USED: A Large-scale Social Event Detection Dataset

Kashif Ahmad, Nicola Conci, Giulia Boato, Francesco G. B. De Natale
DISI - University of Trento, Italy
{kashif.ahmad, nicola.conci, giulia.boato, francesco.denatale}@unitn.it

ABSTRACT

Event discovery from single pictures is a challenging problem that has raised significant interest in the last decade. During this time, a number of interesting solutions have been proposed to tackle event discovery in still images. However, a large scale benchmarking image dataset for the evaluation and comparison of event discovery algorithms from single images is still lagging behind. To this aim, in this paper we provide a large-scale properly annotated and balanced dataset of 490,000 images, covering every aspect of 14 different types of social events, selected among the most shared ones in the social network. Such a large scale collection of event-related images is intended to become a powerful support tool for the research community in multimedia analysis by providing a common benchmark for training, testing, validation and comparison of existing and novel algorithms. In this paper, we provide a detailed description of how the dataset is collected, organized and how it can be beneficial for the researchers in the multimedia analysis domain. Moreover, a deep learning based approach is introduced into event discovery from single images as one of the possible applications of this dataset with a belief that deep learning can prove to be a breakthrough also in this research area. By providing this dataset, we hope to gather research community in the multimedia and signal processing domains to advance this application.

CCS Concepts

•Applied computing → Digital libraries and archives;

Keywords

Event detection, Dataset, CNN, Multimedia Indexing

1. INTRODUCTION

In recent years, user-generated multimedia contents have changed the way in which people consume and communicate through social media. With the advent of webcams and low cost handheld devices, such as digital cams and smartphones, it has become easier to generate multimedia contents to be shared over the network making it possible for users to share their experiences in the form of multimedia contents. The user generated multimedia contents are usually associated with personal experiences, such as fishing and skiing, or collective activities, such as concerts, soccer matches or other sports and social events. On the web, user's personal/collective experiences can be seen as a collection of multimedia contents that can be assembled in the form of events. Events can be defined as real world happenings planned and attended by people; moreover related multimedia data is also captured by people.

In the literature, events have been analyzed following different approaches. Since the very initial work by Jain et al. [9], event-based models for multimedia indexing and retrieval got tremendous attention of the research community. During this time, a number of interesting attempts have been made for an efficient representation of event-related multimedia items, and strategies that can incorporate all the available information to find revealing patterns in unknown multimedia data [2, 4].

However, a large-scale benchmarking image dataset for the evaluation and comparison of event discovery algorithms from single images is still lagging behind.

The current 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.

On this point, there is an ongoing trend of image representation, which derives benefits from deep neural architectures, namely convolutional neural networks (CNNs). 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. Based on these considerations, in this work we are providing a large

MMSys'16, May 10-13, 2016, Klagenfurt, Austria
© 2016 ACM. ISBN 978-1-4503-4297-1/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2910017.2910624

collection of event related images covering 14 different types of social events, selected among the most shared ones in the social network.

The rest of the paper is organized as follow: Section 2 provides a detailed review of the some existing benchmark datasets for event detection in single images. Section 3 provides a detailed description and motivation for collection of the dataset, while Section 4 provides the description of our CNN based approach to event detection along with experimental results on the newly collected dataset. In Section 5, some concluding remarks are presented.

2. RELATED WORK

Over the last decade, event-based models for multimedia analysis have gained an increasing attention of the research community. Since the very first common event-based model for multimedia analysis [9], it has been an area of keen interest for the researchers, and during this time a number of interesting solutions to event detection/classification have been proposed. However, a large scale benchmarking image dataset for the evaluation and comparison of event discovery algorithms from single images is still lagging behind. Some benchmark datasets are available in the literature (e.g., EiMM [6] and SED [7]) but none of them is large enough to be used for training convolutional neural networks. For instance, EiMM [6] provides a total of 32973 images, in which more than half of the images are related to sports. EiMM is organized into two main categories of events, namely social events and sports events. The social event category is further organized into 8 different types of social events. The list of social events contained in EiMM dataset along with the number of images per event-class is given in Table 1.

Multimedia event detection was also included as a part of the benchmarking workshops TRECVID[1] (2010 to 2014) and MediaEval[2] (2011 to 2014). The basic goal of the media event detection task in TRECVID was to develop a system that can search multimedia recordings for user-defined events based on pre-computed metadata. In TRECVID, annotated datasets have been provided for the event detection task in videos every year.

On the other hand, event detection has been part of Mediaeval for four years (2011 to 2014) with slight modification of the task every year. Each year different datasets have been provided for the evaluation of the submissions.

The main focus of the Mediaeval 2013 social event detection challenge [7], was to learn how an event-related multimedia item looks like. In the task, participants were asked to develop a system that can detect and classify social events in a single image. For this purpose, an annotated dataset has been provided containing images from 7 different types of social events. Although, this dataset provides a reasonable amount of images along with metadata for both development as well as test purposes, it is not balanced. As it can be seen in Table 1, some event-classes in the SED dataset have a very large number of images (e.g., concert class has 71556 images) while other have very few images (e.g., exhibition class has just 342 images). Since SED dataset was mainly created for multi-modal analysis, it also contains some images that do not visually represent the underlying event. It is also worth noticing that it has been observed that images in this dataset have strong perceptual correlation with other class images.

More recently in a benchmarking workshop ChaLearn Looking at People Challenge and workshop[3] 2015, cultural events detection has been introduced as a challenge. For this challenge a collection of more than 11000 images has been provided. The collection is composed of images from 50 different types of cultural events. This dataset encourages the research community to exploit garments, human poses, objects and other contextual information for event recognition. Some well-known cultural events from this dataset include Carnival, Oktoberfest, San Fermin, Maha-Kumbh-Mela and Aoi-Matsuri.

3. DATASET

In this section, we discuss about the basic motivation behind the collection of the event-related images along with a detailed description of the collected dataset, collection method and annotations.

3.1 Motivation

As aforesaid, the main limitation in the application of CNNs in the context of event detection is the unavailability of large datasets. Similar to human visual system, CNNs also need a large amount of training data before reaching good recognition capabilities, where the high variability of the represented information can be effectively explored as a valuable asset to ensure better performances in event classification.

To verify the fact that CNNs require a large amount of training before achieving good recognition capabilities, we conducted some experiments with CNN on a publically available dataset, namely EiMM [6]. Initially, we fine-tuned our CNN with event-related images from the original collection of EiMM. In these experiments, we could not achieve adequate validation accuracy due to limited training data. In these experiments, we used a dataset of 13,000 images composed of social-event images from the original EiMM dataset, which was divided into training, validation and test sets with ratios of 60%, 20% and 20%, respectively. Since the dataset is not balanced, we did the division of images for each phase at event-class level. The overall accuracy on the validation set we achieved was just 19%.

As it can be seen in Table 1, the maximum number of images in a class, in the EiMM dataset, is 2253 (sea-holiday), which is not sufficient to be used for training a convolutional neural network. This assumption is also confirmed by the experimental results with original EiMM dataset as discussed above. The same is the case with SED dataset, where most of the event-classes have a few hundreds of images as shown in Table 1. To cope with this problem, in this work we provide a large collection of media using as a basis two public datasets, namely EiMM [6] and SED [7] by gathering images from Flickr. By providing this large-scale dataset, we encourage research community to advance this application.

3.2 Dataset Collection and Annotation

[1] http://trecvid.nist.gov/

[2] http://www.multimediaeval.org/

[3] http://gesture.chalearn.org/2015-looking-at-people-cvpr-challenge/dataTrack3

Table 1: List of events in EiMM [6], SED [7] and our newly collected datasets along with the number of images in each class of events

EiMM Dataset	**Class**	Concert	Graudation	Mountain Trip	Meeting	Picnic	Sea-holiday	Ski-holiday	Wedding
	#Images	1085	1815	2051	795	1627	2253	1817	1776
SED Dataset	**Class**	Concert	Conference	Exhibition	Fashion	Protest	Sports	Theater	-
	#Images	71556	2975	342	1556	1039	403	4342	-
Our Dataset	**Class**	Concert	Graudation	Mountain Trip	Meeting	Picnic	Sea-holiday	Ski-holiday	Wedding
	#Images	35,000	35,000	35,000	35,000	35,000	35,000	35,000	35,000
	Class	Conference	Exhibition	Fashion	Protest	Sports	Theater	-	-
	#Images	35,000	35,000	35,000	35,000	35,000	35,000	-	-

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. Table 1 shows the list of event-classes from our new-collected dataset along with the number of images per event-class. The dataset is downloaded between 7th and 20th of September 2015, based on event-related keywords. 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). As mentioned earlier that one of the main reasons of the failure of conventional handcrafted visual features based approaches is their inefficacy in coping with the variations in event contents. Based on this consideration, in the newly collected dataset we tried our best to cover every aspect of the considered social events by collecting images for 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 [7]. 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 1 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.

3.3 Dataset organization

The collected dataset is made publically available at http://mmlab.disi.unitn.it/USED/. 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 as in single pools of test and training images containing images from all event-classes. Thus, the users will have option 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.

4. VALIDATION

In this section, we provide an experimental validation of our newly collected dataset as a demonstration of one of its possible applications. In particular, we provide a detailed analysis of multimedia contents in the context of social event detection using convolution neural network features.

In our approach to event detection, we follow the same pipeline adopted in [5] for object recognition. The convolutional neural network used in this work is composed of eight weighted layers including five convolutional and three fully connected layers. The first convolutional layer is composed of 96 kernels, each of size 11×11×3, with a stride of 4 pixels. The second, third, fourth and final convolutional layers consist of 256, 384, 384 and 256 kernels with different sizes, respectively. Pooling layers, which help in dealing with variances/translations of image features, are used between first and second, and second and third convolutional layers. Since our dataset covers event-classes from two benchmarking datasets (i.e., EiMM and SED), which have images from 8 and 7 different kinds of social events, respectively. Therefore, the output of the top fully connected layer is fed to an 8 or 7-way softmax function, correspondingly for classification purposes.

From an experimentation point of view, our CNN based approach to event detection is composed of two phases, namely pre-training and fine-tuning phase. In the pre-training phase we acquired learned parameters from a CNN [5], pre-trained on a large-scale object detection dataset [3]. Subsequently, the CNN is fine-tuned (i.e., retrained) on a subset of our newly collected dataset. Both phases of the proposed ap-

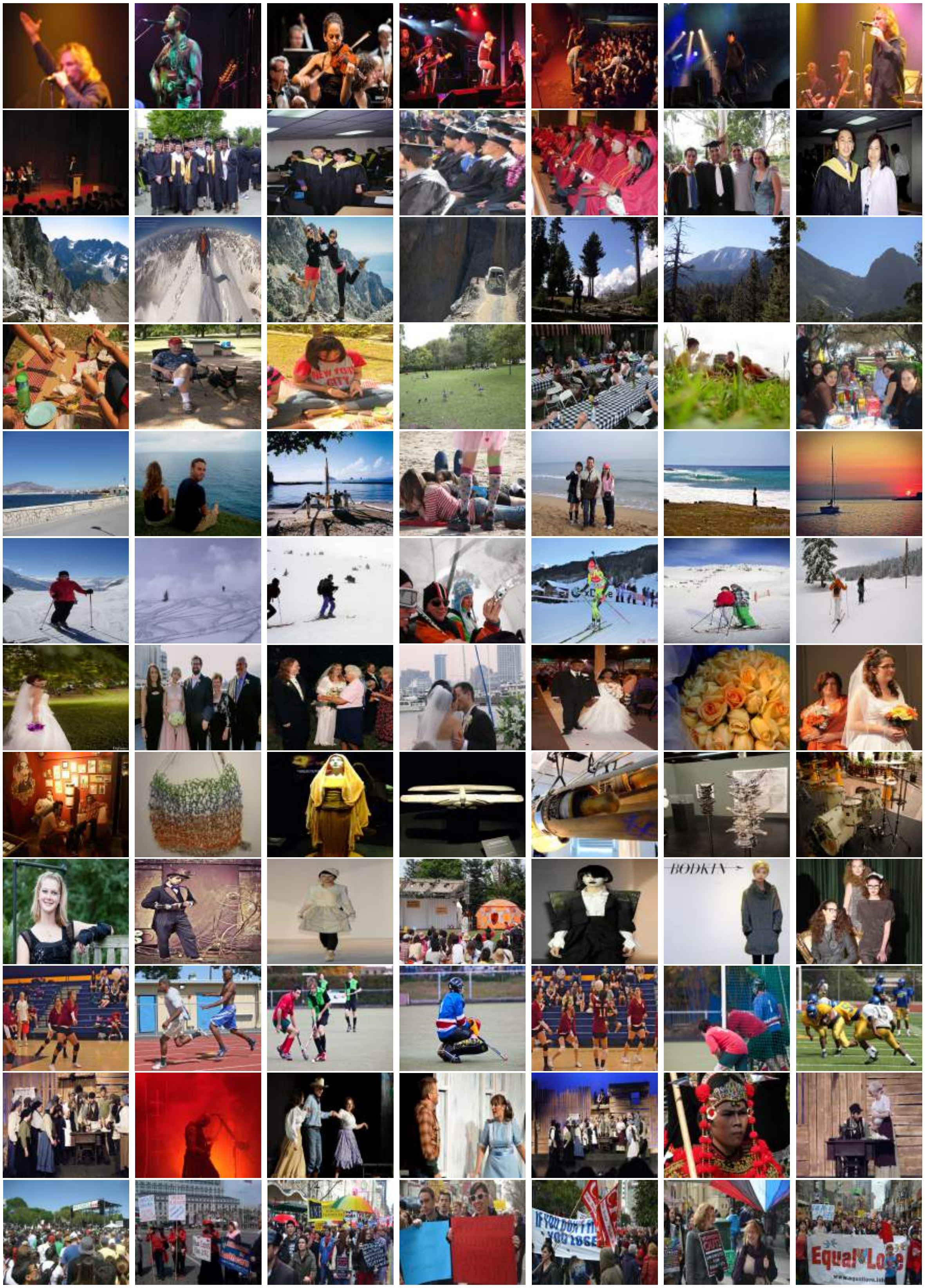

Figure 1: Some sample images from the newly-collected dataset.

proach are described in details in the next subsections.

4.1 Pre-training Phase

The basic idea of pre-training phase is to acquire basic image features, such as edges, corners and lines, on a larger dataset. For this purpose, we used parameters of a pre-trained CNN [5] available in the Caffe toolbox[4], which is an open source framework for deep learning. Starting with the parameters of a pre-trained CNN leads to a faster convergence.

4.2 Fine-tuning Phase

Although the ImageNet dataset [3] has a huge collection of images, using parameters of a CNN pre-trained on ImageNet dataset directly for social event detection is not a proper choice. With ImageNet dataset we can get basic image features, such as edges, corners and lines, however for high-level description/semantics of event-related images we need to fine-tune our CNN on subject specific dataset. Since our target dataset have images with different resolutions, before proceeding with fine-tuning our CNN with target dataset, we down-sampled all the images, in the target dataset to a fixed size resolution of 256×256 as recommended in [5].

As the lower convolutional layers of CNNs response to low-level image features, in order to ensure a fast convergence, we avoided re-training the lower layers of the CNN with new subject specific data. In the experimentation process, we fine-tuned two separate convolutional neural networks, each for the event-classes in EiMM and SED datasets. The basic motivation for fine-tuning separate networks for the each type of event-collections (i.e. SED and EiMM) is to provide an in-depth analysis by providing confusion matrices, and comparison with the state-of-the-art approaches. Since our datasets contain 8 and 7 event-classes from EiMM and SED datasets, respectively, we included a layer at the top of the baseline network [5] with a number of outputs corresponding to the number of classes in the target datasets. Since there is no layer with the same name of the newly included layer in the CNN [5], the top layer of our network is initialized with arbitrary weights, which are then tuned during the training phase. To further speed-up the learning process, we used a low value for the step-size, which helps the learning rate to go down faster.

4.3 Experiments and Results

In this section we provide a detailed description of the experimental evaluation along with distribution of the dataset into training, validation and test sets. Experimental results and comparison of the proposed approach with two baseline approaches is also discussed in detail.

4.4 Data Assemblage

In the experimentation process, the newly created dataset is divided into 3 subsets, namely training, validation, and test by randomly selecting images for each phase. For training (fine-tuning) of the neural network, 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 (20,000*7) and 160,000 (20,000*8) for training/fine-tuning purposes from SED and EiMM related event-classes, respectively. As far as the validation and test collections are concerned, we used 49,000 (7,000*7) from SED, and 56,000 (7,000*8) images from EiMM event-classes, respectively for both phases.

4.5 Results and Analysis

Experimental results of our CNN based approach to event detection are reported in Table 2 and Table 3 on the newly collected dataset. On the event-classes from EiMM dataset (i.e., 8 classes including 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 SED dataset is concerned, we achieved an overall accuracy of 70.03%.

For a thorough analysis of the experimental results, we provide confusion matrices of our CNN on both test sets as shown in Table 2 and Table 3.

In Table 2 it can be seen that our 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 between these event classes. The research community is encouraged to provide novel strategies and efficient representation 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 mountains trip. Similarly, in Table 3 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 except between exhibition and conference, which are 19.5% times confused with each other, due to the high perceptual correlation between these event classes.

In order to show the versatility of our newly collected dataset, We also provide a comparison of the performance of our convolutional neural network, trained on our dataset, with a baseline approach [8], which relies on SURF features [1] in the bag of words model with Support Vector Machine. In [8], an evaluation is carried out on a subset from EiMM dataset [6] (original collection) used in [8]. The overall accuracy of our approach on test collection from EiMM, used in [8], is 71.54%, while the baseline approach [8] has an accuracy of 38.80%. In Figure 2, we can observe an improvement in the performance, with a gain of 32.74% compared to the approach presented in [8]. It is the evident of superior performance of CNN compared to the conventional handcrafted features based approach. It must be noted that in the experiments for comparison, we tested our network, trained on the newly collected dataset, on a subset of original EiMM dataset. The high gain in the performance shows the versatility of our newly collected dataset, which covers every aspect of the included event-classes.

5. CONCLUSIONS

In this work a large collection of event-related images belonging to 14 different types of social events, selected among the most shared ones in the social network, has been made publically available, which is proposed as a valuable support

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

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

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

Table 3: Confusion matrix of our network on the event-classes belonging to SED dataset (accuracy in percentage).

Actual-class	Predicted-class						
	Concert	Conference	Exhibition	Fashion	Protest	Sport	Theater/Dance
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
Sport	.34	5.84	4.61	2.81	10.17	**72.21**	4.02
Theater/Dance	14.78	10.18	8.40	12.20	2.47	.05	**51.90**

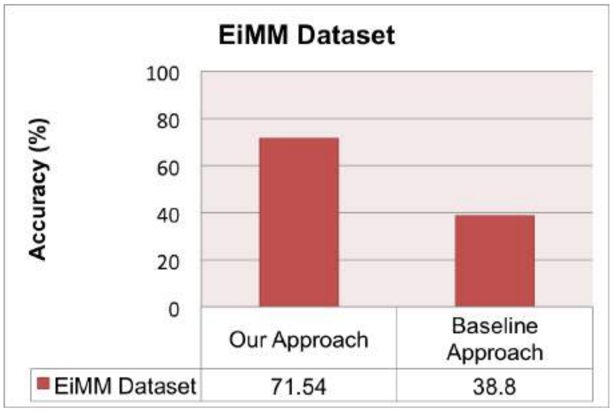

Figure 2: Comparison of the performance of CNN trained on our dataset with a baseline approach [8] on a test collection used in [8].

tool for benchmarking within multimedia analysis domain. Moreover, a deep learning based approach has been introduced into event discovery from single images as one of its possible applications. In this paper, a detailed description of the collection procedure, organization modalities along with the experimental results have been provided.

With this work, we intend to provide a large collection of event-related images, covering different aspects of the considered social event-classes, which can be used for training purposes in deep learning based approaches to event discovery from single images. We believe that deep learning can prove to be a breakthrough 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. Such a large-scale dataset will help the research community to train and test their deep learning algorithms in the context of event detection. In future we aim to further extend the dataset in terms of both: number of images and the number of event-classes.

6. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van G. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

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

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

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

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advan. Neur. Info. proces. sys.*, pages 1097–1105, 2012.

[6] R. Mattivi, J. Uijlings, F. G. De Natale, and N. Sebe. Exploitation of time constraints for (sub-) event recognition. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 7–12. ACM, 2011.

[7] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *MediaEval Workshop*, 2013.

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

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