DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Events based Multimedia Indexing and Retrieval

Kashif Ahmad

SUBMITTED TO THE DEPARTMENT OF
INFORMATION ENGINEERING AND COMPUTER SCIENCE (DISI)
IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF

DOCTOR OF PHILOSOPHY

Advisor:	Prof. Nicola Conci, Università degli Studi di Trento, Italy
Examiners:	Prof. Marco Carli, Università degli Studi di Roma Tre, Italy
	Prof. Pietro Zanuttigh, Università degli Studi di Padova, Italy
	Prof. Giulia Boato, Università degli Studi di Trento, Italy

December 2017

Abstract

Event recognition is one of multimedia applications that has been gaining ground recently. However, it has received scarce attention relatively to other applications. The methodologies presented hereby are aimed at event-based analysis of multimedia content, which is achieved from three perspectives, namely (i) event recognition in single images, (ii) event recognition in personal photo collections and (iii) fusion of social media information and satellite imagery for natural disaster detection. A close look at the relevant literature suggests that more attention has been paid to event recognition in single images. Event recognition in personal photo collection has also received a number of interesting solutions. Natural disaster detection in images from social media and satellite imagery, however, is relatively new. As a matter of fact, many issues remain unsolved mostly due to the heterogeneity, multi-modality and the unstructured nature of the data.

In this dissertation, such open problems are presented and analyzed. New perspectives and approaches are suggested, alongside a detailed experimental validation and analysis. In details, our contribution is multi-fold. On the one hand, we aim at demonstrating that the fusion of different feature extraction and classification strategies can outperform the single methods by jointly exploiting the learning capabilities of individual deep models. On the other side, we analyze the importance of event-salient objects and local image regions in event recognition. We also present a novel framework for event recognition in personal photo collections. Moreover, we also present our system JORD, and our Convolutional Neural Networks (CNNs) and Generative Adversarial Network (GAN) based fusion of social media and satellite images for natural disaster detection. A thorough experimental analysis of each proposed solution is provided on benchmark datasets along with the potential direction of future work.

Keywords

Multimedia Indexing, Multimedia Retrieval, Event Recognition, Event Saliency, Machine Learning, Deep Learning

Acknowledgements

I would like to extend my profound thanks to my supervisor, Prof. Nicola Conci, for his immense guidance and support through out my PhD.

I would also thank Prof. Francesco De Natale and Prof. Giulia Boato for their support.

I would also like to thank our collaborators Mohammed Lamine Mekhalfi, Michael Rieglar and Pogorelov Konstantin for their honest efforts to make our collaborative research possible.

Last but not least, I would always be indebted to my family and friends for their moral support through out my carrier.

Kashif Ahmad
Trento, Italy
December 2017

Contents

List of Tables

List of Figures

List of Acronyms

CNNs	Convolutional Neural Networks
SED	Social Event Detection
WIDER	Web Images Dataset for Event Recognition
LSCOM	Large Scale Ontology for Multimedia
USED	UNITN Social Events Dataset
IOWA	Induced Ordered Weighting Average
PSO	Particle Swarm Optimization
GA	Genetic Algorithms
GANs	Generative Adversarial Networks
PEC	Personal Event Collections
HMMs	Hidden Markov Models
MIL	Multiple Instance Learning
CKNN	Citation K-nearest Neighbours
SVM	Support Vector Machine
RF	Random Forest
LMT	Logistic Model Tree
DIRSM	Disaster Images Retrieval from Social Media
FDSI	Flood Detection in Satellite Images

Chapter 1

Introduction

1.1 Context

The habit of taking pictures of everyday life moments is more and more spreading into the society, especially with the advent of cams and low cost hand-held devices, together with the increasing popularity of social networks, it has become easier to generate and share multimedia contents. These phenomena have changed the way in which people consume and communicate through social media, causing a huge number of images being collected, stored, posted, and shared through the Internet. For instance, according to a recent report[1] based on an analysis conducted on Flickr, in 2016, a total of 612 millions public pictures have been uploaded to the platform at a rate of 1.68 millions photos per day, and these figures keep increasing on a daily basis. As a consequence, there is an ever increasing need for automatic tools able to suitably organize and retrieve image data from large unstructured multimedia archives, relieving content owners from the tedious task of manually arranging their media collections.

Although this domain has been widely investigated in the past [54], no ultimate solution is still available. Looking at the problem from a user's perspective, multimedia collections often refer to personal experiences and

[1] `https://www.flickr.com/photos/franckmichel/6855169886/`

activities, which can be referred to as *events*. The assumption that personal media collections can be viewed as the visual facet of events [134] opened new research directions in media indexing and retrieval, where the personal experience plays a central role. Some interesting applications of such concepts can be found in the area of event summarization and event-based organization of personal photo collections [124].

1.2 Problem Statement

So far, various solutions have been proposed to address the issue of event recognition in videos [58, 23], although event recognition from still pictures remains a more challenging problem, due to the sparsity of data and the absence of a coherent and contiguous information flow. The computer vision literature suggests that conventional paradigms based upon shallow handcrafted visual features are prone to failure, as they cannot fill the gap between the spatial/chromatic content of an optical image and its semantic attributes [32, 95].

Following their immense success in image classification, object recognition and detection [90], Convolutional Neural Networks (CNNs) have demonstrated to perform well also in event recognition tasks (e.g., [106, 131, 40, 93]). However, though very efficient, CNNs are known to require large training sets in order to capture the undergoing spatial/chromatic cues across the images at hand. This is a major problem in media event recognition, as it is extremely difficult and time consuming to collect sufficiently large and significant datasets to meet the training and validation requirements on one side, and to avoid generalization problems on the other. As an alternative, a common option is to fine-tune pre-trained (on large-scale datasets) CNNs to tailor them to event classes [2, 131, 130]. Common datasets used in the literature for this purpose are ImageNet [33] and Places [144].

Another important aspect to be taken into account is the use of scene and object-level information. The works proposed in [130, 93, 131] demonstrated that appropriately blending these two levels of information can provide significant improvements in event recognition. Most approaches tend however to fuse the two types of information (i.e., object-level and scene-level) assigning them equal contributions (e.g., weights). In our view, this is sub-optimal as images of events exhibit a diversified set of chromatic and spatial contexts. Thus, some images may favor scene-level information over object features and vice-versa. In this respect, we believe that, in fusing various CNN models/architectures, personalized weights should be allocated to each model, based on its capacity in representing specific pieces of information and features that are characteristic of the underlying event.

Moreover, little attention has been paid to understanding and analyzing the key visual elements, which are more revealing for a human observer. We think targeting such event salient objects can help in improving the performances of event recognition algorithms.

Furthermore, it is important to note that most of the existing literature on event recognition focuses on the analysis of single images while very few attempts have been made for event recognition in personal photo collections [124, 28]. There are many factors that make event recognition from personal photo albums a very challenging task. Indeed, they may contain ambiguous or irrelevant pictures and are usually annotated at album level. Such problems of the presence of irrelevant photos and the so called weakly-labelled data (at album only) make event recognition in personal photo collections a more challenging task. The existing approaches relying on supervised learning can not cope with such issues.

Another interesting application is to collect and analyze information about natural disasters available on the social network. To this aim, a number of interesting solutions have been proposed to effectively utilize social media for information collection and analyzing the impact of a natural disaster [69, 105, 7]. The literature reveals that Twitter has been heavily exploited for inferring information about different types of events. However, to the best of our knowledge, there is very limited prior work which aims to collect information from multiple platforms for natural disasters, simultaneously. Although collecting information about natural and technological disasters from different platforms is a time consuming task, the combination of different sources (text, images and videos) to one summarized overview can be very useful for users to analyze the impact of a disaster and be a good source of information.

In addition, satellite imagery of the effected area before and after a disaster can be helpful to give a bird's-eye view of the damage incurred. To the best of our knowledge no prior efforts have been subjected by literature to combine information from social media and satellite for event detection.

To sum up, five points are to be highlighted. First, a large scale benchmarking image dataset for the evaluation and comparison of event discovery algorithms from single images is still missing. Second, a thorough analysis of the performances of existing deep models, particularly a proper utilization of object and scene-level information for event recognition, is still missing. Third is that, to the best of our awareness so far, event saliency has been treated rather scarcely in the relevant literature. The fourth point is that very few attempts have been made for event recognition in personal photo collections. The final element regards a proper combination of satellite data and social media information that to the best of our knowledge, has not been properly investigated in prior works. This can help to tell a much broader story of a disaster.

1.3 Research Contributions

The main research goal of this thesis is to move forward the state-of-the-art on event based analysis of multimedia contents for an efficient multimedia indexing and retrieval. Here, we briefly discuss the main research contributions of this work.

1.3.1 Event Recognition in Single Images

UNITN Social Event Dataset (USED)

There is an ongoing trend of image representation, which derives benefits from deep neural architectures. However, the existing benchmark datasets for event discovery in single images are not large enough to be used for training deep learning algorithms. Therefore, it is essential to establish a large-scale bench-mark dataset for the evaluation and comparison of event discovery algorithms in single images. The key contributions in this regard are:

- We provide a large collection (around 490,000 images) of annotated event related images.
- Moreover, a deep learning based approach is introduced into event discovery from single images as one of the potential applications of this dataset and to set a benchmark.

Ensembles of Deep Models

As aforementioned, there is an ongoing trend to jointly utilize object and scene-level information for event recognition. Existing approaches tend to treat both types of information equally. In our view, this is not optimal as different images exhibit different characteristics. Thus, we believe that, in fusing various CNN models/architectures, personalized weights should be allocated to each model. Thus:

- Three efficient late fusion strategies are presented. We show in particular that an ad-hoc fusion of an ensemble of deep models can outperform, considerably, the use of each individual model.
- Stemming from the fact that deep learning is a staple ingredient in many multimedia analysis and computer vision pipelines, we carry out a thorough analysis of four well-known architectures, pre-trained on object and places datasets, both individually and in different combinations.

A Hierarchical Approach to Event Discovery

In event recognition, both object specific and background information play an important role, depending on the nature of the underlying event. In order to provide a detailed analysis of the importance of both earlier components, we contribute with:

- We investigate full images, which usually contain contextual clues of underlying event in the background, as well as event-salient objects to uncover the event depicted in the image.
- We propose a two-step hierarchical approach where initially full images are utilized for event classification, and then event-salient features are exploited to further refine the classification decision.

A Saliency based Approach to Event Recognition

Event-related images usually contain objects that are more revealing for a human observer. We believe that targeting such event salient objects can help in improving the performances of event recognition algorithms. To this aim in this work:

- We propose a novel framework exploiting event-salient regions.
- We propose and conduct a crowd-sourcing activity to extract event-salient objects and regions.

1.3.2 Event Recognition in Personal Photo Collections

The conventional approaches relying on supervised learning methods lack in dealing with non-relevant images in photo albums annotated at album-level only. To this end, in this work:

- We propose a novel pipeline relying on MIL paradigm for event recognition in personal photo collections.
- We provide a detailed analysis of the trade-off between classification performance and computational cost.
- We also collect an image dataset containing a large number of photo albums per event.

1.3.3 Natural Disasters Events

We believe that a proper fusion of social media and satellite imagery can help to provide a more detailed story of a disaster. Key contributions in this regard are:

JORD: A System for collecting Information and Monitoring Natural Disasters

- We present a system that is able to automatically collect information and news items about natural disasters from social media, and links it with satellite imagery in real time.

- We also provide query refinement by automatically generating queries in all local languages that are relevant to the position of a disaster.
- In addition, to ensure the quality of the retrieved multimedia data, we propose a hierarchical content based filtering mechanism.
- It is equipped with a novel method for linking and retrieving satellite imagery with the events by analyzing the tweets text to identify and extract GPS coordinates of the areas struck by the disaster.
- JORD also consists of a novel framework for flood detection in satellite images as a use-case of the disaster event detection in satellite imagery.

Medieval Bench-marking

- We propose Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) based satellite and social media fusion for natural disasters detection.

1.4 Structure of the Thesis

This dissertation is organized as follows:

- Chapter 2 provides a detailed description of the event recognition in single images, our proposed solutions along with conducted experiments, the results achieved, and their description and comparisons against the state-of-the-art.
- Chapter 3 is mainly devoted to event recognition in personal photo collections. It provides the details of proposed solutions, conducted experiments and experimental results. It also covers the details of our newly collected dataset for event recognition in personal photo collections.

- Chapter 4 shifts the analysis from daily life events to natural and technological disaster events. It presents a detailed description of our proposed system JORD and the approaches we proposed for a benchmark competition in MediaEval 2017.

- Chapter 5 draws the conclusions of this research work and discusses the possible future directions.

Chapter 2

Events in Single Images

2.1 Introduction

In recent years, event based analysis of multimedia content got great attention of the research community. To this aim, a number of interesting solutions have been proposed for event recognition. 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 [72] and the recent trend is to shift towards the analysis of visual information.

The state-of-the-art in visual-based event detection has so far revealed considerable uncertainties and poor classification performances. We believe that such limitations can be mostly attributed to the selection of the visual features used for classification. Recently Convolutional Neural Networks (CNNs) have been proven efficient in various application domains (e.g., object recognition and remote sensing). Based on these considerations, we believe that deep learning can prove to be a breakthrough also in this research area, providing a more detailed and complete description of the visual content, and bringing the quality of the analysis one step closer to the performances of a human observer, who, in this area, still demonstrated to outperform automatic systems. The main limitation of CNNs is their requirement of a large number of annotated samples, which is the main hurdle for its applicability in event discovery from single images. The existing benchmark datasets for event discovery in single images are not large enough to be used for training deep learning algorithms, in particular convolutional neural networks.

Another important aspect to be taken into account is the proper use of scene and object-level information. As demonstrated in [130, 93] a blend of these two levels of information can provide significant improvements in event recognition. However, most of the existing approaches tend to treat the models for these different types of information equally by assigning equal weights to each type of information. In our view, this is sub-optimal as images of events exhibit a diversified set of chromatic and spatial contexts, then, some images may favor scene-level information over object features and vice-versa. Thus, we believe that personalized weights should be allocated to each model, based on its capacity in representing specific pieces of information and features that are characteristic of the underlying event.

Moreover, the current research in visual information-based approaches to event recognition mostly focus on defining better representation and classification schemes. However, no significant efforts have been made to understand and analyze the event-salient visual elements in event-related images. We believe, a proper use of such objects can help in improving the performances of event recognition algorithms.

Based on these considerations, we proposed some solutions to mitigate the limitations of state-of-the-art. The rest of the chapter presents our proposed solutions for event recognition in single images.

2.2 Related Work

The mainstream of the literature on event analysis is largely developed along two main streams: (i) evidencing the most adequate representations schemes [130, 49], and (ii) establishing discriminative classification paradigms [124].

In this regard, numerous contributions suggest exploiting additional information associated to the multimedia data (e.g., [32, 95, 24, 92, 39]). This additional information may include user-supplied tags, titles, owner and upload information along with comments from users. Moreover, geolocation and temporal references could play an important role in event recognition. Though such additional data, manifested in the form of metadata, has proven to be effective in event recognition, they also come with many practical limitations, which make their use questionable [124]. These challenges include wrong or no settings of camera's time zone, missing time-stamps and modification of tags. Moreover, the ambiguous meaning of user-supplied tags also affects the performance of event recognition methods relying on metadata [72].

Considering these limitations of metadata, visual content can be regarded as a valuable information for event recognition [32, 42]. For instance, Dao et al. [31] proposed an event-type specific representation for event related images incorporating three different types of features, namely GIST [89], time and visual salient features [51]. Similarly in [42], visual information are used for event detection in videos. However, most of the earlier works in this domain rely on hand-crafted visual features, such as SIFT [74] and SURF [**?**], which cannot cope with the gap between image features and event semantics [124]. To cope with such issues, Tsampoulatidis et al. [123] 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 [67], 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. However, the literature shows that considerable improvements are still possible by attacking the problem from different viewpoints: introducing more discriminative visual features, improving models, mixing content and context [124, 82].

More recently, from the viewpoint of feature extraction tools, deep neural architectures demonstrated cutting-edge performance in the multimedia analysis, and proved to be effective in a variety of application scenarios. Similar to other computer vision applications [57], the mainstream approaches to event recognition tend to capitalize on CNN architectures, exploiting both their capability of offering significant and compact representations of data, and their superior classification performance [93, 106, 129]. Due to the unavailability of large scale event-related datasets, most of the efforts that are being spent in this domain concerning the optimization of fine-tuning strategies to be applied to existing pre-trained models [2, 130]. For instance, in [2], a pre-trained model [66] is fine-tuned on a newly collected dataset covering 14 different social events. In [130], three different network architectures, namely AlexNet [66], VGGNet [110] and bn-inception [56], have been exploited for event recognition.

An ongoing trend is to jointly utilize the prior models pre-trained on ImageNet [33] and Places dataset [144] for event recognition. More in detail, a network pre-trained on ImageNet is expected to focus on object-centric information, while a network pre-trained on the Places dataset shows a better response to scene-level information. Interesting solutions have been proposed to efficiently map CNN models pre-trained on ImageNet and Places datasets onto event recognition [130, 93]. For instance, in [49], object and scene-level information are used for event recognition in a hierarchical way. These approaches adopt simple fusion strategies relying on equal weights for object and scene-level information, which may not be convenient in general, as object cues may dominate over scene information and vice-versa, depending on the dataset. This suggests that allocating optimized weights, tailored to different deep models may significantly improve the performance.

It is also 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 made by Rosani et al. [104], 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. It is important to mention that the concept of event saliency is different from the conventional visual saliency [77].

Table 2.1 summarizes the most relevant literature on event recognition, reporting the features, the datasets and the fusion mechanisms adopted by each approach.

2.3 Solutions

2.3.1 Ensembles of Deep Models

Overview

In this part of the work, we propose to fuse different CNN models pre-trained on ImageNet [33] and Places [144] datasets, exploring the capabilities of three late fusion strategies: (i) Induced Ordered Weighted Averaging (IOWA); (ii) Genetic Algorithms (GA); and (iii) Particle Swarm Optimiza-

Table 2.1: Summary of some relevant works in event recognition: overall objectives, features and dataset used for validation, fusion schemes, and a brief description of the method.

Refs.	**Features**	**Fusion**	**Dataset**	**Notes**
[72]	Meta-data and visual features	Early	MediaEvent [120]	Concatenates different types of information (meta-data and hand-crafted visual features)
[104]	SURF with BoW	N/A	SED2013 (Subset)	The concept of event saliency was introduced with some initial experiments. Provides very few event salient objects for 14 events, only
[88]	SIFT, BoW	Early	SED2013 [102]	Low level features along with meta-data
[20]	Textual and hand-crafted visual features	Hybrid	Soccer and Concerts Event [91]	Firstly, aggregates the textual features which are later used in combination with visual features in a hierarchical way
[49]	CNNs features	Late	PEC [19]	It uses object and scene-level information obtained via a single architecture in a hierarchical way. It also relies on late fusion with equal weights for event recognition in personal photo collections
[100]	CNNs features	Late	USED [2] and WIDER [138]	Equally treats object and scene-level information from two architectures
[130]	CNNs features	Late	UIUC [70], Cultural Events [38] and WIDER	Focuses on other aspects of object and scene-level information obtained from deep and very deep architectures while simply adopting equal weights mechanism for fusion
[138]	CNN features	Early	Cultural Events [38]	Fuses features extracted from different layers of a CNN
[4]	CNN features	N/A	SED [102]	Provides a hierarchical approach with event salient objects and full images. Does not divide the test image into regions
[71]	CNN features	Late	WIDER	A combination of models of an architecture fine-tuned on full images and image regions.

tion (PSO). Furthermore, with respect to the proposed fusion schemes, we assess the impact of various combinations of ten different CNN models, from four commonly used deep architectures. We also evaluate the generalization capabilities of the proposed framework by testing a dataset with the weights learned on the other. Finally, we exhaustively validate the performance of a variety of deep models along with their respective per-class rates. Such rigorous analysis will contribute to creating a benchmark for future deep learning based event recognition research.

In summary, we can synthesize the main contributions of this work as:

(i) Through the introduction of three efficient late fusion strategies, we demonstrate that the ad-hoc fusion of an ensemble of deep models can considerably outperform each individual model.

(ii) Stemming from the fact that deep learning is a staple ingredient in many multimedia analysis and computer vision pipelines, we carry out a thorough analysis of four well-known architectures, pre-trained on object and places datasets, both individually and in different combinations.

(iii) We carry out thorough experiments on three challenging benchmark datasets and show that better scores are achieved as compared to recent literature.

Proposed Methodology for the Ensembles of Deep Models

Figure 2.1 displays the block diagram of the proposed framework. It is essentially composed of two stages: (i) feature extraction and classification, and (ii) score-level fusion. In the first stage, CNN features are extracted by means of a bunch of pre-trained CNN models, and the extracted features are fed into an ensemble of Support Vector Machines (SVM), which provide classification scores in terms of posterior classification probabilities. In the second stage, the obtained posterior probabilities are fused using the proposed late fusion schemes. The first stage is rather standard, and we mainly focus on the second part of the methodology. We opted for SVM on the account of its proven efficiency in many applications, such as object recognition [21] and remote sensing [16]. In the next subsections, a detailed description of both stages is provided.

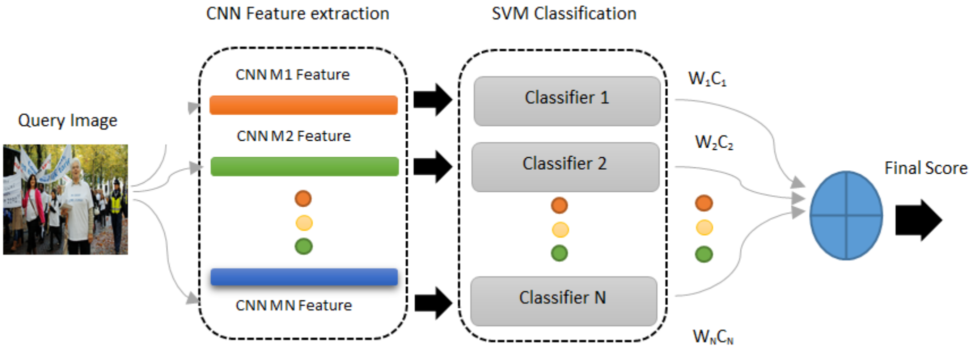


Figure 2.1: Block diagram of the proposed framework for Ensembles of deep models.

Feature Extraction and Classification

In order to conduct a thorough analysis and evaluation of deep features for event recognition, we made reference to the four most commonly used architectures in this domain, namely AlexNet [66], GoogleNet [114], VGGNet [110], and ResNet [53]. Each network was pre-trained on both object and Places datasets. AlexNet consists of 8 weighted layers, while GoogleNet is composed of 22 layers. VGGNet and ResNet are available in different configurations. In this study, we opted for both configurations of VGGNet (VGGNet16 with 16 layers, and VGGNet19 with 19 layers), while for ResNet we evaluated the configurations containing 50, 101 and 152 layers. Table 2.2 summarizes some characteristics of these CNN architectures while a detailed description is available in the literature [66, 114, 110, 53].

For feature extraction with AlexNet, VGGNet16 and VGGNet19, we

Table 2.2: Summary of the properties of the CNN models used in this work

CNN Model	**# Layers**	**# Parameters**	**Properties**
AlexNet	8	60 M	5 convolutional and 3 fully connected layers. GPU-based implementation of the convolution operation. Total of 1.5 billions of floating point operations per second (FLOPS)
GoogleNet	22	102 M	network in a network policy, relies on the Hebbian principles and multi-scale processing
VGGNet-16	16	138 M	Stack of convolutional layers are followed by 3 fully connected layers, convolutional layers are composed of filters with small receptive fields, 19.6 billions of FLOPs
VGGNet-19	19	144 M	More deeper than VGGNet-16 with same configurations/techniques. 19.6 billions of FLOPs
ResNet-50	50	25.6 M	Deeper models with less number of parameters compared to other models. Avoids the vanishing/exploding gradients problem with residual learning framework, 3.8 billion FLOPS
ResNet-101	101	44.5 M	More deeper model with same techniques. 7.6 billion FLOPs
ResNet-152	152	60.2 M	More deeper, more parameters and more FLOPs (11.3 billion)

made use of the Caffe toolbox[1]; for GoogleNet and ResNet we relied on Vlfeat Matcovnet[2]. Overall, we extracted ten feature vectors through four different architectures for each image. AlexNet and VGGNet returned a feature vector of size 4096; GoogleNet and ResNet (all configurations) provided feature vectors of size 1024 and 2048, respectively. These features are then used to train individual SVM classifiers, which provide the classification results in the form of posterior probabilities.

Fusion of CNN Models

Coming to the second stage of the methodology, to jointly utilize the results achieved by the ten classifiers, we propose three different mechanisms for score-level fusion. These methods are based on Induced Ordered Weighted Averaging (IOWA) operators, Genetic Algorithms (GA), and Particle Swarm Optimization (PSO). In the next subsections, we describe each method in detail.

Order-Induced Score Fusion

The proposed fusion strategy is learning-free, which entails that the weights are inferred directly from the posterior classification probabilities. It is inspired by Induced Ordered Weighted Averaging Operators (IOWA) by Yager et al. [141]. The essence of the proposed fusion strategy emerges from the fact that different CNN models yield different classification scores for a given set of event images. Thus, it is convenient to fuse their outcomes assigning higher weights to the models that show higher confidence.

Let us assume that N is the number of the adopted pre-trained models, and that an SVM classifier is associated to each model. The total number of event classes at hand is denoted by M. Therefore, an NxM matrix is built, whose entries correspond to the posterior probabilities of an image with respect to all classes. Suppose p_i, $i = 1, 2, 3, \ldots, N$, is the score array

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

[2] http://www.vlfeat.org/matconvnet/

pointed out by the i^{th} classifier. The proposed IOWA fusion strategy aims at gathering an ensemble of pairs $[p_i, o_i]$. Here p_i is called the argument value, while o_i represents the corresponding order-inducing value, which quantifies how confident the score p_i produced by the i^{th} classifier is (i.e., how correct the classification probabilities are). Hence, the IOWA operator, which represents the final decision as a weighted sum of the reordered posterior probability vectors, is given by:

$$F(p_i, o_i) = \frac{1}{N} \sum_{i=1}^{N} w_i s_i \tag{2.1}$$

where $W = [w_i, w_2, \ldots, w_N]$ denotes the respective weights, and $S = [s_1, s_2, \ldots, s_N]$ denotes to reordered $P = [p_1, p_2, \ldots, p_N]$ in descending order (i.e., posterior probability arrays are sorted according to their associated inducing values o_i).

Regarding the selection of the inducing value (i.e., o_i), which gauges the reliability of its corresponding argument value in the form of a probability array p_i, we adopt the standard deviation of the highest values in the array of posterior classification probabilities p_i returned by the i^{th} classifier in order to address the uncertainties in the (e.g., five) highest scores. Thus, the event class with the closest similarity corresponds to the most probable outcome in F.

Genetic Modeling of Deep Features

The second fusion scheme is based on GAs [108]. A Genetic Algorithm is an optimization strategy that seeks for an optimal value (or a set of values) that contribute to the minimization of a given cost function. Differently from the previous approach, GAs require a training procedure: to seek the optimal solution, the algorithm iteratively minimizes the output of the cost function as depicted in Figure 2.2. As can be seen, the process usually starts with an initial randomly generated population. Subsequently, the fitness of each individual in the current generation is evaluated based on the fitness function, and fit individuals are selected for the next generation with modified genome, iteratively. The process terminates either by reaching to maximum number of generations or achieving a sufficient level of fitness. Moreover, the crossover and mutation are the basic operators of GAs to discover local and better (global) minimum/maximum, respectively.

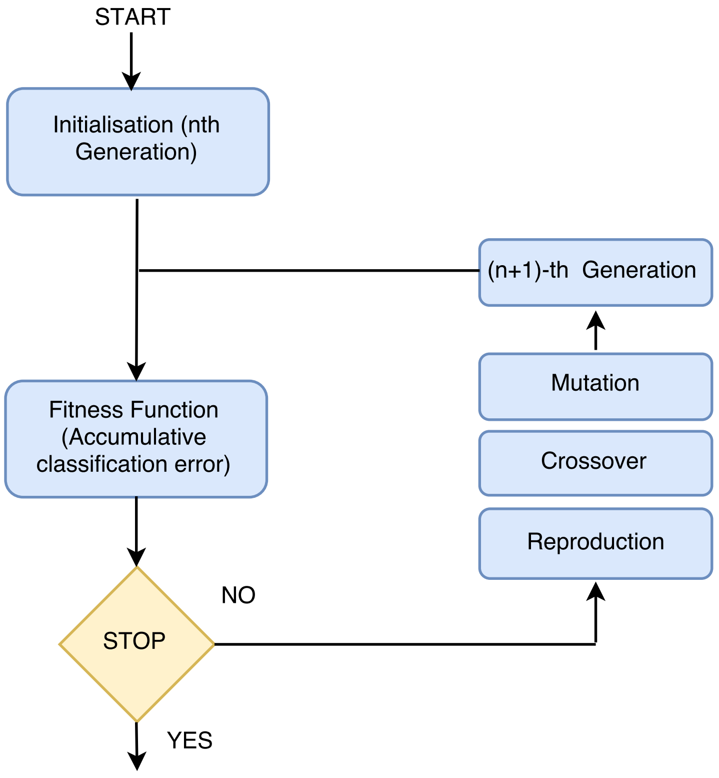

START
Initialisation (nth Generation)
(n+1)-th Generation
Mutation
Crossover
Reproduction
Fitness Function (Accumulative classification error)
NO
STOP
YES

Figure 2.2: Block diagram of the Genetic Algorithm optimization scheme.

Two key components determine the performance of GAs: (i) the definition of a suitable fitness function, and (ii) the structure of chromosomes. In our case, the chromosomes denote the CNN models, while the fitness function aims at minimizing the accumulative classification error. Therefore, we compute the posterior probabilities for each classifier on a validation set in order to infer the weights. Given a test image, the obtained weights are employed to linearly combine posterior probabilities as expressed in Equation 2.2. Here, $w(n)$ represents the weight while p_n represents the probability vector pertaining to the n^{th} model. P is the combined validation posterior probability.

$$p = w(1) * p_1 + w(2) * p_2 + ... + w(n) * p_n \tag{2.2}$$

Based on the combined validation posterior probabilities, we compute the accumulative accuracy on the validation set denoted as A_{acc}. Next, the accumulative classification error is computed according to Equation 2.3, with the goal of finding the combination of weights that minimize the classification error.

$$Y = 1 - A_{acc} \tag{2.3}$$

The weights learned on the training set via cross validation, are then mapped on the test set. Details on GAs parameter selection are provided in the experimental section.

Particle Swarm Optimization

The third proposed fusion method is based on Particle Swarm Optimization (PSO) [15, 43]. PSO, inspired by the social behaviour of birds flocking, is a stochastic optimization technique. Despite the many similarities with GAs, PSO does not involve any evolution operators, such as crossover and mutation, in the optimization process. It rather solves the optimization problem by iteratively trying to improve a candidate solution (particle) with respect to a given criterion. Figure 2.3 shows a block diagram of our PSO-based optimization approach. In total, it is composed of 9 different phases starting from parameter settings, which is followed by a random selection of a candidate solution. In our case, each combination of weights can be referred to as a candidate solution while the fitness function uses the same concept of accumulative classification error used in GAs (Equation 2.3) as a qualitative criteria for a candidate solution. The weight selection encompasses three main steps; (i) evaluation of each candidate solution according to the the fitness function; (ii) updating

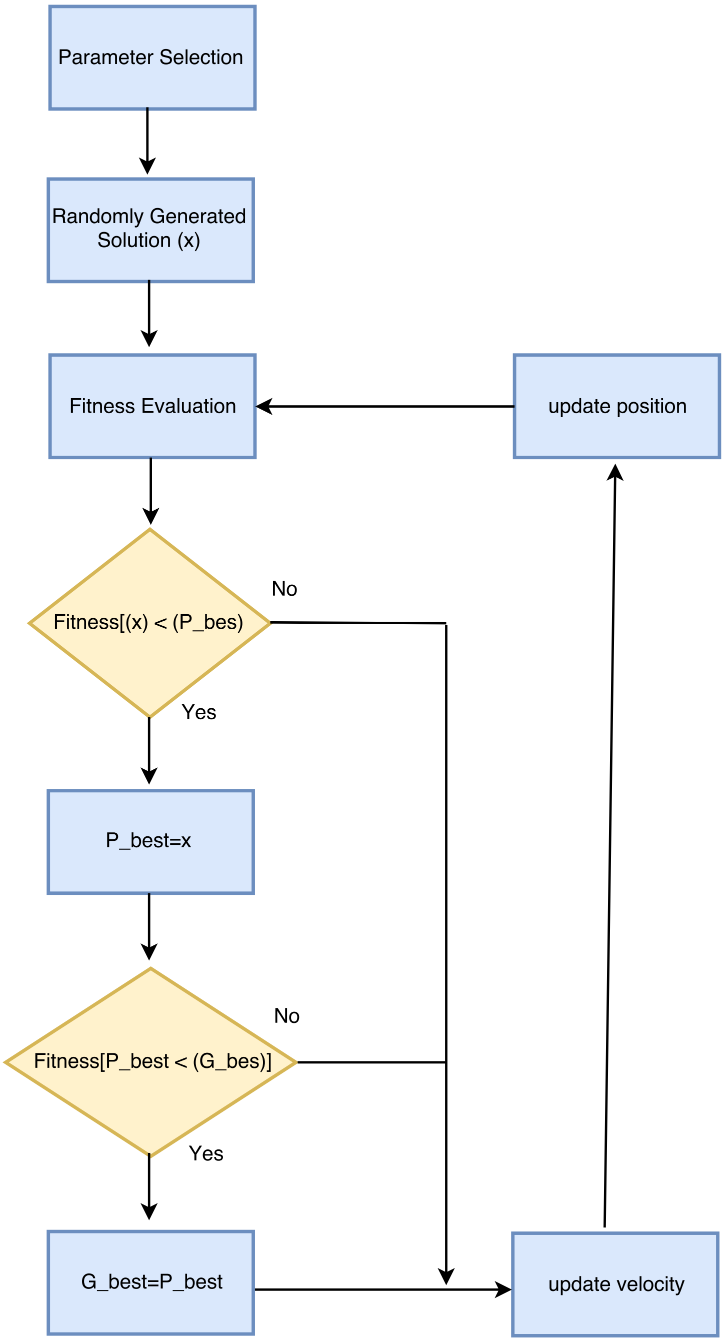


Figure 2.3: Block diagram of the PSO optimization scheme.

the individual best (P_{best}) and global best (G_{best}) fitness and positions by comparing the new fitness output with the previous values; (iii) updating velocity and position of each particle. The final step is mainly responsible for the optimization abilities of PSO, where the velocity of the swarm is updated according to Equation 2.4.

$$v_i(t+1) = w * v_i(t) + c_1 r_1 [x_i'(t) - x_i(t)] + c_2 r_1 [g(t) - x_i(t)] \tag{2.4}$$

Here, $v_i(t)$ and $v_i(t+1)$ represent the velocity of the i^{th} particle at time t and $t+1$, respectively. The position of the i^{th} particle at time t is represented as $x_i(t)$. Similarly, the best individual and global best candidate solutions at time t are represented by $x_i'(t)$ and $g_i(t)$, respectively; w, c_1, and c_2 are user-defined parameters; r_1 $(0 \leq r_1 \leq 1)$ and r_2 $(0 \leq r_2 \leq 1)$ are the randomly generated values for each velocity update.

Also in this case it is necessary to train the optimizer on a pre-defined set of samples, and then transfer the calculated weights to the test set.

2.3.2 A Hierarchical Approach with MIL Framework

Overview

In this work, we provide a detailed analysis of multimedia contents in the context of event recognition in single images focusing on key visual-elements and background information. In details, we investigate full images, which usually contain contextual clues of underlying event in the background, as well as event-salient details, i.e. visual objects critical to understand the underlying event depicted in the image. In particular, we propose a two-step hierarchical approach based on the MIL paradigm where initially full images are utilized for event classification, and then event-salient features are exploited to further refine the classification decision. For event-salient features, event-saliency maps introduced in [104] have been used. The refinement process will benefit from event specific visual objects, which are common in all positive samples of an event. Indeed MIL paradigm will focus only on such features/objects common in all positive samples of a particular event.

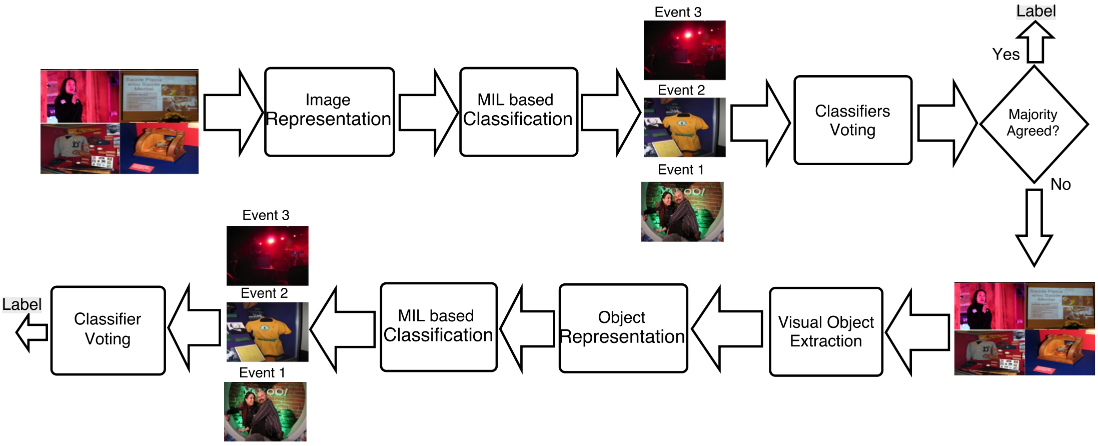


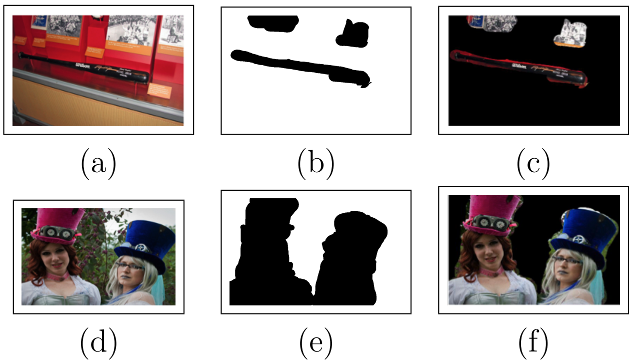
Figure 2.4: Hierarchical approach to event discovery in single images: in the first phase full images are analyzed in Multiple-instance learning framework for event detection. In the second phase event-sailent details extracted as in [104] are exploaited for the refinement of the classification.

Proposed Methodology

Figure 2.4 shows the block diagram of the proposed approach. As shown in the block diagram, both phases (upper and lower rows) differ at the first stage. In the first phase, Colour Structure Descriptor (CSD) [81] is used for the representation of full images with background information while in the second phase, first we used event-saliency maps [104] to extract event-related visual objects. Subsequently, CSD descriptor is used for the representation of these event-related visual objects that is then followed by MIL based classification. To deal with multi-class classification problem, in both phases we adopted one-against-one strategy where decision is made on the basis of majority votes. All steps of the proposed approach are discussed below.

Visual Cues Extraction

As aforesaid, in the refinement step of the proposed method, we exploit visual information not about the whole image but representing the so-called event-salient parts of a picture. Indeed, it has been demonstrated



(a) (b) (c)

(d) (e) (f)

Figure 2.5: Visual objects extraction: (a) and (d) represent original images; (b) and (e) represent their corresponding event-saliency maps; visual objects, extracted using event saliency maps, are shown in (c) and (f).

in [104] that there exist special areas in event-related images which contain key visual elements allowing human observer understanding the depicted event. Such information is coded in event-saliency maps, as reported in the two examples of Figure 2.5. An event saliency map, which is different from the traditional concept of visual saliency [65], does not focus on perceptual prominence but rather on event-related semantics of media. Event-saliency maps have been created using crowd intelligence [104], where a gamification technique with a large number of users has been proposed. In the game, users are involved in competitive roles where one is asked to prevent the other from recognizing the underlying event by hiding the key visual elements, which help the human observer to recognize the underlying event. Subsequently, binary maps are created of the images by highlighting the parts containing event-salient information. Starting from these maps, we can extract specific visual objects in images which contain important clues for the detection of underlying event. In the refinement phase (second phase), we used these event-related visual objects contained in the images as instances of positive and negative bags for training purposes in the MIL paradigm.

Image Representation

This step is common in both phases of the proposed hierarchical approach. To this aim, we use Color Structure Descriptor (CSD) [81]. We use CSD because it encodes both information: about the spatial structure as well as colors occurrence frequencies in an image using structure elements. The number of structure elements used in CSD is usually 64. The biggest advantage of CSD descriptor is that it uses no more space than ordinary color histogram with significant improvement in performance [81].

Classification with MIL

In this work, we treat event recognition as a multiple instance learning problem. The basic motivation comes from the fact that an event-related image may contain multiple event-salient objects, and we do not know which of them is responsible for the label of the image. To this aim, we rely on the Citation K-Nearest Neighbor (C-KNN) based implementation [128] of the MIL paradigm.

In order to map our event recognition task into MIL paradigm, we collect images/event salient objects into bags, where each image (in the case of full images) or event salient objects (in the case of event salient features) is considered as an instance of the bag. Moreover, label is assigned to complete bag instead of individual instances. For the prediction of an event class/bag b, both R-nearest references (bags in its neighborhood) and C-nearest citers bags, which consider b as their own neighbor, are considered. For reference bags simply the k-nearest samples are taken into account while for the selection of citer bags a ranking mechanism is adopted. Suppose n is the number of all samples in a database B_s, represented as $B_s = \{b_1, b_2, b_3....b_n\}$. Then, for a test sample $b_i \in B_s$, the rest of the samples are ranked according to the similarity to the 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, the C-nearest citers are defined as:

$$Citers(b_i, C) = \{b_j | Rank(b_j, b_i) \leq C, b_j \in B_s\}. \quad (2.5)$$

For the similarity measurement, a bag-level distance metric known as Hausdorff distance [128], which is the shortest distance between any two instances from each bag, is used.

After the summation of R-nearest references and C-nearest citers in terms of positive bags (denoted as b_p) and negative bags (denoted as b_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), otherwise as negative as show in Equation 2.6.

$$b_{Label} = \left\{ \begin{array}{ll} 1 & \text{if } b_p > b_n \\ 0 & \text{otherwise} \end{array} \right\} \tag{2.6}$$

Finally, in order to deal with multi-class classification, we adopt the one-against-one strategy where the final classification is made on the basis of majority voting. It must be noted that the only difference between classification with MIL framework in the first and second phase is the type of information used in positive and negative bags. In the first phase full images are used as instances of bags while in the second phase only the event-related visual objects contained in the images are used as instances of bags.

Decision Making Module

In the proposed hierarchical approach, the decision is taken as follows: first, the analysis is done on the whole image where the background information can strongly help the classification of the depicted event. If the

MIL framework recognizes already a clear majority agreement among binary classifiers, then the decision is taken and the image is classified. Otherwise, we proceed with the refinement step which takes into consideration only event-salient visual objects in the image (i.e., bags of event-related visual objects are provided as negative and positive samples to C-KNN) and the final decision is made on the majority test. This way, we aim at first recognizing events utilizing complete images. Since some events may share very common backgrounds (e.g., concert and theater images usually have similar backgrounds) so in such cases in the second step we exploit the new concept of event-saliency to improve the classification based on event-related visual clues (e.g., music instruments are better visual clues for concert events).

2.3.3 A Saliency-based Approach

Overview

In this work, we aim to investigate how the objects that are more revealing for a human observer can be utilized for an automatic event recognition in single images? and how these objects can be extracted from event-related images? To this aim, we propose a novel framework that exploits event salient regions in a Multiple Instance Learning (MIL) paradigm. We propose and conduct a crowd-sourcing 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. Our choice of choosing a crowd-sourcing task is motivated by the need of finding 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 the work is at the classification stage, where a Multiple Instance classification (MIL) framework is adopted for the classification of a test image at the hand on the basis of the extracted regions. Multiple Instance Learning (MIL) and classification has proven to be very effective in many applications [109], and well fits our needs for region-based approach to event recognition.

Proposed Methodology

As can be seen in Figure 2.6, 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 crowd-sourcing task. The crowd-sourcing activity is followed by feature extraction, where we use a pre-trained network VGGNet16 [110] 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 strategy to deal with multi-class classification, where the final decision is made on the basis of majority voting. In the following sub-sections, we provide a detailed description of the each stage of the proposed approach.

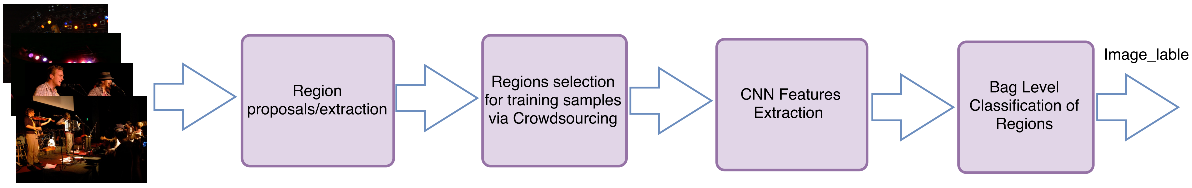

Region proposals/extraction
Regions selection for training samples via Crowdsourcing
CNN Features Extraction
Bag Level Classification of Regions
Image_lable

Figure 2.6: Block diagram of the proposed event-salient regions-based methodology.

Region Extraction and Pre-filtering

Images often contain information provided by the presence of 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 for event recognition, following the data driven selective search approach introduced in [125], 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 that a significant number of the regions obtained through Selective Search [125] are irrelevant with respect to certain events, as well as for the event classification problem itself (i.e., they are not enough discriminative). Moreover, processing more regions per image requires more processing resources and time. To this aim, we first introduce a filtering stage, to 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). The pre-filtering phase helps to reduce the processing time by dropping the less informative regions from both test and training samples. This assumption is verified by our initial experiments on the validation set, which shows that the initial filtering reduces the processing time considerably without any significant impact on the classification results.

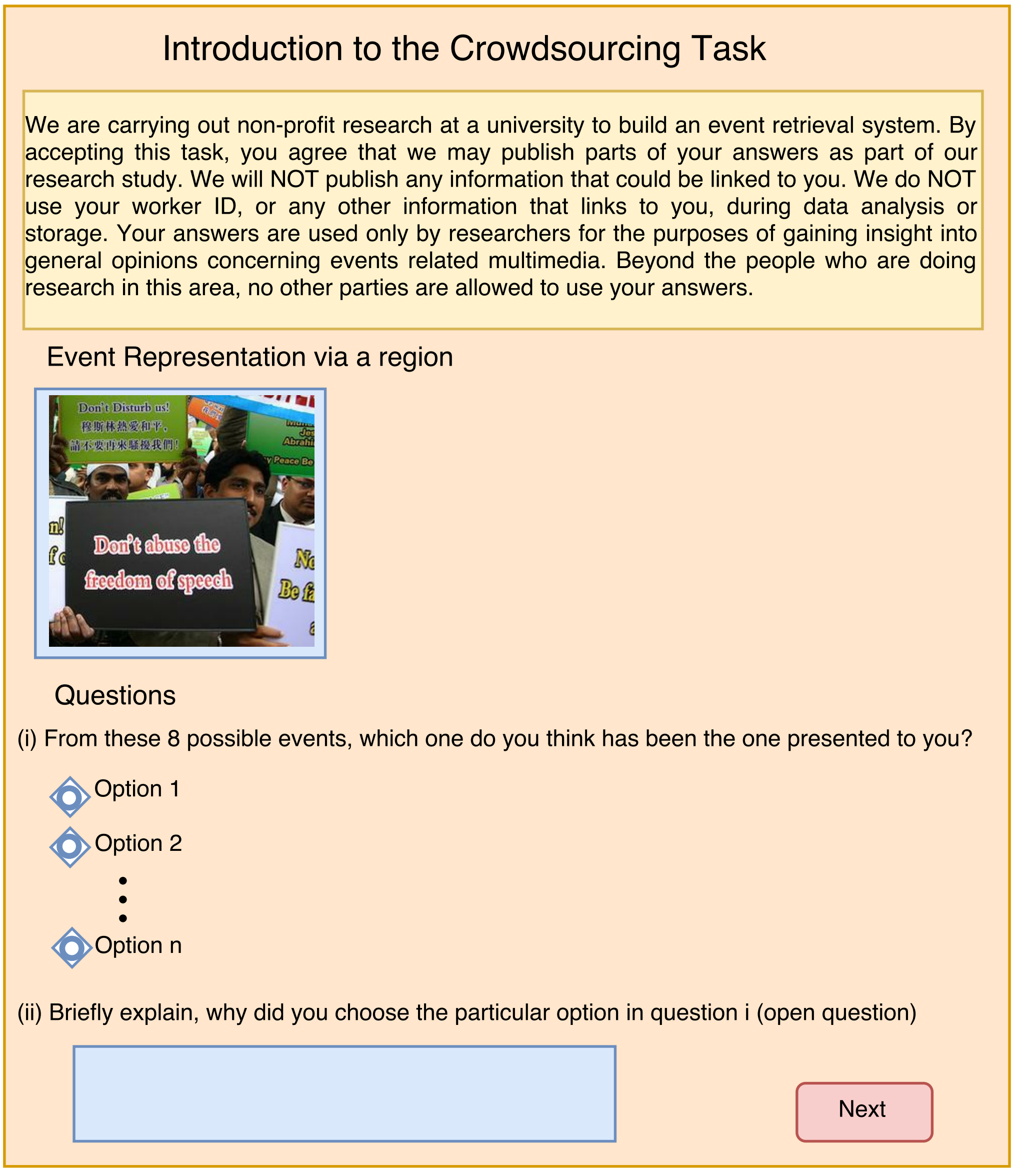
Salient Regions Selection via Crowd-sourcing

In the pre-filtering phase, we remove some regions based on the size and width-height ratio of the regions. However, still there are a number of regions which are either not enough discriminative 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 [104]. To this aim, in order to select more relevant image-regions for training purposes, we propose and conduct a crowd-sourcing study where a large number of volunteers are engaged. In order to reduce efforts in the crowd-sourcing task, we randomly choose candidate regions from the pool of the extracted regions for the crowd-sourcing study.

In the crowd-sourcing 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 [103], we tried to keep the task as simple as possible, making sure that most of the answers can be considered reliable.

Figure 2.7 depicts the design of the proposed crowd-sourcing task. In the proposed 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



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)

Figure 2.7: The design of the crowd-sourcing 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 crowd-sourcing activity for annotation purposes. Two different questions are posed regarding the shown regions: the first one asks to select the event class, the second one is meant to receive a motivation for the selection.

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.

The second question is an open question where the volunteers need to briefly motivate their choice. This question aims to get feedback from the users about the visual contents (i.e., objects and regions) that help humans to perceive the underlying event in an event-related image. Moreover, a selection of users’ answers has been inspected manually. This question has been useful to evaluate the reliability of the volunteers’ answers participating in the study.

In Figure 2.8, we illustrate the whole 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.

Feature Extraction

The current literature in event recognition reveals that image representation schemes deriving benefits from deep neural architectures, namely 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 object (ImageNet [33]) and places [144] datasets 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 datasets [49, 131]. 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 opt for VGGNet16 [110] pre-trained on ImageNet [33], which focuses on object centric information. VGGNet16 is composed of a total of 16 layers. For further details about

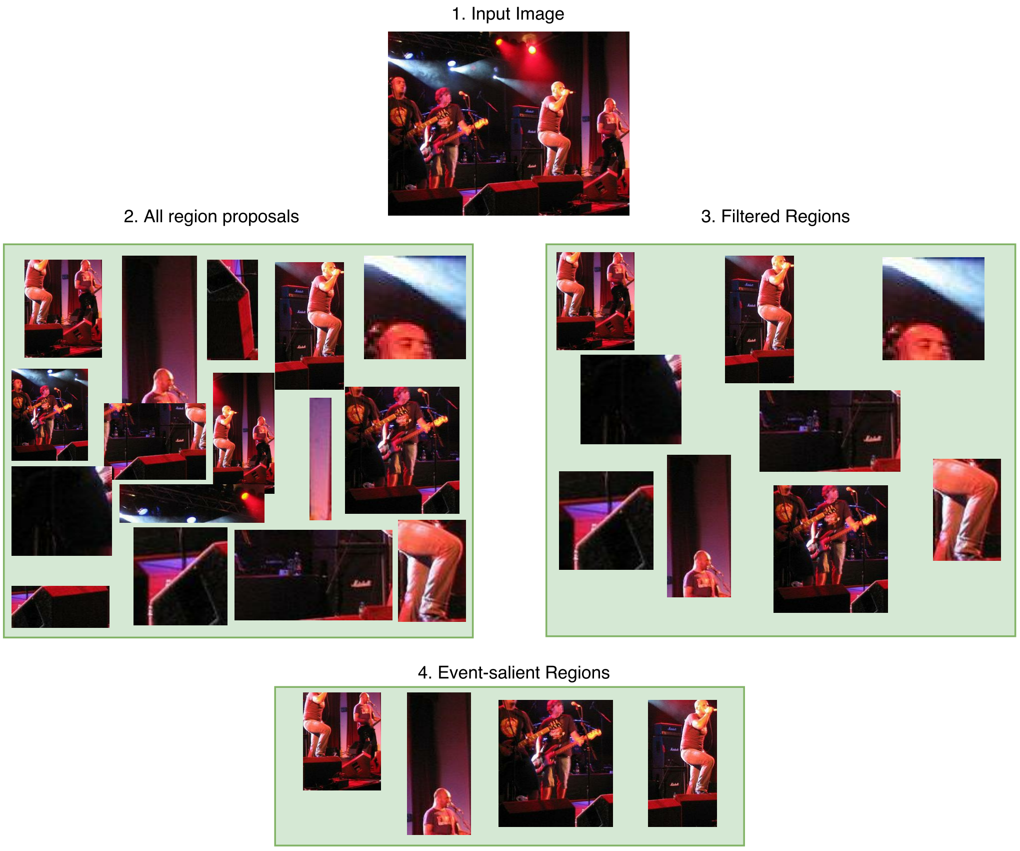


Figure 2.8: Region selection process: (1) represent an input image; (2) shows sample region proposals; (3) provides some sample regions after filtering phase while (4) represents sample event-salient regions obtained through crowd-sourcing study from the input image.

the network architecture, please refer to [110]. From each image region, we extract a feature vector of size 4096.

Multiple Regions based Classification of an Image

As mentioned earlier, in the proposed approach, we divide the training as well as the test images into a number of regions of different sizes, and then classify the image at hand on the basis of the extracted regions. This problem can be easily mapped into a Multiple Instance Classification (MIC) problem. 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 [10]. Similarly, the test bags are also composed of multiple feature vectors. Moreover, labels are assigned to the bags, only.

In order to map our salient regions-based approach into multiple instance classification, we treat each image as a bag of multiple regions, where each region is treated as an independent instance of the bag. It is to be noted that the crowd-sourcing study is carried out for the training samples, only. On the other hand, for the test samples all the extracted regions that pass the pre-filtering stage from each image are gathered into a single test bag. Moreover, in the training bags we put image-regions randomly from the pool of regions obtained in in the crowd-sourcing study, which are not necessary to belong to the same image. On the other hand, the test bags are composed of the image-regions extracted from the same test image, only.

For the bag-level classification of image regions, we use an approach inspired by C-KNN [127], 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 of citer bags is originated from library sciences [41]. The underlying insight is 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 neighbours. However, for the selection of C-citers of the bag a ranking mechanism [127] is adopted. For instance, if n is the number of total samples we have in a database B_s, represented as $B_s = \{b_1, b_2, b_3, \ldots, b_n\}$, 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:

$$Citers(b_i, c) = \{b_j | Rank(b_j, b_i) \leq c, b_j \in B_s\} \tag{2.7}$$

For the similarity measurement among bags, a bag-level distance metric, the Hausdorff distance [127], is used. For instance, for the comparisons of two bags X and Y, the Hausdorff distance is defined as follows:

$$h_k(X, Y) = k^{\text{th}}_{x \in X} min_{y \in Y} \|x_i - y_i\| \tag{2.8}$$

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 [127]. In our case, we opt for the minimal Hausdorff distance (i.e., k = 1) [127].

After the summation of R-nearest references and C-nearest citers in terms of positive bags of image-regions (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), otherwise as negative according to Equation 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 to positive ones in multiple-instance learning paradigm [127]).

$$C_{Label} = \begin{Bmatrix} 1 & \text{if } B_p > B_n \\ 0 & \text{Otherwise} \end{Bmatrix} \tag{2.9}$$

Finally, in order to deal with multi-class classification we choose to 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 votes.

Table 2.3: Event classes covered in USED.

Event Names	**Event Names**
Concert	concert
Graduation	Conference
Mountain Trip	Exhibition
Meeting	Fashion
Picnic	Sports
Sea Holiday	Protest
Ski Holiday	Theater/Dance
Wedding	-

2.4 Datasets

In this section, we provide a detailed description of our self-collected dataset USED along with other different datasets used for the validation of the proposed approaches.

2.4.1 UNITN Social Event Dataset (USED)

Overview

As aforesaid, the existing benchmark datasets for event discovery in single images are not large enough to be used for training deep learning models. To this aim, in this work, we are providing a large collection of event related images covering 14 different types of social events, as shown in Table 2.3, selected among the most shared ones in the social network. Table 2.4 shows the details of different datasets available for the evaluation of event recognition algorithms in single images. As can be seen, our newly collected dataset is the largest among the all in terms of total number of images. The next subsections provide a detailed description of the dataset, its collection and annotation processes and its organization.

Table 2.4: Details of the datasets for event recognition in single images

Dataset	Total Images	Total Event-classes
SED [102]	57,165	7
EiMM [79]	13,219	8
Cultural Events Dataset [38]	11,000	100
WIDER [138]	60,000	61
UIUC [70]	1,579	8
USED (Ours) [2]	490,000	14

Dataset Collection and Annotation

The newly collected dataset is composed of 490,000 images, which are arranged into 14 different types of social events. In order to make it balance, we collected an equal number of images (35,000) per event-class from Flickr using the respective API. The dataset is downloaded between 7th and 20th of September 2015 based on the event-related keywords. Since, in this work, we intend to provide a benchmark dataset for visual analysis of events. Therefore, in order to make sure the quality of the dataset, we removed the outliers and borderline cases manually.

The collected images provide a good variety in terms of contents (e.g., it has indoor as well as outdoor images, single person images and group pictures). We also tried our best to cover every aspect of the considered social events by collecting images for the same events with diverse contents in terms of viewpoints, colours, group pictures vs. single portrait and outdoor vs. indoor images, where the high variability of the represented information can be effectively explored to ensure better performances in event classification. For example, in graduation, sports and wedding event-classes, we collected single person pictures, group pictures and the pictures taken at the time of celebration. Similarly, in ski-holiday and mountain-trip classes, our dataset covers both the pictures taken in green mountains as well as images of white and bare mountains. Another important characteristic of this dataset is the diversity in culture. For example, in wedding image collection, we tried our best to cover diverse cultures by collecting wedding images from different cultures and communities (e.g., we have collected wedding images from both Asian and European countries).

In the context of visual contents, there are certain event-classes which usually overlap with each others. For instance, concert and theater/dance events often have similar visual contents in backgrounds. In such situations, for visual information based approaches to event detection, it becomes difficult to differentiate among such event-classes. Such kinds of images with overlapping contents/concepts have been observed in SED dataset [102]. To cover this aspect of the events, in our dataset, we also provide images having similar contents in the backgrounds with less noise (i.e., images with less resemblance with other classes), where precision in correct association to a class can be achieved by exploiting visual information. Figure 2.9 shows some sample images from the newly collected dataset.

As far as the annotation of the images is concerned, we labeled each image with one of the 14 categories. To facilitate the retrieval and experimentation process, we also provide an event id, representing an event type, to each image in the dataset.

Dataset organization

The collected dataset is made publically available[3]. A user-friendly and attractive interface has been provided to facilitate the research community to download the dataset. As aforesaid the dataset is composed of 14 different social event-classes covering two different benchmarking datasets i.e., EiMM and SED. In order to facilitate the downloading process, we provide each type of event-related images in separate directories as well

[3] http://mmlab.disi.unitn.it/USED/

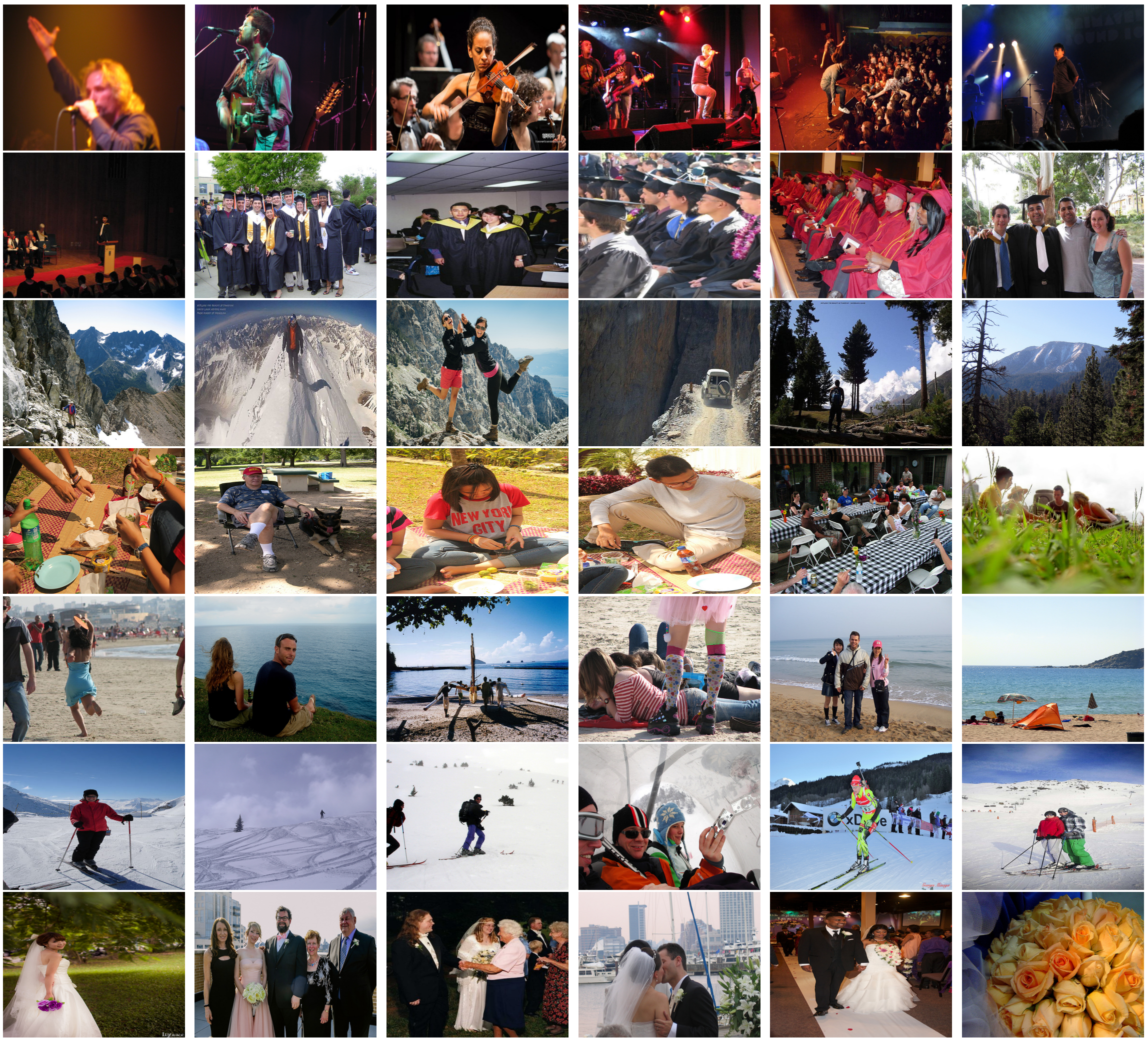


Figure 2.9: Sample images from the newly-collected dataset.

as in single pools of test and training images containing images from all event-classes. Thus, the users have options to download either the whole dataset or selected event-classes according to their needs. We also provide separate CSV files, containing image names and the corresponding event classes and IDs, for each event-class.

2.4.2 Social Event Detection Dataset (SED)

Social Event Detection Dataset (SED) [102] is a large scale benchmark dataset created within the framework of the MediaEval 2013 competition task on social event detection [102]. SED mainly covers 7 different types of social events including: concerts, conference, exhibition, fashion shows, protests, sports events and theater. It is important to mention that sports events are, though composed of different sub events, considered as a single class.

The dataset is provided in two subsets, namely training set and test set. The training set contains a total of 27,754 images, which are collected during 27th to 29th of April 2013. On the other hand, the test set is collected between 7th and 13th of May 2013. It counts a total of 29,411 images. All the images in the dataset are downloaded from Instagram using event-related key words.

SED also provides additional information, such as user's tags, title, description and geo-location information. However, these additional features are not present for all pictures. For example, geo-location information is available only for 27.8% of the pictures. Similarly, 93.4% of images contains title while at least one tag is available for almost all pictures. Figure 2.14 shows some sample images from SED dataset.

2.4.3 EIMM

EiMM Event Detection Dataset [79] mainly targets events related to the personal spher, which are divided into two different categories (i) sport events and (ii) social events. Table 2.5 provides the details of events classes from both categories of EiMM dataset. All the images in the dataset are downloaded from Picasa Web Album service[4], and annotated through hu-

[4] http://picasa.google.com/

Table 2.5: Details of EiMM Event Detection Dataset

Sports Events		**Social Events**	
Baseball	Golf	Concert	Graduation
Basketball	Hockey	Mountain Trip	Meeting
Motor Bike	Rowing	Picnic	Sea Holiday
Cycling	Skatting	Ski Holiday	Wedding
Swimming	-	-	-

man annotators. The dataset also divides some events (e.g., wedding, graduation and sea holiday) into sub-categories by providing sub-category labels. For instance, wedding images are further divided into group pictures, ceremony, party eating and unknown.

2.4.4 Web Images Dataset for Event Recognition (WIDER)

Web Images Dataset for Event Recognition (WIDER) [138] is one of the most recently introduced datasets for the evaluation of event recognition paradigms. WIDER holds a total of around 60,000 images from 61 different event classes. It stems as most complex benchmark for event recognition in still images to date covering diverse event classes. For instance, it contains event categories from sports (such as football, basketball and tennis), daily life events (such as shopping and meeting) and social events (such as concert, celebration and funeral). Moreover, it also covers some specific events, such as demonstration, riots, surgery and soldier marching and drilling. Most of these event classes are taken from Large Scale Ontology for Multimedia (LSCOM) [86]. Table 2.6 lists the event names covered by WIDER dataset. Each event class contains a significant number of images in both training and test sets. All the images are downloaded from Flickr using the corresponding Api. Figure 2.14 shows some sample images from WIDER.

Table 2.6: Events covered by WIDER dataset

Event Names	Event Names	Event Names
Parade	Handshaking	Demonstration
Riot	Dancing	Car Accident
Funeral	Cheering	Election Campaign
Press Conference	People Marching	Meeting
Group	Interview	Traffic
Stock Market	Award Ceremony	Ceremony
Concerts	Couple	Family Group Photo
Festival	Picnic	Shoppers
Soldier Firing	Soldier Patrol	Soldier Drilling
Spa	Sports Fan	Students Schoolkids
Surgeons	Waiter, Waitress	Worker Laborer
Running	Baseball	Basketball
Football	Soccer	Tennis
Ice Skating	Gymnastics	Swimming
Car Racing	Row Boat	Aerobics
Balloonist	Jockey	Matador Bullfighter
Parachutist Paratrooper	Greeting	Celebration, Party
Dresses	Photographers	Raiar Racing
Rescue	Sports Coach Trainer	Voter
Angler	Hockey	People Driving Car
Street Battle	-	-

2.4.5 UIUC Sports Event Dataset

UIUC Sports Event Dataset (UIUC) is comparatively a small dataset released by Li et al. [70]. UIUC is considered as one of the oldest datasets made publically available for the evaluation of event recognition paradigms. It mainly covers 8 sports events, namely badminton, rowing, polo, bocce, snowboarding, croquest, sailing and rock-climbing. The dataset is not balanced; providing different number of images per event class. Table 2.7 provides the details of the number of images in each class of the dataset.

The dataset also provides some additional information in terms of the complexity in recognition on the basis of human subjective judgment for each image. The images from each class are divided into three different categories, namely easy, medium and complex images. Moreover, the distance of the foreground for objects are also provided. Figure 2.14 provides some sample images from UIUC Sports Dataset.

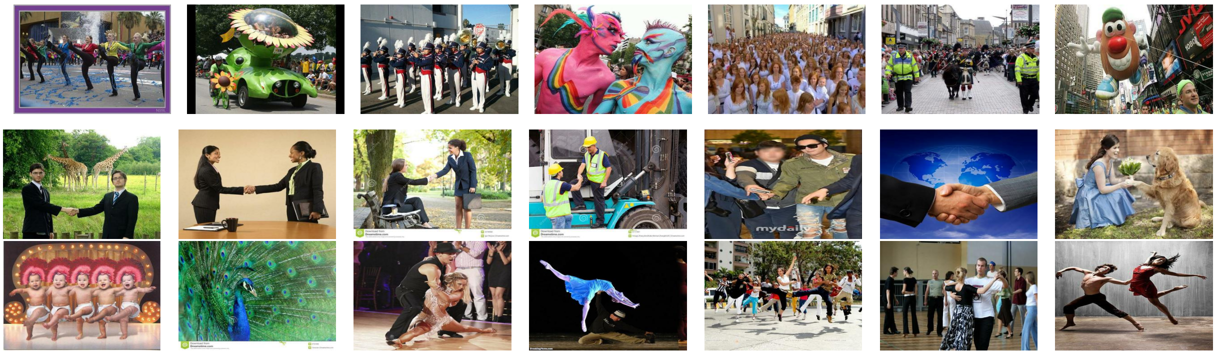


Figure 2.10: Sample images from WIDER dataset

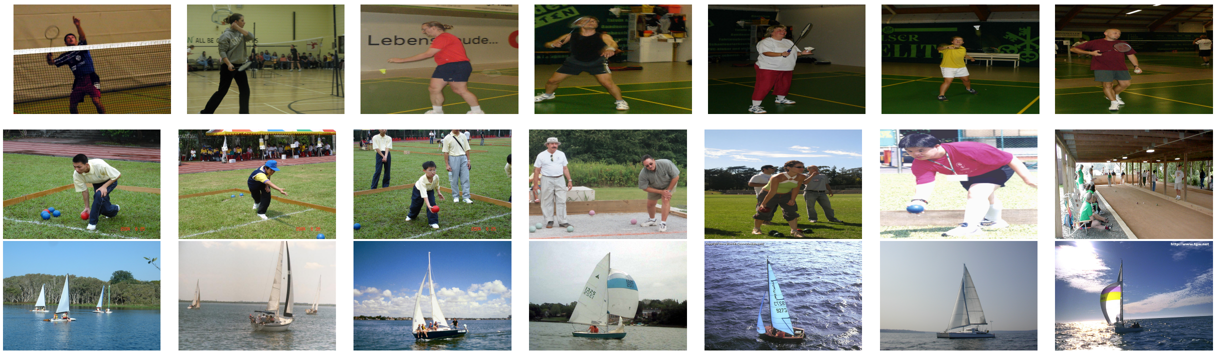


Figure 2.11: Sample images from UIUC Sports dataset

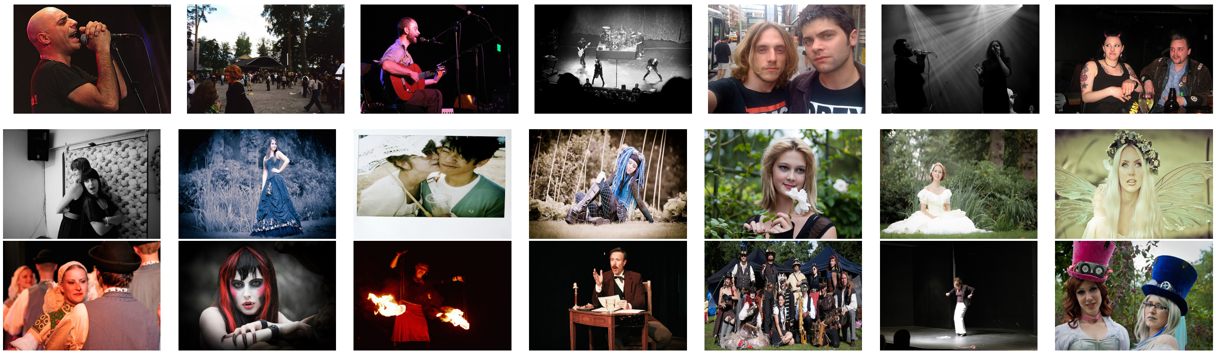


Figure 2.12: Sample images from SED2013 dataset

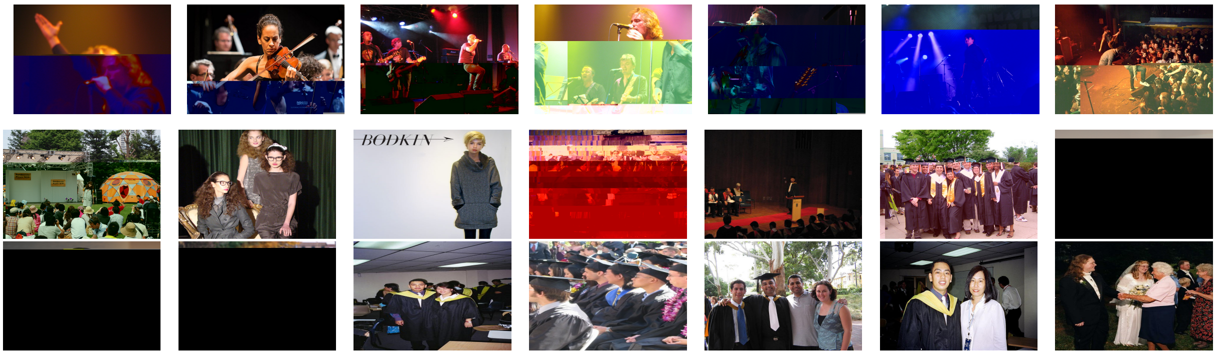


Figure 2.13: Sample images from USED dataset

Figure 2.14: Sample images from all datasets

Table 2.7: Details of UIUC Sports Dataset

Event Name	# Images	Event Name	# Images
Badminton	200	Polo	182
Bocce	137	Rowing	250
Croquet	236	Rock-climbing	194
Snowboarding	190	Sailing	190

2.5 Experiments

2.5.1 UNITN Social Event Dataset (Benchmark)

In this section, we provide a detailed description of the experimental evaluation along with distribution of our self-collected dataset into training, validation and test sets. Experimental results of the basic experiments with a state-of-the-art CNN on the dataset to set a benchmark are also discussed in detail.

Data Assemblage

In the experimentation process, the newly collected dataset is divided into 3 subsets, namely training, validation, and test sets by randomly selecting images for each phase. For training (fine-tuning) of the Convolutional Neural Network (AlexNet [66]), we used 20,000 images per class while for validation and test purposes we used 7,000 images per class for each phase. The validation set is used to estimate how well the model has been trained. Thus, we used a total of 140,000 for training/fine-tuning purposes from each subset. As far as the validation and test collections are concerned, we used 49,000 images from each subset in both phases. It is to be noted that, in order to be consistent with state-of-the-art, in these experiments we train/fine-tune a separate model on each subset (i.e., one of the event classes covered in SED and the other on the events of EiMM).

Table 2.8: Confusion matrix of our network on the event-classes belonging to subset 1 (accuracy in percentage).

	Predicted classes								
		Conc.	Gra.	Meet.	Mount. Trip	Pic.	Sea Holi.	Ski Holi.	Wedd.
Actual-class	Conc.	**74.00**	11.27	8.15	0	6.45	.10	0	.01
	Gra.	.24	**66.00**	18.38	0	15.18	.18	0	0
	Meet.	.94	9.38	**78.70**	2.41	7.98	.47	.07	0
	Mount. Trip	0	4.42	0	**67.00**	15.94	2.18	10.44	0
	Pic.	.98	5.65	12.62	8.68	**54.74**	2.97	.08	14.25
	Sea Holi.	.05	.31	1.10	14.32	10.20	**74.00**	0	0
	Ski Holi.	.21	2.15	13.67	30.22	5.48	.24	**48.00**	0
	Wedd.	.44	19.71	26.15	1.04	1.61	.01	.01	**51.00**

Results and Analysis

Experimental results of our CNN based approach to event recognition are reported in Table 2.8 and Table 2.9 on the subset 1 and subset 2 of the newly collected dataset, respectively. On the subset 1 (i.e., concert, graduation, mountain trip, meeting, picnic, sea-holiday, ski-holiday and wedding), we got an overall accuracy of 67% and 65.96% on validation and test sets, respectively. As far as the performance of our trained CNN on the event-classes from subset 2 is concerned, we achieved an overall accuracy of 70.03% on the test set.

For a thorough analysis of the experimental results, we provide confusion matrices of our CNN on both test sets as shown in Table 2.8 and Table 2.9. In Table 2.8, it can be seen that the proposed approach provides good results on all classes of social events. However, some concepts/events are misclassified. The confusion is typically due to the similarity of visual contents, as an example in the case of graduation and meeting, and ski holiday and mountain trip the backgrounds are visually correlated with each other, which causes significant confusion among these event classes. The research community is encouraged to provide novel strategies and efficient repre-

Table 2.9: Confusion matrix of our network on the event-classes belonging to subset 2 (accuracy in percentage).

	Predicted-class							
		Concert	Conference	Exhibition	Fashion	Protest	Sport	Theater
Actual-class	Concert	**91.98**	2.10	2.00	1.70	0	0	2.3
	Conference	.91	**75.70**	9.80	2.24	7.88	3.47	0
	Exhibition	.98	19.58	**58.54**	7.04	.84	2.95	10.01
	Fashion	2.10	9.34	12.17	**65.34**	.61	2.41	8.01
	Protest	.77	9.90	8.62	2.64	**74.58**	3.47	0
	Sports	.34	5.84	4.61	2.81	10.17	**72.21**	4.02
	Theater	14.78	10.18	8.40	12.20	2.47	.05	**51.90**

sentation schemes to tackle such issues. Best performances are achieved on meeting, concert and sea holiday. We have slightly lower accuracy on ski holiday class, which is most of the time confused with mountain trips. Similarly, in Table 2.9, it can be seen that some events are confused with each others, such as concert is confused with conference, exhibition and protest while conference is confused with exhibition. There is no significant miss-classification among event classes in this test set tough, except between exhibition and conference, which are 19.5% times confused with each other, due to the high perceptual correlation between these event classes.

2.5.2 Ensembles of Deep Models

Experimental setup

In this part of the work, the objective of our experimental testing is manifold. We want to assess the performance of each individual CNN model, pre-trained on object and places datasets, along with the performance of different combinations of such models for event recognition. Moreover, we are interested in analyzing the effect of transferring weights from one dataset to another, in order to prove their generalization capabilities. To attain these goals, we performed a sequence of experiments:

- First, we analyze the performance of ten individual CNN models from 4 different architectures on the large-scale dataset WIDER. This experiment basically builds a basis for the next experiments conducted in this work.
- Then, we investigate the performances of different combinations of the CNN models, where the weights are learned through three methods described in the previous section (IOWA, GA, and PSO) for each combination. For a fair comparison, we also investigate the performance of these combinations of CNN models when treated equally.
- Finally, we assess the generalization abilities of GA and PSO, swapping the weights on two different datasets, and computing the classification results before and after swapping. This further draws an idea about the cross-dataset transferability of the weights of the proposed solution.

Experimental Results

This section reports a detailed description of the conducted experiments, the results achieved, and their description and comparisons against the state-of-the-art. Since GA and PSO require a learning phase, we use a subset of the training set from all datasets for validation purposes. The validation data is used to calculate the error rate in the fitness functions. Although IOWA does not require learning, the same data are used for training and testing to ensure a fair comparison.

Analysis of individual models' performances

Table 2.10 shows the experimental results of our first experiment, where we assess 10 different CNN models from 4 different architectures on WIDER, including 2 different configurations of VGGNet (i.e., 16 and 19 layers). For

CNN Model	**Avg. Acc.**	**CNN Model**	**Avg. Acc.**
AlexNet (ImageNet)	.4220	AlexNet (Places Dataset)	.4154
VGGNet16 (ImageNet)	.477	VGGNet16 (Places Dataset)	.454
VGGNet19 (ImageNet)	**.479**	VGGNet19 (Places Dataset)	.4682
GoogleNet (ImageNet)	.448	ResNet50 (ImageNet)	.3010
ResNet152 (ImageNet)	.3008	**ResNet101 (ImageNet)**	**.3006**

Table 2.10: Performances of individual CNN models on WIDER (High and low performing models are highlighted in bold)

ResNet, we evaluate the performance of all three configurations with 50, 101 and 152 layers. It is observed that deeper VGGNet19 pre-trained on both ImageNet and places datasets performs slightly better than its counterpart of 16 layers, and achieves overall the best performances. ResNet is instead the worst performing, for all the three configurations.

Although there is no significant difference in the performance of models pre-trained on ImageNet and Places datasets, the variation in the performances suggests that in event recognition, object specific information and scene-level information can well complement each other.

To better show the performance variations of different models for each event class, we compute the standard deviation of the per-class performance of all the classifiers in Figure 2.15. The limited performance of ResNet in this task highlights how it strongly contributes in increasing the standard deviation values. The standard deviation values provide evidence about how differently these models respond to the same event classes, and this is the main driver for the next set of experiments, where we want to assess how different classification architectures could contribute to the global classification goal using learned weights.

Besides providing the achieved results following the three optimization methods mentioned above, we include for the sake of clarity and for a more objective evaluation, also the combination of different architectures using equal weights.

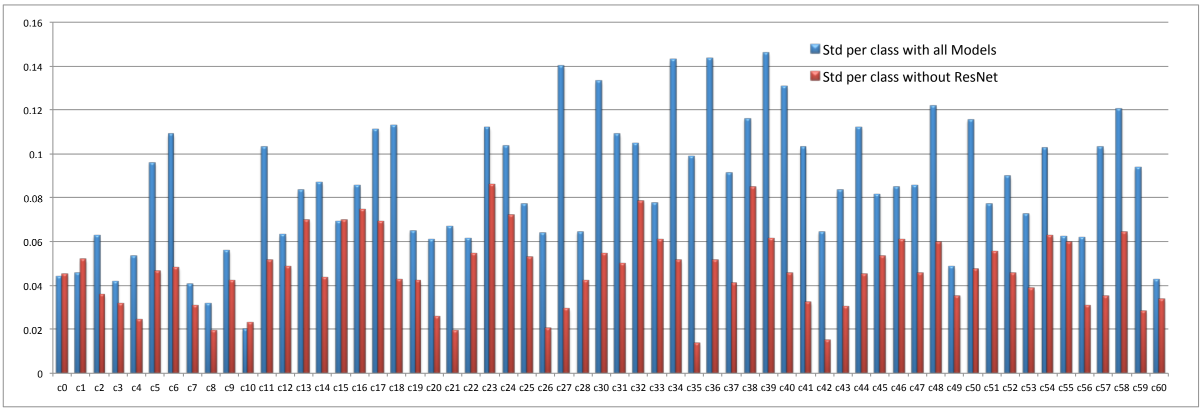


Figure 2.15: Standard deviation of the performances of individual models per class on WIDER.

Experimental results of different combinations of CNN Models

In the next experiment (see Table 2.11), we compare the performance of different combinations of existing CNN models via four different fusion techniques. It is to be observed that, for each pair of CNN models, the weights for both GA and PSO are re-learned. Moreover, to conduct a fair comparison, the parameter configuration is kept unaltered for all tests. In the case of Genetic modeling, we use 1000 generations with a population size of 50 and fitness limit of 0.5. For PSO, the maximum number of iterations is set to 1000, and the upper and lower bounds are set to 1 and 0, respectively.

We initiate our analysis by combining models of the same architecture, pre-trained on two different datasets (ImageNet and Places). We evaluate the performance of 3 different architectures, namely AlexNet and both configurations of VGGNet with 16 and 19 layers. As mentioned previously the idea of having networks pre-trained on ImageNet and Places is to grasp object centric and scene-level information at the same time. As can be noticed in Table 2.11 this blend of object and scene-level information improves the performance, significantly. In fact, the best combination (VGGNet19 ImageNet and VGGNet19 Places via IOWA based fusion) achieves a significant gain of 4% over the individual best model (VGGNet19 pre-trained on ImageNet). This improvement in the performance encourages a further exploration in this direction, where we try different combinations of 10 different models.

In this setup, the most accurate results are obtained with the combination of VGGNet19 and AlexNet, both pre-trained on ImageNet for all the fusion methods adopted in this study. This confirms the results of our initial experiments, where object-centric models provided a higher classification accuracy. The best results are achieved using IOWA; it is also observed that GA and PSO provide comparable results with a slight advantage for PSO. The main strength of IOWA comes from its ability to give more importance to the decision of the most confident models. On the other side, less accurate results are observed when these models are treated equally (i.e., assigned equal weights), confirming that it is convenient to optimize the weights for the CNN models based on their merit in discriminating diverse event classes. The gap in the performance of fusion methods is more evident when a low-performing model (i.e., ResNet in our case) is fused with other models.

We now investigate whether combining different configurations of the same model pre-trained on the same dataset can improve the results. To this aim, we combine both configurations of VGGNet with 16 and 19 layers, and we also test different combinations of the 3 configurations of ResNet. The results reported in Table 2.11 show a slight gain in the performance when combing different configurations of the same architectures, pre-trained on the same datasets. It turns out though that it is more beneficial to use different models instead of combining different configurations of the same architecture.

In order to analyze whether it is feasible (in terms of performance and computational complexity) to combine more models, in our next experi-

Combinations	**Fusion Methods**			
	IOWA	**GA**	**PSO**	**Equal Weights**
AlexNet ImageNet + AlexNet Places	.5074	.4794	.4873	.4767
VGGNet16 ImageNet + VGGNet16 Places	.5149	.5135	.5132	.5110
VGGNet19 ImageNet + VGGNet19 Places	.5191	.5180	.5172	.4999
VGGNet16 ImageNet + VGGNet19 ImageNet	.5003	.5012	.5021	.4891
VGGNet16 Places + VGGNet19 Places	.4935	.4878	.4996	.4701
ResNet50 (ImageNet) + ResNet101 (ImageNet)	.3330	.3310	.3319	.3309
ResNet50 (ImageNet) + ResNet152 (ImageNet)	.3361	.3320	.3329	.3310
ResNet101 (ImageNet) + ResNet152 (ImageNet)	.3274	.3261	.3256	**.3216**
VGGNet19 ImageNet + AlexNet ImageNet	**.5412**	.5396	.5391	.5206
VGGNet19 ImageNet + AlexNet Places	.5310	.5229	.5320	.5192
VGGNet19 ImageNet + GoogleNet (ImageNet)	.5130	.5096	.5073	.4867
VGGNet19 Places+ AlexNet ImageNet	.5251	.5173	.5201	.5130
VGG19 Places + GoogleNet (ImageNet)	.4989	.5029	.4953	.5169
VGG19 Places + ResNet50 (ImageNet)	.4291	.4225	.4285	.3918
AlexNet ImageNet + ResNet50 (ImageNet)	.4110	.4162	.4210	.3903
AlexNet ImageNet + GoogleNet (ImageNet)	.5097	.4974	.5032	.4982
GoogleNet (ImageNet) + ResNet50 (ImageNet)	.4271	.4192	.4187	.3874

Table 2.11: Evaluations of different CNN models on WIDER using three different methods for late fusion in the event recognition context (Highest and low performing combinations are highlighted in bold).

ment we combine the available models for the classification of a test image at hand. We first use all of them, and then we leave out ResNet (according to the experimental results from Table 2.10). This experiment also allows investigating which fusion method is more affected by a relatively low performing model. Although combining more models requires more computational resources, Table 2.12 shows an evident gain in the performances when jointly using more CNNs. Another interesting observation is that, in contrast to previous experiments, PSO-based method performs slightly better than the other two methods. Moreover, although the individual performance of ResNet is much lower compared to other models, it still contributes in improving the overall accuracy for both IOWA and PSO-based fusion methods. One of the possible reasons for this is its capability

Method	**Avg. Acc.**	
	All models	**All except ResNet**
IOWA	.5840	.5798
GA	.5826	.5838
PSO	**.5908**	.5897
Equal Weights	.5593	.5671

Table 2.12: Evaluation results of IOWA, GA, PSO, and equal weights late fusion on WIDER by combining more than 2 models (Highest score is highlighted in bold)

of accurately classifying certain event classes.

Generalization analysis of learned weights

We finally analyze the generalization and transfer capability of the learned weights across different datasets. In our third experiment, we swap the weights learned on WIDER and USED. Considering that IOWA is learning-free, this experiment is performed for GA and PSO, only. Table 2.13 shows the results before and after exchanging the weights learned through both methods. Although the expected reduction in the performance can be seen for both methods, it is worth noting that the combination of more models still guarantees significant performance, thus demonstrating good generalization capabilities.

Dataset	**Avg. Acc. with GA**		**Avg. Acc. with PSO**	
	Before Swapping	**After Swapping**	**Before Swapping**	**After Swapping**
WIDER	.5826	.5522	.5908	.5612
USED	.7875	.7685	.7991	.7848

Table 2.13: Impacts on the performance by swapping the weights learned on two different datasets (WIDER and USED)

Comparisons against State-of-the-art

Although in this comparative study we are mainly interested in a detailed evaluation of different CNN models and their combinations, we want to also demonstrate the absolute performance of the proposed ap-

WIDER Dataset		**UIUC Dataset**	
Method	**Avg. Acc.**	**Method**	**Avg. Acc.**
Baseline Method [138]	.397	Baseline Method [70]	.7340
Deep Channel Fusion [138]	.424	Places CNN Features [144]	.9410
Rachmadi et al. [100]	.4406	GoogleNet GAP [143]	.9500
Wang et al. [130]	.530	Wang et al. [130]	.9880
Our Approach (IOWA)	**.5840**	**Our Approach (IOWA)**	**.9854**
Our Approach (GA)	**.5826**	**Our Approach (GA)**	**.9870**
Our Approach (PSO)	**.5908**	**Our Approach (PSO)**	**.9887**
Our Approach (equal weights)	**.5593**	**Our Approach (equal weights)**	**.9731**

Table 2.14: Comparison against state-of-the-art on WIDER and UIUC Datasets

proach by comparing the achieved results against best performing methods [138, 100, 130, 70, 143, 2, 99] on three different datasets. To demonstrate the superiority of the three fusion methods, we also include the comparisons of the same pool of network models by equally treating all the models in the pool. Table 2.14 summarizes the comparison against best performing methods on WIDER and UIUC. On WIDER, our approach achieves a significant gain of 5.4%, 5.26%, 5.58% and 2.93% over state-of-the-art with IOWA, GA and PSO and fusion by equally treating all the models, respectively. On the other hand, on UIUC the results are similar with the best performing method but the dataset is comparatively smaller, with a slight advantage in favour of the PSO-based fusion. Table 2.15 shows the comparisons of our fusion methods against state-of-the-art on USED dataset [2]. Our approach, for each fusion method, scores higher results compared to the state-of-the-art approaches. In fact, our best combination (PSO) achieves a gain of 7.9% over the best performing method on USED dataset.

2.5.3 A Hierarchical Approach using MIL Framework

Experimental Setup

The event-saliency maps, we used in this work, are available for only 7 and 8 classes of social events from SED and EiMM dataset, respectively

Methods	**Avg. Acc.**
Baseline Method [2]	.700
Rachmadi et al. [99]	.72
Our Approach (IOWA)	**.7883**
Our Approach (GA)	**.7875**
Our Approach (PSO)	**.7991**
Our Approach (equal weights)	**.7833**

Table 2.15: Comparison vs state-of-the-art on USED dataset

[104]. Therefore, in this work, we consider only SED and EiMM datasets for the validation of the proposed hierarchical approach to event discovery in single images. Details of the datsets are provided in Section 2.4.

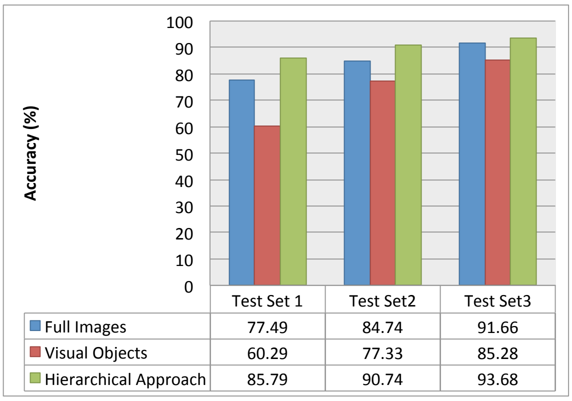
In this work, we conducted three different types of experiments with different information for training purposes:

- First, we use full images with background information as training samples. The basic insight of this experiment is to utilize the revealing contextual information often contained in the background.
- Then, we rely on the event-related visual cues extracted with the so called concept of event saliency [104] for training our classifier. This experiment aims to investigate the importance of event salient visual cues in event recognition.
- Finally, we combine both background information and event-salient details where first full images are analyzed, and then event-related visual objects are used to support the first decision.

During the experimentation process, we used a group of images as a single bag. For the selection of number of images per bag, we tried out different numbers of images per bag on the validation set. Although there was not much variation in the performance of the proposed approach with the number of images per bag, we achieved a slightly better results with five images per bag. Therefore, in all experiments, we used a group of five images per bag in the training phase while the test bag/sample contains only one image. In the case of event-related objects, since an image may have more than one event-salient objects, therefore, we used every event-salient object in an image as an instance of the bag.

Experimental Results

The experimental results are reported in Figure 2.16 where a comparison of all three types of experiments conducted are provided in terms of an overall accuracy on three test collections. As can be seen in Figure 2.16, compared to event-related visual objects only, MIL paradigm performs better with full images as training samples. This shows that background information plays an important role in event classification. Although in event-related images backgrounds usually contain rich contextual information as also demonstrated by experimental results, during the experimentation process we noticed that sometimes certain event-classes have been confused. This confusion is due to the strong visual correlation among certain event-classes. For instance, concert images share similar backgrounds with theater/dance, and conference has visual correlation with exhibition. In order to further investigate this issue, after the analysis of full images we added a refinement phase, where event-salient details are used for training purposes in a hierarchical way to avoid the confusion due to the background information in the early phase. With the refinement phase we achieved significant improvement over our other two types of approaches (i.e., single phase experiments with full images and event-salient details only). As can be seen in Figure 2.16, the hierarchical approach has an overall accuracy of 85.79% on test collection from SED dataset [102] while the MIL framework with full images and event-salient features provides an overall accuray of 77.49% and 60.29%, respectively. Similarly, on test collections from EiMM and SED datasets used in [104] the hierarchical approach provides better



	Test Set 1	Test Set2	Test Set3
🟦 Full Images	77.49	84.74	91.66
🟥 Visual Objects	60.29	77.33	85.28
🟩 Hierarchical Approach	85.79	90.74	93.68

Figure 2.16: Comparison of event detection with full images, event-salient features and hierarchical approach on three different datasets. Test set 1 represents the test collection from [101] while test set 2 and test set 3 represent the test collection used in [104] from SED and EiMM datasets, respectively.

performance compared to the individual single phase approaches demonstrating the effectiveness of the proposed hierarchical approach.

We also provide a comparison of our approach with a baseline methodology presented in [104], which also exploits event-saliency. Figure 2.17 shows a comparison of our approaches with [104] on two test collections from EiMM and SED datasets used in [104]. The overall accuracy of our hierarchical approach on test collections from EiMM and SED datasets, used in [104], is 93.68% and 90.74%, respectively, while the approach used in [104] has an overall accuracy of 41.54% and 45.95% on EiMM and SED datasets, respectively. Similarly, also in the case of single phases (i.e., MIL framework with full images and salient features only) our approach outperforms the approach presented in [104] by a significant margin.

2.5.4 A Saliency-based Approach

Experimental Setup

As mentioned in Section 2.4, each of the datasets devotes a large portion of data for training purposes. However, in order to reduce efforts in the

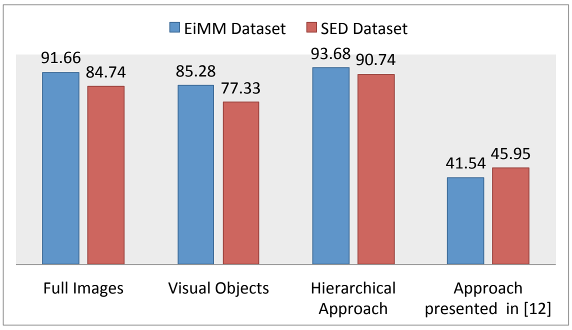


Figure 2.17: Comparison of our approach with the baseline methodology presented in [104] in terms of overall accuracy.

crowd-sourcing study, we use a subset of training samples. From SED2013 and UIUC datasets, 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 task 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 crowd-sourcing volunteers as we observed a higher number of regions tagged as ”others” for WIDER.

As aforesaid, the crowd-sourcing 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 are concerned, we used all the regions except the ones discarded in the pre-filtering phase. As far as the number of citer and reference bags is concerned, we tried different combinations on the validation set to find the best values. At the end, we choose 3 references and 5 citer bags.

Experimental Results

In this section, we provide the analysis of our crowd-sourcing study along with the detailed description of the conducted experiments for event recognition in single images.

Crowd-sourcing Analysis

All in all, in the crowd-sourcing task, we received around 25,000 responses from more than 400 distinct volunteers for 76 different events. On the average each volunteer investigated more than 60 image regions. We discarded 47 responses because the answers in the open question 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 at least 3 responses from 3 different volunteers, and put them into the bags of an event class with majority of the responses. Figure 2.18, shows some sample regions along with the tags provided by the volunteers. There are also a number of image regions for which the majority of volunteers are not sure, and are tagged as ”others”. Figure 2.19 shows some sample regions which are tagged as ”others” by the volunteers. Such regions are discarded from the training samples.

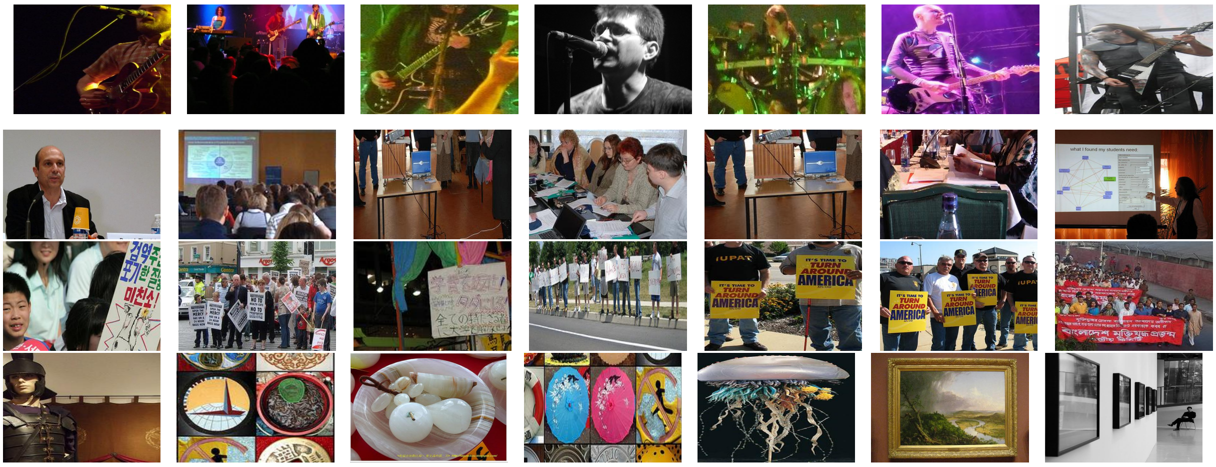


Figure 2.18: Sample regions tagged by the volunteers as concert, conference, protest and exhibition (top to down) during the crowd-sourcing task.

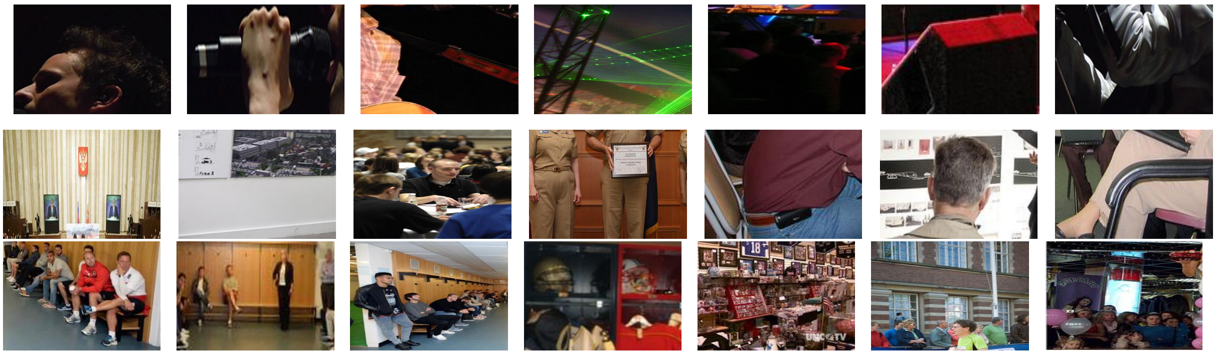


Figure 2.19: Sample regions from concert, conference and sports images (top to down) tagged as *others* by the volunteers.

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[5].

Moreover, having a closer look into the answers of the second question from the volunteers that participated in the study reveals interesting facts about the objects and regions a human thinks should be associated to a specific event. For example, in concert images volunteers tagged the regions

[5] The created dataset of the selected event salient regions will be made publicly available upon acceptance of the paper

Table 2.16: Classification results in terms of accuracy per class on SED Dataset

Event	**Acc.**	**Event**	**Acc.**
Concert	.916	Conference	.856
Exhibition	.868	Fashion	.963
Protest	.892	Sport	.893
Theater	.950	-	-
Overall Acc.		**.911**	

based on the musical instruments and the lighting effects. Similarly, in all sports events, volunteers tend to tag the regions based on the sport goods and kits, such as tennis racket and ball, baseball bat and ball.

Event Analysis

In this section, we analyze the importance of event-salient regions in event recognition. We tested our method with different bag sizes, to determine the best trade-off between computational complexity and classification accuracy. On the validation set, we observed similar performances by using 5, 10 and 15 image regions per training bag, therefore in order to limit the computational requirements in our experiments we use 5 regions per bag in the training samples.

In order to provide a thorough analysis, we validate our approach on four different datasets namely, SED2013, USED, UIUC Sports and WIDER. Table 2.16 provides the experimental results of our approach on SED2013. As can be seen in Table 2.16, our approach provides overall better performance on almost every event-class of the dataset. Exceptions can be noticed on events containing more distinguishable objects and regions, such as concert and theater, where the performance is significantly superior. We also observed that some test samples from exhibition and conference events are mis-classified as fashion, a problem that was evident also when conducting the crowd-sourcing study.

On WIDER our approach achieves an overall accuracy of 55.04%. Table

Table 2.17: Classification results on WIDER dataset [138] in terms of accuracy per class

Event	**Acc.**	**Event**	**Acc.**	**Event**	**Acc.**
Parade	.546	Handshaking	.410	Demonstration	.839
Riot	.260	Dancing	.475	Car_Accident	.743
Funeral	.506	Cheering	.285	Election Campaign	.225
Press Conference	.689	People Marching	.267	Meeting	.666
Group	.499	Interview	.328	Traffic	.597
Stock Market	.563	Award Ceremony	.591	Ceremony	.415
Concerts	.373	Couple	.465	Family_Group	.281
Festival	.382	Picnic	.681	Shoppers	.547
Soldier Firing	.621	Soldier Patrol	.690	Soldier Drilling	.357
Spa	.804	Sports Fan	.269	Students Schoolkids	.238
Surgeons	.631	Waiter, Waitress	.685	Worker Laborer	.425
Running	.716	Baseball	.656	Basketball	.569
Football	.675	Soccer	.550	Tennis	.741
Ice Skating	.718	Gymnastics	.648	Swimming	.726
Car Racing	.747	Row Boat	.798	Aerobics	.506
Balloonist	.463	Jockey	.364	Matador Bullfighter	.732
Parachutist Paratrooper	.684	Greeting	.267	Celebration, Party	.482
Dresses	.735	Photographers	.359	Raiar Racing	.477
54	.427	Sports Coach Trainer	.161	Voter	.346
Angler	.462	Hockey	.550	People Driving Car	.618
Street Battle	.229	-	-	-	-
Overall Acc.			**.5504**		

2.17 shows the results on the individual events. In contrast to SED2013, performance on the event classes from WIDER are very diverse. In fact for certain events, such as riot, spa and some sports events (e.g., tennis and ice skating) the proposed approach performs very well. 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.

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. The gain our approach achieves over the state-of-the-art is reported in Table 2.18.

The comparisons on WIDER dataset [138] are provided in Table 2.19.

Table 2.18: Comparisons against state-of-the-art on SED2013.

Method	**Avg. Acc.**
Schinas et al. [107]	.334
Rosani et al. [104]	.4595
Ahmad et al. [2]	.7003
Ahmad et al. [4]	.8579
Our Approach	**.9115**

Table 2.19: Comparisons against state-of-the-art on WIDER [138].

Method	**Avg. Acc.**
Baseline Method [138]	.397
Deep Channel Fusion [138]	.4204
Rachmadi et al. [100]	.4406
Init. based object-Scene Transferring [130]	.508
Knowl. based object-Scene Transferring [130]	.520
Data based object-Scene Transferring [130]	.526
Data + Knowl. based object-Scene Transferring [130]	.530
Our Approach [3]	**.5504**

As can be seen in Table 2.19, our approach 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 [2] and UIUC Sports Events dataset [70] are provided in Table 2.20 and Table 2.21, respectively. Our approach achieves a significant gain of around 5% against the state-of-the-art on USED. 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.

Table 2.20: Comparison against state of the art on USED.

Methods	**Avg. Acc.**
Baseline Method [2]	.700
Rachmadi et al. [99]	.720
Our Approach [3]	**.771**

Table 2.21: Comparisons against state of art on UIUC Sports Dataset [70].

Methods	**Avg. Acc.**
Baseline Method [70]	.7340
ImageNet CNN Features [144]	.9440
Places CNN Features [144]	.9410
GoogleNet GAP [143]	.9500
Object-Scene Transferring [130]	.9880
Our Approach [3]	**.9838**

2.6 Summary

In this chapter, we presented three different solutions to event recognition in single images along with a benchmark dataset. We conducted a comprehensive analysis of the state-of-the-art deep models and assessed their individual as well as joint performance. We also analyzed the importance of event-salient objects and regions in event recognition where event-salient objects are extracted through a crowd-sourcing study. We demonstrated that it is possible to achieve superior event recognition performance by selecting the best models and combining them in an optimal way through appropriate late fusion strategies. Moreover, we showed that better results can be obtained by targeting the so called event-salient visual objects in event recognition.

Chapter 3

Events in Photo Collections

3.1 Introduction

As aforesaid, most of the existing literature on event recognition focus on the analysis of single images while very few attempts have been made for event recognition in personal photo collections [124, 28]. In contrast to event recognition in single images, there are a number of factors that make event recognition in personal photo collections a more challenging task. In fact, in event recognition from single photos, the visual content of photos is usually strictly related to the labeled event. However, personal photo collections tend to have a large portion of ambiguous and irrelevant photos, which do not necessarily match a particular event tag. Figure 3.1 shows some sample irrelevant photos in the context of event recognition in personal photo collections. In details, there can be face close-ups (which could be part of any event class) or just images where the specific objects (e.g., birthday cake and candles in birthday images) are not present. Moreover, most of the events in personal photo collections are composed of multiple sub-events.

Another important challenge for event recognition frameworks in personal photo collection is the weakly-labeled training data. In fact, in event recognition in personal photo collection training labels are available at

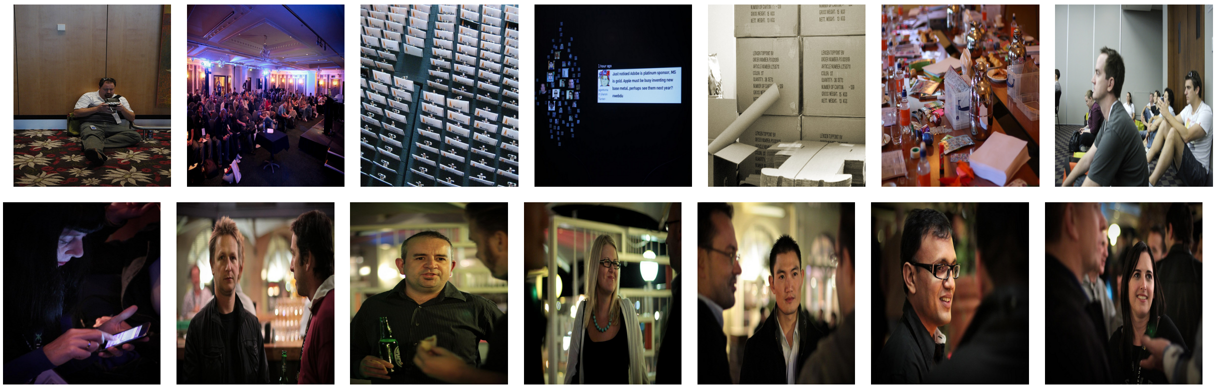


Figure 3.1: Sample irrelevant images from the collections: images not containing objects of interest (top) and images with face close-ups (bottom)

album-level only.

3.2 Related Work

To tackle event recognition in personal collections, some existing works use Hidden Markov Models (HMMs) [36] to consider the time gap between photos[19]. On the other hand, Tang et al. [117] approach the problem using a probabilistic fusion of different classifiers trained on manually-selected pictures from a collection. In another work [121] from the same authors, an object centric approach is proposed where Histogram of Gradient (HoG) [30] are used to capture objects of interest in an image. A similar strategy is adopted by Tsai et al.[122], which mines training samples for the most frequent object patterns; object patterns are then ranked according to their discrimination ability for training purposes.

Similarly, Guo et al.[49] use average and aggregated visual features extracted with multiple Convolutional Neural Networks (CNNs) [66] of all pictures in a collection for event recognition in a hierarchical way. Initially, photo albums are classified via a coarse event classifier trained on features extracted with Convolutional Neural Network (CNN) pre-trained on places datasets [144]. Subsequently, CNN features extracted with AlexNet [66] pre-trained on ImageNet [33] are used to train fine event classifiers. Similar visual features extracted via Convolutional Neural Networks (CNNs) pre-trained on ImageNet and Places datasets are used by Bach et al. [13] in a probablistic graphical model.

In addition to visual features, some works also utilize the additional information available in the form of meta-data for event recognition in personal photo collections [45, 63]. For instance, Namaan et al. [85] rely on temporal and geo-location information for the organization of photo collections. Similarly, in [46] a combination of visual and temporal information are used to train four different classifiers. This work mainly covers a limited number of indoor and outdoor events.

Although a number of interesting solutions have been proposed for event recognition in personal photo collections, most of the existing approaches particularly the ones relying on supervised learning lack in dealing with non-relevant images in photo albums annotated at album-level only, which may significantly affect their performance, as demonstrated in the experimental validation of more recent works in this domain [49, 13, 46].

3.3 Solutions

3.3.1 MIL based Classification of Multiple Images

Overview

Despite the variations in content, there are certain objects or features that are common to all albums of a particular event, and are crucial for the identification of underlying events (e.g., birthday albums usually have images of a cake)[104]. The key to success in event recognition in photo collections consists also of the ability to identify such images across all albums.

Identifying such images within the example/training albums, with a large number of ambiguous images, remains a hard problem for the conventional approaches relying on supervised learning.

In this work, we propose a novel pipeline for event recognition in personal photo collections relying on a Multiple Instance Learning (MIL) paradigm. MIL is a modified form of supervised learning, which fits well in applications with weakly labeled data. This strategy aims at minimizing the effects of ambiguous photos in the learning and prediction phases. The underlying insight of the MIL-based approach, we propose, is its suitability for applications with polymorphism and part-whole ambiguities [12]. Such capabilities make MIL strategies a better choice also in a number of other computer vision tasks [126]. In the context of event recognition in personal photo collections, polymorphism ambiguities refer to the fact that photo albums may contain images representing different sub-events of a particular event, and it is not known which of these images is responsible for the label of the album. Similarly, the part-whole ambiguity represents the annotations at album-level, instead of each image inside the albums.

The important benefits of the proposed approach are: (i) it guarantees higher performances even in the presence of irrelevant photos in weakly-labeled data; (ii) event classification is achieved with a reduced number of training samples per class; (iii) computation complexity is reduced by using a limited number of images per bag as detailed later.

The main contributions of this work are: (1) we propose a novel pipeline for event recognition in personal photo collections relying on a MIL paradigm that outperforms state-of-the-art approaches; (2) we provide a detailed analysis of the trade-off between classification performance and computational cost through extensive experimental evaluation (3) we propose an image dataset containing a large number of photo albums per event.

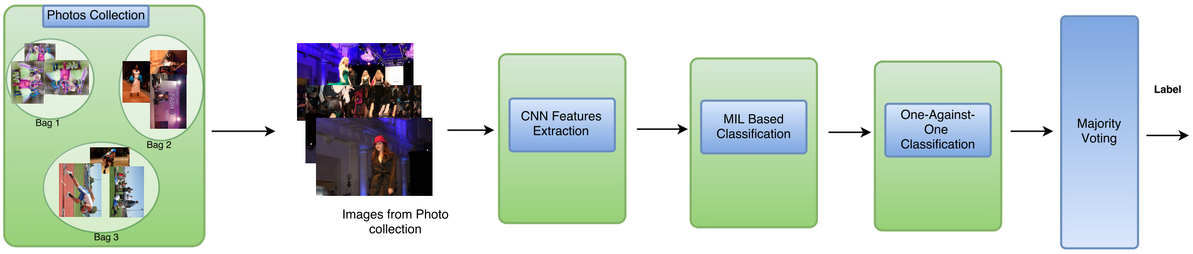


Figure 3.2: Block diagram of the proposed methodology for event recognition in personal photo collections.

Methodology

As can be seen in Figure 3.2, the proposed solution mainly consists of three different steps. In the first step, we use Convolutional Neural Networks (CNNs) features for the representation of images in photo albums. After feature extraction, photo albums are divided into negative and positive bags for MIL-based classification, where each bag contains multiple images from a photo album. To deal with multi-class classification, we adopt a one-against-one strategy, and the final classification decision is made on the basis of majority voting.

Feature Extraction

The state-of-the-art in visual-based event recognition has so far revealed considerable uncertainties with ample room for improvement. We believe that such limitations can be mostly attributed to the selection of the visual features used for representation. In fact, as also reported in the experimental validation of some of the existing works [104], conventional approaches to event recognition in single images, relying on handcrafted visual features, cannot cope with high complexity and variations in event-related multimedia contents. On the other hand, deep models have proven to be very effective in different application areas, such as image and video analysis, as in the *Cultural event recognition* challenge introduced at ChaLearn Looking at People [106, 93], and TRECVID event detection task [40]. Based on these considerations, in this work we choose CNN features for image representation, and we use VGGNet [110], pre-trained on ImageNet object dataset, as a feature extractor. The motivation for the selection of object-level features is due to the typical association that objects have with events. 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. We extract a 4096-dimensional feature vector from each photo in the dataset using Caffe toolbox[2].

Classification Via MIL

Considering the weak labeled and ambiguous data in personal photo collections, for conventional approaches relying on supervised learning it becomes hard to identify which photos in the training albums are relevant. However, this problem fits well in MIL [127, 136, 133], which is a variation of supervised learning, conceived for applications with incomplete or ambiguous knowledge about training labels. In MIL, training labels are assigned to the bags of instances, only.

In this work, in order to map event recognition in personal collections into a MIL problem, each image in an album is treated as an instance of a bag representing the album. For the prediction of a given test album a, we adopt a k-nearest neighbor approach, by considering both R-nearest references (bags in its neighborhood) as well as C-citers bags, which consider album a as their own neighbor [127]. This blend of references and citer bags helps to mitigate the effect of false positive instances in positive bags. The reference bags are simply the R nearest neighbors. However, defining C-citers of an album is slightly more complex, and a ranking mechanism is used to this aim. For instance, if n is the number of all example albums, represented as $A_s = a_1, a_2, a_3, \dots, a_n$, then, for an album $a_i \in A_s$, the rest

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

of the albums are ranked according to the similarity to the album a_i. For instance, the rank of a sample $a_j \in A_s$ with respect to a_i is represented as $Rank(a_j, a_i)$. Subsequently, c-nearest citers are defined as:

$$citers\ (a_i, c) = a_j | Rank(a_j, a_i) \leq c, a_j \in A_s \tag{3.1}$$

where c represents the number of citers to be used.

For similarity measurement, a bag-level distance metric, the modified Hausdorff distance [127], is used. In contrast to the original implementation of Hausdorff distance, the modified version is less sensitive to outliers. For instance, for the comparisons of two bags/albums X and Y, the modified Hausdorff distance is defined as follows:

$$h_k(X, Y) = k^{\text{th}} min_{x \in X y \in Y} \|x_i - y_i\| \tag{3.2}$$

where X and Y represent the two bags/albums, x_i and y_j are the corresponding instances and the k^{th} ranked value decides the value of the overall distance[127]. In our case, we opt for the minimal Hausdorff distance (i.e., k = 1) [127].

After the summation of R-nearest references and C-nearest citers in terms of positive bags/albums (i.e., $S_p = R_p + C_p$) and negative bags/albums (i.e., $S_n = R_n + C_n$), a majority voting approach is used for the prediction of a given test bag/album a according to Equation 2:

$$C_{Label} = \left\{ \begin{array}{ll} 1 & \text{if } S_p > S_n \\ 0 & \text{otherwise} \end{array} \right\} \tag{3.3}$$

Finally, in order to deal with multi-class classification, we adopt the one-against-one strategy where we trained $n * (n - 1)/2$ binary classifiers. Subsequently, the final classification decision is made on the basis of majority voting where a class with more positive labels is selected as the final classification outcome.

3.4 Experiments

3.4.1 Datasets

For the experimental evaluation of the proposed work, we use two large-scale datasets namely (i) Personal Events Collection (PEC) released by Bossard et al. [19] and (ii) the one collected by ourselves. Although PEC [19] provides a sufficient amount of photos (61,000) assembled into 807 albums from 14 different events, a larger portion of them belongs to the training set while the test set contains only 10 albums per event. Table 3.1 provides the details of the benchmark dataset. Moreover, it also lacks a large number of photo albums per event. Therefore, in order to conduct a more thorough evaluation of the proposed approach, we have collected a second dataset, composed of 7 different event classes and 662 albums. Each album contains at least 20 photos, covering different aspects of the underlying event. All the photos are downloaded from Flickr. The details are summarized in Table 3.2.

In contrast to other approaches, our system does not require large amount of data to train the classifier. Therefore, we use a large portion of our dataset for testing purposes. We use 20 albums per event-class for training, 60 for test, and 20 albums for validation, in order to determine the model parameters. As far as the second dataset is concerned, training and test sets have been already defined in the corresponding paper [19].

3.4.2 Experimental Settings

In order to validate the proposed algorithm for event recognition, we test it by changing the number of images per bag (album) to understand the trade-off between classification performance and computational cost. Therefore, the most important parameters to be defined are the number of citers and reference bags as well as the number of images used per bag.

Table 3.1: Details of the dataset released in [19]

Event	**# albums**	**Event**	**# albums**
Birthday	60	Cruise	45
Childbirthday	64	Easter	84
Christmas	75	Exhibition	70
Concert	43	Graduation	51
Halloween	40	Hiking	49
Road-trip	55	Saint Patrick's Day	55
Wedding	69	Skiing	44

Table 3.2: Details of the proposed dataset.

Event	**# Albums**	**Event**	**# Albums**
Concert	100	Fashion	100
Conference	100	Protest	100
Exhibition	62	Sports	100
Theater	100	**Total Albums**	**662**

We validate our approach using 7 different configurations, randomly selecting 1 to 15 images per test bag, and 5 to 15 images per training bag, on the validation set. As far as the number of citers and reference bags is concerned, after an extensive test of the possible combinations, we opted for 3 references and 5 citers. Some of the combinations we tried are provided in Table 3.3.

Table 3.3: Parameter analysis for reference and citer bags on the validation set

No. of References	No. of Citers	Accuracy on the validation set
1	1	.80
2	2	.84
3	3	.84
4	4	.86
5	5	.83
3	5	.87

3.4.3 Experimental Results

According to the results achieved on the validation set, we achieve the highest accuracy using 15 images per bag in training (see Table 3.4). Therefore, in the experiments on the test set we use 15 images per bag in the training samples. Similar to the training phase, we are now required to find an optimal number of instances/images for each bag for the test set. Also in this case, we conducted experiments with 7 different bag-sizes (see Table 3.5).

As can be seen in Table 3.5, there is a significant improvement in the performance of the proposed approach when using multiple images in test bags, with a difference of 12% when using five images instead of a single one. This demonstrates that using multiple images for event recognition reduces the errors in classification, as we have more chances of picking up an image that well represents the underlying event. However, as the number of images in the bag increases, the performances stabilize. This is because similar events also share similar visual features, at least at the bag/album level, and our approach is able to capture such common pattern/features even with a limited number of images from each album, without compromising the performances significantly and, instead, considerably reducing the computational burden. In fact, using 15 images per bag/album turns out to be about 3 times slower, compared to using a bag-size of 5 images.

Table 3.4: Accuracy on the validation set with different number of randomly selected images in training and test bags (in %)

		Test						
		1	2	3	4	5	10	15
Training	5	75.00	77.14	78.57	83.57	86.42	87.14	86.42
	10	76.42	80.71	81.42	85.00	89.28	89.28	89.82
	15	77.85	80.00	82.85	87.85	92.14	91.42	92.85

Table 3.5: Performance of our approach with different number of photos in test bags on the proposed dataset

#photos in Test Bags	**Avg. Accuracy (%)**
1	84.29
2	87.28
3	89.60
4	94.45
5	96.48
10	96.98
15	96.50

Table 3.6: Comparisons against best method in state-of-the-art approaches, tested on the dataset[19]

Method	**Accuracy (%)**
HAS [49]	91.04
SVM-CNN (majority voting)	89.01
MIL-SMO	93.89
Our Approach	**96.48**

We also provide a comparison of our approach against the state-of-the-art. On our dataset, the proposed approach achieves an average accuracy of 96.48% with an improvement of 5.44% over the state-of-the-art method [49], which attains an average accuracy of 91.04% as shown in Table 3.6.

The comparison results on the benchmark dataset [19] are shown in Table 4.1. Our approach has an overall gain of 8.95% over the state-of-the-art. We observed that the state-of-the-art methods especially the ones relying on aggregated score of all photos in an album (e.g., AgS [19]) are badly affected by the irrelevant photos in albums. The average feature vectors adopted in HAS [49] leads to expected mis-classification if an album contains a large number of ambiguous photos. The same pattern has been observed on both datasets, where the existing approaches show poor performance on the albums with more ambiguous photos. For instance, state-of-the-art achieve poor performances on events, such as birthday, children birthday, and Easter collections. We also perform some additional experiments with SVM trained on VGGNet features along side with an alternative implementation of Multiple-instance Learning paradigm with SVM (MIL-SMO). In the experiment with SVM without MIL, we classify each and every image in the test albums, and the final decision is made on the basis of majority voting. We achieve an average accuracy of 82.30% with SVM without MIL paradigm while the MIL-SMO implementation achieves an overall accuracy of 91.53%, which shows the superiority of MIL strategy in this particular application.

Moreover, we also investigate the performance of the reference method [49] with different configurations, although not foreseen in the original work. Instead of all photos in an album, we use 15 randomly selected images from each album to analyze its performance using less data, and achieve a significant reduction in performances (6.25%). Thus, to make accurate predictions, this confirms that the reference method [49] needs a significant number of relevant photos in an album, which is not always the case in personal photo collections.

Table 3.7: Comparisons against state-of-the-art approaches, tested on the dataset[19]

Method	Accuracy (%)	Method	Accuracy (%)
AgS [19]	41.43	ShMM [19]	55.71
Method in [137]	73.43	HAS [49]	86.32
AgS-CNN(Scene) [49]	73.31	AvS-CNN(Scene) [49]	80.61
R-OS-PGM [13]	74.28	SVM with CNN Features	82.30%
Our Approach (MIL-SMO)	**91.53**	**Our Approach (Test Bag = 5 images)**	**95.27**
Our Approach (10 images per bag)	**94.48**	**Our Approach (15 images per bag)**	**95.27**

3.5 Summary

In this chapter, we present a MIL-based approach to event recognition in personal photo collections. We consider each photo album as a single bag, and choose multiple images from each album for classification purposes. We show that, even due to album level annotation and presence of ambiguous photos in the albums MIL can guarantee higher performances. At the same time, the approach achieves good performances also using a limited number of images per bag, thus keeping the computational load acceptable.

Chapter 4

Disasters and Social Media

4.1 Introduction

Natural disasters include the adverse events caused by natural processes (e.g., floods, hurricanes and tornadoes), which might have a negative impact on the environment and people. These adverse events can be more dangerous if they occur in dense populated areas. Recovery from such disasters is a gradual process, and can be made faster and more effective if detailed information about its impact is known. For instance, government and non-government relief and aid organizations can then direct their resources to the areas struck by the disaster, and policies can be amended to speed up the recovery process, accordingly.

During the last few decades, satellite data has been widely used to analyze the impacts of adverse events on the surface of the earth. Being able to cover a large spatio-temporal area, remote sensed data has been proved to be very effective in different applications, such as classification and mapping of vegetation, crop stress detection, and disaster management [59, 52]. More recently, NASA released a dataset from the longest running satellite program called Landsat[1]. The latest Landsat dataset[2] contains images shot

[1] `http://landsat.usgs.gov/`

[2] `https://aws.amazon.com/public-data-sets/landsat/`

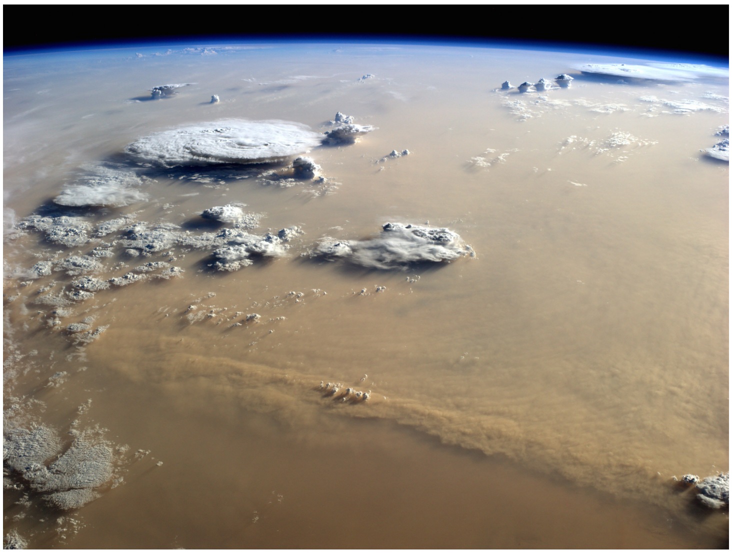


Figure 4.1: A satellite image of a sandstorm in the Sahara. Based on the image, however, it is almost impossible to give a clear statement about its impact on the environment and society (image from NASA).

by the Landsat 8 satellite. Such a dataset holds a lot of opportunities for society, and enables researchers to develop systems that integrate remote sensed data in different applications. However, remote-sensed data also comes with some challenges. For example, satellite imagery usually takes several days to be available after an event, and more importantly, they only give a bird's-eye view of an event [94]. For instance, sandstorms can be detected in the satellite images of Sahara, as shown in Figure 4.1, but how to determine, just by looking to the image at hand, if this sandstorm had any impact on the people and the environment.

On the other hand, over the last few years, social media has emerged as an important source of information and rapid communication in emergency situations. Particularly, Twitter has been proved to be very effective in dissemination of news about natural disasters [11]. Moreover, as demonstrated in [112], there are many situations in which news agencies could not provide information at all or in time simply due to the lack of having reporters spread all over the world. In such cases, social media plays an important role [11].

A rather recent trend is to combine content from social media with remote sensed data, e.g., in the MediaEval benchmark initiative with the ”placing” task[3], where researchers try to predict geo-coordinates for images and videos from Flickr. Additionally, a task to build a system that links social multimedia to events, which can be detected in satellite images, has been introduced as a challenge at ACM MM 2016[4] and MediaEval 2017 [17]. This clearly shows that the multimedia research community is interested in this new direction.

In this work, we propose two different solutions to jointly utilize social media and satellite imagery to provide a better overview of the underlying natural disaster event. Our first work is mainly concerned to develop a system that is able to collect and monitor natural disasters by linking social media and satellite imagery. Our second work is based on Multimedia and Satellite challenge introduced in MediaEval 2017 [5]. In the next sections, we provide detailed descriptions of each of the proposed work.

4.2 Related Work

This section starts with a general overview of the importance of remote sensed data in different applications with a particular emphasis on disaster management, followed by a detailed description of the characteristics of social media and its applicability as a medium of information in emergency situations.

[3] `http://multiMediaEval.org/MediaEval2016/placing/`

[4] `http://www.acmmm.org/2016/wp-content/uploads/2016/03/ACMMM16_GC_Sky_and_the_Social_Eye_latest.pdf`

[5] http://www.multiMediaEval.org/MediaEval2017/

4.2.1 Remote Sensed Data

Since the launch of Landset 1[6], formerly known as Earth Resources Technology Satellite (ERTS), satellite imagery has been used in different application areas [59]. For instance, satellite imagery is widely used in meteorology, fishing industry, agriculture, forestry, landscape, geology, regional planning, education, and warfare [22]. However, the use of satellite imagery in an application depends on a number factors. These factors include spatial and spectral resolution, coverage and cloud cover.

Over the last few years, satellite data has been also widely used in disaster management, and analyzing its impacts on the environment. The wide geographical coverage and multi-spectral resolution make satellite imagery an important source of information and support tool in disaster management activities. For instance, according to the authors in [59], the disaster management process can be roughly divided into 4 different phases, and satellite data is equally useful in all of them. These phases include reduction, readiness, response and recovery. In this regard, a number of international cooperation mechanisms and organizations have been established to help and support in the disaster management, which heavily rely on remote sensed data. For instance, the Disaster Management Support Group (DMSG) [135] is developed to perceive the specifications, basic observations, and monitoring requirements for disaster management systems based on satellite imagery. Following the guidelines of such organizations and mechanisms, a number of interesting systems have been proposed to effectively utilize satellite images in investigating the impact of a disaster on the environment [59].

Literature shows that most of the disaster management systems focus on the acquisition and pre-processing of satellite imagery, however, little

[6] https://landsat.gsfc.nasa.gov/landsat-1/

attention has been paid to develop a system that can detect a disaster in satellite imagery. To analyze satellite imagery for disaster detection, Amit et al. [9] proposed a Convolutional Neural Networks (CNNs) based approach for the detection of certain disasters, such as landslides and floods. Similar approach is adopted in [61], where a deep model is trained on aerial photos captured through an unmanned aerial vehicles (UAV). Similarly, Convolutional Neural Networks (CNNs) features are exploited by Liu et al. [73] for the representation of landslide images.

More recently, in a benchmark challenge in MediaEaval 2017, flooded regions detection in satellite images has been introduced as a separate task [17]. A number of interesting solutions are proposed in the response to the challenge. For instance, Benjamin et al. [18] approach the challenge as a segmentation problem relying on three different variations of a deep model, namely VggNet [110]. In details, the final convolutional layer is replaced with a up-sampling layer relying on bi-linear interpolation to re-scale the down-sampled feature maps into original patch size. Subsequently, a softmax layer is used to classify the pixels into flooded and non-flooded regions. Similarly, in [62], an approach based on the concept of convolutional deep model with dilated convolution is proposed to deal with the segmentation and classification of satellite image patches into flooded and non-flooded regions. In total, four different models with different number of dilated convolutional layers are used. Moreover, all the models are trained with overlapping patches each of size 25x25. The same strategy is used in the prediction phase, where the final result is based on the average probabilities of all patches. On the other hand, in contrast to earlier two methods, Avgerinakis et al. [64] use Mahalanobis distances with stratified co-variance estimates along with morphological post-processing to this aim.

It is to be noted that, satellite data also have some limitations. For example, low temporal frequency is one of the biggest hurdles in different applications, particularly in disaster management and monitoring. However, satellite images before and after a disaster[7] combined with other information can be a better source to get an overview of a disaster's impacts, and monitor the recovery process.

4.2.2 Social Media and Disasters

On the other side, The huge amount of content shared through social networks represents a potential resource for many applications and research studies in different fields, such as economics, sociology and computer science. One question that emerges is why social networks are so attractive? It is possible to find a lot of answers to this question. For instance, the authors in [26] have proved the effectiveness of the social network as a powerful medium for disseminating good practices. Moreover, in [113], a case study has been provided to highlight the importance of social media in e-commerce. Besides being a social and business instrument of influence, a social network can be considered also a medium of mass communication [112, 55]. For instance, Stelter et al. [112], analyzed Twitter as a medium of communication in emergency situations.

In recent years, a common practice is to infer events from the information shared through social media. For instance, Popescu and Pennacchiotti [98] extracted a list containing names of actors, musicians, politicians and sports men from Wikipedia to be used for crawling Twitter to detect controversial events about them. Subsequently, a regression model is used to asses the controversial contents. Similarly, there are several other approaches relying on unsupervised frameworks for the detection of social events, such as concert and theaters etc., in Twitter [11]. As an example, Mathiodakis et al. [78] use clustering techniques on bursty key words

[7] `http://www.satimagingcorp.com/applications/environmental-impact-studies/natural-disasters/`

to detect trends in Twitter. Similarly, Meladianos et al. [80] proposed a methodology for sub-event (i.e., key moments of an event) detection in Twitter streams using the concept of graph degeneracy. In [48], a statistical approach relying on tweets, and the frequency of links, inserted by users in their tweets, has been proposed to detect social events.

A number of works in this regard also exploit Twitter data to detect and analyze emergency situations and disaster events. For instance, Li et al. [69] proposed a method to detect crime and disaster events in Twitter's text streams. Similarly, in [105], tweets are analyzed to detect earthquakes in Japan, where some key words, such as earthquake and typhoons, are used to crawl Twitter. Similarly, in [14], a method for the detection of earthquake in tweets has been proposed. In the proposed method, a graph based clustering technique has been utilized to target geo-located communities in Twitter. Similarly, in [29], Twitter is utilized as a social sensor to capture information about a natural disaster from users in real time. In [35], a concept derived from seismology, originally developed to detect and time seismic phases, is used for earthquake detection in Twitter text streams. The authors monitor a rapid increase in the tweets containing words *earthquake* relying on a short-term-average over long-term-average (STA/LTA) algorithm. More recently, Xu et al. [140] proposed a participatory sensing-based model for collecting information about disaster events in micro-blogs. In [116], the authors examine the use of social media particularly twitter in emergency situations considering a number of factors, such as time and location of the user, and type of users (e.g., general public, journalist and agencies etc.,).

Twitter is of course the most exploited resource, but other works have also tried to exploit other social media platforms to detect such events, as for example Flickr [124]. In this regard, most of the existing works target social events and daily life activities [124, 132, 4]. However, more recently, a benchmark is initiated to detect flood related images in social media [17]. In the response to the task, a number of interesting solutions have been proposed for the classification of flooded and non-flooded images in social media [62, 87, 64, 84, 5]. For instance, in [8], the classification results of different classifiers trained on different Convolutional Neural Networks (CNNs) models are combined in two late fusion methods. Moreover, user tags, geo-location information and description of an image are also used, as an additional information to support the visual features. Benjamin et al. [18] also rely on an image representation scheme deriving benefits from deep architectures. In details, they extract features from two deep architectures, namely DeepSentiBank [25] and X-ResNet [60]. Subsequently, a Support Vector Machine (SVM) is used for the classification purposes. On the other hand, to solve the same problem, some works rely on hand-crafted visual features [83, 142]. For instance, in [87], a combination of CEDD, CL and GABOR features are used along with meta-data.

The analysis of literature on multimedia analysis reveals that social media platforms, particularly Twitter has been heavily exploited for inferring information about natural disasters. However, little attention has been paid to collect information from other platforms of social media. To the best of our knowledge there is no prior work which collects multi-modal information from multiple platforms at a time. Although, collecting and analyzing information from different platforms of social media is a tedious and time consuming job, the combination of different sources to one summarized overview can be very useful for users.

4.3 Solutions

4.3.1 The JORD System

Overview

In this section, we present our system called JORD (after the Norwegian goddess of the earth), which is to the best of our knowledge the first one that is able to automatically collect information and news items about natural disasters from four different social media platforms, and links it with satellite imagery in real-time[8]. It also provides query refinement by automatically generating queries in all local languages that are relevant to the position of a disaster. With such a system, that combines multimedia mining, retrieval, linking and summarization methods [76, 37], we are able to tell a much clearer and more useful story to the users[9]. Moreover, the proposed system retrieves information continuously, which makes it a better source to monitor long term recovery efforts.

In addition, to ensure the quality of the retrieved multimedia data, we propose a hierarchical filtering mechanism. Firstly, temporal and geolocation information are used to filter out irrelevant data, which is followed by a content based filtering and analysis scheme to further filter out less informative content, and provide the more relevant one, only. In the current implementation, we are providing the content analysis for the images and tweets only, and intend to extend it to other types of multimedia contents (videos). In order to obtain positioning information necessary to retrieve and link satellite imagery to the underlying events, we extract the GPS coordinates of the places and city names mentioned in the tweets relying on natural language processing (NLP) techniques. The system is

[8] Real-time in the context of information retrieval in JORD system means that JORD continuously monitors various information sources and retrieves the information as soon as a query match is found.

[9] The potential users of the system are relief workers and aid agencies, who need to know where things are and what has changed, and the general public to get information about the disaster.

also equipped with a novel methodology for identifying the areas hit by the disaster in complex satellite imagery. For evaluation purposes, we have conducted a crowd-sourcing campaign with a large number of users, asking them to share their feedback about the retrieved contents and the system itself.

In summary, we can synthesize the main features of JORD as:

(i) It collects data about events autonomously and automatically in real-time from a disaster database (i.e., when JORD is running and an event occurs, it will continuously gather new information from social media to enhance the event information).

(ii) JORD is able to generate queries in local languages spoken in the area hit by the underlying disaster.

(iii) JORD automatically filters irrelevant information in a hierarchical way relying on temporal information and content analysis of the retrieved data.

(iv) JORD combines social media and satellite imagery in a novel way, and provides a more detailed event description to the users.

(v) It is equipped with a novel method for linking and retrieving satellite imagery with the events by analyzing the tweets text to identify and extract GPS coordinates of the areas struck by the disaster.

(vi) JORD also consists of a novel framework for flood detection in satellite images as a use-case of the disaster event detection in satellite imagery.

Proposed System

As shown in Figure 4.2, JORD consists of four main components (highlighted in different colors with corresponding labels): (i) query refinement,

(ii) multimedia data retrieval from social media, (iii) temporal and content-based filtering of the retrieved multimedia content, and (iv) linking social media data with remote-sensed data. In the query refinement phase, we generate new queries in local languages spoken in the areas struck by the disaster. Subsequently, we crawl different social media platforms to collect as much information as possible. The data retrieval phase is followed by a filtering stage, where we analyze and process the retrieved content. Next, we extract the geo-location information from images and tweets, which are then used to retrieve the satellite images. Finally, the satellite images are analyzed and processed to detect the underlying disaster event. In the next subsections, we provide a detailed descriptions of these phases.

Sensing New Events (Natural Disasters)

As aforesaid, one of the main advantages of JORD is the capability to collect information about natural and technological disasters in real time, which basically means that if JORD is once started, it will continue collecting and linking events as long as they occur. To this aim, the proposed system extracts a list of natural and technological disaster events from the EM-DAT database [47] in real-time. This means that as soon as a new event occurs in the database, JORD starts collecting and linking information about it. EM-DAT is an international disaster database (supported by the World Health Organization - WHO) that provides information of natural and technological disasters that have occurred all over the world. Table 4.1 provides a list of some samples events sensed and analyzed by our system. It is to be noted that JORD is able to collect, link and analyze an unlimited number of events depending on the processing and storage resources, and can operate live as an quasi autonomous system, and on demand, namely controlled by a user.

Query Refinement and Translation

JORD utilizes the time and location information, provided in the EM-

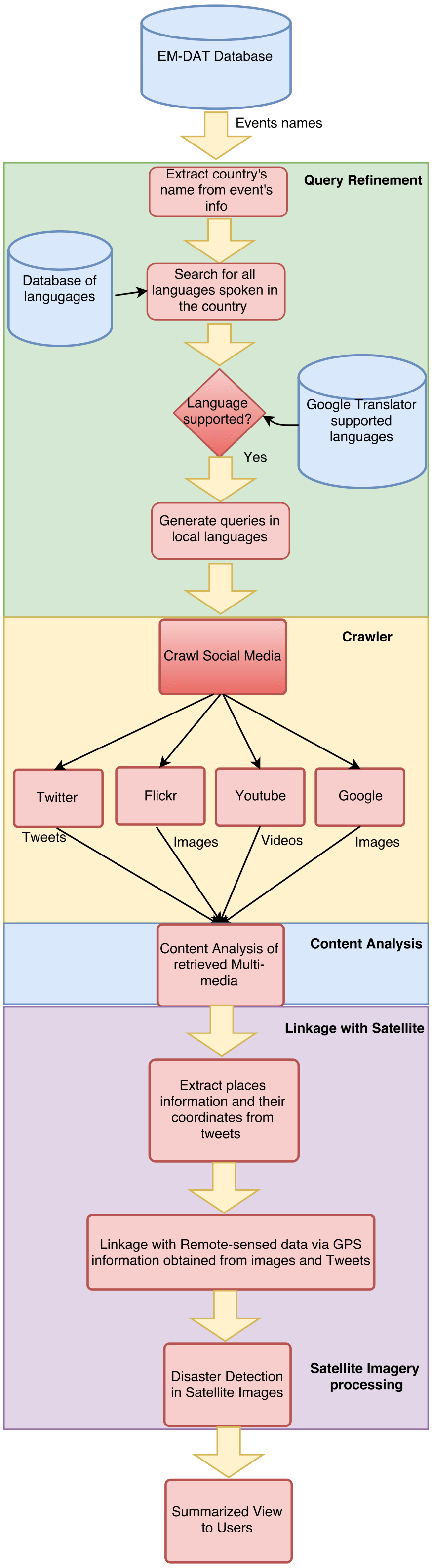

EM-DAT Database
Events names
Query Refinement
Extract country's name from event's info
Database of langugages
Search for all languages spoken in the country
Language supported?
Google Translator supported languages
Yes
Generate queries in local languages
Crawler
Crawl Social Media
Twitter
Flickr
Youtube
Google
Tweets
Images
Videos
Images
Content Analysis
Content Analysis of retrieved Multi-media
Linkage with Satellite
Extract places information and their coordinates from tweets
Linkage with Remote-sensed data via GPS information obtained from images and Tweets
Disaster Detection in Satellite Images
Satellite Imagery processing
Summarized View to Users

Figure 4.2: Block diagram of the proposed JORD system.

Table 4.1: A list of examples for natural and technological disasters retrieved by JORD.

Event	Location	Time Period	Event	Location	Time Period
Earthquake	Italy	August 2016	Floods	Laos	August 2016
Earthquakes	Esmeraldas, Ecuador	May 2016	Landslides	Kegalle district, Sri Lanka	May 2016
Cyclone Roanu	Bangladesh	May 2016	Landslides	West regions of Uganda	May 2016
Tornadoes	Oklahoma, United States	May 2016	Floods	Kilinochchi district, Sri Lanka	May 2016
Thunderstorms	Bangladesh	May 2016	Landslide	Sibolangit, Indonesia	May 2016
Landslide	Rwanda	May 2016	Floods	Ethiopia	April 2016
Landslide	Uganda	May 2016	Mudslide	Taining district	May 2016
Severe weather	Haiti	May 2016	Wildfires	Alberta province, Canada	May 2016
Thunderstorms	Uruguay	April 2016	Flash flooding	Texas, United States	April 2016
Floods	Port-au-Prince, Haiti	April 2016	Floods	Southern China	April 2016
Thunderstorms	Myanmar	April 2016	Floods	Assam, Nagaland, India	April 2016
Thunderstorms	China	April 2016	Drought	India	April 2016
Drought	Timor-Leste	April 2016	Floods	Saudi Arabia	April 2016
Earthquake	Kumamoto, Japan	April 2016	Storm	Dolores, Uruguay	April 2016
Earthquake	Ecuador	April 2016	Floods	Santiago region, Chili	April 2016
Flash floods	Yemen	April 2016	Earthquake	Kumamoto, Japan	April 2016
Earthquake	Pakistan	April 2016	Floods	Ethiopia	April 2016
Storm Katie	France and UK	March 2016	Floods	KpK Pakistan	April 2016
Severe weather	United States	March 2016	Drought	India	March 2016
Floods	Kashmir, Pakistan	March 2016	Severe weather	United States	March 2016
Floods	Indonesia	March 2016	Floods	China	March 2016
Coal mine explosion	Lougansk, Ukraine	May 2016	Shipwreck	Libya	April 2016
Thunderstorms	Uruguay	April 2016	flooding	Texas, United States	April 2016
Floods	Haiti	April 2016	Floods	Southern China	April 2016
Thunderstorms	Myanmar	April 2016	Floods	Assam, India	April 2016
Drought	Timor-Leste	April 2016	Explosion in a plant	Mexico	April 2016
Shipwreck	Lybia	April 2016	Shipwreck	Mynamar	April 2016
Plane crash	Papua New Guinea	April 2016	Storm Katie	France and UK	March 2016
Floods and landslides	Pakistan	March 2016	Earthquake	Tainan, Taiwan	Feb. 2016
Earthquake	Spain and Morocco	Jan. 2016	Floods	China	Jan. 2016
Snowstorm	East coast, United States	Jan. 2016	Earthquake	Qinghai province, China	Jan. 2016
Wildfires	Spain	Dec. 2015	Tornadoes	South of United States	Dec. 2015
Floods	Kenya	Dec. 2015	Cyclone Chapala	Yemen	Nov. 2015
Plane crash	South Sudan	Nov. 2015	Floods	Somalia	Oct. 2015
Floods	Nigeria	Sept. 2015	Wildfires	California, United States	Sept. 2015
Floods	Ibaraki (Japan)	Sept. 2015	Landslides	Kaski, Nepal	July 2015
Earthquake	Pakistan	Oct. 2005	Cyclone Winston	Fiji	Feb. 2016
Wildfires	Greece	July 2015	Floods	Myanmar	July 2015

DAT database for each event, in the retrieval of related social multimedia data including videos, images and text. JORD supports query refinement by collecting additional tags for the underlying events by determining and translating them to the local languages of the region where the disaster occurred. By looking at news or social media posts, one can easily observe that during natural and technological disasters the local community usually initiates the process of sharing news about a disaster. This normally happens by commenting, posting and sharing information using different channels and media types just after an event occurs. Furthermore, geographically close people usually tend to post and share information in their local languages. Based on these observations, our system automati-

Table 4.2: Sample queries of events translated by JORD using the Google Translate Api.

Original Query	**System-generated Query**	**Translated to**
Floods Saudi Arabia	الفيضانات المملكة العربية السعودية	Arabic
Storm and flood Dolores Uruguay	Tormentas e inundaciones Dolores Uruguay	Spanish
Earthquake Ecuador	terremoto de Ecuador	Spanish
Cyclone Chapala Yemen	الإعصار تشابالا اليمن	Arabic
Floods Ibaraki Japan	洪水茨城日本	Japanese
Floods Kashmir Pakistan	پاکستان کشمر سلاب	Urdu

cally generates queries in the local languages that are relevant to the position of a disaster. This step is achieved by including Google Translator[10] in combination with a database of spoken local languages per country in our pipeline. The list of local languages along with the country names are retrieved from InfoPlease[11], which is a free encyclopedia almanac, atlas, dictionary, and thesaurus. Table 4.2 provides some sample queries generated by our system in local languages.

Subsequently, the translated queries are used for the multimedia data collection. This blend of translated and original queries results in a larger amount of retrieved data per query (we observed that for some events, a search based on only English queries results in very little or none results). Moreover, the information retrieved with translated queries are more relevant to underlying events. As an example, Table 4.3 shows the top 5 tweets retrieved by our system for recent floods in Saudi Arabia. Both the original and translated queries along with meanings of the tweets in the local language are provided. It can be seen in Table 4.3 that most of the tweets retrieved with original query are irrelevant (e.g., most of them are

[10] https://cloud.google.com/translate/docs/

[11] http://www.infoplease.com/ipa/A0855611.html

reporting about Saudi Arabia's aid to flood victims in different parts of the world). On the other hand, the tweets retrieved with translated queries, which are posted by local community in the local language, provide more accurate and relevant information about the recent floods in Saudi Arabia.

Table 4.3: Top 5 Tweets retrieved with original English query and JORD generated query in Arabic language

Tweets with English query	**Tweets with JORD generated query**	**Meaning of the Arabic tweets**
Saudi Arabia provides 200 tons of relief and aid to those affected by the floods in Sudan	أمطار غزيرة تغسل جدة ا لمملكة العربية السعودية، شار	Heavy rain washed Jeddah, Saudi Arabia, street flooding spread
This is my pic of the day: Camels stuck in #SaudiArabia floods.	قناة تقهبر التركية الفيضانا ت تضرب المملكة العربية السعودية.	tvahaber channel Turkish Floods hit Saudi Arabia.
#SaudiArabia Sends Aid to Citizens Affected by Floods in #Sudan	فرانس برس: ارتفع عدد ضحايا الفيضانات في المملكة العربية السعودية التي تضر ب البلاد خلال الأيام الثلاثة الأخيرة الى ٧	AFP: The number of flood victims in Saudi Arabia, which hit the country during the last three days to 7
#Venezuela could collapse and take much of its #oil production with it or summer can end and #SaudiArabia really floods us. Good Luck!	إشعاعات واجهزه لتفريق السحب من إيران تت سبب فى الفيضانات فى المملكه العربية الس عودية مراجعلهاتسقط رافعة الحرم المكى	Radiation devices to disperse the clouds of Iran cause floods in Saudi Arabia Mrajolhatsagt crane Haram
Saudi Arabia floods global market destroys American jobs. #OPEC	بداية الفيضانات المفاجئة أمس في المملكة العربية السعودية، ومياه الفيضانات تغمر الجسر في ثوان!!	The beginning of the flash floods yesterday in Saudi Arabia, and flood waters submerged the bridge in seconds !!

Social Media Platforms

Social networks have emerged as important sources of information that report events in real-time, and provide a much broader story [34]. In this regard, Twitter has been widely used by commenting, posting and sharing information just after an adverse event occurs [11]. Similarly, Flickr and YouTube allow users to share audio-visual contents about an event whenever it happens. Therefore, to get an overview of an event, it is a convenient option to look for multimedia data in social media. To this aim, we crawl four different platforms, i.e., Twitter, YouTube, Flickr and Google image search. However, the queries in local languages, generated in query refinement phase, are supported by three platforms of media, namely

Twitter, YouTube and Google. Flickr supports English queries only.

Content based Filtering

In order to investigate the importance of the content based methods to filter and analyze the collected content for the end users, we conduct extensive experiments on retrieved multimedia data (images and tweets) collected by JORD. In this section, we provide a detailed description of the methodologies, we propose, for the content analysis of the collected multimedia data.

Content Analysis of Retrieved Images

The basic motivation of the content analysis of retrieved images is to filter out irrelevant or less informative images from content point-of-view, and provide the ones which visually well represent the underlying events. To this aim, we perform explorative multi-class recognition experiments on the images collected by JORD. To this aim, from the most common adverse events related crawled images, we created a dataset containing 14 classes of events. The classes are cyclone, drought, earthquake, flood, thunderstorm, tornado, wildfire, not-relevant-cyclone, not-relevant-drought, not-relevant-earthquake, not-relevant-flood, not-relevant-others, not-relevant-thunderstorm, not-relevant-tornado, not-relevant-wildfire. All in all, the dataset contains $20,934$ images.

We experimented using various configurations of two main approaches including: (i) classification using global features (GF), and (ii) classification using concepts extracted with a pre-trained deep learning model.

In the GF approach, we extract global features from the images using the latest version of the Lire open source software [75]. We extracted JCD features with a feature vector of size 167, representing texture and color in an image.

In the implementation of the deep learning based approach, we use Keras [27] with Google Tensorflow [1]. The extraction of concepts is based on the Inception-v3 model [115], which allows the extraction of $1,000$ concepts learned from the ImageNet dataset. In this case, we simply use the pre-trained model to extract all possible concepts per image, and use them as a feature vector of size $1,000$ as input in different classification configurations.

Content Analysis of the retrieved tweets

The objective of tweets analysis is two-fold. On one hand, we assess the quality of the retrieved tweets and filter out the less informative and irrelevant tweets. On the other hand, we are interested in collecting the coordinates of the areas affected by the disaster. To do so, we perform the following two experiments on the collected tweets.

- We perform binary and multi-class classification to identify and filter-out the irrelevant tweets. This experiment is intended to improve users' experience providing them with more appropriate data.
- We also analyze and extract the places and city names mentioned in the tweets' text to retrieve and link satellite imagery with the events. The basic motivation for this experiment comes from the fact that the GPS coordinates associated with tweets do not necessarily match with the location of the disaster. Moreover, the presence of GPS information in tweets is not always guaranteed. Instead, users tend to mention the exact places affected by the underlying disaster in the text.

Similarly to image analysis, we started with the collection of a dataset by choosing tweets related to eight common natural disasters from the pool of tweets retrieved by JORD. These disasters include: cyclone, drought, earthquake, floods, landslides, snow-storm, thunder-storm and wildfires. We also populate the dataset with 8 additional classes including: not-relevant-cyclone, not-relevant-drought, not-relevant-earthquake, not-relevant-floods, not-relevant-landslides, not-relevant-snowstorm, not-relevant-thunderstorm and not-relevant-wildfires. For the labeling of tweets with positive and negative samples (i.e., relevant and irrelevant tweets), annotation is performed manually. To further populate the negative samples, we crawled twitter with additional queries containing the names of the countries affected by the disaster.

To discard irrelevant tweets, we have explored two different solutions: (i) binary classification (i.e., relevant vs non-relevant), and (ii) multi-class classification with 9 classes: 8 of them refer to disaster events, while the 9th represents the non-relevant tweets. As far as the text analysis is concerned, we rely on a state-of-the-art library[12], used both for tweets' classification and to retrieve places and city names. Initially, text is broken into tokens, followed by identifying the places and city names in the extracted tokens. Some sample tweets, where the places (e.g., states, districts, city and local areas names) affected by the underlying disaster are mentioned, include : ”FIF Pakistan distribute Relief goods of Drought victims in Tharparkar (city name)”, ”The EU supports livelihoods nutrition in drought-stricken Sindh (province name) Pakistan”, ”11 dead 50 wounded in Bundibugyo (District name) landslide Uganda”. Moreover, the GPS coordinates of the identified places and cities are crawled for remote sensed imagery, which are then processed for disaster detection.

Linkage with remote sensed data

In this section, we detail how the geo-location information is used to retrieve and link remote-sensed images to the underlying events. To this aim, JORD relies on Google Earth, which provides satellite images continuously; this allows to retrieve series of images before and after a disaster. JORD extracts GPS information from the retrieved data (images and tweets) and crawls Google Earth over a time window centered in the event date. Figure

[12] https://textblob.readthedocs.io/en/dev/

4.3 shows sample satellite images of the national palace of Haiti retrieved through Google Earth before and after the earthquake. Without loss of generality, other sources of remote sensed data can be crawled and integrated in JORD. Figure 4.4 shows a sample output of our system for a query about recent floods in Kenya, where the retrieved images, tweets, videos and the satellite data from Google Earth are shown.

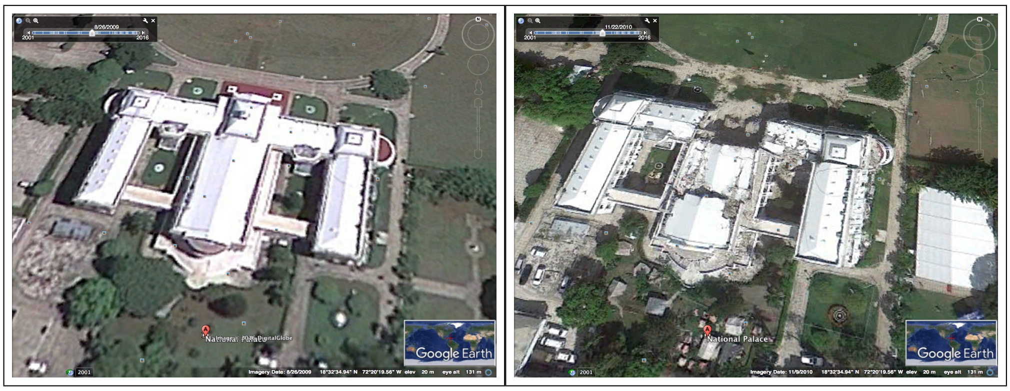
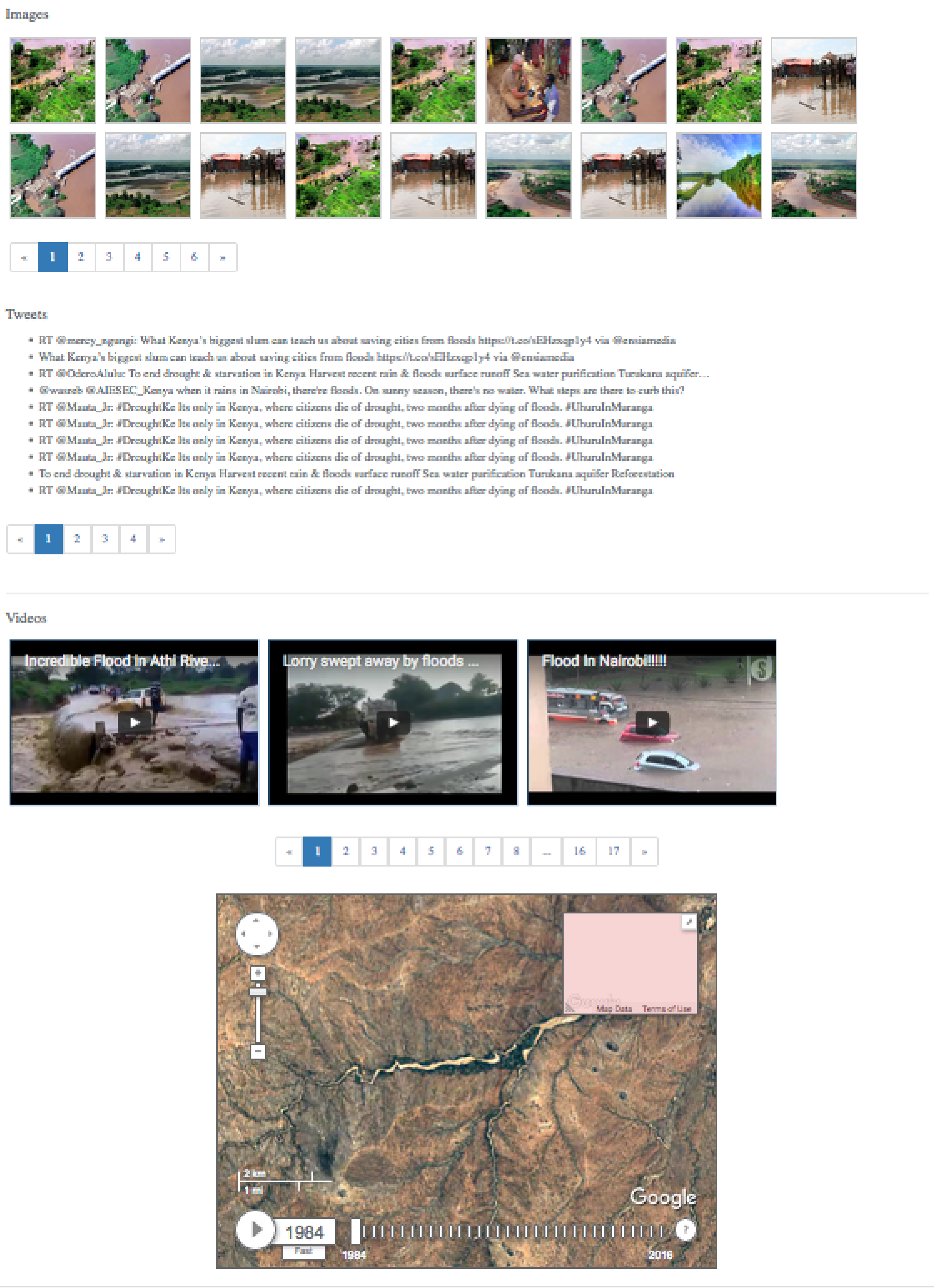


Figure 4.3: Sample Google Earth images before and after Haiti earthquake.

Flood detection in satellite images: a use-case

In this section, we present the application of our method to the use case of flood detection in satellite images. The proposed method is designed to process image patches of satellite images covering a wide spatial areas of multiple instances of flooding events. The satellite image patches are usually recorded during (or shortly after) the flooding event at different locations. The basic satellite imagery used in this work has been taken from Planet's 4-band satellites [118]. The whole dataset and the corresponding data usage instructions are publicly available [17] and consists of a set of image patches with corresponding pixel-level segmentation masks of the flooded areas. The image patches are stored in the non-normalized 4-channel 16-bit TIFF file format while the corresponding segmentation masks are stored in the 1-channel 8-bit PNG file format.

The image patches consist of fours 16-bit channels: Red (R), Green (G),



- RT @mercy_ngangi: What Kenya's biggest slum can teach us about saving cities from floods https://t.co/sEHzxgp1y4 via @ensiamedia
- What Kenya's biggest slum can teach us about saving cities from floods https://t.co/sEHzxgp1y4 via @ensiamedia
- RT @OderoAlulu: To end drought & starvation in Kenya Harvest recent rain & floods surface runoff Sea water purification Turukana aquifer…
- @wasreb @AIESEC_Kenya when it rains in Nairobi, there're floods. On sunny season, there's no water. What steps are there to curb this?
- RT @Mauta_Jr: #DroughtKe Its only in Kenya, where citizens die of drought, two months after dying of floods. #UhuruInMuranga
- RT @Mauta_Jr: #DroughtKe Its only in Kenya, where citizens die of drought, two months after dying of floods. #UhuruInMuranga
- RT @Mauta_Jr: #DroughtKe Its only in Kenya, where citizens die of drought, two months after dying of floods. #UhuruInMuranga
- RT @Mauta_Jr: #DroughtKe Its only in Kenya, where citizens die of drought, two months after dying of floods. #UhuruInMuranga
- To end drought & starvation in Kenya Harvest recent rain & floods surface runoff Sea water purification Turukana aquifer Reforestation
- RT @Mauta_Jr: #DroughtKe Its only in Kenya, where citizens die of drought, two months after dying of floods. #UhuruInMuranga

Figure 4.4: A sample output of JORD in terms of retrieved images (at the top), Tweets, Videos and the satellite data.

Blue (B) and Infrared (IR). None of the existing satellite image visualization software was able to display such data correctly. Moreover, most of the existing image processing softwares are designed to be used with standard three-channel RGB images. To overcome this issue, we decided to convert each image patch into a pair of images, namely three-channel RGB and single-channel IR images. After the extraction of raw channels data, we performed the normalization for both image components, independently. For the RGB images, we use the joint three-channel normalization, which fits all the R, G and B pixel values of the input geo-image into the standard 0-255 RGB values region. It has to be noted that the normalization coefficients are kept same for all three channels, which helps to achieve real color balance even in cases of low variations in one of the three components. The normalization of the IR component is performed separately, as shown below:

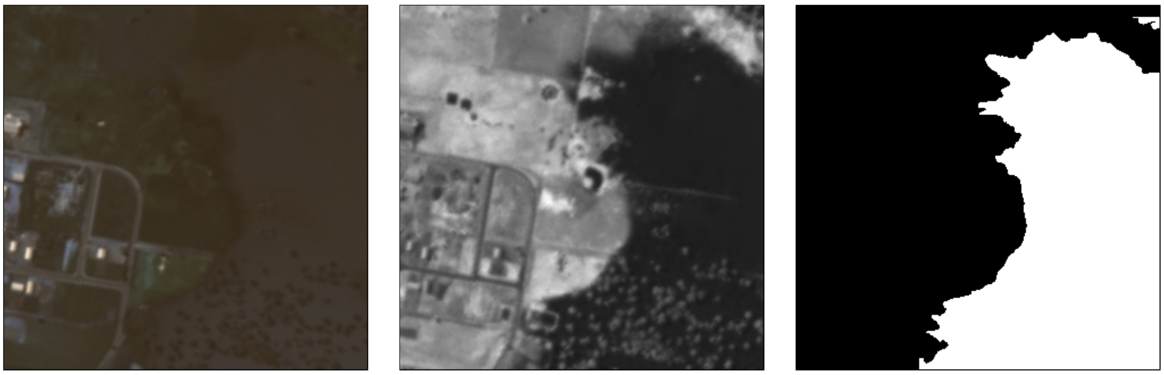
$$rgb_{min} = \min(\min_{i \in R} r_i, \min_{i \in G} g_i, \min_{i \in B} b_i)$$

$$rgb_{max} = \max(\max_{i \in R} r_i, \max_{i \in G} g_i, \max_{i \in B} b_i)$$

$$ir_{min} = \min_{k \in IR} ir_k, \quad ir_{max} = \max_{k \in IR} ir_k$$

$$\forall i \in \{R|G|B\} \quad \{r|g|b\}_i^* = \frac{(\{r|g|b\}_i - rgb_{min}) * 255}{rgb_{max} - rgb_{min}}$$

$$\forall k \in IR \quad ir_i^* = \frac{(ir_k - ir_{min}) * 255}{ir_{max} - ir_{min}}$$



(a) Three-channel normalized RGB image. (b) Single-channel normalized IR image. (c) Flooding area segmentation mask.

Figure 4.5: Example of the converted image patch from the original satellite imagery.

After the conversion to RGB and IR image pairs (see example in figure 4.5), we performed visual analysis of the converted images in order to assess the resulting image quality, the correctness of the conversion and the contents of the dataset. We found the images to be non-contrast, blurry and significantly color-range-limited. During our initial experiments, we realized that it is not possible to use off-the-shelf image segmentation frameworks due to the nature of the provided satellite imagery. Based on our previous experience [97], we decided to use GANs for the segmentation task. GANs [44] are machine learning algorithms used in unsupervised learning, and implemented via two neural networks, namely Generator and Discriminator, contesting with each other in a zero-sum game framework. They achieved promising results both in terms of performance and data processing speed in image segmentation tasks.

As the basis for our method, we use a neural network architecture originally developed for the retinal vessel segmentation in fundoscopic images with GANs (V-GAN) [111]. The V-GAN architecture [111] is designed for the processing of retinal images that have comparable visual properties, and provides the required output with one-class per-pixel image segmentation output. The basic insight behind V-GAN is to treat the vessels detection/segmentation as an image translation task, where the generator network is supposed to generate the segmentation map of the input fundoscopic image. On the other hand, discriminator network tries to refine the output of the generator. V-GAN proved to be very effective in the detection of fine vessels, and at the same time less affected by false positive compared to existing approaches [111].

In order to adapt V-GAN to our flood detection approach, we modified the network architecture by changing the top-layers configuration in order to support both standard three-channel RGB and four-channel RGB+IR geo-image-compatible input. Furthermore, the final layer of the generator network is extended with a threshold activation layer to generate the binary segmentation maps.

During our initial experiments with our model, we observed that, though the modified V-GAN is able to perform the segmentation of the provided satellite images, the estimated performance metrics were below the expected level. Additional visual analysis of the converted RGB and IR images showed that sometimes the IR component of the sourced geo-images is irrelevant to the flooding areas, which is one of the possible reasons that caused our model to be biased during the training process, preventing it from the extraction of the properties of the flooding areas. Based on these considerations, we decided to exclude the IR component from the model input, and process the RGB components only, which resulted in a good detection performance. Furthermore, we continued to investigate deeper into the multi-channel approach and, after debugging of our model, we realized that the used normalization scheme is causing problems. Despite the good results obtained from the detection using the RGB-normalized images only, the independent normalization procedure of IR channel was resulted in the significantly base-value-shifted output images, mostly because of the high variations in the IR channel caused by the significant difference of the value of the reflected IR light depending on the day time and cloud coverage for the area. To resolve this problem, we redesigned the data preparation and augmentation code as well as the input layer of our model in order to support direct input of the raw satellite imagery data. This resulted in significant improvements in the model training behaviour and allowed us to perform experimental evaluations using both RGB and RGB+IR channels configurations.

System Evaluation via Crowd-sourcing

To evaluate the system, in terms of if the retrieved multimedia contents about the events are correct and useful for the users, we conducted a crowd-sourcing-study on Microworkers. We asked workers to give their opinion about the retrieved multimedia contents, including images, tweets and videos, related to underlying events. We paid each worker 1.50 USD and tried to be as fair as possible regarding the discarding of workers. As shown in [103, 6], controlling and discarding workers too much can lead to an undesired outcome of the study, which we tried to avoid by accepting almost every worker if they did the task in a reasonable way. We asked the crowd-workers five different questions:

1. Do you think the system provided information was useful? This question aims to get feedback from workers about the usefulness of the collected multimedia data using a scale from one (not useful) to five (very useful).

2. From three possible events, which one do you think has been the one presented to you? This question was used to evaluate if JORD can help the user to understand the retrieved event.

3. How useful was each type of information for you? Here, the worker had to scale the usefulness of different type of multimedia content (images,

tweets and videos) from one (not useful) to five (very useful).

4. If such a system would exist, would you use it? This was a simple yes or no question where we asked the workers if they would use such a system or not.

5. Why would you use or not use it? This was an open question where the crowd-workers had to reason their yes or no from the previous question. We used this question to filter out workers who did the task in a wrong way. We checked each answer manually for each worker. If the answer made sense and showed that the worker was thinking about it, we accepted it. If not, we did not include it in the final evaluation.

4.3.2 Multimedia Satellite: A Benchmark Task

Overview

In this section, we present the solution proposed for the MediaEval 2017 Multimedia and Satellite challenge [17]. The basic insight of the challenge is to jointly utilize satellite imagery and social media to provide a detailed story of the underlying disaster event. The challenge is mainly composed of two sub-tasks namely (i) Disaster Image Retrieval from Social Media (DIRSM) and (ii) Flood Detection in Satellite Images (FDSI). The basic insight of the first task is to design and develop a system that is able to identify flood related images in a collection of images along with meta-data from social media. Although additional information, such as user's tags, date on which the image is taken along with geo-location information are provided, only the images having a visual evidence of a flood are considered to be true positive samples. The main challenge of DIRSM task is to differentiate in images of lakes and flooded streets as well as among the different types of flooding, such as coastal flooding and river flooding.

In order to efficiently utilize all of the available information, the participants were asked for three different runs using:

- Visual Information, only
- Meta-data, only
- A combination of meata-data and visual information

The second task is mainly concerned to develop a system that is able to identify flooded regions in satellite images. The participants are provided with satellite image patches of multiple instance flooding events along with the segmentation masks in the development set to train their model.

Proposed Approach

Methodology for DIRSM Task

To tackle challenge (i), we rely on Convolutional Neural Network (CNN) features. In detail, first we extract CNN features for seven different models from state-of-the-art architectures pre-trained on the ImageNet [33] and places datasets [144]. These models include AlexNet [66] (pre-trained on both ImageNet and places datasets), GoogleNet [114] (pre-trained on ImageNet), VGGNet 19 [110] (pre-trained on both ImagNet and places datasets) and different configurations of ResNet [53] with 50, 101 and 152 layers. In total, Alexnet is composed of 8 weighted layers , Google Net contains 22 layers while VGGNet19 has 19 layers. For feature extraction from Alexnet and VGGNet19, we use the Caffe toolbox[13] while in the case of GoogleNet and Resnet we exploited Vlfeat Matcovnet[14].

All in all, we extract eight feature vectors through four different network architectures from the same image. AlexNet and VGGNet16 provide a feature vector of size 4096 while GoogleNet and Resnet provide feature

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

[14] http://www.vlfeat.org/matconvnet/

vectors of sizes 1024 and 2048, respectively. Subsequently, the extracted features are fed into ensembles of Support Vector Machines (SVMs), which provide classification scores in terms of posterior classification probabilities. We also consider user's tags, date taken along with GPS information from the available meta-data. For the classification of meta-data, we rely on the Random Tree classifier provided by the WEKA toolbox [50]. We opted for SVM on the account of its proven efficiency in many applications, such as object recognition [21] and remote sensing [16]. On the other hand, for textual features, we tried different classification techniques and choose Random Forest based on its better performance on the development set. Finally, the classification scores obtained through Random Trees and SVM trained on meta-data and visual features are fused using a late fusion mechanism. For the late fusion, we propose two different methods, namely, (i) Induced Ordered fusion scheme inspired by Induced Ordered Weighting Averaging Operators (IOWA) by Yager et al. [141] and (ii) Particle Swarm Optimization (PSO). Figure 4.6 provides a block diagram of the proposed methodology for the Disaster Images Retrieval from Social Media (DIRSM) task.

Methodology for FDSI Task

For the challenge (ii), we used the methodology described in Section 4.3, where we rely on a neural network architecture originally developed for the retinal vessel segmentation in fundoscopic images with generative adversarial networks (V-GAN) [15]. The only difference is that in this particular challenge we used only RGB components of the satellite imagery. On the other side, we perform additional analysis in Section 4.3.

[15] `https://bitbucket.org/woalsdnd/v-gan`

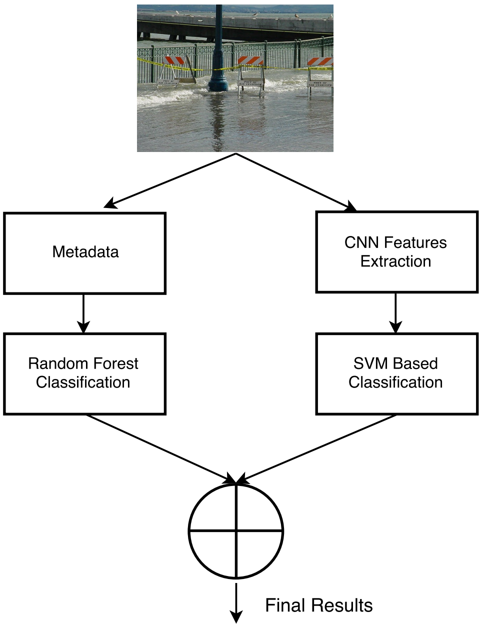


Figure 4.6: Block diagram of the proposed methodology for DIRSM task.

4.4 Experiments

4.4.1 JORD

Content Analysis

As mentioned earlier, based on the collected data by JORD, we conducted experiments with the goal of exploring if the content information of the collected data can help to improve the results for the users. As a first step, in this section, we present the experimental results of our content based analysis of the JORD collected images and tweets.

Content Analysis of the Retrieved Images

In this work, for the classification of the retrieved images, we use Random Forest (RF) and Logistic Model Tree (LMT) classifiers provided in the Weka machine learning library [50]. It is worth mentioning that in the current implementation, we do not perform any data augmentations, such as cropping, for any of the approaches. For the evaluation, we use 10-fold cross validation to get a robust and representative results. We tested

Table 4.4: Classification results of our content based analysis

Features	Classifier	Precision	Recall	F-Measure
Concept	Random Forrest	0.564	0.51	0.452
Global Feature	Random Forrest	0.544	0.493	0.466
Global Feature	Simple Logistic	0.37	0.385	0.346
Concept	Simple Logistic	0.426	0.42	0.394
Baseline	ZeroR	0.06	0.244	0.096

several different classifiers, and the baseline is calculated with the ZeroR classifier that finds and uses the majority class in the dataset for classification. Table 4.4 shows the experimental results of our content analysis. Over all, in terms of precision and recall, we get better results with Random Forest on the concepts extracted through Inception-v3 [115]. During the experiments, we observed that certain disasters, such as earthquakes, wildfires, floods and cyclone, have specific textures and patterns, and thus are easy to be identified and recognized through visual content. However, we also noticed that for certain disasters, such as tornadoes and droughts, it is very difficult to recognize or differentiate among them through visual content. Nevertheless, we observe that using the visual content for filtering the results can be a promising step as also confirmed by our results.

Content Analysis of the Retrieved Tweets

As mentioned earlier, for filtering out the irrelevant tweets, in the proposed work we perform two different experiments (i) binary and (ii) multiclass classification. In both experiments, we rely on two different classifiers namely, Navie Bayes and Decision Tree classifier provided in TextBlob toolbox[16]. Table 4.5 provides experimental results of our binary classification, where tweets are investigated to be relevant or non-relevant to a particular disaster event, with Naive Bayes and Decision Tree classifiers. We report the results of our tweet classification experiments in terms of accuracy, precision, recall and F1 score. As can be seen, though we achieve en-

[16] https://textblob.readthedocs.io/en/dev/

Table 4.5: binary tweet-classification results with Naive Bayes and Decision Tree classifiers

Disasters	**Naive Bayes**				**Decision Tree**			
	Acc.	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**
Cyclone	.9322	.9189	1.0	.9557	.9661	.9705	.9705	.9705
Drought	.6451	.6956	.8421	.7640	.774	.973	.77	.859
Earthquake	.8521	.7972	.9833	.7618	.9565	.933	1.0	.965
Floods	.8971	.851	.9273	.8876	.92	.8659	.9491	.9056
Landslides	.9271	.9176	.975	.9457	.8344	.851	.862	.8569
Snowstorm	.8785	.8387	1.0	.9126	.8598	.8431	.347	.886
Thunderstorm	.8095	.7441	.8648	.8003	.8095	.729	.843	.7818
Wildfires	.6744	.75	.78	.7352	.9065	.92	.92	.92

Table 4.6: Multi-class classification with Decision Tree classifier

Event	**Accuracy**
Cyclone	.9491
Drought	.7903
Earthquake	.9565
Floods	.901
Landslides	.8741
Snowstorm	.8037
Thunderstorm	.75
Wildfires	.9069
Non-relevant	.912

couraging results with both classifiers, Decision Tree classifier has a slight improvement over the Naive Bayes classifier.

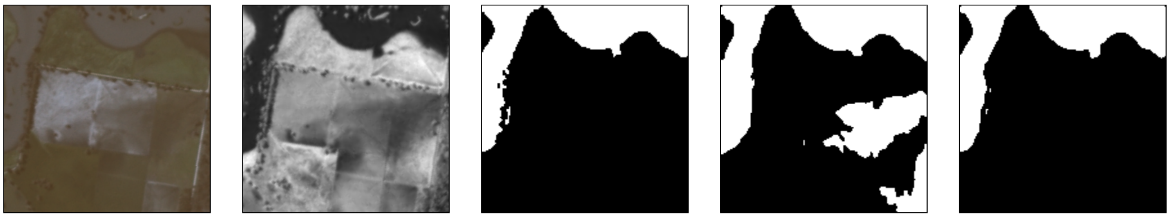
In order to further investigate the performance of our proposed content based filtering scheme, we also perform multi-class classification using Decision Tree classifier. Table 4.6 reports the experimental results of our multi-class classification experiment. Overall, we achieve better results on each event motivating the fact that it will lead to an improved experience for the users.

Flood Detection in Satellite Images

For the detailed experimental evaluation of our algorithm of flood detection in satellite images, we used the publicly available dataset [17] of the Multimedia Satellite Task, which was a part of the 2017 MediaEval Benchmarking Initiative for Multimedia Evaluation [17]. The dataset consists of development and validation sets of satellite image patches with corresponding pixel-level segmentation masks of the flooded areas. The development set consists of 463 image patches with corresponding flooding segmentation masks. The test set contains 260 image patches along with the flooding segmentation masks. The dataset covers seven different flooding events occurred in the different regions of the world. In order to evaluate the generalization properties of our detection algorithm, we mixed all the images from the different events and regions interpreting the image sets as the data sources with unknown geographical, temporal and event-related information. In this experimental study, we evaluate our method with a two-fold cross-validation strategy. During a first evaluation run, we used the dataset in the original order: the development set is used as a training set, and the validation set as a test set. In the second evaluation run, we used the dataset in the flipped order: the development set is used as a test set, and the validation set is used as a training set. The non-equal splitting of the number of images in the training and test sets can be seen as an additional test for the redundancy and the efficiency of the proposed detection algorithm. Moreover, we also performed the evaluation of both detection approaches: three-channel normalized RGB and four-channel raw RGB+IR, which gave us four different evaluation runs in total.

The proposed neural network model performs flooded areas detection on the pixel-level, and provides the output in the form of a binary segmen-

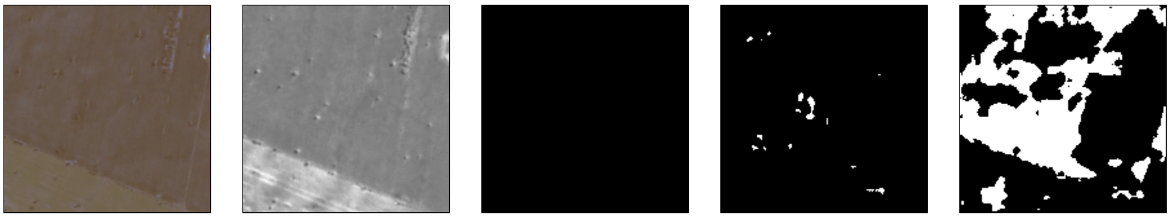
[17] http://www.multimediaeval.org/mediaeval2017/



(a) Input RGB channels (b) Input IR channel (c) Ground truth mask. (d) Segmentation of RGB. (e) Segmentation of RGB+IR.

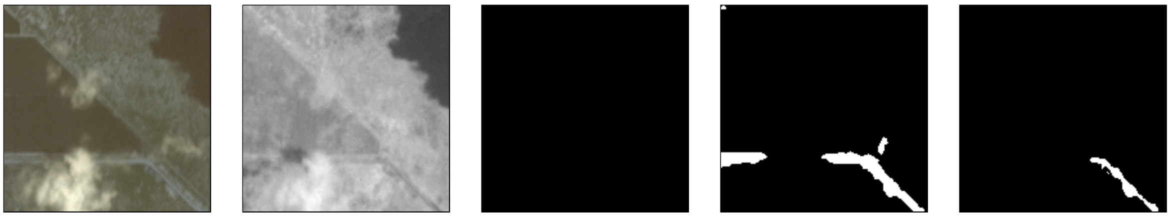
Figure 4.7: Example of the correctly found flooded area. This example shows that detection was performed better for combination of RGB and IR channels.

tation map, which contains true values (white pixels) for the pixels belongs to the detected flooded areas and false values (black pixels) for the areas without flooding detected. The examples of the model's segmentation output together with the source RGB and IR images as well as corresponding ground truth masks are presented in figures 4.7, 4.8 and 4.9. As one can see, the RGB and IR channels provide a different information about the region being analyzed. In most of the cases, the four-channel combination of RGB and IR channels results in the better detection performance and can increase the detection accuracy significantly (see figure 4.7 for an example). Nevertheless, in some cases when the IR channel contains data that confuses the detection algorithm and leads to a mis-detection with a tendency to increase number of false-positive pixels (see figure 4.8 for an example). Moreover, in some quite rare cases (at least within the dataset used) no combination of the channels are sufficient to perform a distinctive and accurate detection of the flooded areas because of a presence of a water that is "legal" (see figure 4.9), for example in the irrigation channels, normal rivers and lakes, etc. To be able to deal with such cases the comparative time-based analysis must be added to the detection algorithm utilizing as many satellite images of the same region taken in different periods as possible. The time-based analysis will be a subject of a future work of our research.



(a) Input RGB channels (b) Input IR channel (c) Ground truth mask. (d) Segmentation of RGB. (e) Segmentation of RGB+IR.

Figure 4.8: Example of the correctly found flooded area. This example shows that in some cases detection was performed better for three RGB channels.



(a) Input RGB channels (b) Input IR channel (c) Ground truth mask. (d) Segmentation of RGB. (e) Segmentation of RGB+IR.

Figure 4.9: Example of the false positive detection of the flooding. The water in this image patch is ”legal” water in the irrigation channels. To be able to deal with such cases the comparative time-based analysis must be added to the detection algorithm utilizing many satellite images of the same region taken in different periods of time.

Our proposed model for the flood detection includes the top layer with an adjustable threshold parameter, which is used for the final output segmentation map binarization. The value of this threshold parameter defines a border line for each pixel to be counted as belonging to flooding area depending of the model’s output probability value, and it has a direct effect on the number of flooded pixels detected. Thus, in order to perform a complete model evaluation, we have repeated all four evaluation runs with different values of the threshold parameter. For an overall performance evaluation of this threshold-value-effect evaluation experiments, we selected the Matthews correlation coefficient (MCC) which is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives, and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. The MCC value lies in the region between -1 and +1. Literature [96] also confirms that MCC is the most convenient metric for the binary classification tasks.

The results of the threshold value evaluation are depicted in figure 4.10. As one can see, the low < 0.1 and the high > 0.9 values of the threshold have strong negative effect on the performance of the proposed model. The optimal threshold value lies, as it was expected before the experimental studies, in between 0.4 and 0.6 depending on the exact order of the samples in the dataset and the number of channels used. The best value of the threshold parameter is 0.42 for the three-channel RGB model regardless of the sets order. For the four-channel RGB+IR model, best value is 0.518. Despite the fact that the best values of the threshold parameter are slightly different depending on the dataset and number of channels used, the resulting difference of the performance is small for the threshold values within the interval from 0.4 to 0.6 (which includes the found best values) and for the future work we will use the threshold value of 0.5 for all the cases. Nevertheless, in this work we have performed an evaluation of the performance metrics of the developed detection method using the best found threshold values. The interesting finding is that RGB and RGB+IR approaches perform almost equally for the original dataset, but RGB+IR performs better for the flipped datasets. That can be caused by a significantly reduced size of the training dataset in the flipped runs, which makes one additional information channel important for the proper model generalization during training process.

The results of the performance evaluation for the best threshold values are presented in Table 4.7. The first two runs was performed by the three-channel RGB model and the original and flipped datasets. The threshold

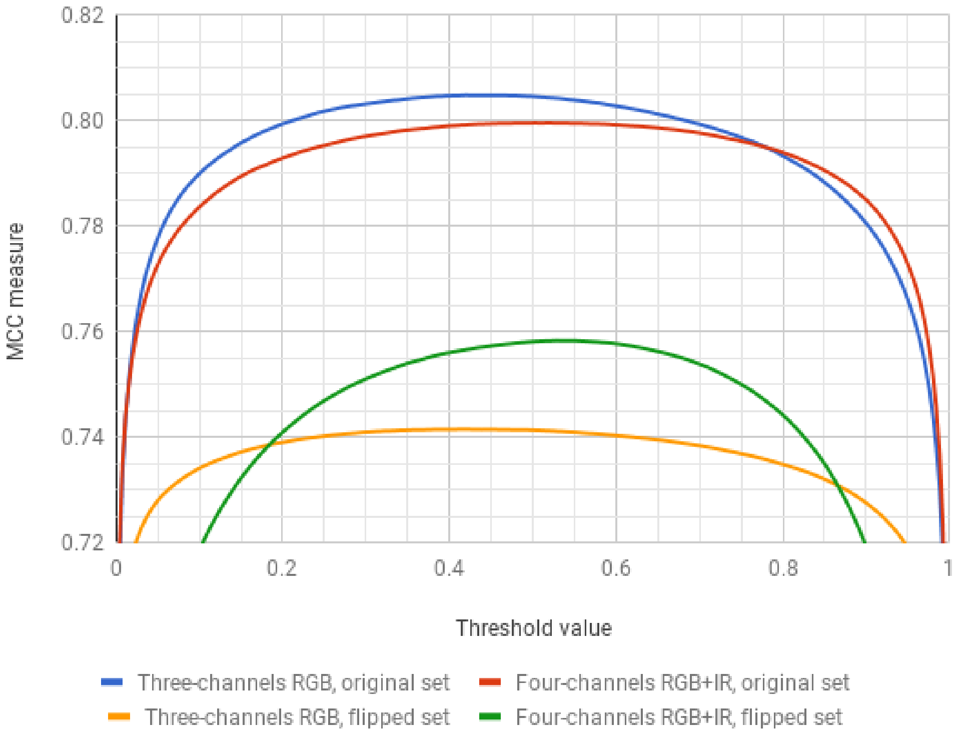


Figure 4.10: The comparison of the flooding detection performance in terms of MCC measure computed with the different probability threshold values for three- and four-channel satellite imagery data for the original and the flipped datasets.

value used was 0.42 for both runs. The results show that the best MCC evaluation performance metrics value of 0.805 was achieved for the original datasets order. For the flipped datasets order, MCC metrics was a slightly lower with a value of 0.742 which can be caused by the significantly lower number of the training images in the flipped datasets that caused a less level of the model generalization. Nevertheless, the MCC values as well as other common performance characteristics depicted in table 4.7 confirms the validity and usability of the model developed together with the high adaptation rate and ability to learn even on the limited training dataset size.

The performance results computed for the four-channel RGB+IR model

Table 4.7: Two-fold cross-validation results for the two presented flooding detection approaches. The performance numbers of accuracy (ACC), precision (PREC), sensitivity or recall (REC), specificity (SPEC), F-Measure (F1) and Matthews correlation coefficient (MCC) are presented in the original / flipped order regarding to the original dataset [17] for the selected values of the probability threshold value (THRESH).

Input		**Thresh**	**Acc.**	**Pre.**	**Rec.**	**Spe.**	**F1**	**MCC**
RGB	Orignal set	0.420	0.913	0.879	0.862	0.940	0.870	0.805
	Flipped set		0.883	0.835	0.827	0.913	0.831	0.742
RGB+IR	Orignal set	0.518	0.911	0.883	0.849	0.943	0.865	0.800
	Flipped set		0.889	0.827	0.862	0.904	0.844	0.758

runs (see table 4.7) with the threshold value of 0.518 shows almost the MCC performance values of 0.8 and 0.758 respectively. As one can see, the original datasets run has the same performance as the RGB model (the difference is not significant). For the flipped run the RGB+IR model perform noticeable better, thus using of four-channel RGB+IR model can be considered as preferable for the flooding detection tasks. Moreover, visual inspection of the datasets provided showed that in some cases IR channel may provide distinctive clues for distinguishing between flooded areas and ”normal” water areas, but this should be investigated deeper using more datasets of bigger sizes.

Crowd-sourcing Analysis

In the crowd-sourcing-study, we had 349 distinct valid responses. We discarded 36 because their answers in the open question showed that they did not understand the task or did it in a wrong way. We did not discard workers, if they did not choose the right event in question 3 since this question was intended to evaluate how useful and informative the retrieved content is.

The first question (i) where the workers had to state if they find the system useful or not, had an average score of 4.47 for all workers. This is a clear indication that workers find the system and the provided information useful.

For the second question (ii) where the workers had to chose the correct event out of three possible ones, only 19 workers out of 349 failed to correctly recognize the event presented to them. A closer investigation showed that all of them had just the country wrong but gave a correct answer about the disaster. This shows two important things. Firstly, that the retrieved information of JORD is accurate and can help users to get more information about events and secondly, that connecting it to satellite images and showing on a map is important to improve understanding of the event in terms of location.

For the third question (iii) where the workers had to report how useful they find the different types of multimedia content we got an average of 4.23 for images, 4.08 for tweets and 4.44 for videos. Based on this, we can see a tendency that users find videos most useful and tweets least. We think that this might be due to the fact that a video usually contains more information than a text or image and that it helps people more to understand and experience the current situation.

The last two questions (iv) and (v) are evaluated together. For the first question (iv), 336 from 349 workers (around 96%) stated that they find the system useful, which is a promising indicator that such a system would be useful and used by users. Having a closer look into the answers of the last question (v) from crowd-workers that did not find it useful revealed the following reasons: they find the system scary; can use Google; would use a system that can predict event; or not do and never will face such an event. Examples from users who would like to use the system are: ”I was a victim of Katrina...”, ”I like how it provides different forms of media, from different perspectives of the world”. ”It is an interesting way to view news;

Videos are always more impressive than images and tweets; I would like to use it to get better trusted info; and It is informative and gives a very COMPLETE view of what is happening". "I love the use of ALL forms of information, as in, photos, videos, and tweets".

Based on our evaluation using crowd-sourcing, it appears that such a system would be interesting and useful for users.

4.4.2 Multimedia Satellite: A Benchmark Task

In this section, we provide a detailed description of the conducted experiments, and their analysis along with the details of the dataset.

Dataset

For the DIRSM task, a total of 6,600 Flickr images along with the additional information in the form of meata-data are provided from YFCC100M-Data [119]. The meta-data include user's tags, user's id, user's nickname, title, description, longitude and latitude. The dataset is provided in two separate parts namely (i) development set and (ii) test set. The development set contains a total of 5,280 images along with the labels while the test set is composed of 1,320 images. The images are collected with relevant tags, such as flooding, floods and flood. However, human annotators are used to rate the collected images based on their relevance with the events.

On the other hand, satellite image patches obtained from Planet's 4-band satellites with ground-sample distance (GSD) of 3.7 meters [118] have been provided for the FDSI task. The dataset mainly contains images from 8 different flood events and is collected during 01.06.2016 and 01.05.2017. All the patches have been projected in the UMT projection using WGS84 datum, and are provided in the GeoTiff format having shape of 320x320x4 pixels. Moreover, each patch is composed of 4 channels, namely Red (R), Green (G), Blue (B) and Infrared band (IF) information. Similar to DIRSM, FDSI dataset is divided into development and test test. The development set contains 462 image patches from 6 different locations. However, the test set has been further divided into two subsets. The first test set contains unseen image patches from the same locations covered in the development set while the test set 2 contains unseen image patches from different location not covered in the development dataset. Figure 4.13 provides some sample images from both datasets.

Runs Description in DIRSM Task

For DIRSM, we submitted five different runs. Table 4.8 provides the official results of our methods in terms of average precision at the cut-off 480 and mean over precision at different cutoffs (50, 100, 250, 480). Run 1 and run 4 are mainly based on visual information extracted with seven different CNN models, which are jointly utilized in PSO and IOWA based fusions, respectively. As can be seen in Table 4.8, the PSO based fusion method outperforms IOWA with a significant gain of 3.79% and 5.34%. On the other hand, run 2 is based on meta-data achieving the worst results among the all runs. This lower performance with meta-data reveals that the additional information available in the form of meta-data is not much useful in this particular case. Similarly, run 3 and run 5 represents two different variations of our method used for combining meta-data and visual information. Run 3 is based on IOWA while run 5 represents our PSO based fusion of meta-data and visual information. Again, PSO based fusion performs better. One of the main limitations of IOWA based fusion is its mechanism of assigning more weight to a more confident model. In this particular case, we noticed that our classifier trained on meta-data provides more confident decisions with high probabilities causing significant

Figure 4.11: Sample images from the dataset for DIRSM task.

Figure 4.12: Sample images from the dataset for FDSI task.

Figure 4.13: Sample images from the Multimedia and Satellite challenge.

Table 4.8: Evaluations of the proposed approach in terms of precision at 480 and mean over average precision at different cutoffs (50, 100, 250 and 480).

Run	**Features**	Precision at 480	**Mean precision**
1	Visual only	84.94%	95.11%
2	Meta-data only	25.88%	31.45%
3	Meta-data and Visual	54.74%	68.12%
4	Visual only	81.15%	89.77%
5	Meta-data and Visual	73.83%	82.68%

Table 4.9: Evaluations of the proposed approach (team 2) in terms of precision at 480 and mean over average precision at different cutoffs (50, 100, 250 and 480).

Run	**Features**	**Precision at 480**	**Mean precision**
1	Visual only	86.81%	95.73%
2	Meta-data only	22.83%	18.23%
3	Meta-data and Visual	83.73%	92.55%

reduction in the performance. This can also be concluded from the results on run 2 where the meta-data obtain worst results.

During our experiments on the development set, we noticed some CNN models providing better results compared to others. Particularly, better results are obtained for AlexNet. As a separate team, we submitted 3 different runs with visual information extracted with AlexNet pre-trained on ImageNet and Places datasets only. Table 4.9 provides the experimental results of our second team participated in the task.

We also provide the comparison of our method against the methods proposed by other participants. As can be seen in Table 4.10, overall, on visual information we got first and second places on both at cutoff 480 and mean average precision at 50, 100, 250 and 480. Although, the lower performance on meta-data affects the performance of our fusion methods, we achieved higher performance at cutoffs 50, 100, 250 and 480. Moreover, better results of other participants on meta-data shows a hint of improvement in this direction. In future, we aim to develop better schemes to

Table 4.10: Comparison against other participants on DIRSM task

Team	**Cutoff 480**			**Cutoff (50,100, 250 and 480)**		
	Visual	Meta-data	Visual + Meta-data	Visual	Meta-data	Visual + Meta-data
WISC [87]	.5095	.6678	.8087	.6275	.7437	.7226
Lopez et al. [68]	.6158	.6754	.6840	.6638	.7016	.8396
ELEDIA [83]	.7762	.5707	**.8541**	.8787	.5712	**.9039**
Konstantinos et al. [64]	.7882	.3615	.6857	.9227	.3990	.8337
Keiller et al. [62]	.7460	**.7671**	**.9584**	.8788	.6253	.8563
Zhengyu et al. [142]	.5146	.6370	.7316	.6470	.7574	.8543
Hanif et al. [84]	.65	.649	.646	.8098	**.7179**	.8484
BMC [139]	-	-	-	.1921	.1284	.1830
UTAOS (ours) [5]	**.8494**	.2585	**.7383**	**.9511**	.3145	.8268
MLDCSE (ours) [8]	**.8681**	.2283	**83.73**	**.9573**	.1823	**.9255**

make better use of the additional information.

Runs Description in FDSI Task

Table 4.11 represents the experimental results of our proposed method for Floods Detection in Satellite Images (FDSI) task. In total, we submitted 5 different runs for 7 different locations. We have used different binarization threshold level for the different runs with the same model in order to find the optimum balance in the number of false-positive and false-negative pixels in the segmented images. The best results are reported for location 03 (which have the best ground visibility without clouds and proper lighting with strong light reflections from the water surface in the flooded areas) in all runs. Overall better results are obtained at runs 3 and 4 with threshold .5 and .35 achieving mean IoU of 0.83 each. For the new location (07) better results are obtained at runs 3 and 4.

4.5 Summary

In this chapter, we presented two approaches to jointly utilize information from social media and satellite imagery to provide a more detailed story

Table 4.11: Evaluations of the proposed approach for Floods Detection in Satellite Images (FDSI) task

Run(Thresh.)	**Mean IoU per Location**							
	01	**02**	**03**	**04**	**05**	**06**	**Overall**	**07 (new)**
1 (0.78)	0.79	0.81	0.88	0.78	0.75	0.80	0.82	0.73
2 (0.94)	0.77	0.78	0.86	0.74	0.72	0.78	0.80	0.70
3 (0.5)	0.79	0.82	0.88	0.79	0.76	0.81	0.83	0.74
4 (0.35)	0.79	0.82	0.87	0.79	0.77	0.80	0.83	0.74
5 (0.12)	0.78	0.80	0.86	0.78	0.77	0.78	0.81	0.73

about a natural disaster. In the first part, we presented our system JORD which is able to autonomously and automatically retrieve social media data from various social platforms about natural disasters and links it to remote-sensed images. Moreover, a hierarchical filtering mechanism relying on temporal and content analysis. We also demonstrated that queries in local languages that are relevant to the exact position of natural disasters retrieve more accurate information about a disaster event. We also presented a novel approach to extract places and city names to find the coordinates of flood affected area, which are used to retrieve satellite imagery and link it with the underlying events.

The evaluation of the JORD system was carried out through a crowdsourcing-study where workers were asked to evaluate the usefulness of the system and to identify an event presented to them with collected images, tweets and videos. The evaluation indicates that JORD works very accurate without human input, and it can be used to collect and merge a large number of event based data for technological and environmental disasters from different sources.

In the second part of the chapter, we presented our approach based on Convolutional Neural Network (CNN) and Generative Adversarial Networks (GANs) for disaster image retrieval and flood detection in satellite images, respectively. In the first task, we also utilized meta-data along with visual features using two different fusion methods. We observed that visual features perform better compared to meta-data.

Chapter 5

Conclusions and Future Directions

In dissertation, three open issues in event-based analysis of multimedia contents are studied including: (i) event recognition in single images; (ii) event recognition in personal photo collections; and (iii) joint utilization of social media and satellite imagery for natural disaster events. For event recognition in single images, three different novel solutions have been proposed. Similarly, to tackle the challenges of event recognition in personal photo collections, in this work, we presented a novel pipeline relying on a multiple instance-learning (MIL) strategy. Moreover, we also presented our system JORD, and our CNN and GAN based fusion of social media and satellite images for natural disaster detection. In the rest of this chapter, we discuss conclusions drawn from each individual part of the work and their corresponding future directions.

Event Recognition in Single Images:

In this part, we argued about the problems of event recognition in single images, addressing the problem from two perspectives. On the one hand, we aimed at demonstrating that the fusion of different feature extraction and classification strategies can outperform the single methods by jointly exploiting the learning capabilities of individual deep models. To this aim, we conducted a comprehensive analysis of renown deep models and assessed their individual as well as joint performance. On the other side, we analyzed the importance of event-salient objects and regions in event recognition. For the selection of event salient objects and regions, a crowd-sourcing study was conducted with a large number of volunteers.

We showed that it is possible to achieve superior event recognition performance by selecting the best models and combining them in an optimal way through appropriate late fusion strategies. All the proposed strategies articulate over the fact that different CNN architectures show diverse and complementary image characterization capabilities. Thus, we can conclude that fusing different CNN models stands out as a reliable choice to tackle the event recognition problem. This opens the possibility of relying on different models or combination of models depending on the task at hand, splitting the computational burden across a distributed architecture.

We also 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. A possible future direction for this part of the work is to enrich the framework with better schemes to automatically filter out the less informative and irrelevant image regions, which will ultimately lead to more accurate classification and significant reduction in the processing time and resources.

Event Recognition in Personal Photo-collections:

In this part of the work, we presented a MIL-based approach to event recognition in personal photo collections. We showed that, even at album-level annotation and presence of ambiguous photos in the albums, MIL can still guarantee higher performances. At the same time, the approach achieves good performances also using a limited number of images per bag, thus keeping the computational load acceptable. Moreover, considering the performance of the proposed approach, we did not observe any major limitation of the approach. The performance may degrade in the case of false positives in positive bags. However, considering both the references and citer bags helps to overcome this issue of false positive.

In the current implementation, we are relying on single network features (i.e., object information) only. As a future development, we aim to utilize a better scheme to fuse the scene and object-level information obtained through deep models pre-trained on ImageNet and Places datsets, without increasing the dimensionality of features, which may further improve the results. Moreover, active learning is an other potential direction of research to tackle the problems associated with event recognition in personal photo collections due ambiguous and irrelevant pictures therein.

Disasters and Social Media:

This part is devoted to an interesting application, namely natural disaster recognition. We approached the problem from two complementary perspectives. On the one side, we collect information from social media while on the other side we rely on satellite imagery to give a bird's eye-view of the natural disaster. We presented our system JORD, which is able to autonomously and automatically retrieve social media data from various social platforms about natural disasters and links it to remotely sensed images, to ensure the quality of retrieved data, we perform temporal and content based filtering and analysis. We have shown that the combination of social media data and satellite images provides a better story of a disaster. We also demonstrated that queries in local languages that are relevant to the exact position of natural disasters retrieve more accurate information about a disaster event.

The evaluation of the JORD system was carried out through a crowdsourcing-study where workers were asked to evaluate the usefulness of the system and to identify an event presented to them in the form of collected images, tweets and videos. The evaluation indicates that JORD works very accurate without human input, and it can be used to collect and merge a large number of event based data for technological and environmental disasters from different sources.

In another work, presented in this chapter, we rely on Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) for disaster image retrieval and flood detection in satellite images, respectively. In the first task, we utilized meta-data along with visual features, and we observed better performances with visual features only. Moreover, VGAN, originally developed for medical images, can be easily mapped into such applications.

Closing Remarks

In this dissertation, we established an argument about the widespread use of event-based analysis of multimedia content in a number of applications, such as multimedia indexing, retrieval, and organization and management of multimedia collections. We argued about how user-generated data are usually associated with personal experiences or collective activities, and how multimedia data can be assembled in the form of events. Most of the work presented in this dissertation has already been published in international journals and conference proceedings. A list of the publications can be found in Appendix A.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco GB De Natale. Used: a large-scale social event detection dataset. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 50. ACM, 2016.

[3] Kashif Ahmad, Nicola Conci, FGB De Natale, et al. A saliency-based approach to event recognition. *SIGNAL PROCESSING-IMAGE COMMUNICATION*, 60:42–51, 2018.

[4] Kashif Ahmad, Francesco De Natale, Giulia Boato, and Andrea Rosani. A hierarchical approach to event discovery from single images using mil framework. In *Signal and Information Processing (GlobalSIP), 2016 IEEE Global Conference on*, pages 1223–1227. IEEE, 2016.

[5] Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Nicola Conci, and Holversen Pal. Cnn and gan based satellite and social media data fusion for disaster detection. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[6] Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Nicola Conci, Pål Halvorsen, and Francesco 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*, page 12. ACM, 2017.

[7] Kashif Ahmad, Michael Riegler, Ans Riaz, Nicola Conci, Duc-Tien Dang-Nguyen, and Pål Halvorsen. The jord system: Linking sky and social multimedia data to natural disasters. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 461–465. ACM, 2017.

[8] Sheharyar Ahmad, Kashif Ahmad, Nasir Ahmad, and Nicola Conci. Convolutional neural networks for disaster images retrieval. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[9] Siti Nor Khuzaimah Binti Amit, Soma Shiraishi, Tetsuo Inoshita, and Yoshimitsu Aoki. Analysis of satellite images for disaster detection. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 5189–5192. IEEE, 2016.

[10] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[11] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[12] Boris Babenko. Multiple instance learning: algorithms and applications. *View Article PubMed/NCBI Google Scholar*, 2008.

[13] Siham Bacha, Mohand Saïd Allili, and Nadjia Benblidia. Event recognition in photo albums using probabilistic graphical models and fea-

ture relevance. *Journal of Visual Communication and Image Representation*, 40:546–558, 2016.

[14] Mohamed Bakillah, Ren-Yu Li, and Steve HL Liang. Geo-located community detection in twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon haiyan. *IJGIS*, 29(2):258–279, 2015.

[15] Alec Banks, Jonathan Vincent, and Chukwudi Anyakoha. A review of particle swarm optimization. part ii: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications. *Natural Computing*, 7(1):109–124, 2008.

[16] Yakoub Bazi and Farid Melgani. Toward an optimal svm classification system for hyperspectral remote sensing images. *IEEE Transactions on geoscience and remote sensing*, 44(11):3374–3385, 2006.

[17] Bischke Benjamin, Helber Patrick, Schulze Christian, and Srinivasan Venkat. The multimedia satellite task at mediaeval 2017. *Mediaeval Challenges 2017, Dublin Ireland*, 29(2):13–15, 20.

[18] Bischke Benjamin, Bhardwaj Prakriti, Gautam Aman, Helber Patrick, Borth Damian, and Dengel Andreas. Detection of flooding events in social multimedia and satellite imagery using deep neural networks. In *Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland.

[19] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Event recognition in photo collections with a stopwatch hmm. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1193–1200, 2013.

[20] Markus Brenner and Ebroul Izquierdo. Social event detection and retrieval in collaborative photo collections. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 21. ACM, 2012.

[21] Hyeran Byun and Seong-Whan Lee. Applications of support vector machines for pattern recognition: A survey. *Pattern recognition with support vector machines*, pages 571–591, 2002.

[22] James B Campbell and Randolph H Wynne. *Introduction to remote sensing.* Guilford Press, 2011.

[23] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[24] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *ACM Conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[25] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[26] Marc Cheong and Vincent Lee. Twittering for earth: A study on the impact of microblogging activism on earth hour 2009 in australia. In *Springer ACIIDS*, pages 114–123. Springer, 2010.

[27] François Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras.io/k*, 2015.

[28] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. *ACM*

Transactions on Multimedia Computing, Communications, and Applications (TOMM), 1(3):269–288, 2005.

[29] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.

[30] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[31] M. Dao, G. Boato, and F.G.B. DeNatale. Discovering inherent event taxonomies from social media collections. In *ICMR*, page 48. ACM, 2012.

[32] Minh-Son Dao, Duc-Tien Dang-Nguyen, and Francesco GB De Natale. Robust event discovery from photo collections using signature image bases (sibs). *Multimedia Tools and Applications*, 70(1):25–53, 2014.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[34] Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. Multiscale event detection in social media. *IJDMKD*, 29(5):1374–1405, 2015.

[35] Paul S Earle, Daniel C Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

[36] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[37] Berna Erol, Jonathan J Hull, and Dar-Shyang Lee. Linking multimedia presentations with their symbolic source documents: algorithm and applications. In *Proc. of ACM MM*, pages 498–507, 2003.

[38] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[39] C. S Firan, M. Georgescu, W. Nejdl, and R. Paiu. Bringing order to your photos: event-driven classification of flickr images based on social knowledge. In *ACM International Conference on Information and knowledge management*, pages 189–198. ACM, 2010.

[40] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

[41] Eugene Garfield and Robert King Merton. *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. Wiley New York, 1979.

[42] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *CBMI*, pages 85–90. IEEE, 2011.

[43] Yue-jiao Gong and Jun Zhang. Real-time traffic signal control for roundabouts by using a pso-based fuzzy controller. In *Evolutionary Computation (CEC), 2012 IEEE Congress on*, pages 1–8. IEEE, 2012.

[44] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[45] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 326–335. ACM, 2002.

[46] Jia-Min Gu, Yi-Leh Wu, Wei-Chih Hung, and Cheng-Yuan Tang. Personal photo organization using event annotation. In *Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on*, pages 1–4. IEEE, 2013.

[47] Debarati Guha-Sapir, Regina Below, and Philippe Hoyois. Emdat: International disaster database. *Catholic University of Louvain: Brussels, Belgium*, 2015.

[48] Adrien Guille and Cécile Favre. Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach. *Social Network Analysis and Mining*, 5(1):1–18, 2015.

[49] Cong Guo and Xinmei Tian. Event recognition in personal photo collections using hierarchical model and multiple features. In *Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on*, pages 1–6. IEEE, 2015.

[50] Mark Hall, Eibe Frank, Geoffrey Holmes, et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[51] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[52] Chunyang He, Qiaofeng Zhang, Yuechen Li, Xiaobing Li, and Peijun Shi. Zoning grassland protection area using remote sensing and cellular automata modeling—a case study in xilingol steppe grassland in northern china. *IJAE*, 63(4):814–826, 2005.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[54] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):797–819, 2011.

[55] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *IJEM*, 6(3-4):248–260, 2009.

[56] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the ICML*, pages 448–456, 2015.

[57] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[58] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International journal of multimedia information retrieval*, 2(2):73–101, 2013.

[59] Karen E Joyce, Stella E Belliss, Sergey V Samsonov, Stephen J McNeill, and Phil J Glassey. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 2009.

[60] Martin Jung, Kathrin Henkel, Martin Herold, and Galina Churkina. Exploiting synergies of global land cover products for carbon cycle modeling. *IJRSE*, 101(4):534–553, 2006.

[61] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Disaster monitoring using unmanned aerial vehicles and deep learning.

[62] Nogueira Keiller, Fadel Samuel, Dourado Ícaro, Werneck Rafael, Muñoz Javier, Penatti Otávio, and Calumby Rodrigo. Data-driven flood detection using neural networks. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[63] Slava Kisilevich, Daniel Keim, Natalia Andrienko, and Gennady Andrienko. Towards acquisition of semantics of places and events by multi-perspective analysis of geotagged photo collections. In *Geospatial visualisation*, pages 211–233. Springer, 2013.

[64] Avgerinakis Konstantinos, Moumtzidou Anastasia, Stelios Andreadis, Michail Emmanouil, Gialampoukidis Ilias, Vrochidis Stefanos, and

Kompatsiaris Ioannis. Visual and textual analysis of social media and satellite images for flood detection @ multimedia satellite task mediaeval 2017. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[65] Petros Koutras and Petros Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015.

[66] A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[67] Z. Lan, L. Bao, S. Yu, W. Liu, and A. G Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia tools and applications*, 71(1):333–347, 2014.

[68] Lopez-Fuentes Laura, Weijer Joost, Bolaños Marc, and Skinnemoen Harald. Multi-modal deep learning approach for flood detection. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[69] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: segment-based event detection from tweets. In *Proc. of ACM IKM*, pages 155–164. ACM, 2012.

[70] L. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, pages 1–8, 2007.

[71] Mengyi Liu, Xin Liu, Yan Li, Xilin Chen, Alexander G Hauptmann, and Shiguang Shan. Exploiting feature hierarchies with convolutional neural networks for cultural event recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 32–37, 2015.

[72] X. Liu and B. Huet. Heterogeneous features and model selection for event-based media classification. In *ICMR*, pages 151–158. ACM, 2013.

[73] Ying Liu and Linzhi Wu. Geological disaster recognition on optical remote sensing images using deep learning. *Procedia Computer Science*, 91:566–575, 2016.

[74] D. G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[75] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. Lire: open source visual information retrieval. In *Proc. of ACM MMSys*, 2016.

[76] TN Manjunath, Ravindra S Hegadi, and GK Ravikumar. A survey on multimedia data mining and its relevance today. *IJCSNS*, 10(11):165–170, 2010.

[77] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2232–2239. IEEE, 2009.

[78] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proc. of ACM SIGMOD*, pages 1155–1158, 2010.

[79] R. Mattivi, G. Boato, and F.G.B. De Natale. Event-based media organization and indexing. *Infocommunication Journal*, 3(3):9–18, 2011.

[80] Polykarpos Meladianos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Degeneracy-based

real-time sub-event detection in twitter stream. In *Proc. of AAAI ICWSM*, pages 248–257, 2015.

[81] Dean S Messing, Peter Van Beek, and James H Errico. The mpeg-7 colour structure descriptor: image description using colour and local spatial information. In *Proceedings of the ICIP*, volume 1, pages 670–673. IEEE, 2001.

[82] Vasileios Mezaris, Ansgar Scherp, Ramesh Jain, and Mohan S Kankanhalli. Real-life events in multimedia: detection, representation, retrieval, and applications. *Multimedia Tools and Applications*, 70(1):1–6, 2014.

[83] Dao Minh-Son, Pham Quang-Nhat-Minh, and Dang-Nguyen Duc-Tien. A domain-based late-fusion for disaster image retrieval from social media. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[84] Hanif Muhammad, Atif Muhammad, Khan Mahrukh, and Rafi Mohammad. Flood detection using social media data and spectral regression based kernel discriminant analysis. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[85] Mor Naaman, Yee Jiun Song, Andreas Paepcke, and Hector Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, pages 53–62. IEEE, 2004.

[86] Milind Naphade, John R Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE multimedia*, 13(3):86–91, 2006.

[87] Tkachenko Nataliya, Zubiaga Arkaitz, and Rob Procter. Wisc at mediaeval 2017: Multimedia satellite task. In *Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland.

[88] T. Nguyen, M. Dao, R. Mattivi, E. Sansone, F.G.B. De Natale, and G. Boato. Event clustering and classification from social media: Watershed-based and kernel methods. In *MediaEval Workshop*, 2013.

[89] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. j. of computer vision*, 42(3):145–175, 2001.

[90] Kaoru Ota, Minh Son Dao, Vasileios Mezaris, and Francesco GB De Natale. Deep learning for mobile multimedia: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3s):34, 2017.

[91] Symeon Papadopoulos, Raphael Troncy, Vasileios Mezaris, Benoit Huet, and Ioannis Kompatsiaris. Social event detection at mediaeval 2011: Challenges, dataset and evaluation. In *MediaEval*, 2011.

[92] Symeon Papadopoulos, Christos Zigkolis, Yiannis Kompatsiaris, and Athena Vakali. Cluster-based landmark and event detection for tagged photo collections. *IEEE MultiMedia*, 18(1):52–63, 2011.

[93] Sungheon Park and Nojun Kwak. Cultural event recognition by subregion classification with convolutional neural network. In *Proceedings of the CVPR*, pages 45–50, 2015.

[94] Frank Paul and Liss M Andreassen. A new glacier inventory for the svartisen region, norway, from landsat etm+ data: challenges and change assessment. *Journal of Glaciology*, 55(192):607–618, 2009.

[95] G. Petkos, S. Papadopoulos, V. Mezaris, R. Troncy, P. Cimiano, T. Reuter, and Y. Kompatsiaris. Social event detection at mediaeval: a three-year retrospect of tasks and results. In *ACM ICMR Workshop on Social Events in Web Multimedia (SEWM)*, 2014.

[96] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM.

[97] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. Efficient disease detection in gastrointestinal videos–global features versus neural networks. *Multimedia Tools and Applications*, pages 1–33, 2017.

[98] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proc. of ACM IKM*, pages 1873–1876. ACM, 2010.

[99] R. F. Rachmadi, K. Uchimura, and G. Koutaki. Spatial pyramid convolutional neural network for social event detection in static image. *arXiv preprint arXiv:1612.04062(presented in ICAST 2016)*, 2016.

[100] Reza Fuad Rachmadi, Keiichi Uchimura, and Gou Koutaki. Combined convolutional neural network for event recognition. In *Korea-Japan Joint Workshop on Frontiers of Computer Vision*, pages 85–90, 2016.

[101] Timo Reuter and Philipp Cimiano. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 22. ACM, 2012.

[102] Timo Reuter, Symeon Papadopoulos, Giorgos Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher de Vries, and Shlomo Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013*, 2013.

[103] M. Riegler, V. R. Gaddam, M. Larson, R. Eg, P. Halvorsen, and 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)*, pages 1–6, June 2016.

[104] Andrea Rosani, Giulia Boato, and Francesco GB De Natale. Eventmask: A game-based framework for event-saliency identification in images. *IEEE Transactions on Multimedia*, 17(8):1359–1371, 2015.

[105] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of ACM WWW*, pages 851–860, 2010.

[106] Amaia Salvador, Matthias Zeppelzauer, Daniel Manchon-Vizuete, Andrea Calafell, and Xavier Giro-i Nieto. Cultural event recognition with visual convnets and temporal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–44, 2015.

[107] E. Schinas, G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Certh@ mediaeval 2012 social event detection task. In *MediaEval Workshop*. Citeseer, 2012.

[108] Luca Scrucca. Genetic algorithms for subset selection in model-based clustering. In *Unsupervised Learning Algorithms*, pages 55–70. Springer, 2016.

[109] Vijay K Sharma and KK Mahapatra. Mil based visual object tracking with kernel and scale adaptation. *Signal Processing: Image Communication*, 53:51–64, 2017.

[110] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[111] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318*, 2017.

[112] Brian Stelter and Noam Cohen. Citizen journalists provided glimpses of mumbai attacks. *The New York Times*, 30, 2008.

[113] Gayatri Swamynathan, Christo Wilson, Bryce Boe, Kevin Almeroth, and Ben Y Zhao. Do social networks improve e-commerce?: a study on social marketplaces. In *Proc. of the workshop on Online social networks*, pages 1–6. ACM, 2008.

[114] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[115] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, et al. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[116] Bruno Takahashi, Edson C Tandoc, and Christine Carmichael. Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. *Computers in Human Behavior*, 50:392–398, 2015.

[117] F. Tang, D. R Tretter, and C. Willis. Event classification for personal photo collections. In *Proceedings of the ICASSP*, pages 877–880. IEEE, 2011.

[118] Planet Team. Planet application program interface: In space for life on earth. san francisco, ca, 2016.

[119] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[120] Raphaël Troncy, Bartosz Malocha, and André TS Fialho. Linking events with media. In *Proceedings of the 6th International Conference on Semantic Systems*, page 42. ACM, 2010.

[121] S. Tsai, T. S Huang, and F. Tang. Album-based object-centric event recognition. In *Proceedings of the ICME*, pages 1–6. IEEE, 2011.

[122] Shen-Fu Tsai, Liangliang Cao, Feng Tang, and Thomas S Huang. Compositional object pattern: a new model for album event recognition. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1361–1364. ACM, 2011.

[123] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *ICMR*, page 68. ACM, 2011.

[124] Christos Tzelepis, Zhigang Ma, Vasileios Mezaris, Bogdan Ionescu, Ioannis Kompatsiaris, Giulia Boato, Nicu Sebe, and Shuicheng Yan. Event-based media processing and analysis: A survey of the literature. *Image and Vision Computing*, 53:3–19, 2016.

[125] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[126] G. Vanwinckelen, D. Fierens, H. Blockeel, et al. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*, 30(2):313–341, 2016.

[127] J. Wang and J. Zucker. Solving multiple-instance problem: A lazy learning approach. 2000.

[128] J. Wang and J. D. Zucker. Solving multiple-instance problem: A lazy learning approach. In *Proceedings of the International Conference on Machine Learning*, pages 1119–1125, 2000.

[129] L. Wang, Z. Wang, S. Guo, and Y. Qiao. Better exploiting os-cnns for better event recognition in images. In *CVPR Workshops*, pages 45–52, 2015.

[130] L. Wang, Z. Wang, Y. Qiao, and L. V. Gool. Transferring object-scene convolutional neural networks for event recognition in still images. *arXiv preprint arXiv:1609.00162*, 2016.

[131] Limin Wang, Zhe Wang, Wenbin Du, and Yu Qiao. Object-scene convolutional neural networks for event recognition in images. In

Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 30–35, 2015.

[132] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. Transferring deep object and scene representations for event recognition in still images. *International Journal of Computer Vision*, pages 1–20, 2017.

[133] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *arXiv preprint arXiv:1610.02501*, 2016.

[134] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, 14(1):19–29, 2007.

[135] H Wood. The use of earth observing satellites for hazard support: Assessments and scenarios. *Final Report of the CEOS Disaster Management Support Group. Available from¡ http://disaster. ceos. org/legal. cfm*, 2002.

[136] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2015.

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

[138] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *CVPR*, pages 1600–1609, 2015.

[139] Fu Xiyao, Yi Bin, Peng Liang, Zhou Jie, Yang Yang, and Shen Heng. Bmc@mediaeval 2017 multimedia satellite task via regression random forest. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[140] Zheng Xu, Hui Zhang, Vijayan Sugumaran, Kim-Kwang Raymond Choo, Lin Mei, and Yiwei Zhu. Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016(1):44, 2016.

[141] R. R Yager and D. P Filev. Induced ordered weighted averaging operators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(2):141–150, 1999.

[142] Zhao Zhengyu and Martha Larson. Retrieving social flooding images based on multimodal information. In *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland.

[143] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE CVPR*, pages 2921–2929, 2016.

[144] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Proceedings of the NIPS*, pages 487–495, 2014.

Appendix A

Publications

A.0.1 Journal Publications

- **Kashif Ahmad**, Nicola Conci, Francesco GB De Natale, ”A Saliency-based Approach to Event Recognition”, International Journal of Signal Processing and Image Communication, v. 60, pages: 42-51, 2017, DOI: https://doi.org/10.1016/j.image.2017.09.009.

- **Kashif Ahmad**, Nicola Conci, Giulia Boato, Francesco GB De Natale, ”Event Recognition in personal photo collections via MIL-based Classification of Multiple Images”, **Accepted (in press)** in International Journal of Electronic Imaging.

- **Kashif Ahmad**, M. Lamine, Nicola Conci, Farid Melgani, Francesco GB De Natale, ”Ensembles of Deep Models for Event Recognition”, **under review** in ACM Transactions on Multimedia (TOMM)

- **Kashif Ahmad**, Konstantin Pogorelov, Michael Reiglar, Nicola Conci, Pal Hoversin, ”Social Media and Satellites: Disaster event detection, linking and summarization”, **under review** in International Journal of Multimedia Tools and Applications (MTAP)

A.0.2 Conference Publications

- **Kashif Ahmad**, Nicola Conci, Giulia Boato, Francesco GB De Natale, ”USED: A LArge-scale Social Event Dataset”, Proceedings of the 7th International ACM Conference on Multimedia Systems (MMSys), 2016.

- **Kashif Ahmad**, Francesco GB De Natale,Giulia Boato, Andrea Rosani, ”A Hierarchical Approach to Event Discovery from Single Images using MIL Framework”, Proceedings of the IEEE Conference on Signal and Information Processing (GlobalSIP),pages 1223-1227,2016.

- **Kashif Ahmad**, M. Lamine, Nicola Conci, Giulia Boato, Farid Melgani, Francesco GB De Natale, ”A Pool of Deep Models for Event Recognition”, Proceedings of International Conference on Image Processing (ICIP), 2017.

- **Kashif Ahmad**, Michael Rieglar, Ans Riaz, Nicola Conci, Duc-Tien Dang-Nguyen, Pal Hoversen, ”The JORD System: Linking Sky and Social Media to Natural Disasters”, Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR), pages 461-465, 2017.

- **Kashif Ahmad**, Michael Rieglar, Konstantin Pogorelov, Nicola Conci, Pal Hoversen, Francesco De Natale, ”JORD: A System for Collecting Information and Monitoring Natural Disasters by Linking Social Media with Satellite Imagery”, Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI), 2017.

- **Kashif Ahmad**, Konstantin Pogorelov, Michael Rieglar, Nicola Conci, Pal Hoversen, ”CNN and GAN based Satellite and Social Media Fusion for Disaster Detection”, Proceedings of the MediaEval 2017 Workshop, Dublin Ireland, 2017.

- Sheharyar Ahmad, **Kashif Ahmad**, Nasir Ahmad, Nicola Conci ”Convolutional Neural Networks for Disaster Images Retrieval”, Proceedings of the MediaEval 2017 Workshop, Dublin Ireland, 2017.

- **Kashif Ahmad**, Mohammed Lamine Mekhalfi, Nicola Conci ”Event Recognition in Personal Photo Collections: An Active Learning Approach”, **accepted** for presentation at the Visual Information Processing and Communication Conference, at IS&T Electronic Imaging 2018.