https://doi.org/10.2352/ISSN.2470-1173.2018.2.VIPC-173

Event Recognition in Personal Photo Collections: An Active Learning Approach

Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci
DISI, University of Trento; Trento, Italy

Abstract

In this paper, we propose an active learning based approach to event recognition in personal photo collections to tackle the challenges due to the weakly labeled data, and the presence of irrelevant pictures in personal photo collections. Conventional approaches relying on supervised learning can not identify the relevant samples in training albums, often leading to wrong classification. In our work, we aim to utilize the concepts of active learning to choose the most relevant samples from a collection and train a classifier. We also investigate the importance of relevant images in the event recognition process, and show how the performance degrades if all images from an album, containing the irrelevant ones, are included in the process. The experimental evaluation is carried out on a benchmark dataset composed of a large number of personal photo albums. We demonstrate that the proposed strategy yields encouraging scores in the presence of irrelevant images in personal photo collections, advancing recent leading works.

Introduction

The advent of smart-phones and digital cameras as well as the popularity of photo sharing platforms resulted in a huge number of multimedia content being collected, locally stored, and shared through the social networks. Similarly, cheap and dense storage devices also encourage users to generate and share more and more multimedia content. As a consequence, there is an ever-increasing need of automatic tools to collect, organize and retrieve multimedia data from huge unstructured archives. Recently, events emerged as a viable solution and powerful tool to organize and manage multimedia archives [1].

Literature on event-based analysis of multimedia content can be categorized into three groups, namely (i) event recognition/analysis in single images [2, 1]; (ii) event recognition in personal photo collections [3], and (iii) content analysis of multimedia related to natural disasters [4, 5]. In contrast to the other categories, there are many factors that make event recognition in personal photo collections/albums a more challenging task. These challenges are mostly due to the intrinsic nature of personal photo collections. For instance, photo albums are likely to contain ambiguous or irrelevant pictures, as face-close ups, which can be for example part of any event collection. Similarly, there can be images which can not be directly associated with the events as they do not include specific event-related objects [6]. Figure 1 shows some sample irrelevant images in the context of event recognition in personal photo collections. Moreover, they are annotated at album level, notwithstanding the fact that collections typically accommodate multiple sub-events. All these factors in combination make it difficult for the conventional approaches relying on supervised learning to identify the most relevant images in the albums from the training samples.

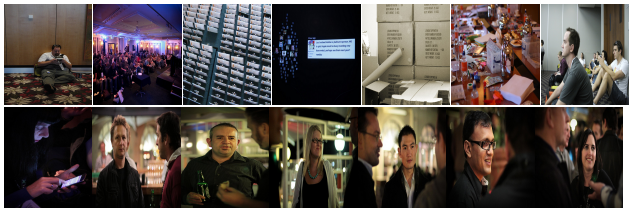

Figure 1. *Sample irrelevant images from the collections: images not containing objects of interest (top) and images with face close-ups (bottom)*

To tackle these challenges, in this paper, we propose a novel pipeline for event recognition in personal photo collections relying on Active Learning (AL). To the best of our awareness, no prior work leveraged on AL for event recognition in personal photo collections. The underlying insight of the proposed approach is to mitigate the limitations of the state-of-the-art methods based on conventional supervised learning, where it becomes hard to identify which images in the training albums are relevant.

In the proposed AL method, we aim to make use of the most relevant images for event recognition purposes where we first divide the training set into two separate subsets, namely (i) learning set and (ii) an initial training set containing a small number of samples from each sub collection of an event. Subsequently, the initial training set is iteratively augmented by adding relevant samples from the learning set, as detailed later. In other terms, the core idea is to start with a small training set annotated at image-level, and populate it with the help of Support Vectors (SVs) reflecting the most sensitive/informative training samples (i.e., closest to the separation hyperplane) from the training set annotated at bag-level.

To summarize our contribution, in this paper, we propose a novel active learning based framework for event recognition in personal photo collections to deal with irrelevant photos in photo collections. Moreover, we investigate the importance of relevant images in the classification process through the proposed active learning method. We also show how the performance degrades if all images from an album, most likely containing the irrelevant one, are included in the classification process.

The rest of the paper is organized as follows: Section 2 provides a detailed overview of the relevant literature. In Section 3, we present our framework. Section 4 is devoted to experimental setup, the conducted experiments and the analysis of the obtained results. At the end, Section 5 provides some concluding remarks.

Related Work

Over the last few years, event-based analysis of multimedia content has been widely utilized in a number of applications, such as event summarization, multimedia indexing and retrieval, and organization of personal photo collections [1]. In literature, events have been analyzed from three different perspectives including event recognition in single images, event recognition in personal photo collections, and analysis of extreme events (natural disasters) related multimedia content.

Most of the literature focuses on event recognition in single images, and a number of interesting solutions have been proposed to this aim [1, 7, 8, 2]. Similarly, the literature also provides some interesting solutions for the retrieval of extreme events (disasters) images from social media [9, 10, 11]. However, in contrast to the earlier two categories, there are number of challenges that make event recognition in personal photo collections a tougher problem to deal with.

In literature, very few solutions have been proposed for event recognition in personal photo collections. Most of the existing works focus on a joint use of visual data, and the additional information available in the form of meta-data. In this regard, temporal information along with visual features are mostly exploited. For instance, in [12], temporal information along with Convolutional Neural Networks (CNNs) features extracted through two different models pre-trained on ImageNet and Places datasets are used in a hierarchical way. Moreover, to deal with the irrelevant images in an album, they rely on average and aggregated features of all the images in an album. Similarly, Bossard et al. [3] use an image representation scheme relying on a joint use of temporal information and visual features. For the visual information they rely on low-level visual features (SURF [13]) while for the temporal information, time of day, and date are considered to form global temporal features. Besides time and date, additional information from meta-data, such as flash and exposure details are also exploited for event recognition in personal photo collections in combination with visual features in both late and early fusion schemes [14].

On the other side, there are some works mainly relying on visual information only. For instance, in [15], an object-centric approach relying on Histogram of Gradient (HoG) features for the extraction of object-centric information in an event-related image is proposed. A similar object-centric concept is also adopted in [16], where the training set is mined for the most frequent objects in the images from personal photo collections. More recently, in [17], a deep model is trained on images from personal photo collections to capture the co-occurrences, and frequencies of images features. In [18], features extracted through a deep model are used in a graphical model to capture a link among the pictures in a photo album.

However, most of the existing approaches rely on conventional supervised learning strategies, where it becomes difficult to identify relevant and more significant images in photo albums annotated at album-level only. Such limitations affect the performance of the existing approaches, as demonstrated in the experimental validation of more recent works in this context.

Methodology

As mentioned earlier, in this paper, our goal is to deal with irrelevant images in personal photo collections relying on Active Learning (AL). AL is a modified form of supervised learning, which involves the user in an interactive manner to obtain output at different data points. AL is generally used in applications with less labeled and abundant unlabeled data [19]. The paradigm aims to obtain higher accuracy with fewer labeled samples by involving an oracle to annotate the unlabeled data iteratively. In the proposed framework, we utilize the active learning concept to choose the most relevant samples from the personal photo collections, where the annotation is available at album level only, to train a classifier. Thus, the core idea is to select the most relevant images from the photo albums in the training samples, and avoid the ambiguous samples in the training data.

Figure 2 shows the block diagram of the proposed active learning method for the selection of most relevant photos from the photo collections for training a classifier. As can be seen, the proposed framework is composed of 3 main phases. In the first phase, we divide the training set into two separate subsets, namely (i) learning set and (ii) training set. Initially, a small portion of the samples from each album of an event are selected as an initial training set while a larger portion is dedicated to the learning set. The second phase of the proposed framework is rather standard, where we use standard Convolutional Neural Networks (CNNs) features for the representation of the photos in the collection. The next phase with a closed loop represents our proposed active learning method for the selection of relevant samples from the learning set. The active learning is further composed of three different steps as labeled with different alphabets in Figure 2 .

In the next subsections, we provide detailed description of feature extraction and proposed active learning method.

Feature Extraction

In this phase, our approach benefits from the state-of-the-art deep model for feature extraction. Literature suggests that the existing models pre-trained on ImageNet [20] and places dataset [21] perform better in event recognition with generally a slight improvement with the models pre-trained on ImageNet over the ones pre-trained on places dataset. It is important to mention that the models pre-trained on ImageNet correspond to object level information while the one pre-trained on places dataset extract scene level information.

In this work, our choice for features selection is motivated by better performance of VggNet [22] pre-trained on ImageNet in event recognition in single images, reported in [23]. VggNet is available in two different configurations; with 16 and 19 layers. In the current implementation, we rely on VggNet with 16 layers, where a 4096 dimensional feature vector is extracted from the last fully connected layer (Fc7) from each image using Caffe toolbox[1].

Selection of relevant photos via an Active Learning Method

The basic insight of the proposed active learning method is to augment the initial training set by adding relevant samples from the learning set, iteratively. It is important to mention that the learning set is much larger compared to the initial training set.

In order to populate the initial training set, as a first step (labeled as A in Figure 2) in our proposed active learning approach,

[1] http://caffe.berkeleyvision.org/

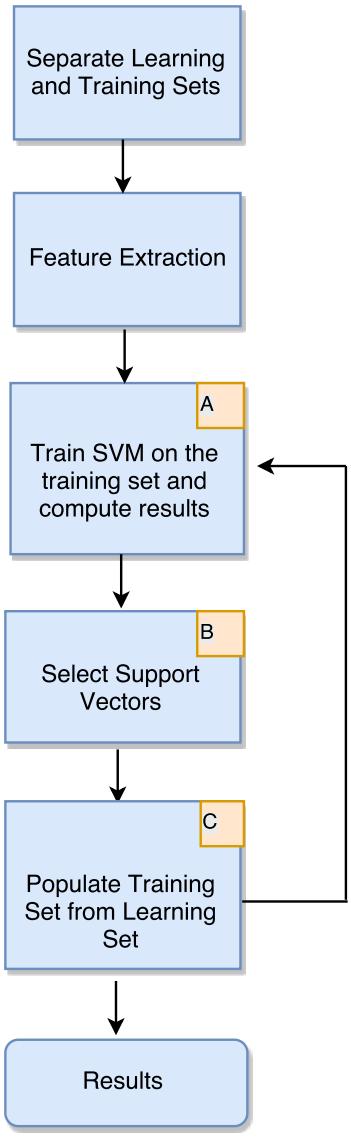

Figure 2. *Block diagram of the proposed methodology.*

we train a Support Vector Machine (SVM) classifier on the initial training set, and separate the significant samples from each album by selecting the Support Vectors (Svs) of the SVM classifier. Subsequently, in the final step (C), these significant samples are used to populate the initial training set by collecting their nearest neighbors from the learning set. In other terms, the core idea is that the SVs pertain to the most sensitive/informative training samples (i.e., closest to the separation hyperplane) for the classification task, which are employed to capture the most relevant samples from the learning ensemble in the Euclidean space. Thus, following the proposed method, we iteratively increase the training set with the relevant images from the collections, and evaluate the performance of the trained classifier at each iteration until a sufficient level of the performance is achieved.

The sample selection process can be repeated for any number of iterations; however, to prevent the irrelevant samples to be included in the training set, a limited number of iteration must be chosen to ensure the performance of the proposed method. In the current implementation, we use a threshold value on the similarity measure between Svs and their nearest the neighbours.

Experiments and Results

In this section, we provide description of the dataset used for evaluation, the experimental setup, the conducted experiments, and a detailed analysis of the experimental results and comparisons against state-of-the-art.

Dataset

The experimental validation of the proposed active learning approach to event recognition in personal photo collections is carried out on a large scale benchmark dataset known as Personal Events Collections (PEC) [3]. The dataset mainly covers 14 different events. Most of the events are related to personal sphere and collective activities. In total, the dataset contains 61,364 images from 807 different albums. The number of albums and photos inside the albums vary event to event. The details of the photos albums in each class are provided in Table 1.

The photo albums are downloaded from Flickr. The dataset also provides additional information, such as the timestamps (i.e., taken time in terms of day, month and year) and geo-location information, in the form of meta-data. Since the dataset is collected by downloading the complete albums from Flickr, it also contains a significant number of images which are irrelevant and ambiguous with respect to the event classification task. These irrelevant images consist of pictures not displaying event-specific objects or do not allow to perceive the underlying event, for example, face close-ups which can be part of any event. Such characteristics make the dataset a very challenging one. Figure 3 shows some sample images from the dataset.

Event	#Albums	Event	#Albums
Birthday	60	Graduation	51
Child. Birthday	64	Halloween	40
Christmas	75	Hiking	49
Concert	43	Road Trip	55
Cruise	45	St. Patricks Day	55
Easter	84	Skiing	44
Exhibition	70	Wedding	69

Details of Personal Events Collections Dataset

Experiments and Results

For the experiments, we divide the original training set into two subsets namely initial training set and learning set. Initially, we choose on the average 5 images per album for training purposes. These images are selected manually, where the idea is to have an initial training set with a limited number of reliable images. The initial training set is populated later on iteratively, by using the proposed active learning framework. The goal is to achieve better accuracy by populating the initial training set with the most relevant samples from the dataset, where annotation is available at album level only.

In order to show the effectiveness of the proposed active learning framework in this particular application, we report the performance of the proposed approach, in Table 2, by progressively increasing the number of images at each iteration. As can

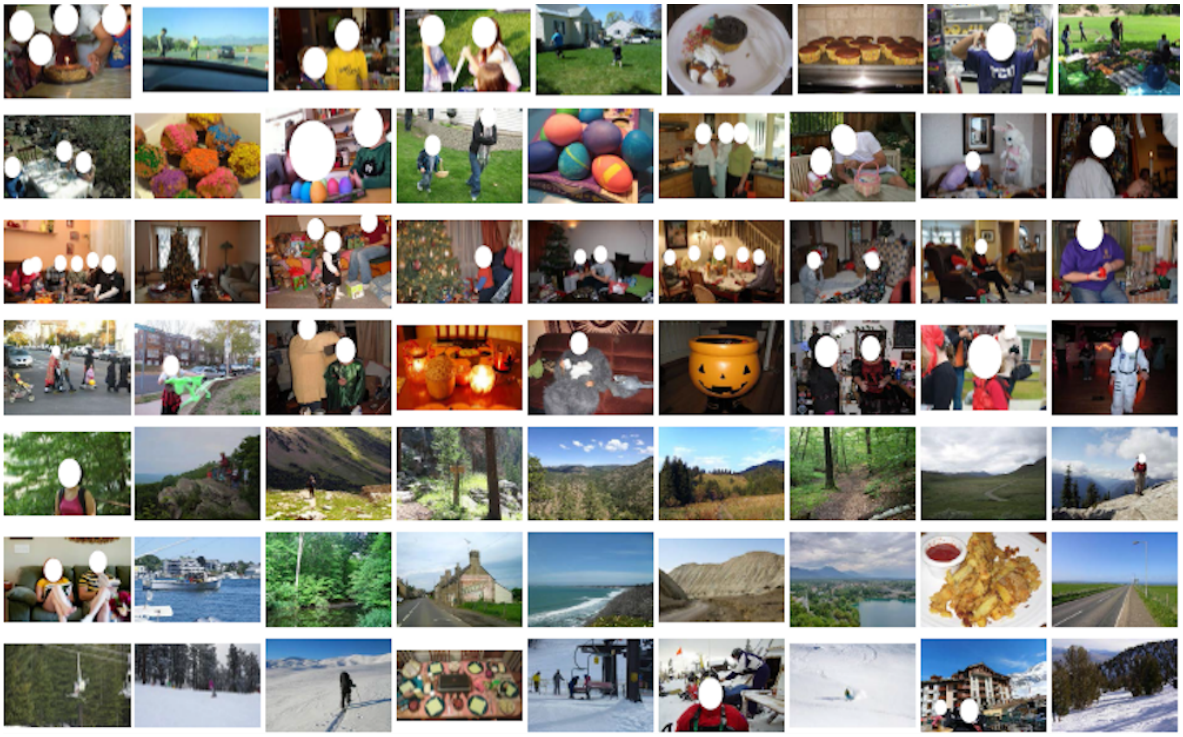

Figure 3. *Sample images from PEC dataset.[3].*

Iteration	Accuracy
1	.785
2	.814
3	.835
4	.842
5	.850
6	.850
7	.835
All images	.821

Experimental results at different iterations

be seen, the performance of the framework improves by populating the initial training set with the most relevant images from the learning set through the proposed active learning method at each iteration. However, it has been observed that after a certain number of iterations the performance stabilizes, and then starts decreasing, in fact, as we increase the number of iterations (i.e., including a higher number of images from the learning set). This reduction in the performance is due to the inclusion of the irrelevant images in the training samples because as we increase the number of iterations the probability of the inclusion of irrelevant images in the training samples increases. To further analyze this phenomenon, we compute the performance of the proposed framework by considering all the images from the learning set into the training set, which is basically equivalent to supervised learning. In this setting the performance decreses significantly, which shows the effectiveness of the proposed active learning framework.

To further confirm the suitability of our approach, we also provide comparisons of our proposed active learning based method against some known state-of-the-art solutions to event recognition in personal photo collections (see Table 3). The proposed active learning based approach, which helps to deal with the irrelevant images in the collections, achieves promising results against state-of-the-art, and ensures better performances in the presence of irrelevant images and weakly labeled data in personal photo collections.

Method	Accuracy (%)
AgS [3]	41.43
ShMM [3]	55.71
Method in [17]	73.41
R-OS-PGM [18]	74.28
Our method (best iteration)	85.0

Comparisons against state-of-the-art methods

Conclusions

In this paper, we proposed an active learning method to cope with the challenges in event recognition due to the collection-level annotation and presence of irrelevant images in personal photo collections. Through extensive experimentation, we show that compared to approaches relying on conventional supervised learning, the proposed active learning method is very promising in dealing with the irrelevant images in the photo collections. We also show that starting with a small training set annotated at image level can be easily populated with most relevant samples from the album level annotated personal photo collections, and better results can be achieved by considering all the images from the albums for training purposes.

References

[1] C. Tzelepis, Z. Ma, V. Mezaris, *et al.*, "Event-based media processing and analysis: A survey of the literature," *Image and Vision Computing* **53**, 3–19 (2016).

[2] K. Ahmad, N. Conci, G. Boato, *et al.*, "Used: a large-scale social event detection dataset," in *Proceedings of the 7th International Conference on Multimedia Systems*, 50, ACM (2016).

[3] L. Bossard, M. Guillaumin, and L. Gool, "Event recognition in photo collections with a stopwatch hmm," in *Proceedings of the ICCV*, 1193–1200 (2013).

[4] K. Ahmad, M. Riegler, A. Riaz, *et al.*, "The jord system: Linking sky and social multimedia data to natural disasters," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 461–465, ACM (2017).

[5] K. Ahmad, M. Riegler, K. Pogorelov, *et al.*, "Jord: A system for collecting information and monitoring natural disasters by linking social media with satellite imagery," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, 12, ACM (2017).

[6] K. Ahmad, N. Conci, G. Boato, *et al.*, "Event recognition in personal photo collections via multiple instance learning-based classification of multiple images," *Journal of Electronic Imaging* **26**(6), 060502 (2017).

[7] L. Wang, Z. Wang, Y. Qiao, *et al.*, "Transferring deep object and scene representations for event recognition in still images," *International Journal of Computer Vision* , 1–20 (2017).

[8] K. Ahmad, N. Conci, and F. De Natale, "A saliency-based approach to event recognition," *Signal Processing: Image Communication* **60**, 42–51 (2018).

[9] B. Bischke, P. Helber, C. Schulze, *et al.*, "The multimedia satellite task at mediaeval 2017,"

[10] K. Ahmad, P. Konstantin, M. Riegler, *et al.*, "Cnn and gan based satellite and social media data fusion for disaster detection," in *Working Notes Proc. MediaEval Workshop*, 2 (2017).

[11] S. Ahmad, K. Ahmad, N. Ahmad, *et al.*, "Convolutional neural networks for disaster images retrieval,"

[12] C. Guo and X. Tian, "Event recognition in personal photo collections using hierarchical model and multiple features," in *Proceedings of the MMSP*, 1–6, IEEE (2015).

[13] H. Bay, A. Ess, T. Tuytelaars, *et al.*, "Speeded-up robust features (surf)," *Computer vision and image understanding* **110**(3), 346–359 (2008).

[14] F. Tang, D. R. Tretter, and C. Willis, "Event classification for personal photo collections," in *Proceedings of the ICASSP*, 877–880, IEEE (2011).

[15] S. Tsai, T. S. Huang, and F. Tang, "Album-based object-centric event recognition," in *Proceedings of the ICME*, 1–6, IEEE (2011).

[16] S. Tsai, L. Cao, F. Tang, *et al.*, "Compositional object pattern: a new model for album event recognition," in *Proceedings of the ACM MM*, 1361–1364, ACM (2011).

[17] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Transactions on Multimedia* **17**(11), 1960–1968 (2015).

[18] S. Bacha, M. S. Allili, and N. Benblidia, "Event recognition in photo albums using probabilistic graphical models and feature relevance," *IJVCIR* (2016).

[19] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison* **52**(55-66), 11 (2010).

[20] C. Gan, N. Wang, Y. Yang, *et al.*, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *Proceedings of the CVPR*, 2568–2577 (2015).

[21] B. Zhou, A. Lapedriza, J. Xiao, *et al.*, "Learning deep features for scene recognition using places database," in *Proceedings of the NIPS*, 487–495 (2014).

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[23] K. Ahmad, M. L. Mekhalfi, N. Conci, *et al.*, "A pool of deep models for event recognition," (2017).