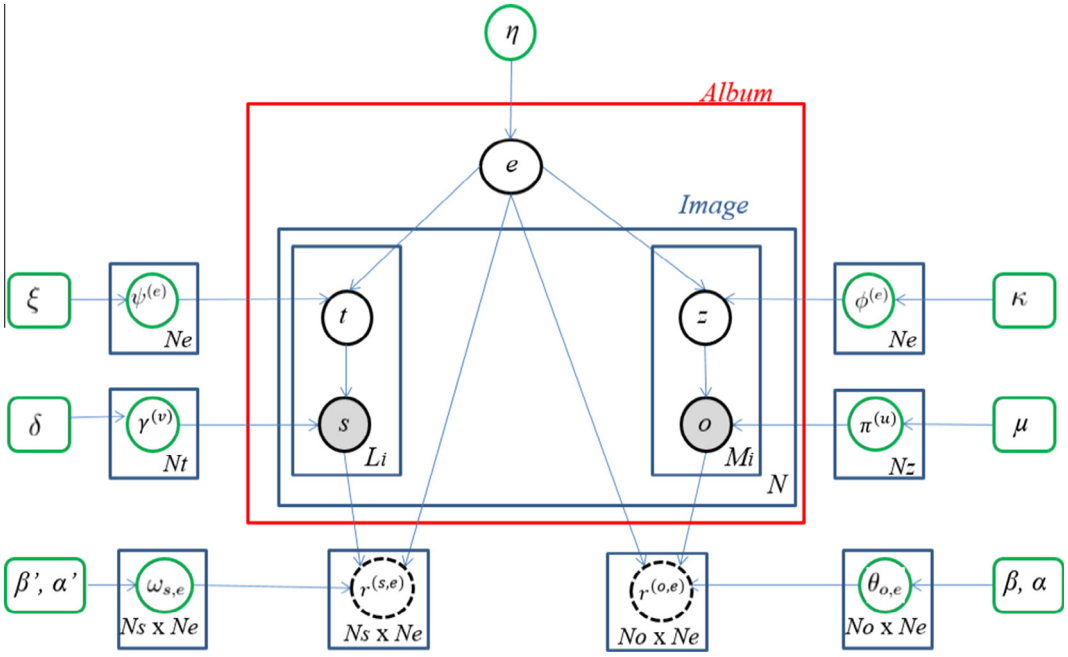

Fig. 1. Graphical representation of our model. Boxes denote replication of the corresponding random variables: there are N images in an album $\mathcal{A}$, with M_i observed objects and L_i observed scenes in image I_i. The variables e, o and s represent the event, object and scene classes, respectively. The latent variables for generating object and scene are z and t, respectively.

recognition. For example, [47] propose a statistical method selecting representative objects for event categories. For each object category, a detector is built and its response used to predict event categories using SVM. Another work from the same authors [48] propose to mine frequent object pattern to determine the most discriminative ones. A photo album is then expressed as frequencies of this discriminative pattern, which they called Compositional Object Pattern Frequency. However, since this work does not use scene information, classes containing similar objects but different scenes can be confused. For instance, *City-tour* and *Motor-show* events can contain the 'car' object but the 'car' appears in different contexts. In other words, it is difficult to generate accurate event semantics based only on object representation. Contextual information is often required to understand the role and dynamics of objects in the events [26].

Although several approaches have been proposed to use either scene categorization or object detection for event recognition, the possibility of jointly using these two tasks to recognize events has not been well investigated. In [26], a statistical model integrating scene and object information for event recognition on single images is proposed. Recently, [50] propose using features derived from CNNs [22] to perform event recognition in groups of images. In the same vein, [28] employ CNNs features to extract scene and object information. Then, event categories in images are predicted using discriminant analysis. It remains, however, that these methods are more adapted on detecting events on single images which provide less rich and complete information about events compared to entire albums. On the one hand, photos constituting albums are usually taken at times-tamps reflecting important moments of the events [3] and, therefore, can be more informative about the events. On the other hand, different photos can give a more exhaustive set of objects/scenes involved in the events.

In this paper, we propose a combined object/scene-based approach for event recognition in personal photo albums. More specifically, a probabilistic graphical model (PGM) is proposed to infer event categories by leveraging high-level information about objects and scenes in sets of images. Given an album of images, we use CNNs first to extract object and scene information in each image of the album which constitute our high-level features for event recognition. We propose a PGM to model relationships between events and object/scene categories. In the same model, we propose to integrate object/scene relevance to boost event recognition. Finally, to infer event categories of new albums, we combine features of their images in a penalized-likelihood function and use the maximum a posteriori probability (MAP) to estimate their event categories.

Contrary to recent work for event recognition in albums [3,8], our approach relies only on visual information, although using time or localization can be included in later versions of this work to improve performance. Moreover, we combine object/scene cues and incorporate their discriminative power (i.e., relevance) for more efficient event recognition. For instance, a 'stage' object may help to determine the event *Concert*, but not the event *skiing*. Similarly, a 'Christmas tree' can tell much about a the event *Christmas* than other objects. Therefore, including object/scene relevance can be a very useful tool to yield more accurate and interpretable event recognition. Our key contributions in this paper can be summarized as follows:

(1) We propose a probabilistic graphical model (PGM) for event recognition in photo albums by combining scene and object cues. Since Rand [35], several works have discussed how event perception occurs in the human vision system (HVS) [4,13,29]. The advent of Bayesian methods has offered a new an attractive tool to explore the potential of modeling visual inference in a closer way to the HVS. Therefore, using PGMs is a promising tool to explore for event recognition on sets of images.
(2) We introduce a feature relevance (FR) scheme incorporating the predictive power of each object/scene for event recognition. Using FR is motivated by several findings of psychological studies in human cognition, where real-world scene understanding at large extent relies on analyzing objects and their contexts [7,12,43,56]. Most often, people can

directly infer their understanding about scene and events after seeing some objects occurring at specific contexts [2,11,30]. This logic is taken to support the consideration of using FR for boosting event recognition.

(3) We propose a MAP approach using Bayesian inference to predict event categories for new albums. We have performed extensive evaluation of the proposed approach on the well-known PEC dataset, where detailed results are presented for event prediction. Obtained results have demonstrated the performance or the proposed approach and has confirmed the potential of Bayesian methods for visual recognition.

The remainder of this paper is organized as follows. Section 2 describes our graphical model for event recognition in photo albums. Section 3 provides details of parameter estimation of our model. Section 4 describes how to label a new albums. Section 5 presents experimental results validating our approach. We end the paper with a conclusion and future work perspectives.

2. A generative model for event recognition in photo albums

The goal of our method is to label albums with event categories based on the analysis of their images in terms of objects and scenes. We argue that event categories in albums can be characterized at a great extent by the objects and scenes composing their images [23,30]. Therefore, we use objects and scenes as building blocks for album classification. For this goal, we propose a generative model for learning and representing relationships and dependencies between events, scenes and objects. Generative models provide a powerful tool to combine several related variables and use Bayesian inference for prediction. In what follows, we provide the details of the proposed graphical model for album generation as well as the probabilistic feature relevance scheme used to enhance discrimination between different event categories.

2.1. Model structure and album generation

The structure of our generative model corresponds to a directed Bayesian network (DBN) and is depicted in Fig. 1. DBNs are suitable for representing causal relationships and dependencies between variables, learning from incomplete data and combining data and domain knowledge. They are also efficient for handling uncertainties originating from incomplete knowledge of dependencies in real-world situations [20]. In the Album box of Fig. 1, gray nodes are observed variables in both training and testing phases. Nodes with no shading are not observed in the testing phase. In the outer box, nodes with dotted lines are calculated variables. Finally, the green nodes and the rounded green boxes are the parameters and hyper-parameters of the model, respectively.

Album generation in our model starts by selecting an event category label, say a *Round-trip* event. In our case, we suppose an album is a set of (unordered) images which are independent from each other given the event category. Moreover, to facilitate our modeling, we assume objects and scenes are independent in each image given the event category. Finally, we suppose that an event category can exhibit multiple object and scene *latent topics* which represent semantic contexts in which the object and the scene will appear in the event, respectively. For example, in a *Round trip* event, we can consider 'scholar road trip' and 'non-scholar road trip' object topics, and 'littoral road trip' and 'mountain road trip' as scene topics. Selecting a 'scholar road trip' object topic will privilege objects that occur frequently in this theme (e.g., scholar bus, etc.). Likewise, selecting a 'littoral road trip' scene topic will favor scenes that occur frequently in this topic (e.g. beach, sea, etc.).

To generate an image album for a particular event category, we first generate scene and object topics from mixtures of available topics. Then, given the generated topics, we generate the object and scene instances that are contained in the image. More formally, let an album $\mathcal{A}$ of a particular event containing a set of N images denoted by $\mathcal{A}=\{I_1,\ldots,I_N\}$, where I_i is the i-th image of the album. To better understand our model, we go through the generative process of each album for the event category.

Let N_e be the number of event categories and let N_o and N_s be the numbers of different object and scene classes, respectively. Moreover, let $e\in\{1,2,\ldots,N_e\}$ be a discrete random variable representing the event category. A category label e is generated according to the distribution $e\sim p(e|\eta)$, where η is N_e-dimensional parameter vector of a Multinoulli distribution. We suppose that an album belongs to only one event category, and therefore, we generate one event e. For each image $I_i\in\mathcal{A}$, contained scene and object instances can be generated as follows.

Suppose a scene occurrence follows a Multinoulli distribution and is represented by an N_s-dimensional vector s, with components $s^{(k)}\in\{0,1\}, k\in\{1,2,\ldots,N_s\}$, and only one component is equal to one. The image I_i can contain L_i scene instances $L_i\in\{1,\ldots,L_{max}\}$, which are represented using a set of random vectors $\mathbf{s}_i=\{s_{il}\}_{l=1,\ldots,L_i}$. A scene instance s_{il} in the i-th album image is generated using the following steps:

(1) Choose a scene *topic* t_{il}, where t_{il} follows a Multinoulli distribution $\mathrm{Mut}(\psi^{(e)})$ with N_t-dimensional parameter vector $\psi^{(e)}$. N_t is the number of topics in the scene latent space and $t_{il}^{(v)}=1$ indicates that the v-th topic is selected. The vector $\psi^{(e)}$ has a Dirichlet prior with hyper-parameter ξ.

(2) Once the scene topic $v\in\{1,\ldots,N_t\}$ is selected, a scene instance s_{il} is generated according to a Multinoulli distribution $\mathrm{Mut}(\gamma^{(v)})$ with N_s-dimensional parameter vector $\gamma^{(v)}$. The variable s_{il} is an N_s-dimensional vector, where $s_{il}^{(k)}=1$ indicates that the scene instance belongs to the k-th scene class. Finally, the parameters $\gamma^{(v)}$ has a Dirichlet prior with hyper-parameter δ.

Similarly to scenes, object occurrence follows a Multinoulli distribution and is represented by an N_o-dimensional vector o, with components $o^{(q)}\in\{0,1\}, q\in\{1,2,\ldots,N_o\}$, and only one component is equal to one. The image I_i can contain M_i object instances $M_i\in\{1,\ldots,M_{max}\}$, which are represented using a set of random vectors $\mathbf{o}_i=\{o_{im}\}_{m=1,\ldots,M_i}$. An object instance o_{im} in the i-th album image is generated using the following steps:

(1) Choose an object *topic* z_{im}, where z_{im} follows a Multinoulli distribution $\mathrm{Mut}(\phi^{(e)})$ with N_z-dimensional parameter vector $\phi^{(e)}$. N_z is the number of topics in the object latent space and $z_{im}^{(u)}=1$ indicates that the u-th topic is selected. The vector $\phi^{(e)}$ has a Dirichlet prior with hyper-parameter κ.

(2) Once the object topic $u\in\{1,\ldots,N_z\}$ is selected, an object instance o_{im} is generated according to a Multinoulli distribution $\mathrm{Mut}(\pi^{(u)})$ with N_s-dimensional parameter vector $\pi^{(u)}$. The variable o_{im} is an N_o-dimensional vector, where $o_{im}^{(q)}=1$ indicates that the object instance belongs to the q-th class. Finally, the parameters $\pi^{(u)}$ has a Dirichlet prior with hyper-parameter μ.

In order to consider object/scene relevance for event recognition, we introduce new variables $r^{(o,e)}$ and $r^{(s,e)}$ to encode the discrimination power of each object o and scene s with regard to each event e. Once all scenes and objects are generated for all

considered events, we can formulate the relevance parameters $r^{(o,e)}$ and $r^{(s,e)}$ as follows:

(1) Let $r^{(o,e)}$ be binary discrete random variable, where $r^{(o,e)}=1$ if the object o is relevant to event category e and $r^{(o,e)}=0$, otherwise. Then, the probability $p(r^{(o,e)}=1|e,o)=\theta_{o,e}$ (i.e., the object o is relevant to the event e) is a Bernoulli distribution $\mathrm{Ber}(\theta_{o,e})$ with parameter $\theta_{o,e}$. The parameter $\theta_{o,e}$ has a Beta prior with hyper-parameters α and β.

(2) Let $r^{(s,e)}$ be binary discrete random variable, where $r^{(s,e)}=1$ if the scene s is relevant to event category e and $r^{(s,e)}=0$, otherwise. Then, the probability $p(r^{(s,e)}=1|e,s)=\omega_{s,e}$ (i.e., the scene s is relevant to the event e) is a Bernoulli distribution $\mathrm{Ber}(\omega_{s,e})$ with parameter $\omega_{s,e}$. The parameter $\omega_{s,e}$ has a Beta prior with hyper-parameters α' and β'.

Putting all the above steps together in a set of model variables $\mathcal{X}=\{e,\mathbf{z},\mathbf{t},\mathbf{o},\mathbf{s},\mathbf{r_o},\mathbf{r_s}\}$, where $e\in\{1,\ldots,N_e\}$, $\mathbf{o}=\{\mathbf{o}_{i=1:N}\}$, $\mathbf{s}=\{\mathbf{s}_{i=1:N}\}$, $\mathbf{z}=\{\mathbf{z}_{i=1:N}\}$, $\mathbf{t}=\{\mathbf{t}_{i=1:N}\}$, $\mathbf{r_o}=\{r_{o,e|o=1:N_o,e=1:N_e}\}$, $\mathbf{r_s}=\{r_{s,e|s=1:N_s,e=1:N_e}\}$, the joint probability of album generation given an event e can be expressed as follows:

$$p\left(\mathcal{X}|\eta,\left\{\phi^{(e)},\psi^{(e)},\pi^{(u)},\gamma^{(v)},\theta_{o,e},\omega_{s,e}\right\}\right)=p(e|\eta)\prod_{i=1}^{N}L_{i,o}\times L_{i,s} \tag{1}$$

where $L_{i,o}$ and $L_{i,o}$ are the likelihoods associated with objects and scenes contained the i-the image. These are given as follows:

$$L_{i,o}=\prod_{m=1}^{M_i}p\left(z_{im}|\phi^{(e)},e\right)p\left(o_{im}|z_{im},\pi^{(u)}\right)p\left(r^{(o,e)}|o_{im},\theta_{o,e},e\right) \tag{2}$$

$$L_{i,s}=\prod_{l=1}^{L_i}p\left(t_{il}|\psi^{(e)},e\right)p\left(s_{il}|t_{il},\gamma^{(v)}\right)p\left(r^{(s,e)}|s_{il},\omega_{s,e},e\right) \tag{3}$$

2.2. Probabilistic feature relevance for event recognition

In this section, we define the criteria to consider a feature, whether a scene or an object, is relevant for the discrimination of event categories. In other words, we go through the learning process to estimate the parameters $\{\theta_{o,e},\omega_{s,e}\}$, respectively. A feature is relevant for an event category if it contributes well to is discrimination. More formally, for each event category $e\in\{1,\ldots,N_e\}$, we define binary variables $r^{(x,e)}$, with $x\in\{o,s\}$ and $o\in\{1,\ldots,N_o\}$ and $s\in\{1,\ldots,N_s\}$. The variable $r^{(x,e)}$ takes value 1 if the feature x is relevant to event category e and takes value 0, otherwise. Therefore, the discrete variable $r^{(x,e)}$ as a Bernoulli distribution.

Feature relevance can be determined by a measure of dependence between a random variable and class category. Among these measures, *mutual information* (MI) quantifies information embodied in a given input feature for predicting a target class variable [34]. In fact, a relevant object/scene can reduce uncertainty and bring knowledge about an event category. For example, viewing 'Christmas tree' or 'ocean liner' objects, respectively, is more informative for the *Christmas* and *Cruise* events than other objects. In our case, given a sample of albums drawn from multiple events, we can compute the MI for each object/scene with respect to each event category, using the following formula:

$$MI(x,e)=\sum_{x\in\{0,1\}}\sum_{a=e,a\neq e}p(x,a)\log\left(\frac{p(x,a)}{p(x)p(a)}\right), \tag{4}$$

where $p(x,a),p(x)$ and $p(a)$ are estimated empirically in a training dataset. When the mutual information between an input feature and the target class e is small, the input feature is not relevant for the target class, regardless of the classification algorithm. By opposite, higher values of MI mean that the input feature is relevant for the target class variable.

Note that to recognize events, people often count on representative/discriminative objects (e.g., 'graduation dress' to identify *graduation* event, etc.), whereas absence of objects is not very informative about the event (i.e, absence of 'boat' does not mean necessarily a *graduation* event). To include this aspect in determining feature relevance for event recognition, we use correlation analysis between objects/scenes and event categories. The correlation coefficient $\rho(x,e)\in[-1,+1]$ measures the strength of this relationship, $\rho(x,e)>0$ (resp. $\rho(x,e)<0$) indicates positive (resp. negative) correlation relationship between an object/scene and an event category. It follows that higher values of $\rho(x,e)$ and MI indicate a strong relevance for the object/scene in predicting the event category. Therefore, we propose the following lenient function to approximate the observation of the variable $r^{(x,e)}$ in a given sample of albums:

$$r^{(x,e)}\approx 1-\exp[-MI(x,e)H(\rho(x,e))], \tag{5}$$

where $H(.)$ is the Heaviside function that considers only cases of positive correlation values to assign higher relevance values. Using the above formula, positive correlation values and high values for MI yield $r^{(x,e)}\to 1$. For negative correlation values and/or low MI values, $r^{(x,e)}\to 0$.

2.3. From images to high-level features

Recognizing events in sets of images relies on concepts with high-level semantics constituted mainly of objects and their context represented by scenes. Therefore, event understanding is intimately linked to two other computer vision problems: object detection and scene recognition. In our work, these two tasks are addressed in two separated phases and their results are combined by our PGM. There is a large body of research work in the areas of object detection and scene recognition [1,2,24,57]. Recently, convolutional neural networks (CNNs) have emerged as an efficient tool and has achieved great success in solving visual recognition problems [36], more particularly for objects [14,27,58] and scenes [59].

There are mainly two types of CNNs applied for object/scene recognition: deep convolutional networks such as AlexNet [22], and the very deep convolutional networks such as GoogLeNet [44] and VGGNet [40]. Deep learning models have been widely used in object recognition and obtained good results on the most challenging datasets [51]. Some recent works have proposed very large networks with deep structures in order to achieve better results for object detection and scene recognition [40]. Motivated by these promising results, Wang et al. [50] have used the GoogLeNet for event recognition in single images. Among others, they produced a good performance on the cultural event recognition dataset in "Looking At People" in a CVPR workshop. Therefore, we have adopted the GoogLeNet architecture to build our object and scene CNNs. The details about the network architecture are provided in its original paper [44].

3. Estimation of model parameters

3.1. Estimation of event, object and scene parameters

In this section, we describe the learning strategy for the parameters $\{\phi^{(e)},\psi^{(e)},\pi^{(u)},\gamma^{(v)},\omega_{s,e},\theta_{o,e}\}$ of our model as depicted in Fig. 1. For convenience, we assume equal prior probabilities for event categories, which is to be a uniform distribution, where $p(e)=1/N_e$. We suppose also that objects and scenes are independent given

the event. Therefore, the object parameters $\{\phi^{(e)}, \pi^{(u)}\}$ and the scene parameters $\{\psi^{(e)}, \gamma^{(v)}\}$ can be learned separately.

Starting by the object branch, the N_z-dimensional parameter vector $\phi^{(e)}$ has Dirichlet distribution prior with the hyper-parameter κ. It governs the distribution of topics z_{im} given an event category e. The Bayesian estimation of the entries of the vector $\phi^{(e)}$ is given as follows [37]:

$$\phi_u^{(e)} = p\left(z_{im}^{(u)} = 1|e\right) = \frac{n_{u,e} + \kappa_u}{\sum_{j=1}^{N_z} n_{j,e} + N_z \times \kappa_u} \tag{6}$$

where $n_{u,e}$ is the number of occurrences of object topic $z_{im}^{(u)}$ in the event category e and κ_u is the u-th entry of the vector κ.

The parameter $\pi^{(u)}$ governs the distribution of an object o given that an object topic $z_{im}^{(u)} = 1$. We first suppose a Dirichlet prior with hyper-parameter μ for the parameter vector $\pi^{(u)}$. The Bayesian estimation for $\pi^{(u)}$ is given as follows:

$$\pi_q^{(u)} = p\left(o_{im}^{(q)} = 1 \middle| z_{im}^{(u)} = 1\right) = \frac{n'_{q,u} + \mu_u}{\sum_{j=1}^{N_o} n'_{j,u} + N_o \times \mu_u}, \tag{7}$$

where $n'_{q,u}$ is the count of occurrences of object class $o_{im}^{(q)}$ given the topic $z_{im}^{(u)}$ and μ_u is the u-th entry of the vector μ.

Similarly to objects, in the scene branch wee have the parameter vector $\psi^{(e)}$ with has a Dirichlet prior with hyper-parameter ξ. It governs the distribution of scene topic t given the event e, which can be computed as:

$$\psi_v^{(e)} = p\left(t_{il}^{(v)} = 1 \middle| e\right) = \frac{m_{v,e} + \xi_v}{\sum_{j=1}^{N_t} m_{j,e} + N_t \times \xi_v} \tag{8}$$

where $m_{v,e}$ is the number of occurrences of the scene topic $t_{il}^{(v)}$ given the event e and ξ_u is the v-th entry of the vector ξ.

The parameter $\gamma^{(v)}$ governs the distribution of scenes s given that a scene topic $t_{il}^{(v)} = 1$. We first suppose a Dirichlet prior with hyper-parameter δ for the parameter vector $\gamma^{(v)}$. The Bayesian estimation for $\gamma^{(v)}$ is given as follows:

$$\gamma_k^{(v)} = p\left(s_{il}^{(k)} = 1 \middle| t_{il}^{(v)} = 1\right) = \frac{m'_{k,v} + \delta_v}{\sum_{j=1}^{N_s} m'_{j,v} + N_s \times \delta_v} \tag{9}$$

where $m'_{k,v}$ is the count of scene type $s_{il}^{(k)}$ when the scene topic $t_{il}^{(v)}$ is realized and δ_v is the v-th entry of the vector δ. Finally, according to [37], the intuitive choice for the entries hyper-parameters $\{\kappa, \mu, \xi, \delta\}$, respectively, is the uniform prior.

3.2. Estimation of relevance parameters

The goal of this phase is to estimate the parameters $\theta_{o,e}$ and $\omega_{s,e}$ of the Bernoulli distributions $Ber(\theta_{o,e})$ and $Ber(\omega_{s,e})$ expressing object and scene relevances with regard to event category e. Without loss of generality, we develop the formulation for object relevance estimation. The same steps can be followed for estimating scene relevance. The Bernoulli distribution for the variable $r^{(o,e)}$ is formulated as follows:

$$p(r^{(o,e)}) = \theta_{o,e}^{r^{(o,e)}} (1 - \theta_{o,e})^{(1 - r^{(o,e)})}. \tag{10}$$

Next, we assume that we have made T samples in multiple datasets, where each sample contains several albums taken in different events. The aim is the estimate through multiple samples the relevance probability of each object and scene with respect to an event category. In other words, to estimate $\theta_{o,e}$, we make several independent draws of the variable $r^{(o,e)}$ according to Eq. (5) and we use Bayesian estimation for the parameter in light of its Beta prior. More formally, let $\mathcal{D} = \left\{r_1^{(o,e)}, \ldots, r_T^{(o,e)}\right\}$ be T *iid* observations for the variable $r^{(o,e)}$. It can readily be shown that the maximum likelihood estimation for $\theta_{o,e}$ can be given as follows:

$$\hat{\theta}_{o,e} = \arg\max_{\{\theta_{o,e}\}} p(\mathcal{D}|\theta_{o,e}) \tag{11}$$

where $p(\mathcal{D}|\theta_{o,e})$ is the likelihood of the parameters given by the following factorization:

$$\begin{aligned} p(\mathcal{D}|\theta_{o,e}) &= \prod_{i=1}^{T} p\left(\left\{r_i^{(o,e)}\right\} \middle| \theta_{o,e}\right) = \prod_{i=1}^{T} \theta_{o,e}^{r_i^{(o,e)}} (1 - \theta_{o,e})^{1 - r_i^{(o,e)}} \\ &= \theta_{o,e}^{N_1} (1 - \theta_{o,e})^{N2}, \end{aligned} \tag{12}$$

where $N_1 = \sum_i^T r_i^{(o,e)}$ and $N_2 = T - N_1$ gives the number of samples where the object is found relevant for the event category. To consider a Bayesian estimation for the parameter $\theta_{o,e}$, we suppose a Beta distribution prior for $\theta_{o,e}$ with hype-parameter α and β. This conjugate prior gives us analytic convenience. The MAP estimation for $\theta_{o,e}$ is then given as follows:

$$\hat{\theta}_{o,e} = \arg\max_{\{\theta_{o,e}\}} [Beta(\alpha + N_1 - 1, \beta + N_2 - 1)] = \frac{\alpha + N_1 - 1}{\alpha + \beta + T - 2} \tag{13}$$

4. Event prediction for new albums

Event recognition from album is considered as an inference problem in the probabilistic graphical model. The aim to estimate the values of target (unobserved) nodes from the values of observed nodes. Given a new album $\mathcal{A} = \{I_1, I_2, \ldots, I_N\}$, the goal is to find its event category e. For this purpose, we compute the maximum posterior probability (MAP) according to each event category $e \in \{1, 2, \ldots, N_e\}$ and we assign the label $\hat{e}$ to $\mathcal{A}$ where:

$$\hat{e} = \arg\max_{e} p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) \tag{14}$$

where $\Theta = (\eta, \left\{\phi^{(e)}, \psi^{(e)}, \pi^{(u)}, \gamma^{(v)}, \omega_{s,e}, \theta_{o,e}\right\})$ denotes the set of parameters used in our PGM. Without loss of generality, we assume the images in $\mathcal{A}$ are independent and that the probability of an event category depends only on the parameters the event class. By writing the album in terms on images, Eq. (14) becomes:

$$\begin{aligned} p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) &= p(e|I_1, I_2, \ldots, I_N, \mathbf{r_s}, \mathbf{r_o}, \Theta) \\ &\propto \prod_{i=1}^{N} p(e|I_i, \mathbf{r_s}, \mathbf{r_o}, \Theta, e) \end{aligned} \tag{15}$$

The basic units of an image I_i are scenes and objects, where $I_i = \{\mathbf{o}_i, \mathbf{s}_i\}$. Thus, we can expand the right term of Eq. (15), and express the probability of an image given its object and scene components:

$$p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) \propto \prod_{i=1}^{N} p(e|\mathbf{o}_i, \mathbf{s}_i, \mathbf{r_s}, \mathbf{r_o}, \Theta) \tag{16}$$

Furthermore, we assume that the objects and scenes of each image are detected separately in independent phases. We assume also that objects $\mathbf{o}_i = \{o_{im}, \text{ where } m \in \{1, \ldots, M_i\}\}$ and scenes $\mathbf{s}_i = \{s_{il}, \text{ where } l \in \{1, \ldots, L_i\}\}$ of an image I_i are generated independently one of each other. Finally, we assume r_s and r_o are independent. Therefore, we obtain:

$$\begin{aligned} p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) &\propto \prod_{i=1}^{N} p(e|\mathbf{o}_i, \mathbf{r_s}, \mathbf{r_o}, \Theta) p(e|\mathbf{s}_i, \mathbf{r_s}, \mathbf{r_o}, \Theta) \\ &= \prod_{i=1}^{N} \prod_{m=1}^{M_i} p(e|o_{im}, r^{(o,e)}, \Theta) \prod_{l=1}^{L_i} p(e|s_{il}, r^{(s,e)}, \Theta) \end{aligned} \tag{17}$$

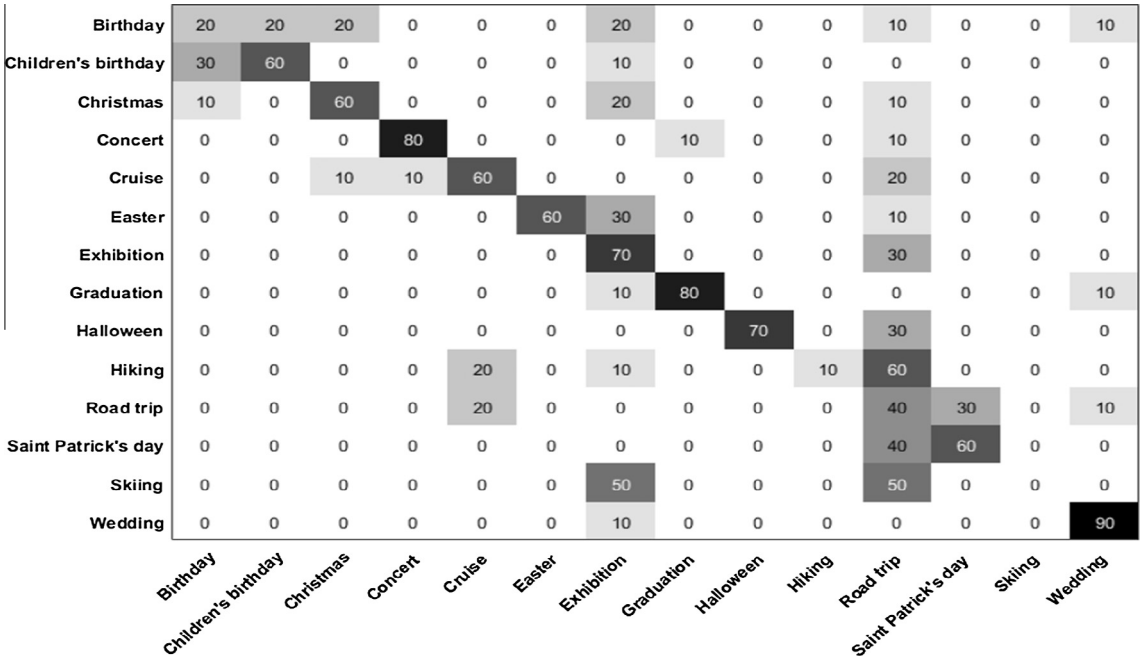

Fig. 2. Confusion matrix for the 14 events using only objects. The overall accuracy is 54.28%. Each column corresponds to ground-truth label of one event class. Each row corresponds to class labels predicted by algorithm. All the numbers are percentage numbers.

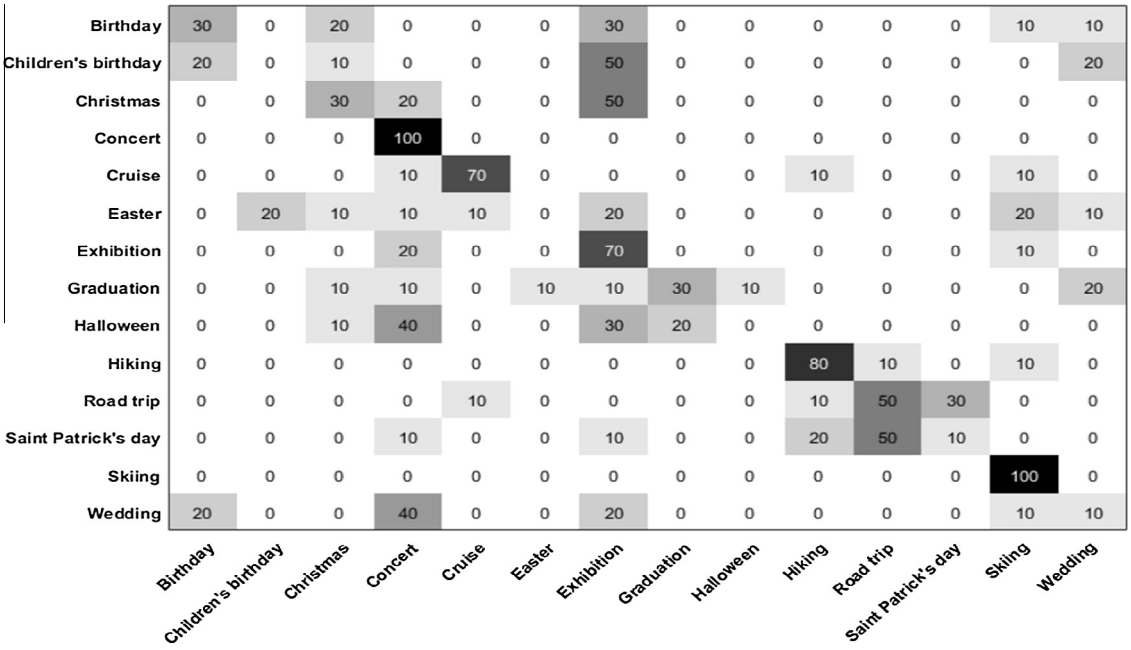

Fig. 3. Confusion matrix for the 14 events using only scenes. The overall accuracy is 41.42%.

Next, note that the first term in Eq. (17), which represent the likelihood of album at the object level, can be marginalized over all object topic values, as follows:

$$\begin{aligned} p(e|o_{im}, r^{(o,e)}, \Theta) &\propto p(o_{im}, r^{(o,e)}, \Theta|e)p(e|\eta) \\ &= p(r^{(o,e)}|o_{im}, e, \Theta)p(o_{im}|e, \Theta)p(e|\eta) \\ &= p(r^{(o,e)}|o_{im}, e, \Theta)\sum_{z_{im}} p(o_{im}|z_{im}, e, \Theta)p(z_{im}|e, \Theta)p(e|\eta) \end{aligned} \tag{18}$$

The expansion of individual terms in Eq. (18) depends on the structure of our model depicted in Fig. 1. The term $p(r^{(o,e)}|o_{im}, e, \Theta)$ is the probability that object o_{im} is relevant to event class e. The probability of co-occurrence of object o_{im} in event e is represented by $p(o_{im}, e, \Theta) = p(o_{im}|e, \Theta)p(e|\eta)$, where $p(e|\eta)$ is the prior probability of event e. Similarly to the first term in Eq. (17), the second term of the same equation can be marginalized over all scene topic values, as follows:

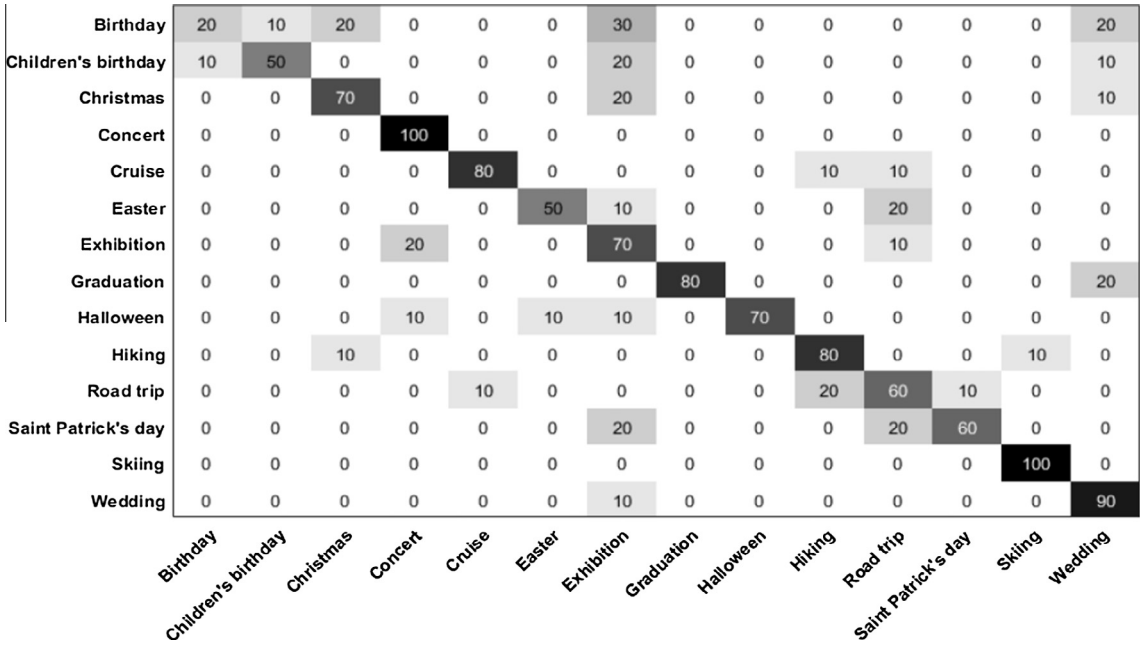

Fig. 4. Confusion matrix for the 14 events by combining scene and object cues but without using relevance feature. The average accuracy is 70%.

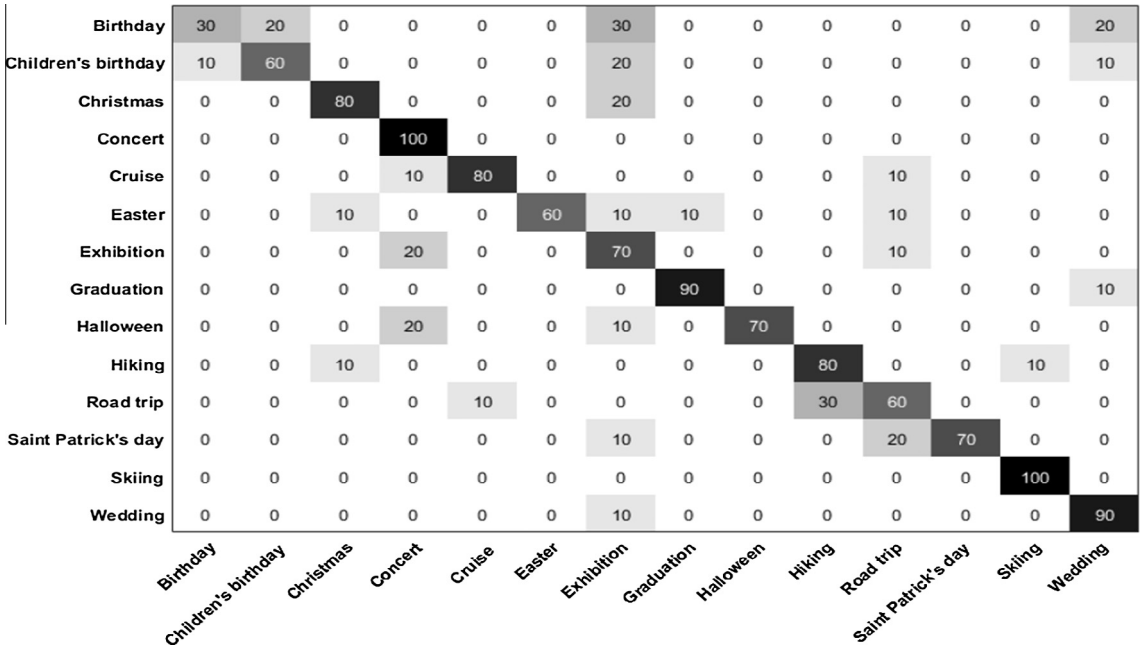

Fig. 5. Confusion matrix for the 14 events by combining scene and object cues and using relevance feature. The average accuracy is 74.29%.

$$\begin{aligned} p\left(e|s_{il}, r^{(s,e)}, \Theta\right) &\propto p\left(s_{il}, r^{(s,e)}, \Theta|e\right)p(e|\eta) \\ &= p\left(r^{(s,e)}|s_{il}, e, \Theta\right)p(s_{il}|e, \Theta)p(e|\eta) \\ &= p\left(r^{(s,e)}|s_{il}, e, \Theta\right)\sum_{t_{il}} p(s_{il}|t_{il}, e, \Theta)p(t_{il}|e, \Theta)p(e|\eta) \end{aligned} \tag{19}$$

where $p(r^{(s,e)}|s_{il}, e, \Theta)$ is the probability that scene s_{il} is relevant to event class e. The probability of co-occurrence of scene s_{il} in event e is represented by $p(s_{il}, e, \Theta) = p(s_{il}|e, \Theta)p(e|\eta)$, where $p(e|\eta)$ is the prior probability of event e.

5. Experimental results

We conducted several experiments for validating the proposed approach. We first describe the datasets used for validation as well as parameter setting for the implementation of our method. We then separately validate the different modules composing our method, namely modules using objects and scenes for event recognition, as well as the impact of using feature relevance. Finally, we evaluate the performance of the proposed approach by comparing it with recently-proposed methods in the literature.

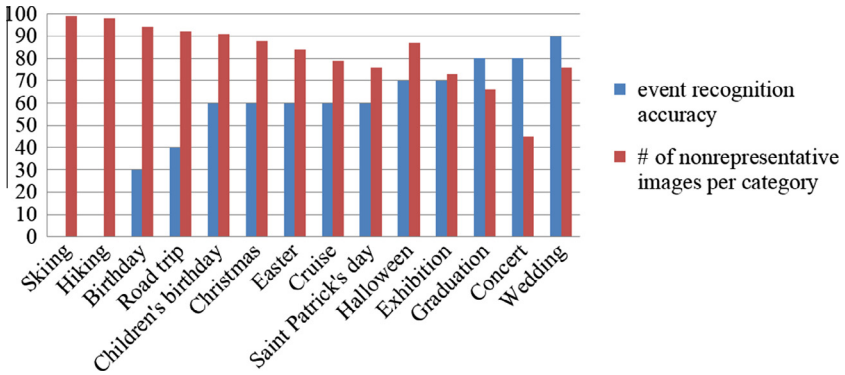

Fig. 6. Influence of the number of nonrepresentative images in albums on the event prediction accuracy. For each event, the first bar represents the percentage of images that do not contain key objects for the event, whereas the second bar represents the prediction accuracy for album events.

Fig. 7. An example from the PEC dataset illustrating the sporadic user-taking behavior in a *Birthday* event.

In the experiments, we evaluate our method on the Personal Event Collections (PEC) available on the Computer Vision laboratory website.[1] Because of the lack of datasets for standard evaluation and comparison for personal photo collections that contain event-related information, Bossard et al. [3] have proposed a Personal Event Collections (PEC) dataset. This dataset contains 807 collections from Flicker composed of 61,000 images that cover 14 event classes of interest, namely: *Birthday, Children's birthday, Christmas, Concert, Cruise, Easter, Exhibition, Graduation, Halloween, Hiking, Road-trip, Saint Patrick's day, Skiing and Wedding*. The annotation of events is defined at the album level.

5.1. Parameters setting

The same experimental protocol suggested in PEC dataset is employed for our evaluations, where 10 albums per class have

[1] https://www.vision.ee.ethz.ch/datasets_extra/pec/.

Fig. 8. Diversity of content inside an *Exhibition* event.

Fig. 9. A misclassified *Wedding* event that does not contain any representative image.

been used for testing (140 albums in total). To learn the parameters of the model, we randomly selected six albums for each event class (84 albums in total) from the proposed training set. In order to build our object net, we have constructed a new object dataset composed of 68% objects taken from the ImageNet dataset, whereas the remaining 32% of objects were added using a supplementary dataset that we have constructed.

We use the Caffe toolbox [19] implementation to detect object/scene instances. We consider GoogLeNet convolutional network architecture for both scene and object nets. The GoogLeNet [44] model pre-trained on the ImageNet dataset [9] is used to initialize the object net. We perform model fine tuning on the new constructed object dataset. During the training phase, all images have been resized to 256×256 pixels. The object net training is done using a mini-batch stochastic descent with momentum value set to 0.9. In order to avoid overfitting problem, a dropout procedure [22] is applied to fully connected layers with *dropout ratio* set to 0.5, and the learning rate is adaptively reduced during the course of training. For scene recognition, we use the GoogLeNet model pre-trained on the Places205 dataset. The database Places205 contains 205 scenes class and 2.5 million images [59]. In the testing

Fig. 10. A misclassified *Christmas* album.

Table 1
Comparison of our method with state-of-art methods using the PEC dataset. Shown numbers are average precision values obtained for each event category. Highest scores for each category are put in bold.

Events	O-PGM	S-PGM	OS-PGM	R-OS-PGM	Bossard et al. [3]	Wu et al. [52]	Tsai et al. [47]	Kwon et al. [23]
Birthday	20%	30%	20%	**30%**	10%	12%	0%	0%
Children's birthday	**60%**	0%	50%	**60%**	30%	57%	**60%**	10%
Christmas	60%	30%	70%	80%	70%	**89%**	60%	40%
Concert	80%	100%	100%	**100%**	**100%**	**100%**	80%	100%
Cruise	60%	70%	80%	80%	50%	**82%**	70%	40%
Easter	**60%**	0%	50%	**60%**	50%	44%	**60%**	20%
Exhibition	70%	70%	70%	70%	70%	**75%**	50%	50%
Graduation	80%	30%	80%	**90%**	40%	69%	70%	40%
Halloween	70%	00%	70%	70%	30%	**82%**	70%	10%
Hiking	10%	**80%**	**80%**	**80%**	**80%**	52%	40%	70%
Road trip	40%	50%	60%	60%	40%	**91%**	10%	30%
Saint Patrick's day	60%	10%	60%	70%	30%	**98%**	40%	90%
Skiing	0%	**100%**	**100%**	**100%**	100%	**100%**	**100%**	60%
Wedding	**90%**	10%	**90%**	**90%**	80%	77%	**90%**	10%
Average accuracy	54.28%	41.42%	70%	**74.28%**	55.71%	73.43%	57.14%	41.71%
Average F_1 measure	55.88%	41.72%	71.01%	**74.82%**	56.16%	57.68%	60.11%	38.62%

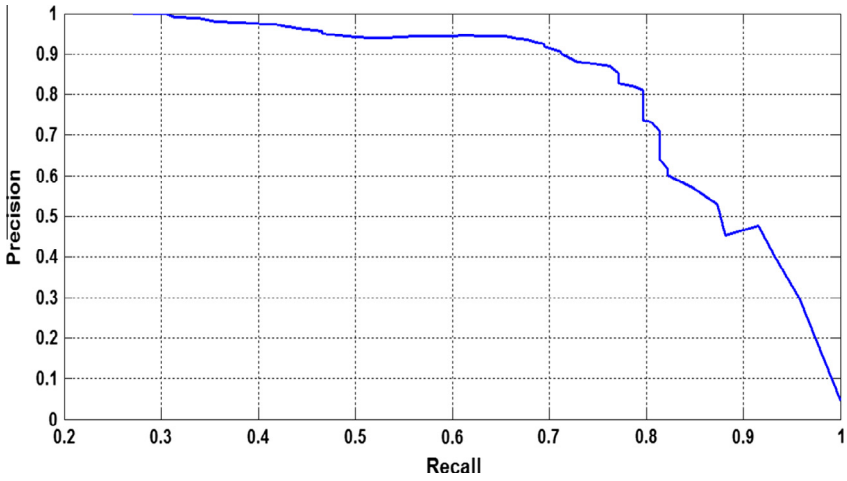

Fig. 11. 'Boat' detection performance in term of *precision* and *recall*.

phase, for both scene and object nets, we resize images to 256×256 pixels before feeding them to the networks.

5.2. Evaluation of the object/scene cues for event recognition

In this section, we evaluate individually the contribution of scene and object cues for event recognition. For this purpose, we implemented two versions of Eq. (17). The first version (O+PGM) contains only information about objects and discards scene terms of Eq. (17). The second version (S+PGM) contains only information about scenes and discards object terms of Eq. (17). We measure the performance of event recognition using confusion matrices. Figs. 2 and 3 present the obtained matrices for the two versions of our model, where columns correspond to event class predictions and rows to ground truth labels, respectively. We can note that O +PGM produce more reliable results (54.28%) than S+PGM (41.42%).

In general, both objects and scenes constitute good cues for event recognition when events can be characterized either by key objects or key scenes. For example, a *Wedding* event is easily recognized when white 'wedding dress' is found in an album. Note, however, that some pairs of event classes can be easily confused, e.g., *Hiking* and *Road trip*, which is likely due to high visual similarity of these events. Moreover, some event misclassifications are due to missed detection of key features. For example, we analyzed the number of missed detections for the object 'boat' in a sample of 403 images taken from PEC albums in the *Cruise* event. Fig. 11 shows the precision-recall curve for 'boat' detection. The best obtained *precision*, *recall* and *F-measure* values are 0.72, 0.78 and 0.75, respectively. We noticed that the missed detections are mainly due to our representation based on GoogLeNet. We believe that using better CNN models, or combing multiple independently CNNs as in [50], can lead to drastic improvements.

Note that there are a few very difficult classes, such as *Birthday*, where most of their album photos do not contain relevant objects/scenes. Our system has failed to categorize correctly these albums, which are even difficult for humans. Fig. 6 shows the influence of the number of *representative* images on the accuracy of labels prediction. An image is *representative* if it contains key scene(s) and/or key object(s) that characterize specific event classes. Overall, the classification results increases gradually with the number of representative images. Note that generally the lack of representative photos is a direct consequence of the user-taking behavior. Sometimes, photo acquisition is sporadic when the user is not focused on key objects/scenes. Fig. 7 shows an example of this sporadic behavior for a *Birthday* event with photos not informative enough for discriminating the event. The final problem consists in the variability and cultural diversity of some classes content. Two

Fig. 12. Misclassified *Halloween* albums. Each row shows some images from different *Halloween* albums.

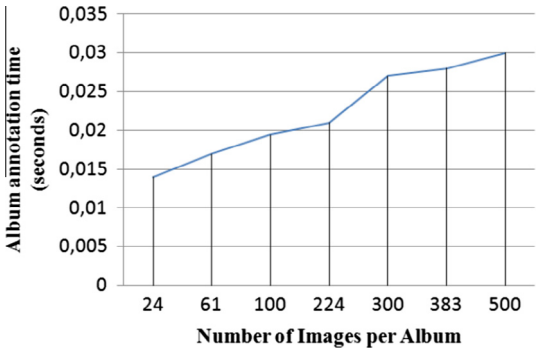

Fig. 13. Event prediction time as a function of album size.

examples of this case are shown in Figs. 8 and 9 which represent an *Exhibition* event with very diversified content and a *Japanese wedding* event with bride and groom not wearing a 'white wedding dress' and 'suit', respectively.

5.3. Evaluation of object/scene relevance for event recognition

To validate the advantage of using feature relevance for event recognition, we further implemented two other versions of our model. The first version, OS+PGM use combination of objects and scenes whiteout using feature relevance. The second version, R +OS+PGM reflects the complete model as proposed in Fig. 1. Figs.4 and 5 show the obtained confusion matrices for the 14 events using the two versions of our model, respectively.

We can note that even without using feature relevance, the majority of event classes have been correctly labeled by OS+PGM, where the obtained average accuracy is 70%. Some event classes have achieved very high accuracies, e.g, *Concert* and *Skiing*. This is likely results from: (1) using excellent features that can characterize specific classes, such as 'snowy mountain' scenes for *Skiing*, (2) the important number of *representative* images in the album, which is the case in *Concert* albums. The combination of objects and scenes helps to distinguish between events with similar visual content and produces more reliable results compared to the use of scenes and objects separately. For example, in the *Graduation* event, objects greatly improve classification accuracy. Indeed, the key feature for the *Graduation* event is the presence of 'mortarboard' and 'academic gown', regardless of the scene of the event (indoor, outdoor or their combination). In the *Cruise* event, scenes help to improve event categorization performance. In this event, people tend to spend more time inside the 'boat', and most photos contain 'sea' and 'coast' scenes.

The introduction of feature relevance in R+OS+PGM has improved the results of object and scene combination by 4.28%. This can be observed especially when the accuracy of some classes has slightly dropped when combining scene and object cues. For example, in *Children's birthday* and *Easter* events, there is a difference between the obtained accuracies using individual cues and their combination. Feature relevance has alleviated the problem by enhancing the contribution of key features and decreasing the contribution of noisy ones, which allowed for better album classification. This also has a good effect in enhancing classification accuracy when the frequencies of key scenes and/or objects are low.

Not surprisingly, when good event categorization can not be achieved using either scenes or objects, combination of both features can not achieve correct results. Furthermore, we have noticed that despite a feature is shared by multiple event classes, it can help to select a subset of possible target classes from the 14 initial ones. For example, the object 'wine bottle' can be present in multiple events such as *Christmas*, *Birthday* and *Wedding*, and absent in *Children's birthday*. The object 'wine bottle', therefore, shrinks the set of predicted target events (containing the object) by enhancing their maximum likelihood.

5.4. Comparison of our method with previous works

To investigate more thoroughly the effectiveness of the proposed method, we compared it to the most recent methods proposed in the literature [3,23,47,52]. Note that [52] have

presented results using different configurations of their method. To have a fair comparison, we used the best results obtained by these configurations. In [23], authors have reported that the combination of multiple features achieves the best results. Therefore, to compare with our work, we have implemented their method combining spatial pyramid matching (SPM) [25] using SIFT+Color features and Regularized Max Pooling (RMP) [16] using CNN features. The vocabulary size used for both SIFT and Color codebooks is 2000. For CNN features extraction, we used Caffe implementation with publicly available trained CNN described by Krizhevsky et al. [22]. This gives a vector of 4096 dimensions. Both SPM and RMP are based on Least-Squares Support Vector Machines (LSSVM [42]). The evaluation has been carried on PEC dataset and obtained results are reported in Table 1.

In terms of classification accuracy, our method achieves an average of 74.29%, exceeding the best average accuracies obtained by [52,23,47,3] by 0.85%, 32.57%, 17.14% and 18.57%, respectively. More specifically, our method outperforms the others in the events *Birthday*, *Children's birthday*, *Easter*, *Graduation* and *Wedding*. For other event categories, we have achieved a close performance to the compared methods. In terms of the F_1 score, we have obtained an average score of 74.82%, exceeding the best F_1 scores obtained by [52,23,47,3] by 17.17%, 36.2%, 14.71% and 18.66%, respectively. This can be explained by the fact that the F_1 score is a compromise measure between the *precision* and *recall*, whereas the accuracy is calculated only from *recall* information. In other words, the obtained *precision* for our method is generally higher than those of the compared ones. This confirms, among other things, that integration of feature relevance and the combination of scene and object cues is more suitable for event classification.

From the results obtained by [23], it is clear that the representation based on low-level features is not sufficient to achieve good event recognition in personal photo collections. In this work, using SPM with SIFT+Color features produces descriptors of $42K$ dimensions, where $K = 2000$ represents the number of histogram bins. By adding the RMP with CNN features, the final features vectors will have 88096dimensions. However, despite the large size of features, the method has limited discriminative power for recognizing event categories with various visual content. Our O-PGM model and [47] have achieved a close classification performance except for the event *Skiing*. This is due to the fact that the detection scores of all objects of the *Skiing* event, such as 'snowmobile' and 'ski mask', were under the confidence threshold. Therefore, no object was considered for this class, which leads to 0% accuracy using our object probabilistic model. However, all images with an empty object vector are classified in the *Skiing* event using SVM, which yields to 100% accuracy, but gives also a bad F_1 score (34, 48% for *Skiing* category).

Finally, to reduce the number of used images, [52] propose to pick randomly from each album a restricted number of images. However, the random selection does not guarantee the presence of representative photos among the chosen ones. Moreover, random sampling can lead to over-representation of albums due to redundancy of visual content between images. The introduction of feature relevance allows to strengthen the contribution of discriminative features and, therefore, reduce the effect of redundant and non-representative images. Furthermore, our method yields more semantically interpretable since relevant objects/scenes can be used to characterize event information about albums. This can be very useful for application such as automatic album annotation and retrieval.

5.5. Computational time

We evaluated the computational time of our method. Since the albums are represented using scenes and objects, we do not consider the time of object detection and scene recognition phases because they depends on the Caffe implementation [19]. We use the 140 albums suggested for test in PEC dataset to obtain the processing time. Our code has been written in Matlab, and it was run on an Intel I7 Core and used 8 GB of RAM. Our model has a fast inference time with an average of 0.5 s for the 140 albums. It is clear that the average processing time per album depends on the number of contained images. The computational cost of the proposed method with album with number of photo from 24 to 500 is shown in Fig. 13. We can note that the computational cost is approximately linear with regard to the number of images per album. Finally, to speed-up computation, object/scene extraction in images can be performed in parallel using multiple-core platforms [53,54].

6. Conclusion and discussion

We have introduced a probabilistic graphical model for album classification in social event categories. Our method takes advantage of recent developments on object/scene recognition in images to build high-level features for event recognition in photo albums. Moreover, we introduced a probabilistic feature relevance scheme that tunes the contribution of features according to their power of event discrimination. This allows to boost event recognition accuracy and obtain more interpretable results. Experiments on the challenging PEC dataset have showed that the proposed model outperforms recent state-of-art methods for event recognition on sets of images.

Despite the obtained performance, there are several ways of improvement for the proposed method. Future extensions of this work will investigate the integration of contextual information for predicting difficult events characterized by the variability of their content. In Fig. 12, we show an example of a misclassified *Halloween* event, where it is difficult to find for this event category a common representation based on objects and scenes. Finally, event recognition in album images will be certainly influenced by advances in object/scene recognition in images. In fact, we have observed in our current implementation that missed object detections of key attributes increases album misclassification (see Fig. 10 for example). We have analyzed, for example, the number of missed detections for the object 'boat' in a sample of 403 images taken from albums in the *Cruise* event. Fig. 11 shows the precision-recall curve for 'boat' detection. The highest *precision*, *recall* and *F-measure* values obtained are 0.72, 0.78 and 0.75, respectively. The missed detections are mainly due to our representation based on GoogLeNet. Using better CNN models, or by combing multiple independent CNNs, as proposed in [50], can lead to improvements of our method.

Acknowledgements

This work has been achieved thanks to the support of the University Saad Dahlab of Blida1 (Algeria) and partially support of Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] M.S. Allili, Wavelet modeling using finite mixtures of generalized gaussian distributions: application to texture discrimination and retrieval, IEEE Trans. Image Process. 21 (4) (2012) 1452–1464.
[2] M.S. Allili, D. Ziou, Object of interest segmentation and tracking using feature selection and active contours, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[3] L. Bossard, M. Guillaumin, L. Van Gool, Event recognition in photo collections with a stopwatch HMM, in: IEEE Int'l Conference on Computer Vision, 2013, pp. 1193–1200.

[4] J.S. Bowers, C.J. Davis, Bayesian just-so stories in psychology and neuroscience, Psychol. Bull. 138 (2012) 389–414.
[5] M. Brenner, E. Izquierdo, Joint people recognition across photo collections using sparse Markov random fields, in: MultiMedia Modeling, Lecture Notes in Computer Science, vol. 8325, Springer, 2014, pp. 340–352.
[6] L. Cao, J. Luo, H. Kautz, T.S. Huang, Image annotation within the context of personal photo collections using hierarchical event and scene models, IEEE Trans. Multimedia 11 (2) (2009) 208–219.
[7] H. Cohen, C. Lefebvre, Handbook of Categorization in Cognitive Science, Elsevier Science Ltd, 2005.
[8] M.-S. Dao, D.-T. Dang-Nguyen, F.B. De Natale, Robust event discovery from photo collections using signature image bases, Multimedia Tools Appl. 70 (1) (2014) 25–53.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[10] L. Duan, D. Xu, S.-F. Chang, Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1338–1345.
[11] M. Dyck, M.B. Brodeur, ERP evidence for the influence of scene context on the recognition of ambiguous and unambiguous objects, Neuropsychologia 72 (2015) 43–51.
[12] C. Galleguillos, S. Belongie, Context-based object categorization: a critical survey, Comput. Vis. Image Underst. 114 (6) (2010) 712–722.
[13] W.S. Geisler, R.L. Diehl, A Bayesian approach to the evolution of perceptual and cognitive systems, Cogn. Sci. 27 (3) (2003) 3797402.
[14] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 142–158.
[15] A. Graham, H. Garcia-Molina, A. Paepcke, T. Winograd, Time as essence for photo browsing through personal digital libraries, in: ACM/IEEE Joint Conference on Digital Libraries, 2002, pp. 326–335.
[16] M. Hoai, Regularized max pooling for image categorization, in: British Machine Vision Conference, 2014, pp. 1–12.
[17] N. Imran, J. Liu, J. Luo, M. Shah, Event recognition from photo collections via pagerank, in: ACM Int'l Conference on Multimedia, 2009, pp. 621–624.
[18] W. Jiang, A.C. Loui, Semantic event detection for consumer photo and video collections, in: IEEE Int'l Conference on Multimedia and Expo, 2008, pp. 313–316.
[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: ACM Int'l Conference on Multimedia, 2014, pp. 675–678.
[20] D. Kasper, G. Weidl, Thao Dang, G. Breuel, A. Tamke, W. Rosenstiel, Object-oriented Bayesian networks for detection of lane change maneuvers, in: IEEE Intelligent Vehicles Symposium, 2011, pp. 673–678.
[21] S. Kisilevich, D. Keim, N. Andrienko, G. Andrienko, Towards acquisition of semantics of places and events by multi-perspective analysis of geo-tagged photo collections, in: Geospatial Visualisation, Lecture Notes in Geoinformation and Cartography, Springer, 2013, pp. 211–233.
[22] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Neural Information Processing Systems (2012) 1106–1114.
[23] H. Kwon, K. Yun, M. Hoai, D. Samaras, Recognizing cultural events in images: a study of image categorization models, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2015, pp. 51–57.
[24] G. Larivière, M.S. Allili, A learning probabilistic approach for object segmentation, in: IEEE Canadian Conf. on COmputer and Robot Vision, 2012, pp. 86–93.
[25] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags-of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2006, pp. 2168–2178.
[26] L.J. Li, L. Fei-Fei, What, where and who? Classifying events by scene and object recognition, in: IEEE Int'l Conf. on Computer Vision, 2007, pp. 1–8.
[27] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 3367–3375.
[28] M. Liu, X. Liu, Y. Li, X. Chen, A.G. Hauptmann, S. Shan, Exploiting feature hierarchies with convolutional neural networks for cultural event recognition, in: IEEE Int'l Conf. on Computer Vision Workshop (ICCVW), 2015, pp. 274–279.
[29] W.J. Ma, Organizing probabilistic models of perception, Trends Cogn. Sci. 16 (10) (2012) 511–518.
[30] L. Mudrik, D. Lamy, L.Y. Deouell, ERP evidence for context congruity effects during simultaneous object-scene processing, Neuropsychologia 48 (2) (2010) 507–517.
[31] M. Naaman, Y.J. Song, A. Paepcke, H. Garcia-Molina, Automatic organization for digital photographs with geographic coordinates, in: ACM/IEEE Joint Conference on Digital Libraries, 2004, pp. 53–62.
[32] W.W.Y. Ng, T.M. Zheng, P.P.K. Chan, D.S. Yeung, Social relationship discovery and face annotation in personal photo collection, in: Int'l Conf. on Machine Learning and Cybernetics, 2011, pp. 631–637.
[33] N. O'Hare, A.F. Smeaton, Context-aware person identification in personal photo collections, IEEE Trans. Multimedia 11 (2) (2009) 220–228.
[34] L. Paninski, Estimation of entropy and mutual information, Neural Comput. 15 (2003) 1191171253.
[35] B. Rand, The Classical Psychologists: Selections Illustrating Psychology From Anaxagoras to Wundt, Houghton, Mifflin and Company, Boston, MA, US, 1912.
[36] J.S.J. Ren, L. Xu, On vectorization of deep convolutional neural networks for vision tasks, in: AAAI Conf. on Artificial Intelligence, 2015, pp. 1–8.
[37] C.P. Robert, The Bayesian Choice, Springer, 2001.
[38] R. Rothe, R. Timofte, L. Van Gool, DLDR: deep linear discriminative retrieval for cultural event classification from a single image, in: IEEE Int'l Conf. on Computer Vision Workshop (ICCVW), 2015, pp. 295–302.
[39] A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, X.G. Nieto, Cultural event recognition with visual convnets and temporal models, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2015, pp. 36–44.
[40] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Int'l Conf. on Learning Representations, 2015, pp. 1–14.
[41] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1349–1380.
[42] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.
[43] K.M. Swallow, J.M. Zacks, R.A. Abrams, Event boundaries in perception affect memory encoding and updating, J. Exp. Psychol. 138 (2) (2009) 236.
[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CoRR, 2014. abs/1409.4842.
[45] F. Tang, D.R. Tretter, C. Willis, Event classification for personal photo collections, in: IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, 2011, pp. 877–880.
[46] I. Tankoyeu, J. Paniagua, J. Stöttinger, F. Giunchiglia, Event detection and scene attraction by very simple contextual cues, in: Joint ACM Workshop on Modeling and Representing Events, 2011, pp. 1–6.
[47] S.F. Tsai, T.S. Huang, F. Tang, Album-based object-centric event recognition, in: IEEE Int'l Conf. on Multimedia and Expo, 2011, pp. 1–6.
[48] S.F. Tsai, L. Cao, F. Tang, T.-S. Huang, Compositional object pattern: a new model for album event recognition, in: ACM Int'l Conf. on Multimedia, 2011, pp. 1361–1364.
[49] A. Ulges, M. Worring, T. Breuel, Learning visual contexts for image annotation from flickr groups, IEEE Trans. Multimedia 13 (2) (2011) 330–341.
[50] L. Wang, Z. Wang, W. Du, Y. Qiao, Object-scene convolutional neural networks for event recognition in images, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2015, pp. 30–35.
[51] Z. Wang, L. Li, H. Qingming, Cross-media topic detection with refined CNN based image-dominant topic model, in: ACM Int'l Conf. on Multimedia, 2015, pp. 1171–1174.
[52] Z. Wu, Y. Huang, L. Wang, Learning representative deep features for image set analysis, IEEE Trans. Multimedia 17 (11) (2015) 1960–1968.
[53] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, Efficient parallel framework for HEVC motion estimation on many-core processors, IEEE Trans. Circ. Syst. Video Technol. 24 (12) (2014) 2077–2089.
[54] C. Yan, Y. Zhang, J. Xu, F. Dai, L. Li, Q. Dai, F. Wu, A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors, IEEE Signal Process. Lett. 21 (5) (2014) 573–576.
[55] J. Yuan, J. Luo, Y. Wu, Mining compositional features from GPS and visual cues for event recognition in photo collections, IEEE Trans. Multimedia 12 (7) (2010) 705–716.
[56] J.M. Zacks, N.K. Speer, K.M. Swallow, T.S. Braver, J.R. Reynolds, Event perception: a mind-brain perspective, Psychol. Bull. 133 (2) (2007) 273–293.
[57] X. Zhang, Y.-H. Yang, Z. Han, H. Wang, C. Gao, Object class detection: a survey, ACM Comput. Surv. 46 (1) (2013) 53 Artcile 10.
[58] Y. Zhang, K. Sohn, R. Villegas, G. Pan, H. Lee, Improving object detection with deep convolutional networks via bayesian optimization and structured prediction, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 249–258.
[59] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Neural Information Processing Systems (NIPS), 2014, pp. 487–495.
[60] C. Zigkolis, S. Papadopoulos, G. Filippou, Y. Kompatsiaris, A. Vakali, Collaborative event annotation in tagged photo collections, Multimedia Tools Appl. 70 (1) (2014) 89–118.