



# A saliency-based approach to event recognition

Kashif Ahmad <sup>\*,1</sup>, Nicola Conci, F.G.B. De Natale

*DISI-University of Trento, Italy*



## ARTICLE INFO

### Keywords:

Event recognition  
MIL  
Event saliency  
Multimedia indexing and retrieval

## ABSTRACT

Over the last few years, a number of interesting solutions covering different aspects of event recognition have been proposed for event-based multimedia analysis. Existing approaches mostly focus on an efficient representation of the image and advanced classification schemes. However, it would be desirable to focus on the event-specific information available in the image, namely the so-called event saliency. In this paper, we propose a novel approach based on multiple instance learning (MIL) to learn the visual features contained in event-salient regions, extracted through a crowd-sourcing study. In total, we collect the salient regions for 76 different events from 4 large-scale datasets. The experimental results demonstrate the efficacy of using only event-related regions by achieving a significant gain in performance over the state-of-the-art.

© 2017 Elsevier B.V. All rights reserved.

## 1. Introduction

The availability of low-cost hand-held devices, together with the increasing popularity of social networks, has contributed to the proliferation of shared multimedia contents, drastically changing the way in which people consume and communicate through social media. According to a recent report<sup>2</sup> based on an analysis conducted on Flickr, in 2016, a total of 612 millions public pictures have been uploaded to the platform at the rate of 1.68 millions photos per day, and these figures are increasing on a daily basis.

User-generated data are usually associated with personal experiences or collective activities, which can be seen as a collection of multimedia data that can be assembled in the form of events. Recent research studies demonstrate that the organization of multimedia data on the basis of underlying events facilitates effective annotation, synchronization, summarization, indexing and browsing [1,2].

The existing works on the subject mostly focus on an efficient representation of multimedia contents, and on strategies to exploit all of available information to achieve better performance in event recognition. In this regard, metadata, such as tags, title, and temporal and geo-location information, have been heavily exploited. However, metadata is not always reliable [3] and the recent trend is to shift towards the analysis of visual information, especially thanks to the learning capabilities offered by Convolutional Neural Networks (CNNs) [4,5], which have shown outstanding generalization capabilities in event recognition from visual contents [2].

Current research in visual information-based approaches to event recognition mostly focus on defining better representation and classification schemes. However, little attention has been paid to understanding and analyzing the salient visual elements, which are more revealing for a human observer. Targeting such event salient objects can help improving the performances of event recognition algorithms.

To this aim, we propose a novel framework for event recognition that exploits event-salient regions in a Multiple Instance Learning (MIL) framework. The underlying insight of the proposed framework is to target only the regions containing event-specific information. As demonstrated in [6,7], the concept of event saliency is different from the conventional visual saliency [8–10] where salient regions are highlighted based on visual properties like brightness, contrast, and position of the region. On the other hand, event saliency does not necessarily emerge by analyzing low-level features, which makes the detection a more challenging task. Based on these considerations, we propose and conduct a crowdsourcing activity for the selection of event-salient regions from a bundle of images for different types of events. The ultimate goal is to choose a set of event-salient regions for different events that can be used to train a classifier. In the proposed approach, image regions are first extracted at different scales via a selective search approach inspired by the method proposed in [11]. Next, after automatic pre-filtering of the less informative image-regions, a large number of volunteers are engaged in a crowdsourcing task to select the most salient regions among the ones extracted automatically. Our choice of conducting a crowdsourcing task is motivated by the need of finding

<sup>\*</sup> Corresponding author.

E-mail address: [kashif.ahmad@unitn.it](mailto:kashif.ahmad@unitn.it) (K. Ahmad).

<sup>1</sup> Member EURASIP.

<sup>2</sup> <https://www.flickr.com/photos/franckmichel/6855169886/>.

out the most significant image patches from a human perspective, being sufficiently generic also in terms of cultural and societal background.

The second contribution of this paper is at the classification stage, where a Multiple Instance classification (MIL) framework is adopted for the classification of a test image on the basis of the extracted regions. Multiple Instance Learning (MIL) and classification has proven to be very effective in many applications [12], and well fits our needs. To map our region-based approach into a multiple instance classification problem, we gather the extracted regions into bags, where each region is treated as an independent instance of the bag, as detailed later. After the bag-level classification of image-regions, we follow a one-against-one strategy to deal with multi-class classification. The final decision is based on majority voting.

The rest of the paper is organized as follows: Section 2 provides a detailed survey of the related work. Section 3 details the proposed methodology, and Section 4 reports the experimental evaluation and the obtained results. In Section 5, we provide some concluding remarks along with some directions of future work.

## 2. Related work

The notion of *event* has emerged as a rich source of contextual information that can be utilized in several applications, including multimedia indexing, retrieval, event summarization and event discovery [2]. Since the seminal work by Jain et al. [13], which proposed a common event model for multimedia analysis, events and their relationship with associated media has been an active area of research.

Most research on the subject targets an efficient representation of event-related multimedia contents and proposes strategies that can incorporate all of the available information to find revealing patterns in unknown multimedia data. To this aim, the associated metadata has been widely utilized for event recognition. Metadata usually include relevant information such as title, owner, and upload information along with temporal and geo-location. A number of research studies demonstrate the efficacy of metadata in event-based models for multimedia indexing and retrieval [14–17]. However, metadata also comes with limitations. In fact, in real world data, the presence of metadata is not guaranteed [3]. It may be either absent or corrupted. Possible sources of errors include wrong or no settings of the cameras time zone, missing time-stamps, as well as post-processing of the media with modification and removal of tags. Moreover, ambiguous meanings of tags, the existence of synonyms, and the ambiguities in different languages also affect its reliability.

Another important issue related to metadata is that geo-location and temporal information, although helpful in the recognition of an event instance (e.g., a particular person's wedding held at a specific time and at a particular location), is not much useful in differentiating among generic event classes (i.e., wedding).

Considering these limitations of metadata, visual content can be regarded as a valuable information for event recognition [18,19]. However, most of the earlier works in this domain rely on hand-crafted visual features, which cannot cope with the gap between image features and event semantics [2]. To cope with such issues, Tsampoulatidis et al. [20] proposed a multi-concept detection approach that combines different visual concept detectors for classification of event-related multimedia contents. A similar approach is used in [21], where a fusion strategy is adopted to combine different types of handcrafted visual features for a better representation of event-related multimedia contents. This joint approach of multi-concept detection through handcrafted features, partially solves the problem.

More recently, Convolutional Neural Networks (CNN) have shown promising performance in different computer vision applications [4,22]. As far as event recognition is concerned, most of the approaches rely on CNNs for a better representation of event-related images [23–27]. However, most of the existing image datasets for event recognition are not large enough to satisfy training requirements of CNNs. Consequently,

existing approaches tend to fine-tune pre-trained CNNs to fit them into event related images. To this aim, most of the existing approaches rely on the earlier models [28–31] pre-trained on the ImageNet dataset [32]. For example in [33], a pre-trained AlexNet [28] is fine-tuned on social event images. In [23], VGGNet-19 is adopted based on its success in object recognition. Similarly, in [34], AlexNet [28], VGGNet [29] and bn-inception network [30] are used to this aim.

Other works rely on the combination of existing models pre-trained on both ImageNet [32] and Places dataset [35]. The gain in the performance by combining object and scene-level information has been demonstrated in a number of works on the subject [23,34,36]. For example, in [36], object and scene-level information are used in a hierarchical way for event recognition in personal photo collections. Similar to our work, the method in [34] adopts a region-based approach to event recognition. The authors propose four different methods to jointly utilize object and scene-level information for event recognition, where the final classification decision is based on the average score of all the regions of the test image. A similar strategy is adopted in [25], where a combination of different CNN models is fine-tuned on image regions.

It is to be noted that little attention has been paid to understand, which are the key-visual elements in an event-related media item that help an observer recognizing the underlying multimedia event. An attempt in this direction is proposed by Rosani et al. [6], where a gamification technique is used to extract event-salient objects from event-related images. In the paper, a limited number of event-saliency samples (35 samples per class) are collected for 14 classes of social events.

In Table 1, we provide a summary of the most relevant approaches in the domain of event analysis, in terms of type of features, dataset used for evaluation along with some general comments.

## 3. Proposed approach

In this section, we provide a detailed description of our solution. As can be seen in Fig. 1, there are four different stages of the proposed approach. We start by extracting regions from event-related images at different scales. Next, in order to select event-salient regions from the different region proposals extracted in the first stage, we conduct a crowdsourcing task. The crowdsourcing activity is followed by feature extraction, where we use a pre-trained network VGGNet16 [29] for a better representation of the selected regions. Event-salient regions are then assembled into positive and negative bags for bag-level classification of the regions obtained from the test images. At the end, we adopt one-against-one classification, where the final decision is made on the basis of majority voting. In the following sub-sections, we provide a detailed description of each stage of the proposed approach.

### 3.1. Region extraction and pre-filtering

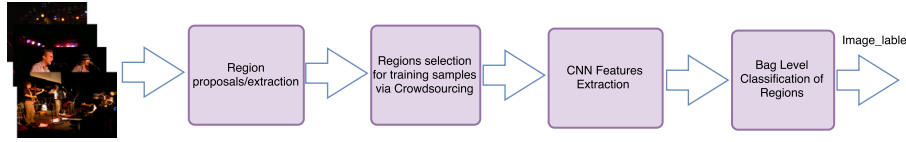
Images often contain objects or details that make them unique and give humans the capability of understanding the underlying event. For example, concert images usually contain musical instruments (e.g., microphones, guitars, etc.). Similarly, birthday images are often characterized by the presence of a cake and candles. A proper use of these event-salient regions may help improving the performances in event recognition. Therefore, we propose to divide images into different regions, and propose bag-level classification of the regions instead of classifying the whole image. The basic motivation of this approach is to target only the event-related objects and regions in the classification.

As we do not have any information of the exact location and scale of the salient regions, following the data driven selective search approach introduced in [11], we obtain a number of region proposals at different scales by combining exhaustive search and segmentation from an image at hand. A detailed analysis of the region proposals shows, however, that a significant number of the regions is irrelevant with respect to certain

**Table 1**

Summary of the most relevant works in terms of objectives, type of features and datasets used for validation along with some general comments.

Refs.	Objective	Features	Dataset	Approach
[37]	Single images	Image tags, gist	SED2013 [38]	Low level image features are used in the BoW model for event recognition in personal collections
[39]	Photo collection	SURF, BoW and temporal	PEC [39]	HMM is used for collection level classification
[40]	Single images	SIFT, BoW	SED2013 [38]	Low level features along with meta-data
[41]	Single images	low level features	UIUC sports [41]	An initial work on event based indexing on low level features
[6]	Single images	SURF with BoW	SED2013 (Subset)	The concept of event saliency was introduced with some initial experiments. Provides very few event salient objects for 14 events, only
[36]	Photo collection	CNNs	PEC	A hierarchical approach relying on supervised learning with ambiguous training data, which can lead to mis-classification
[33]	Single images	CNNs	SED, USED [33]	Pre-trained CNN is fine-tuned on a new dataset
[25]	Single images	CNNs	Cultural events dataset [42]	Pre-trained CNN is fine-tuned on all extracted image regions including irrelevant ones
[23]	Single images	CNNs	Cultural events dataset	Fusion of object and scene information for event recognition
[43]	Single images	CNNs	Cultural events dataset	Combination of networks fine-tuned on full images and image regions. They consider all regions where usually a higher number of regions is irrelevant
[34]	Single images	CNNs	WIDER, UIUC, cltural events	A supervised learning of all regions, majority of them are irrelevant or have strong visual correlation with others, extracted from images. Supervised learning on the ambiguous regions may lead to mis-classification
[7]	Single images	CNNs	SED, EIMM, USED dataset	A hierarchical approach with event salient objects and full images. Does not divide the test image into regions
[44]	Single images	CNNs	WIDER	Fusion of different layers of CNN

**Fig. 1.** Block diagram of the proposed methodology.

events, as well as for the event classification problem itself (i.e., they are not enough discriminative) [11]. Moreover, processing more regions per image requires considerably high resources and time. To this aim, we first filter out the less informative region proposals on the basis of their size and width–height ratio; in particular we propose to remove very small and thin regions, leading to a reduced set of image regions (on the average 15 regions per image). Our initial experiments on the validation set, show that the filtering stage reduces the processing time without significant impact on the classification results.

### 3.2. Salient regions selection via crowdsourcing

After pre-filtering there is still a number of regions, which are either not discriminative enough or have strong visual correlation with regions from other event classes. For example, concert and theater images may have similar backgrounds. Similarly, fashion images, although containing domain-specific objects, they usually contain elements that have strong correlation with the images from other classes like exhibition and conference. Moreover, event salient objects can be anywhere in an image, and are very difficult to be identified automatically through conventional visual saliency approaches [6]. To this aim, in order to select more relevant image-regions for training purposes, we have conducted a crowdsourcing study.

In the crowdsourcing task, we ask the volunteers to give their opinion about the candidate regions extracted after pre-filtering. In order to ensure a correct outcome of the study, and according to the literature [45,46], we tried to keep the task as simple as possible, making sure that most of the answers can be considered reliable.

Fig. 2 depicts the design of the proposed crowd-sourcing task. The extracted regions are presented to the users independently and randomly (regions are randomly shuffled, thus users do not know the order of appearance of the event classes). This strategy helps to make sure that the volunteers make a decision on the basis of the current region, only.

We asked the volunteers two different questions: (i) From these “N” possible events, which one do you think has been presented to you?

In the case of WIDER dataset, which contains 61 event classes, only a limited list of possible classes is presented to the user, where the user has to choose the relevant event, including the correct one. Another event class under the name of “Others” is included in the list, so users may select this option in the case they are not sure about the region class.

Next, the volunteers are asked to briefly motivate their choice. This description aims to get feedback from the users about the visual contents (i.e., objects and regions) that help humans to perceive the underlying event. Moreover, a selection of users’ answers has been inspected manually. This question has been useful also to evaluate the reliability and the engagement of the volunteers’ participating in the study.

In Fig. 3, we illustrate the process of region selection by providing a sample input image, sample region proposals and regions selected after filtering phase along with sample event-salient regions.

### 3.3. Feature extraction


The current state-of-the-art in event recognition reveals that image representation schemes deriving from Convolutional Neural Networks (CNNs), has shown a significant improvement over the conventional hand-crafted visual features. On this point, there is an ongoing trend of utilizing existing deep models pre-trained on objects (ImageNet [32]) and places [35] for the representation of event-related images. However, the earlier works in the domain demonstrate better performance for deep learning models pre-trained on ImageNet compared to the ones pre-trained on places [23,36]. Moreover, in the proposed work we are mainly interested in event-specific objects and regions instead of the complete scene. Therefore, we need an image descriptor that can well represent such event-specific objects in the extracted image regions.

To this aim, for feature extraction, we use VGGNet16 pre-trained on ImageNet [32]. From each image region, we extract a feature vector of size 4096 from layer fc7 (please refer to [29] for further details on the network architecture).

### Introduction to the Crowdsourcing Task

We are carrying out non-profit research at a university to build an event retrieval system. By accepting this task, you agree that we may publish parts of your answers as part of our research study. We will NOT publish any information that could be linked to you. We do NOT use your worker ID, or any other information that links to you, during data analysis or storage. Your answers are used only by researchers for the purposes of gaining insight into general opinions concerning events related multimedia. Beyond the people who are doing research in this area, no other parties are allowed to use your answers.

Event Representation via a region



Questions

(i) From these 8 possible events, which one do you think has been the one presented to you?

☐ Option 1

☐ Option 2

⋮

☐ Option n

(ii) Briefly explain, why did you choose the particular option in question i (open question)

Next

Fig. 2. The design of the crowdsourcing task developed for the selection of the event salient regions for training samples. At the top, an introduction to the task is provided with details of the proposed system. Then, regions extracted from the event-related images are provided one by one to the users involved in the crowdsourcing activity for annotation purposes. Two different questions are posed regarding the shown regions: selection of the event class, and motivation for the selection.

### 3.4. Multiple regions based classification of an image

Multiple instance learning and classification is a modified version of supervised learning, where a classifier is trained on a set of bags containing multiple feature vectors [47,48]. The test bags are also composed of multiple feature vectors, and labels are assigned at the bag-level.

In order to map our salient regions-based approach into multiple instance classification, we treat each image as a bag and each region is an independent instance of the bag. It is to be noted that the crowdsourcing study is carried out for the training samples, only. On the other hand, for the test samples all the extracted regions from each image that pass the pre-filtering stage are grouped into a single test bag. Thus, the regions in a test bags are always from same images while in the case of training bags, they are not necessarily to belong to a single image. For the bag-level classification of image regions, we use an approach inspired by C-KNN [49], by considering  $R$ -nearest references (bags in the neighborhood of the test sample) and  $C$ -citers bags, which consider the test sample as their own neighbor. The concept derives from library sciences: if a paper cites a previous paper (reference) both are considered to be related. Similarly, if a paper is cited by another paper (citer) the paper is said to be related to its citer. Thus, both citers and references are considered to be related to a paper. This blend of references and citer bags helps to mitigate the effect of false positive instances. The reference bags are simply the  $R$ -nearest neighbors. However, for the selection of  $C$ -citers of the bag a ranking mechanism [49] is adopted. For instance, if  $n$  is the number of total samples we have in a database  $B_s$ , represented as  $B_s = \{b_1, b_2, b_3, \dots, b_n\}$ , then, for a test bag  $b_i$ , the training samples are ranked according to the similarity to the test sample  $b_i$ . For instance, the rank of a sample  $b_j \in B_s$  with respect to  $b_i$  is represented as  $Rank(b_j, b_i)$ . Subsequently,  $C$ -nearest citers with threshold  $c$  (i.e., the number of total

citers to be selected) are defined as:

$$Citers(b_i, c) = \{b_j | Rank(b_j, b_i) \leq c, b_j \in B_s\}. \quad (1)$$

For the similarity measurement among bags, a bag-level distance metric, the Hausdorff distance [49], is used. For the comparisons of two bags  $X$  and  $Y$ , the Hausdorff distance is defined as follows:

$$h_k(X, Y) = k^{\text{th}}_{x \in X} \min_{y \in Y} \|x_i - y_i\| \quad (2)$$

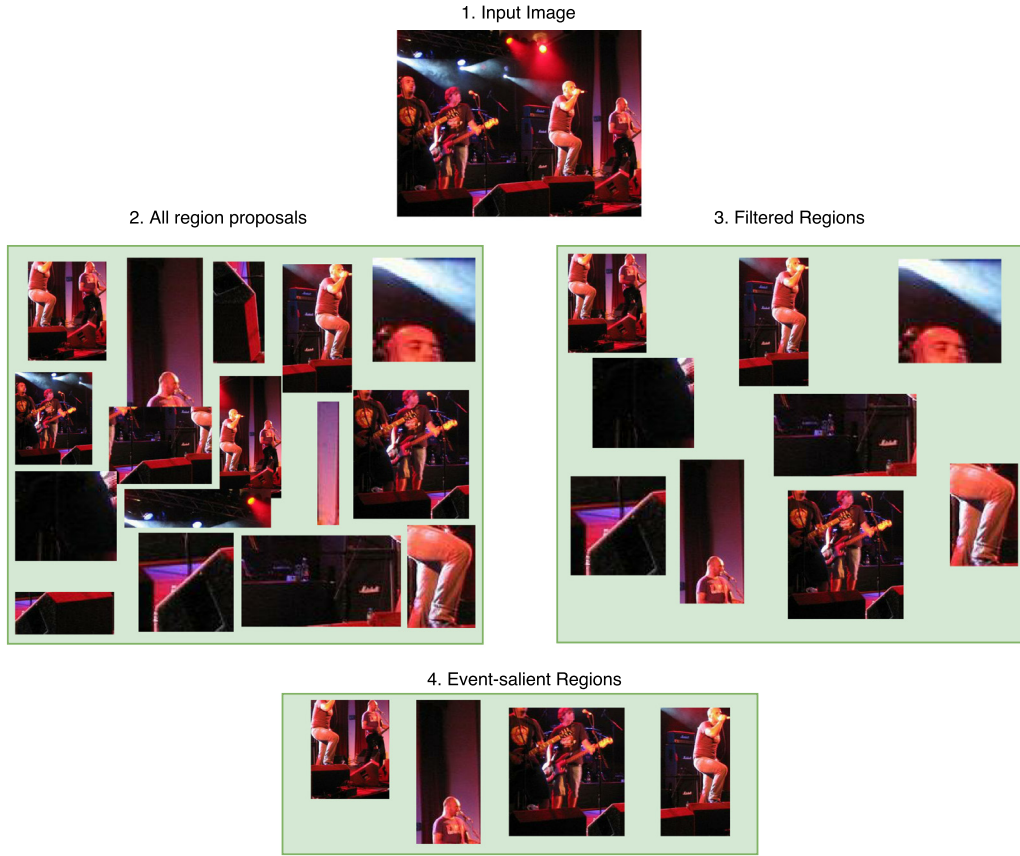
where  $x_i$  and  $y_j$  are the corresponding instances (i.e., image-regions in our case) and  $k$ th is the  $k$ th ranked value, which decides the value of the overall distance [49]. In our case, we opt for the minimal Hausdorff distance (i.e.,  $k = 1$ ) [49].

After the summation of  $R$ -nearest references and  $C$ -nearest citers in terms of positive bags (i.e.,  $B_p = R_p + C_p$ ) and negative bags (i.e.,  $B_n = R_n + C_n$ ), a majority voting approach is used for the prediction of a given test bag  $b_i$ . The bag  $b_i$  is classified as positive if  $B_p$  (# positive bags)  $> B_n$  (# negative bags); it is classified as negative otherwise (Eq. (3)). In the case of a tie, we assign a negative label to the test sample as more weight is given to negative samples compared positive ones in multiple-instance learning paradigm [49]).

$$C_{\text{Label}} = \begin{cases} 1 & \text{if } B_p > B_n \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

Finally, in order to deal with multi-class classification we adopt the one-against-one strategy where results are obtained from all binary classifiers. Subsequently, the final classification decision is made on the basis of majority voting.





**Fig. 3.** Region selection process: (1) Represent an input image; (2) Shows sample region proposals; (3) Provides some sample regions after filtering phase while (4) Represents sample event-salient regions obtained through crowdsourcing study from the input image.

## 4. Experimental setting

### 4.1. Dataset

For the experimental validation of the proposed approach, we use four large-scale events datasets, namely SED2013 [38], UIUC sports events dataset [41], USED [33] and Web Images for Event Recognition (WIDER) [44]. SED2013 covers 7 different social events with a large number of images per event; UIUC Sports Events Dataset is composed of 8 different sports events; USED covers 14 different social events, including the 7 from SED2013 (in this work we consider only these 7 events namely concert, conference, exhibition, fashion, protest, sports and theater); WIDER contains 61 different events, covering sports events, as well as daily life events. Moreover, it also provides images from some social events, like concerts, celebrations, dancing shows and funerals. In addition, there are also some unusual event classes, such as street battle, demonstration, riot and parade. Most of these event classes are taken from Large Scale Ontology for Multimedia (LSCOM) [50].

All in all, we use images from 76 different event classes for the validation of the proposed approach. Fig. 4 shows some sample images from all of these datasets.

Each dataset devotes a large portion of data for training purposes. However, in order to reduce efforts in the crowdsourcing study, we use a subset of training samples. From SED2013 and UIUC, we extract image regions from 150 and 30 randomly selected images per class, respectively. Since USED and SED2013 share similar event classes, we use the same regions as training samples for both datasets. In the case of WIDER, we extract image regions from 200 randomly selected images per class, mostly due to the complexity and the dynamic nature of the events. Moreover, the event classes from WIDER dataset, compared to the other datasets, have closer visual correlation with each others. For example, we have close resemblance among soldier firing, soldier

drilling, and soldier marching events. This strong similarity makes the recognition more challenging, and therefore a higher number of samples would be desirable to allow a better discrimination among classes. This is also confirmed by the crowdsourcing volunteers as we observed a higher number of regions tagged as “others” for WIDER.

### 4.2. Setup

As aforesaid, the crowdsourcing study has been conducted for the selection of regions in training samples only, which are then randomly assembled into bags. The number of regions per bag has a significant impact on the processing time (i.e., using a higher number of image regions per bag will take more time to be processed). Therefore, the most important parameters to be defined in the proposed approach are the number of images per training bag as well as the number of citers and reference bags. To this aim we validate our approach using 3 different configurations with 5, 10, and 15 regions per bag on the validation set. As far as the test bags is concerned, we used all the regions except the ones discarded in the pre-filtering phase in Section 3.1. The number of citer and reference bags is obtained testing different combinations on the validation set to find the best values. At the end, we choose 3 references and 5 citer bags.

## 5. Experimental results

In this section we provide the analysis of the crowdsourcing study along with the detailed description of the conducted experiments.

### 5.1. Crowdsourcing analysis

We received around 25 000 responses from more than 400 distinct volunteers. On average, each volunteer investigated more than 60 image

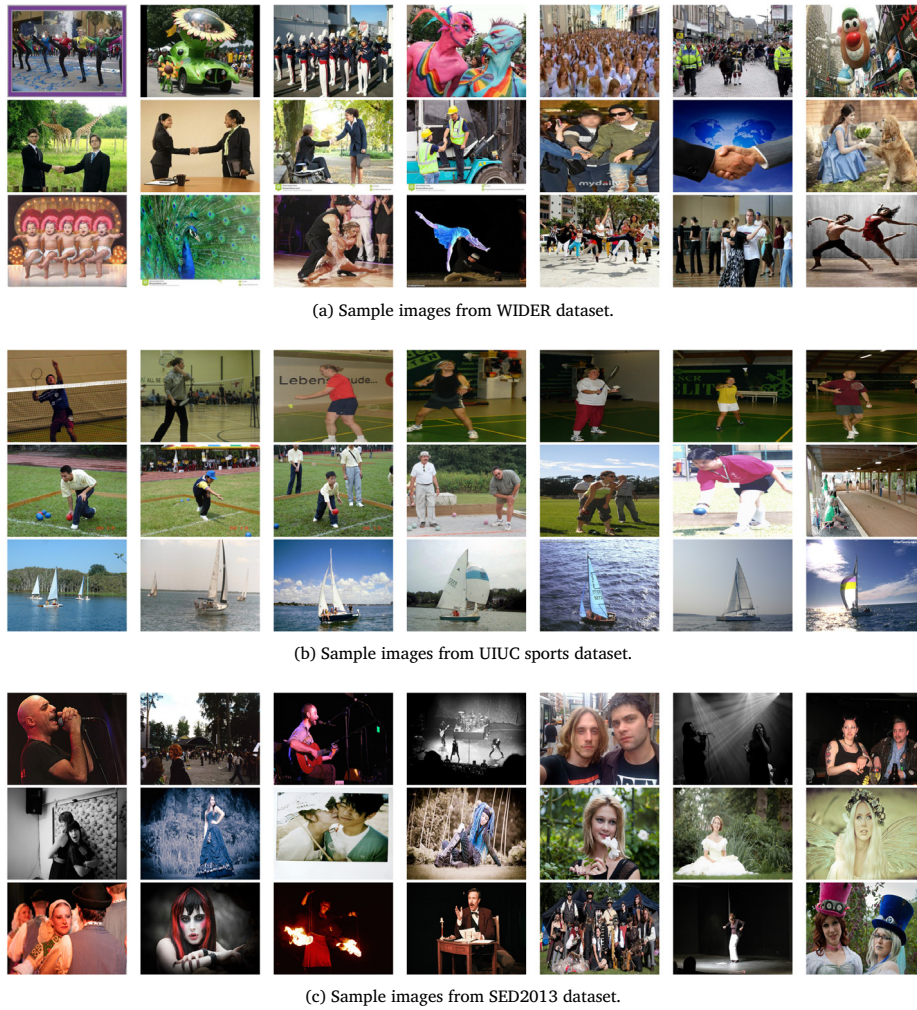


Fig. 4. Sample images from the datasets used in the experiments.

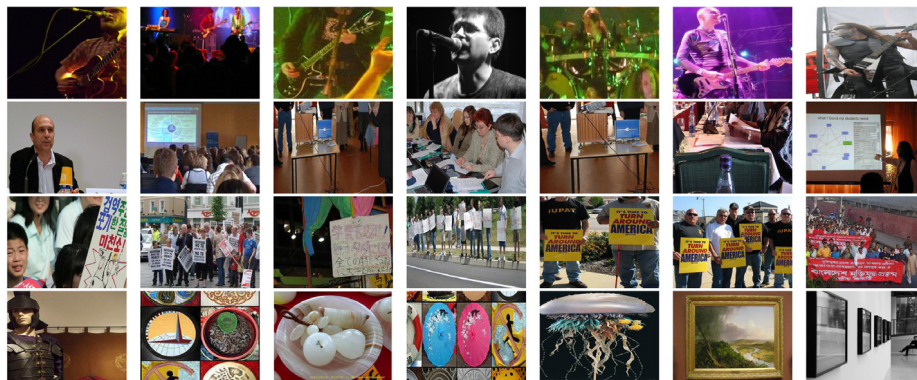


Fig. 5. Sample regions tagged by the volunteers as concert, conference, protest and exhibition (top to down) in the crowdsourcing task.

regions. We discarded 47 responses because the answers in the open question either demonstrated the difficulty of the user in understanding the task, or revealed inconsistencies between the question and the answer. For each image region, we have collected at least 3 responses from 3 different volunteers, and put them into the bags of an event class according to the majority of the responses. Fig. 5 shows some sample regions along with the tags provided by the volunteers. There is also a number of image regions, for which the majority of volunteers are not

sure, tagged as “others” (see Fig. 6). Such regions are discarded from the training samples.

Overall in SED2013 and UIUC Sports Events datasets, we observed a higher precision in the answers of the volunteers. This cannot be said instead for WIDER, where we observed a certain degree of uncertainty in the event class assignment.<sup>3</sup>

<sup>3</sup> The created dataset of the selected event salient regions will be made publicly available upon acceptance of the paper.





**Table 6**  
Experimental results of our approach on SED dataset.

	Predicted-class							
		Concert	Conference	Exhibition	Fashion	Protest	Sport	Theater
Actual-class	Concert	<b>.916</b>	.003	.018	.014	.004	0	.041
	Conference	.004	<b>.856</b>	.027	.104	.001	0	.005
	Exhibition	.002	.020	<b>.868</b>	.102	.005	0	0
	Fashion	.004	.008	.016	<b>.963</b>	.002	.004	0
	Protest	0	.042	.007	.032	<b>.892</b>	.027	0
	Sports	0	0	.062	0	.027	<b>.893</b>	.017
	Theater	.016	.005	.002	.025	0	0	<b>.950</b>

**Table 7**  
Experimental results of our approach on USED dataset.

	Predicted-class							
		Concert	Conference	Exhibition	Fashion	Protest	Sport	Theater
Actual-class	Concert	<b>.817</b>	.011	.004	.042	.013	0	.111
	Conference	.014	<b>.721</b>	.045	.135	.033	.011	.037
	Exhibition	.005	.088	<b>.717</b>	.153	0	.034	0
	Fashion	0	0	.044	<b>.828</b>	.013	0	.114
	Protest	.007	.127	.042	.035	<b>.729</b>	.021	.036
	Sports	.002	.041	.117	.029	.050	<b>.740</b>	.018
	Theater	.008	0	.034	.114	0	0	<b>.843</b>

**Table 8**  
Experimental results of our approach on UIUC sports dataset.

	Predicted-class								
		Badminton	Bocce	Croquet	Polo	Rowing	Rock-climbing	Snow-boarding	Sailing
Actual-class	Badminton	<b>.976</b>	.015	.007	0	0	0	0	0
	Bocce	.013	<b>.958</b>	.027	0	0	0	0	0
	Croquet	0	.035	<b>.953</b>	.011	0	0	0	0
	Polo	0	0	0	<b>1.00</b>	0	0	0	0
	Rowing	0	0	.007	0	<b>.984</b>	0	0	.007
	Rock-climbing	0	.005	0	0	0	<b>.989</b>	.005	0
	Snow-boarding	0	0	0	0	0	0	<b>1.0</b>	0
	Sailing	0	0	0	0	0	0	0	<b>1.0</b>

This is mostly due to the distinctive visual features and image patterns that appear in the scene. On other classes the performances decrease significantly, mostly because of the complexity of the events themselves. On the other hand, the proposed approach provides better results on each class of USED dataset. It is interesting to note that, though SED and USED datasets share similar event classes, the performance on USED is slightly lower than SED dataset as shown in Table 4. One of the main reasons of the lower performance is the nature of the dataset itself. This fact can also be concluded from the lower performances of the state of the art on USED dataset. On the other hand, the proposed approach achieve better performance on UIUC dataset, despite the dataset is comparatively small (Table 5).

In order to provide a more thorough analysis, we also report the confusion matrices of experiments performed on SED, USED and UIUC Sports datasets, as shown in Tables 6–8, respectively. As can be observed, our approach provides good results on all events covered in the datasets. In Tables 6 and 7, noticeable uncertainty can be found for conference and exhibition events with fashion images, a problem that was evident also when conducting the crowdsourcing study. Similarly, we also observed that some test samples from concert events are misclassified as theater and vice-versa. On UIUC no significant errors can be observed, except the few misclassification in badminton, bocce and croquet events.

We also compare our approach against state-of-the-art on 4 different datasets. To show the significance of event-salient features, we provide the comparison of our approach against the best performing methods on each dataset. Table 9 provides the comparisons against state-of-the-art on SED 2013. The gain our approach achieves over the state-of-the-art is reported in Table 9.

On the other hand, the comparisons on WIDER dataset [44] are provided in Table 10. As can be seen in Table 10, our approach

**Table 9**  
Comparisons against state-of-the-art on SED2013.

Method	Avg. acc.
Method in [51]	.334
Method in [6]	.4595
Method in [33]	.7003
Method in [7]	.8579
Our approach	.9115

**Table 10**  
Comparisons against state-of-the-art on WIDER [44].

Method	Avg. acc.
Baseline method [44]	.397
Deep channel fusion [44]	.4204
Method in [52]	.4406
Init. based object-scene transferring [34]	.508
Knowl. based object-scene transferring [34]	.520
Data based object-scene transferring [34]	.526
Data + knowl. based object-scene transferring [34]	.530
Our approach	.5504

shows promising results on these complex event classes. As mentioned earlier, in order to reduce the efforts in crowd-sourcing study, instead of complete training samples, from each event class we use a subset of training data. We have an overall gain of 2.04% over the state-of-the-art, using only a subset of the training set, which shows the significance of the proposed approach.

The comparisons on USED [33] and UIUC Sports Events dataset [41] are provided in Tables 11 and 12, respectively. Our approach achieves a significant gain of around 5% against the state-of-the-art on USED.



**Table 11**

Comparison vs state of the art on USED.

Methods	Avg. acc.
Baseline method [33]	.700
Method in [53]	.720
Our approach	.771

**Table 12**

Comparisons against state of art on UIUC sports Dataset [41].

Methods	Avg. acc.
Baseline method [41]	.7340
ImageNet CNN features [35]	.9440
Places CNN features [35]	.9410
Googlenet GAP [54]	.9500
object-scene transferring [34]	.9880
Our approach	.9838

The performance obtained on UIUC sports dataset, are instead comparable with the state-of-the-art, possibly due to the limited size of the dataset.

## 6. Conclusions

In this paper, we investigate the importance of event salient objects and regions in event recognition through extensive experimentation on four large scale datasets. For the selection of event salient regions, a crowdsourcing study is conducted with a large number of volunteers. Moreover, we propose the adoption of a MIL framework of the extracted regions instead of classifying them individually. We show that better results can be obtained by involving only event-salient regions in event recognition. Moreover, the multiple instance learning and classification scheme better suits the region-based approach to event recognition.

Performance can be further improved by combining scene and object-level information. In future, we aim to propose a better fusion scheme that can utilize multiple CNNs models for a better representation of extracted regions. Moreover, a better scheme for pre-filtering of extracted regions especially for the test samples can also improve the performances both in terms of processing time and classification accuracy.

## References

- [1] E. Sansone, K. Apostolidis, N. Conci, G. Boato, V. Mezaris, F.G. De Natale, Automatic synchronization of multi-user photo galleries, *IEEE Trans. Multimedia* 19 (6) (2017) 1285–1298.
- [2] C. Tzelepis, Z. Ma, V. Mezaris, B. Ionescu, I. Kompatsiaris, G. Boato, N. Sebe, S. Yan, Event-based media processing and analysis: A survey of the literature, *IVC* 53 (2016) 3–19.
- [3] X. Liu, B. Huet, Heterogeneous features and model selection for event-based media classification, in: *ICMR*, ACM, 2013, pp. 151–158.
- [4] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, S. Tubaro, Deep convolutional neural networks for pedestrian detection, *Proceedings of the SPIM* 47 (2016) 482–489.
- [5] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, X. Xu, A multi-scene deep learning model for image aesthetic evaluation, *Signal Processing: Image Communication* 47 (2016) 511–518.
- [6] A. Rosani, G. Boato, F.G. De Natale, Eventmask: A game-based framework for event-saliency identification in images, *TMM* 17 (8) (2015) 1359–1371.
- [7] K. Ahmad, F. De Natale, G. Boato, A. Rosani, A hierarchical approach to event discovery from single images using mil framework, in: *Proceedings of the GlobalSIP*, IEEE, 2016, pp. 1223–1227.
- [8] D. Chen, T. Jia, C. Wu, Visual saliency detection: From space to frequency, *Signal Processing: Image Communication* 44 (2016) 57–68.
- [9] P. Koutras, P. Maragos, A perceptually based spatio-temporal computational framework for visual saliency estimation, *Signal Processing: Image Communication* 38 (2015) 15–31.
- [10] K. Ahmad, N. Ahmad, R. Khan, A. Khalil, Saliency based skin detection in complex scenes, in: *Sixth International Conference on Machine Vision (ICMV 13)*, International Society for Optics and Photonics, 2013 90671U–90671U.
- [11] J.R. Uijlings, K.E. van de Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, *IJCV* 104 (2) (2013) 154–171.
- [12] V.K. Sharma, K. Mahapatra, Mil based visual object tracking with kernel and scale adaptation, *Signal Processing: Image Communication* 53 (2017) 51–64.
- [13] U. Westermann, R. Jain, Toward a common event model for multimedia applications, *Proceedings of the ACM MM* (1) (2007) 19–29.
- [14] M.-S. Dao, G. Boato, F.G. De Natale, Discovering inherent event taxonomies from social media collections, in: *Proceedings of the ICMR*, ACM, 2012, p. 48.
- [15] L. Chen, A. Roy, Event detection from flickr data through wavelet-based spatial analysis, in: *Proceedings of the CIKM*, ACM, 2009, pp. 523–532.
- [16] T. Rattenbury, N. Good, M. Naaman, Towards automatic extraction of event and place semantics from flickr tags, in: *Proceedings of the ACM SIGIR*, ACM, 2007, pp. 103–110.
- [17] S. Papadopoulos, C. Zgkalis, Y. Kompatsiaris, A. Vakali, Cluster-based landmark and event detection for tagged photo collections, *Proceedings of the ACM MM* (1) (2010) 52–63.
- [18] M.-S. Dao, D.-T. Dang-Nguyen, F.G. De Natale, Robust event discovery from photo collections using signature image bases (sibs), *MTAP* 70 (1) (2014) 25–53.
- [19] N. Gkalelis, V. Mezaris, I. Kompatsiaris, High-level event detection in video exploiting discriminant concepts, in: *Proceedings of the CBMI*, IEEE, 2011, pp. 85–90.
- [20] I. Tsapoulatis, N. Gkalelis, A. Dimou, V. Mezaris, I. Kompatsiaris, High-level event detection system based on discriminant visual concepts, in: *Proceedings of the ICMR*, ACM, 2011, p. 68.
- [21] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, A.G. Hauptmann, Multimedia classification and event detection using double fusion, *MTAP* 71 (1) (2014) 333–347.
- [22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *Proceedings of the CVPR* (2016) 770–778.
- [23] L. Wang, Z. Wang, W. Du, Y. Qiao, Object-scene convolutional neural networks for event recognition in images, *Proceedings of the CVPR* (2015) 30–35.
- [24] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, A.G. Hauptmann, Devnet: A deep event network for multimedia event detection and evidence recounting, *Proceedings of the CVPR* (2015) 2568–2577.
- [25] S. Park, N. Kwak, Cultural event recognition by subregion classification with convolutional neural network, *Proceedings of the CVPR Workshops* (2015) 45–50.
- [26] A. Salvador, M. Zeppezauer, D. Manchon-Vizuet, A. Calafell, X. Giro-i Nieto, Cultural event recognition with visual convnets and temporal models, *Proceedings of the CVPR* (2015) 36–44.
- [27] L. Wang, Z. Wang, S. Guo, Y. Qiao, Better exploiting os-cnns for better event recognition in images, *Proceedings of the ICCV* (2015) 45–52.
- [28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Proceedings of the NIPS* (2012) 1097–1105.
- [29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.
- [30] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167*.
- [31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, *Proceedings of the CVPR* (2015) 1–9.
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: *Proceedings of the CVPR*, IEEE, 2009, pp. 248–255.
- [33] K. Ahmad, N. Conci, G. Boato, F.G. De Natale, Used: a large-scale social event detection dataset, in: *Proceedings of the MMSys*, ACM, 2016, p. 50.
- [34] L. Wang, Z. Wang, Y. Qiao, L. Van Gool, Transferring object-scene convolutional neural networks for event recognition in still images, *arXiv preprint arXiv:1609.00162*.
- [35] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, *Proceedings of the NIPS* (2014) 487–495.
- [36] C. Guo, X. Tian, Event recognition in personal photo collections using hierarchical and multiple features, in: *Proceedings of the MMSP*, IEEE, 2015, pp. 1–6.
- [37] M. Brenner, E. Izquierdo, Social event detection and retrieval in collaborative photo collections, in: *Proceedings of the ICMR*, ACM, 2012, p. 21.
- [38] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, S. Geva, Social event detection at mediaeval 2013: Challenges, datasets, and evaluation, in: *MediaEval Workshop*, 2013.
- [39] L. Bossard, M. Guillaumin, L. Van Gool, Event recognition in photo collections with a stopwatch hmm, *Proceedings of the ICCV* (2013) 1193–1200.
- [40] T.-V. Nguyen, M.-S. Dao, R. Mattivi, E. Sansone, F.G. De Natale, G. Boato, Event clustering and classification from social media: Watershed-based and kernel methods, in: *MediaEval*, 2013.
- [41] L.-J. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: *Proceedings of the ICCV*, IEEE, 2007, pp. 1–8.
- [42] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H.J. Escalante, D. Misevic, U. Steiner, I. Guyon, Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results, *Proceedings of the ICCV* (2015) 1–9.
- [43] M. Liu, X. Liu, Y. Li, X. Chen, A.G. Hauptmann, S. Shan, Exploiting feature hierarchies with convolutional neural networks for cultural event recognition, *Proceedings of the ICCV* (2015) 32–37.
- [44] Y. Xiong, K. Zhu, D. Lin, X. Tang, Recognize complex events from static images by fusing deep channels, *Proceedings of the CVPR* (2015) 1600–1609.
- [45] M. Riegler, V.R. Gaddam, M. Larson, R. Eg, P. Halvorsen, C. Griwodz, Crowdsourcing as self-fulfilling prophecy: Influence of discarding workers in subjective assessment

- tasks, in: 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), 2016, pp. 1–6. doi: [10.1109/CBBI.2016.7500256](https://doi.org/10.1109/CBBI.2016.7500256).
- [46] K. Ahmad, M. Riegler, K. Pogorelov, N. Conci, P. Halvorsen, F. De Natale, 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, ACM, 2017, p. 12.
- [47] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, *Artificial Intelligence* 201 (2013) 81–105.
- [48] B. Babenko, Multiple instance learning: Algorithms and applications, [View Article PubMed/NCBI Google Scholar](#) (2008) 1–19.
- [49] J. Wang, J.-D. Zucker, Solving multiple-instance problem: A lazy learning approach.
- [50] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, Large-scale concept ontology for multimedia, *IEEE Multimedia* 13 (3) (2006) 86–91.
- [51] E. Schinas, G. Petkos, S. Papadopoulos, Y. Kompatsiaris, Certh@ mediaeval 2012 social event detection task, in: MediaEval Workshop, Citeseer, 2012.
- [52] R.F. Rachmadi, K. Uchimura, G. Koutaki, Combined convolutional neural network for event recognition, Korea–Japan Joint Workshop on Frontiers of Computer Vision (2016) 85–90.
- [53] R.F. Rachmadi, K. Uchimura, G. Koutaki, Spatial pyramid convolutional neural network for social event detection in static image, *arXiv preprint arXiv:1612.04062*.
- [54] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, *Proceedings of the CVPR* (2016) 2921–2929.