How Deep Features Have Improved Event Recognition in Multimedia: a Survey

KASHIF AHMAD, University of Trento, Italy, Italy
NICOLA CONCI, University of Trento, Italy, Italy

Event recognition is one of the areas in multimedia that is attracting great attention of researchers. Being applicable in a wide range of applications, from personal to collective events, a number of interesting solutions for event recognition using multimedia information sources have been proposed. On the other hand, following their immense success in classification, object recognition and detection, deep learning has demonstrated to perform well also in event recognition tasks. Thus, a large portion of the literature on event analysis relies nowadays on deep learning architectures. In this paper, we provide an extensive overview of the existing literature in this field, analyzing how deep features and deep learning architectures have changed the performance of event recognition frameworks. The literature on event-based analysis of multimedia contents can be categorized into four groups, namely (i) event recognition in single images; (ii) event recognition in personal photo collections; (iii) event recognition in videos; and (iv) event recognition in audio recordings. In this paper, we extensively review different deep learning-based frameworks for event recognition in these four domains. Furthermore, we also review some benchmark datasets made available to the scientific community to validate novel event recognition pipelines. In the final part of the manuscript, we also provide a detailed discussion on basic insights gathered from the literature review, and identify future trends and challenges.

CCS Concepts: • **Information systems** → **Information retrieval**.

Additional Key Words and Phrases: Information Retrieval, Event Detection, Deep Learning, Deep Features, Social Events Detection, Natural Disaster, Social Media, Video Analysis, Audio event analysis

ACM Reference Format:
Kashif Ahmad and Nicola Conci. XXXX. How Deep Features Have Improved Event Recognition in Multimedia: a Survey. 1, 1 (XX XXXX), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Social media and smartphones, have incredibly changed the way, in which people generate and consume multimedia contents. As a consequence, there is an ever increasing demand for tools to automatically collect, organize and retrieve multimedia contents from unstructured repositories, relieving users from manual arrangement of their data.

Looking at the problem from a user's perspective, user-generated multimedia on the web often refer to personal or collective experiences and activities, which can be generically referred to as events. Events are real world happenings that are planned and attended by people; events are captured and shared by the people. Event-based analysis of multimedia contents has been one of the areas of keen interest for researchers and a number of interesting solutions have been proposed in different application domains including detection, summarization, retrieval and indexing. In this

Authors' addresses: Kashif Ahmad, University of Trento, Italy, Trento, Italy, kashif.ahmad@unitn.it; Nicola Conci, University of Trento, Italy, Trento, Italy, nicola.conci@unitn.it.

XXXX-XXXX/XXXX/XX-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

regard, the definition of *event* plays a very important role. In literature a number of definitions have been used for an event depending on the type of media and available information. Some sample definitions from the literature are: "Events represent a change of state in a multimedia item planned and attended [52]"; "An event is a collections of actions performed among different agents [71]"; "Events are real world happening planned and attended by people [110]". Events generally involve multiple objects and characters, which can be in foreground as well as in backgrounds. Such characteristics make event recognition a more challenging task compared to object recognition. In a broad sense, also event analysis can be seen as a kind of monitoring application [129, 163, 171].

In literature, various solutions have been proposed to address the issue of event recognition in multimedia, with diverse classification and feature extraction strategies. To this aim, different types of information including visual, audio, and text, have been widely utilized, where visual features are probably the most widely adopted ones. However, the literature suggests that traditional paradigms based upon shallow handcrafted visual features are prone to failure, as they cannot fill the gap between the spatial content and semantic attributes of multimedia contents. Instead, and similarly to other application domains, deep architectures have demonstrated better performances.

In this paper, we provide a detailed survey of different approaches relying on deep architectures for event recognition. We focus on deep learning-based approaches for event analysis in four different domains, namely single images, photo-collections, videos, and audio recordings. In addition, we also provide a detailed survey of the existing datasets used for evaluation purposes in all four domains. The basic insight of this survey paper is to analyze how deep features have improved the performance of event recognition.

The rest of the paper is organized as follows: Section 2 describes some basic schemes used for event recognition; Section 3 provides a detailed description of the state-of-the-art methods for event recognition in single images; Section 4 details event recognition in personal photo-collections; Section 5 reports literature on event analysis in videos; Section 6 targets event recognition in audio recordings. Section 7 draws some concluding remarks and discusses future directions of research on the subject.

2 BASIC DEEP LEARNING BASED SCHEMES USED FOR EVENT RECOGNITION

In literature, deep architectures have been used in different ways for event recognition. In the next subsections, we provide a detailed description of the main schemes used for event recognition using deep learning.

2.1 Training a deep model from the scratch

One of the possible ways of employing deep architecture for event recognition is to train the architecture from scratch. There are two challenges associated with training a deep architecture for event recognition. Firstly, a large collection of event-related images/videos is required to fulfill the training requirements of deep architectures, they require considerably more training data compared to conventional approaches. The other challenge associated with this scheme is the processing power requirements. Though a number of benchmark datasets are available, as detailed later on, none of them is currently large enough to be used for training a deep architecture from scratch. To solve the challenges associated with the training data, the two approaches at the forefront are transfer learning and synthetic data generation. In transfer learning an existing model pre-trained for a different application is fine-tuned on event-related images as detailed in Section 2.2, while in synthetic data augmentation training sets are populated by generating synthetic images of the training set through different techniques, such as cropping and rotation [39, 43, 155]. On the other hand, and to deal with memory limitations, a number of tricks, such as using a smaller batch size, distributing the model on several machines, or reducing the model size, are used.

2.2 Fine-tuning an existing model on event-related images

Fine-tuning an existing model on event-related images can solve the challenges associated with training data requirements. In fine-tuning, also termed as transfer learning, existing models pre-trained on large datasets, such as ImageNet [42] and Places dataset [170], are tuned on event-related images by starting the learning process from the parameters learned on large collections of images. The fine-tuning process can be initiated by changing the name and number of outputs of the last layers. Moreover, the learning rate of the lower layers is reduced and increased for the newly included layer to let it learn faster compared to the lower layers as detailed in [3]. In addition, a lower step-size is usually set to let the learning rate go down faster.

2.3 Deep models as feature descriptors

In this scheme, existing deep models are used as feature descriptors where the parameters learned on the generic datasets, such as ImageNet and Places datasets, are used to extract features from event-related images. The models pre-trained on ImageNet correspond to object-level information, while the ones pre-trained on Places dataset extract scene-level features. Both types of information are widely used for event recognition purposes in literature. Though the choice of features depends on the nature of the application, deep features have been proven more effective compared to handcrafted visual features in different domains, such as analysis of natural disaster-related images from social media [12], person re-identification [44] and image retrieval [166].

Generally features are extracted from the last fully connected layer (i.e., Fc7 for AlexNet [82] and VGGNet [127]; Fc1000 for all configurations of ResNet [68] and GoogleNet [131]) by removing the top layer responsible for classification. However, features can in principle be extracted from any layer of a model. The length of the feature vectors obtained from varies with the network architectures. For instance, AlexNet and VGGNet return a feature vector of size 4096; GoogleNet and ResNet (all configurations) provide feature vectors of size 1024 and 1000, respectively. These features are then used to train classifiers, such as Support Vector Machines (SVM), Random Forest (RF) and Softmax.

3 EVENT RECOGNITION IN SINGLE IMAGES

Event recognition in single images is made difficult by the complex nature of the events, involving multiple objects and the absence of a consistent and contiguous information flow. The main concern seems to capitalize on either evidencing the most adequate representations or establishing a discriminating classification paradigm [139]. Nonetheless, the literature suggests exploiting subordinate information associated to the multimedia data ([41, 113]), as users' tags, titles, owner, upload date, and other information therein. Though such extra information available in the form of meta-data has proven to be very effective in event recognition, it also comes with lots of challenges and limitations, making its use questionable [91]. These challenges include wrong or no settings of camera time zone, missing time-stamps and modification and ambiguous meaning of tags. These challenges have been subject of different benchmarking activities, including MediaEval[1]. Cultural and disaster event recognition have been introduced as challenges in benchmarking competitions ChaLearn[2] and MediaEval 2017[3], respectively. It is well known that visual information is in general very effective in event recognition [139], owing to the rich (and somewhat self-contained) chromatic/spatial information that captures the peculiarities of the underlying event. This has instigated a number of interesting vision-based solutions. Thanks to the rise of deep

[1]http://www.multimediaeval.org/
[2]http://chalearnlap.cvc.uab.es/
[3]http://www.multimediaeval.org/mediaeval2017/

neural architectures, Convolutional Neural Networks (CNN) models demonstrated cutting-edge performance in the vision field, and proved to rectify the inefficacy of conventional approaches relying on low-level handcrafted visual features [6, 27, 110].

Similar to other computer vision applications [61, 73], the mainstream approaches to event recognition tend to capitalize on CNN architectures [112, 124, 146]. In the next subsections, we provide a detailed survey of the approaches relying on deep architectures for event recognition in single images from social media and satellite imagery.

3.1 Event Recognition in Images from Social Media

Due to the unavailability of large-scale event-related datasets, most of the efforts in this respect fine-tune existing pre-trained models on event images. In the fine-tuning process, usually a higher learning rate is used for the top layers compared to the lower layers of a model, which helps the top layer learning faster. In this regard, the models pre-trained on ImageNet [42] have been mostly exploited. In [3], a model [82] pre-trained on ImageNet is fine-tuned on a new self-collected dataset covering 14 different types of social events. In [90], existing pre-trained models, namely VGGNet [127] and GoogleNet [131] are fine-tuned for cultural event recognition.

Other methods re-train existing and novel CNN architectures for event recognition. In [119], AlexNet [82] and Network in Network (NIN) [89] are re-trained on image datasets (UIUC [87] and WIDER [157]), as well as video datasets, in particular Sport Video in the Wild (SVW) [122]. Classification scores obtained with both models are then combined in different late fusion methods. Apart from the late fusion, features extracted through the trained models are also fused in an early fusion scheme. Park et al. [112] trained a CNN model on a large number of distinctive regions extracted from cultural events-related images. The approach is mainly inspired by [60], and targets the most distinctive image regions for event recognition in single images. Regions are extracted through Selective Search [141], originally developed for object localization and segmentation. A CNN model is then trained on the extracted regions, and the final classification decision is made on the basis of majority voting.

Other methods rely on existing pre-trained CNN models for general feature extraction purposes. To this aim, most of them are pre-trained on ImageNet [42] and Places dataset [170]. Ahmad et al. [5] rely on features extracted through VGGNet-16 [127] pre-trained on ImageNet, for their salient regions-based approach to event recognition. For the selection of the event-salient regions, an image is divided into a number of regions through Selective Search [141] followed by a crowd-sourcing study to choose the most relevant ones among all the extracted regions. A Multiple-instance Learning (MIL) paradigm [145] is then used for the classification of the regions extracted from a test image. In [124], features are extracted from a network pre-trained on ImageNet along with a fine-tuned version of the model on cultural events-related images. For the final classification of an image, the scores obtained through both networks are combined. Wei et al. [154] jointly utilize the existing model pre-trained on ImageNet and Places datasets in both early and late fusion schemes. A similar strategy of combining object and scene-level information for event recognition is adopted by Wang et al. [147]. It has been observed that, in event recognition, object and scene-level information well complement each other. In order to investigate this phenomena, Ahmad et al. [7] provide a detailed comparative analysis of the performances of different CNN models, from 3 different deep architectures, pre-trained on ImageNet and Places datasets. These architectures include AlexNet [82], GoogleNet [131] and VGGNet [127]. Classification scores of these models are combined using IOWA-based late fusion [160]. In an other work from the same authors [9], three different late fusion methods, namely IOWA, Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) [126], have been used to combine classification scores from different combinations of pre-trained models.

One of the ongoing trends in event analysis is related to disaster-related user-generated images. Flood events recognition in single images from social media is also introduced as a task in the benchmark MediaEval 2017 challenge [24]. The majority of the approaches proposed in response to the challenge, are based on deep architectures. For instance, in [13], two models of AlexNet [82] pre-trained on ImageNet and Places datasets are used as feature descriptors for the representation of flood-related images. A late fusion method is then used to combine the classification scores obtained from the classifiers. Meta-data is also used to support visual information in the classification task. However, experimental results have shown the superiority of visual information over the meta-data. Benjamin et al. [22] extract features through two different deep architecture, namely DeepSentiBank [35] and X-ResNet [78] pre-trained on a visual sentiment analysis dataset [35], and ImageNet, respectively. The basic motivation for using DeepSentiBank is to extract disaster-specific information, such as broken roads and trees. A Support Vector Machine (SVM) is then used for the classification purposes. A more detailed analysis of the performance of CNN models individually and in combinations is provided in [11]. Features are extracted through seven different CNN models pre-trained on ImageNet and Places datasets. SVMs are then trained on the features extracted through individual models for the classification purposes. In [16], features are extracted from the last fully connected layer of GoogleNet pre-trained on ImageNet. A classifier is then trained on the extracted features to classify user-generated images into flooded and non-flooded categories. Keiler et al. [80] approach the problem with two different strategies relying on an individual fine-tuned CNN, and a newly trained model by combining two different networks into a single one. Initially, GoogleNet is fine-tuned on flood related images. In the second step, GoogleNet and ResNet are combined through a concatenation layer into a single network, which is then trained on the flood related images. Lopez et al. [95] rely on Inception v3 [132] pre-trained on ImageNet for flood detection in social media images. Other works submitted in the challenge [104, 105] rely on hand-crafted visual features, such as CEDD [33], JCD [32] and PHOG [25]. However, better results are reported for the frameworks relying on deep features.

3.2 Event Recognition in Remote Sensed Data

Over the last few decades, satellite imagery has been widely used in a diversified set of applications [53, 121], and, of particular interest to us, the detection and analysis of natural disasters and other adverse events using remote sensing data. Similarly to event recognition in images from social media, deep architectures have shown outstanding performance also in this domain. Amit et al. [14] proposed a CNN-based approach for adverse event detection in satellite imagery. The proposed deep architecture is composed of a total of 8 layers including three convolutional and max-pooling layers followed by two fully connected layers. Kamilaris et al. [79] rely on VGGNet [127] pre-trained on ImageNet [42], which is fine-tuned on remotely sensed images of natural disasters captured through Unmanned Aerial Vehicles (UAV). Nazr et al. [15] adopt a deep architecture for damage assessment of natural disasters in aerial images from UAV. Liu et al. [92] use deep models along with wavelet transformation for the automatic detection of natural disasters in satellite imagery. At first, wavelet transformation is proposed to enhance the satellite images in the pre-processing step, followed by a deep auto-encoder with several hidden layers to extract high-level features from the satellite images. A softmax classifier is then used for the prediction purposes.

More recently, flood detection has been introduced as a separate task at the MediaEval 2017 challenge on multimedia and satellite [24]. Participants were asked to identify flood related regions/image patches in satellite imagery. A number of interesting solutions, mostly relying on deep architectures trained on both RGB and IR components, have been proposed for the task. Bischke et al. [22] approach the task as a segmentation problem relying on a deep model [127] with three different strategies combining the RGB and IR components of the satellite imagery. In [80], the

Table 1. Summary of some relevant works in event recognition in single images: event types, dataset used, modality (Single, multi-modal) and a brief description of the method.

Ref.	Events	Datasets	Mod.	Method
[119]	Social and daily life events	USED [3], WIDER [157]	S	Relies on features extracted through two different models pre-trained on ImageNet and places datasets along with a late fusion method.
[147]	Cultural, sports and daily life events	UIUC [87], Cultural Events [45]	S	Proposes three different transfer learning methods, namely initialization based transfer learning, knowledge and data based transfer learning, focusing on more relevant objects and scenes.
[157]	Cultural events	Cultural Events [45]	S	formulates a multi-layer framework taking into account both visual appearance and the interactions among humans and objects, and combines them via semantic fusion.
[5]	Social, sports and daily life events	USED,WIDER, UIUC	S	Relies on event-salient regions, extracted through a crowd-sourcing study, for event recognition in single images. Moreover, a MIL paradigm is used for the classification of regions extracted from the test images by considering an image as a bag and the extracted regions as instances of the bag.
[90]	Daily life events	WIDER	S	Proposes a framework by combining models fine-tuned on full images as well as image regions for the classification of cultural events related images.
[112]	Cultural events	Cultural Events	S	Initially regions containing useful information are extracted from images, then CNNs are trained to classify the extracted regions, where the final classification decision is made on the basis of extracted regions.
[124]	Cultural Events	Cultural Events	S	Features are extracted through three different models, and SVMs are then trained on the extracted features. Moreover, a late fusion method with equal weights is used for the final decision.
[7]	Social, sports and daily life events	USED,UIUC, WIDER	S	Features are extracted through 10 models from 4 different deep architectures. SVMs are then trained on features extracted through individual models, and then IOWA based late fusion method is used for the final classification
[80]	Natural disaster events	DIRSM	M	Uses an early fusion, where features extracted through two different models including a fine-tuned and a trained network are combined in a single network through a concatenation layer.
[11]	Natural disaster events	DIRSM	M	Extracts deep features through 10 different CNN models pre-trained on both ImageNet and places datasets, and relies on 3 different late fusion methods to combine the classification scores obtained through individual models.
[23]	Natural disaster events	Self-collected	S	Relies on multiple CNN models as feature descriptors with a late fusion mechanism. In addition, a crawler and a filtering scheme is proposed to retrieve relevant images from social media and LandSat dataset
[11]	Natural disasters	FDSI [24]	S	Relies on GAN with with a threshold controlling mechanism for flood detection in satellite imagery

concept of dilated convolution is proposed to deal with the segmentation of satellite image patches, and classifying the regions into flooded and non-flooded ones. Ahmad et al. [11] approach the task by treating it as a generative problem, exploiting a Generative Adversarial Network (GAN) [62]. Several experiments are conducted to evaluate the performance of the network with both RGB and RGB+IR components at different threshold values. Although the IR component generally contributes to improving the classification performance, it is observed that in certain situations it might lead to false positives. A more detailed analysis of how the IR component can help in the prediction process, is provided in [10].

Table 1 summarizes some relevant works in event recognition in single images reporting the event types, datasets used for the experiments, modality (Single, Multimodal) and a brief description of the method.

3.3 Datasets

A number of benchmark datasets for event discovery in single images have been proposed. These datasets cover a variety of events ranging from social events, to sport, cultural and daily life events. The Social Event Detection Dataset (SED) [120], created within the framework of the MediaEval 2013 competition task on social event detection [120], covers 7 different social events. The training set contains a total of 27,754 images and the test set counts a total of 29,411 images. All the images in the dataset are downloaded from Flickr using event-related key words. SED also provides additional information, such as user's tags, title, description and geo-location information, although not present for all pictures. For example, geo-location information is available only for 27.8% of the pictures, while 93.4% of images contain titles, and at least one tag is available for almost all pictures. EiMM [98] also targets social events along with sports events. All the images in the dataset are downloaded from Picasa Web Album service[4], and annotated manually. The dataset also divides some events into sub-categories with corresponding labels. Ahmad et al. [3] proposed UNITN Social Events Dataset (USED), which includes 490,000 records and covers social event-classes from both SED and EiMM datasets.

With regard to cultural event recognition in single images a benchmark dataset has been released for the challenge Chalearn Looking at People 2015 [45]. The challenge aims to investigate the performance of event recognition frameworks on the basis of visual cues, such as garments, human poses, objects, and backgrounds. In order to cover such aspects, the proposed dataset includes events with diverse cultural backgrounds, having significant variability in terms of clothes, actions, and illumination. In total, the dataset covers 100 cultural events celebrated in different parts of the world.

Li et at. [87] proposed UIUC, a dataset for sports event recognition in single images. UIUC is comparatively small, and is one of the oldest datasets made publicly available for event recognition. It covers 8 sports events and it features some additional information about 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. Moreover, the distance of the foreground for objects are also provided.

Web Images Dataset for Event Recognition (WIDER) [157] covers 61 different event categories. These event categories include sport events (such as football, basketball and tennis), daily life events (such as shopping and meeting) and social events (such as concert, celebration and funeral). It also covers some specific events, such as demonstration, riot, surgery, soldier marching, and drilling. Most of these event classes are taken from the Large Scale Ontology for Multimedia (LSCOM) [106]. WIDER contains a total of around 60,000 with a significant number of images per each class. It stems as the most complex benchmark for event recognition in still images to date, mostly due to the complex nature of events, and less inter and intra class variation among the event classes.

For the disaster-related events, a benchmark dataset [24] has been introduced at MediaEval 2017, including both user-generated images taken from social media and satellite imagery. The challenge is composed of two sub tasks, namely (i) Disaster Image Retrieval from Social Media (DIRSM) and (ii) Flood Detection in Satellite Imagery (FDDI). For the DIRSM task, a total of 6,600 Flickr images along with the additional information in the form of meta-data are provided from YFCC100M-Data [136]. The meta-data include user's tags, id, nickname, together with title, description, longitude and latitude. The development set contains a total of 5,280 images along with the labels, while the test set is composed of 1,320 images. Human annotators are used to rate the collected images based on their relevance with the events. On the other hand, the satellite image patches obtained from Planet's 4-band satellites with ground-sample distance (GSD) of 3.7 meters [135] have been

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

Table 2. Summary of the features of the datasets for event recognition in single images, where "B" indicates whether the dataset has been part of a benchmark competition or not.

Refs.	Name	Features	B	Comments
[110]	SED-2011	2 classes and 73,645 images	Y	Introduced for MediaEval-2011 challenge on social events. covers 2 events, namely (i) soccer matches in Barcelona and Rome, and (ii) concerts in Paradiso; Meta-data is available.
[125]	SED-2012	3 classes and 1,67,332 images	Y	Introduced for MediaEval-2012 challenge on social events. covers 3 events, namely (i) technical events in Germany, and (ii) soccer events in Hamburg and (iii) Indignados events in Madrid; Meta-data is available.
[120]	SED-2013	7 classes and 57,165 images	Y	Introduced for MediaEval-2013 challenge on social events; Meta-data is available.
[3]	USED	14 classes and 490,000 images	N	A large collection of images aiming to fulfill the training requirements of deep learning algorithms. No mata-data available.
[87]	UIUC	8 classes and 1,579	N	Comparatively a small datset, and mainly covers 8 sports events with complexity-level information on the basis of human subjective judgment in terms of easy, medium and complex images. Meta-data not available
[157]	WIDER	61 classes and 60,000	N	One of most complex datasets; covers various types of events, such as sports, daily life events and different festivals. No meta-data.
[45]	Cultural Events	100 classes and 60,000 images	Y	Targets cultural events only; 99 cultural events from worldwide; mainly proposed for visual information based approaches with more focus on garments, human poses and objects
[24]	DIRSM	2 classes and 6,600 images	Y	The first publicly available dataset for disaster analysis in images; covers floods events only; Meta-data is available; Also provides a collection of global features extracted through several feature descriptors [97].
[24]	FDDI	2 classes and around 700 image patches	Y	The first publicly available dataset for disaster analysis in satellite images; covers floods events only; image patches are provided from 8 different locations in 4-channels, namely R, G, B and IR; Also provides temporal information.

provided for the FDSI task. The dataset mainly contains images from 8 different flood events. All the patches have been projected in the UMT projection using WGS84 datum, and are provided in the GeoTiff format and size of 320x320x4 pixels. Moreover, each patch is composed by 4 channels, namely RGB and Infrared. Similar to DIRSM, the FDSI dataset is divided into development and test sets.

In Table 2 we provide a summary of the different properties along with some statistics of the datasets available for event recognition in single images.

4 EVENT RECOGNITION IN PERSONAL PHOTO-COLLECTIONS

Personal photo-collections posses a number of characteristics that make event recognition a more challenging problem compared to single images. One of these characteristics is the presence of irrelevant images. For example, photo-collections may contain face-close ups, images that are not related to the event, or that can be in principle associated to any event. Figure 1 illustrates some relevant and non-relevant images from a wedding album. Furthermore, another challenging aspect of event recognition in personal photo-collection is the album-level annotation.

Similarly to event recognition in single images, deep architectures have shown outstanding performance in event recognition also when dealing with personal photo-collections. In [63], a hierarchical approach comibining a coarse and a fine classifier has been proposed. Initially, a SVM classifier is trained on scene-level features (coarse classifier), followed by a fine event-classifier trained on object-level features, extracted through CaffeNet [75], pre-trained on ImageNet, along with temporal information. To deal with irrelevant images in photo-collections, both classifiers are trained on global average and aggregated feature vectors of all images in a collection. The same

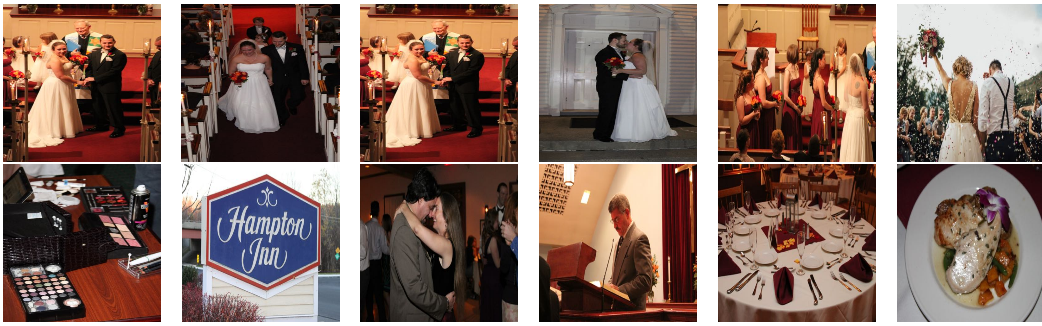

Fig. 1. Some sample relevant (top) and irrelevant (bottom) images from a wedding event.

authors, in [64], rely on three different CNN models pre-trained on ImageNet, Places and on user-contributed data gathered from Flickr [54] for feature extraction. To deal with the irrelevant images, a weighted average features strategy has been proposed in place of global average and aggregated features, for improved robustness. In [4], features extracted through VGGNet [127] pre-trained on ImageNet are used in a Multiple-instance Learning (MIL) paradigm [145], where each album is treated as a bag and the images in an album are considered as instances of the bag. MIL is a semi-supervised approach, and suitable for applications with ambiguous or weakly labeled data. Bach et al. [19] proposed a Probabilistic Graphical Model (PGM) on object and scene-level features extracted through CNN models pre-trained on ImageNet and Places datasets. Moreover, a feature relevance scheme is proposed to predict the relevance of object and scene-level features for event recognition in photo-collections. Wu et al. [156] developed a deep architecture to combine/accumulate features extracted from all images in an album into an album-level representation. Moreover, features are extracted through two different models pre-trained on ImageNet and Places datasets, where both models are combined into a single network. The network is then trained on images from personal photo-collections. Similarly, object and scene-level information are utilized in [147] using three different transfer learning techniques, namely, initialization-based transfer learning, knowledge and data-based transfer learning. In the initialization-based transfer learning method, pre-trained models are fine-tuned on a new dataset. In the knowledge-based approach, existing pre-trained models are used to predict the likelyhood of object and scene classes, whose output scores are used as soft codes to guide the fine-tuning of the models on event datasets. In data-based method, two models are fine-tuned on two different datasets; one on a subset of ImageNet or Places, while the other on event-related images. Both networks have their separate data, fully connected layers and loss function, while the rest of the networks share weights. Overall, the best performances are reported for data-based transfer learning. In Table 3, we report a summary of some of the relevant works in event recognition in personal photo-collections, listing the event types, datasets used, and a brief description of the method.

4.1 Datasets

In contrast to other event recognition strategies, comparatively less literature and datasets are available for event recognition in personal photo-collections. Bossard et al. [26] released a benchmark dataset, namely Personal Events Collection (PEC). The dataset provides a significant number of photos (around 61,000) assembled into 807 albums from 14 different types of events. The dataset also provides additional information including date and geo-location information, which is used to support visual information in the classification. Moreover, the number of images per photo album

Table 3. Summary of some relevant works in event recognition in personal photo collections: event types, datasets used, modality (Single, multi-modal) of information used and description of the method.

Refs.	Events	Dataset	Mod.	Method
[63]	Social	PEC [26]	M	Uses object and scene-level information in a hierarchical way, where firstly SVMs are trained on deep features extracted through a model pre-trained on ImageNet to classify coarse events. Subsequently, a combination of models pre-trained on ImageNet and places datasets along with meta-data are used with equal weights for each type of classifiers for final classification.
[64]	Social	PEC	M	Combines object and scene-level information along with user-contributed attributes [54] in a weighted average features strategy to deal with irrelevant images in photo collections. Subsequently, SVMs are trained on the weighted aggregated features for the classification of photo collections.
[19]	Social	PEC	M	proposes a Probabilistic Graphical Model (PGM) on object and scene-level features extracted through CNN models pre-trained on ImageNet and Places datasets. Moreover, a feature relevance scheme is proposed to predict the relevance of object and scene-level features for event recognition in photo-collection.

Table 4. Summary of the features of the datasets for event recognition in personal photo collections where "B" indicates whether the dataset has been part of a benchmark competition or not.

Refs.	**Name**	**Features**	**B**	**Comments**
[26]	PEC	14 classes, 807 albums and 61,000 images	N	Annotation at collection level only; Meta-data including temporal and geo-location information along with tags and user's ID is also available; devotes a limited number of albums in test set
[138]	Holidays	12 classes, 565 albums and 46609 images	N	Mainly covers holidays events in USA; annotation is provided at album level; object level tags are also provided as additional information along side meta-data, which may help in the classification process

varies from album to album. Another dataset for event recognition in photo collections is collected by Tsai et al. [138]. The dataset mainly covers holidays events. In total, 12 different events are considered, counting 565 albums and 46,609 photos in total. A summary of the datasets is reported in Table 4. The datasets are annotated at album level.

5 EVENT RECOGNITION IN VIDEOS

In contrast to still images, videos tend to be a richer source of information providing visual and motion information. Several interesting solutions have been proposed for video analysis, such as 3D-CNN based methods [137], non-local Neural Networks [148], attention-based approaches [93, 94], graph-based approaches [149], and motion learning approaches [46, 130]. Event-based analysis of videos offers the possibility of exploring the interactions among humans and objects in complex scenes. Some sample events analyzed in literature include attempting a bike trick, fixing musical instrument, and parking a vehicle. Event recognition in videos is generally composed of two complementary phases, namely (i) feature extraction, and (ii) classification. As far as feature extraction is concerned, several interesting algorithms have been proposed to extract significant data in the form of text, audio and visual information. They can be broadly categorized into (i) static frame-based visual features and (ii) motion-based spatio-temporal features. The initial efforts in this regard mainly focus on handcrafted static visual features, such as SIFT [96] and SURF [21], spatio-temporal features, such as Motion SIFT (MoSIFT) [34], and dense trajectories [143, 144].

As far as the use of deep architectures in event recognition in videos is concerned, the literature is rapidly growing [139]. In [159], a CNN-based framework is proposed for event detection in videos by targeting static frame-based visual features, only. Two contributions are made in the paper, firstly an appropriate encoding method is used for aggregating frame-level static features.

Secondly, a set of latent concept descriptors is proposed as a frame descriptor, which not only enriches the extracted visual features/information but also reduces the computational cost. Zha et al. [167], provide a detailed analysis of how CNN models pre-trained for image classification can be utilized for event detection in videos. To this aim, spatial and temporal pooling, feature normalization, along with different choices of CNN layers for feature extraction as well as different choices for classifiers have been incorporated.

In order to better utilize CNN models pre-trained on static images (ImageNet), Mettes et al. [102] succeeded in improving the performances in recognition, by training a model on the complete ImageNet dataset (21,814 classes and more than 14 million images) instead of relying on the conventional 1,000 classes mostly used in the literature. Ye et al. [162] proposed a large-scale event-specific concept library, namely, "EventNet" covering a large number of daily life events and associated concepts (around 500). After data collection, a CNN model is trained on the collected videos (95,321) from 500 different events. The model is then used for feature extraction. Subsequently, binary SVM classifiers are trained to build a concept library, which is used to generate a concept-based representation of the videos. Gan et al. [56] proposed DevNet, a flexible deep model that detects and classifies events, simultaneously. The model takes key frames of videos as input, and classifies the underlying events at video level by aggregating the features extracted at frame-level. The key frames are identified by creating a saliency map. Another interesting solution for static frame-based event detection in videos is proposed in [150], where a deep architecture is used to extract contextual features for video classification. Initially, two different types of contextual features representing event neighbourhood are extracted, followed by training a deep model to learn and combine middle level semantic features. A fusion-based approach with features extracted from static frames has been proposed in [117]. Initially, features are extracted through several deep models and fed into separate SVM classifiers. The classification scores are then fused by assigning different weights learned through a deep architecture to each model. Another fusion method for event detection in videos is proposed in [169], where a deep architecture is proposed to combine multiple semantic cues. The work jointly considers the semantic features, namely actions, objects, and scenes. Initially, each type of features is modeled by feeding it into a corresponding multi-layer feature abstraction pathway, followed by a fusion layer. The interaction/correlation among the semantic cues is modeled through an unsupervised auto-encoder. Finally, the deep architecture is fine-tuned on the event datasets to answer how the semantic cues of who, what, and where, compose a complex event. In [118], an ensemble deep learning framework is proposed to combine feature extracted through different CNN models, including AlexNet [82], GoogleNet [131], RCNN [60], and ResNet [69]. The framework aims to overcome the issues associated with unbalanced data and over-fitting. The ensemble approach is developed based on the performance of each weak learner (SVM) trained on features extracted through individual deep model where the weights are assigned to models considering a metric for imbalanced data.

Motion has also been exploited in a number of works to complement static frame-based information. Xu et al. [158] proposed Appearance and Motion DeepNet (AMDN) for anomaly event detection in videos. The framework relies on deep neural networks to learn features from static frames and motion information. To combine appearance and motion, a novel double fusion method is proposed, by merging the capabilities of early and late fusion. Stacked autoencoders are used to learn the individual and joint representation of appearance and motion. Next, one-class SVMs are trained on the extracted features to predict anomaly scores for each individual feature set, which are finally integrated in a late fusion scheme to compute the final anomaly score. Another approach relying on deep features for abnormal event detection in videos is proposed by Feng et al. [47]. The model is able to extract and fuse different types of features, including appearance, texture, and short-term motion in an unsupervised manner. In order to fuse and learn the correlation among the features, stacked denoising autoencoders are used. Long-term temporal cues are then modeled with a long short-term memory (LSTM) recurrent network to better model the regularities of video events. To combine motion and static frame information for event detection in videos, Wu et al. [156] also rely on a deep learning-based framework. Both types of features are learned through separate convolutional neural networks. A regularized feature fusion network is then used to learn a combined representation of spatial and motion features extracted through individual networks for the final classification. Moreover, Long Short Term Memory (LSTM) networks are used on both types of features to better model the long-term temporal cues. In [165], a multistage hybrid fusion scheme has been used to jointly employ motion and static frame-level features. For static features, a CNN model [82] pre-trained on ImageNet has been used while improved trajectories [143] are used to extract motion information. In order to incorporate the temporal information in videos, Yao et al. [161] propose to consider both the local and global temporal structure of videos through a deep architecture with 3-D convolution. Initially, short temporal information is modeled with a 3-D CNN pre-trained on action recognition. A temporal attention mechanism is then proposed for the representation of complete videos by selecting the most relevant temporal segments through a RNN. Jian et al. [76] rely on a CNN model alongside a RNN to capture spatial and temporal information for event detection in soccer videos. The target events in this work include goals, corners and goal attempts. Initially, event boundaries are determined through Play-Break (PB) segments followed by feature extraction from key frames from the PB segment through a pre-trained CNN. A RNN is then used to map the spatial features to underlying soccer events to consider the temporal information. Spatial and temporal information are also jointly exploited by Yu et al. [164] in a two-step training process. In the first phase, frame-level visual features and temporal information are extracted for the representation of event-related videos. For frame-level visual features, pre-trained CNN models are used, while a RNN is used to learn the temporal properties of the event-related videos in an unsupervised way. Both feature vectors are then aggregated, followed by an activation layer with labels to obtain the final event detection model.

The authors in [77] propose a framework relying on deep learning to effectively use both features and class relationship for events analysis in videos. Chang et al. [31] propose a semantic pooling-based approach focusing on the most relevant parts of videos for event recognition in user-generated videos shared over the Internet. In contrast to conventional pooling strategies that aggregate the video shots, and result in a great loss of information, this method relies on semantic saliency to localize more relevant segments. Segments are then prioritized based on their saliency scores. Finally, a novel isotonic regularizer is used to exploit the constructed semantic ordering information. Some solutions proposed in literature rely on multi-modal features. For instance, in [74], a novel deep learning architecture is proposed to jointly utilize audio-visual information with local contrast normalization and spatial maximum pooling to each type of information, which make the features robust against local variances. In order to model the correlation between audio and visual features, an auto-encoder is used at each layer of the network. There are also some approaches relying on web data for training purposes [55, 57]. For instance, Singh et al. [128] propose an event recognition framework relying on pairs of concepts automatically discovered from the web for the retrieval of event-related videos.

In Table 5, a summary of some of the most relevant approaches discussed in this section is provided.

5.1 Datasets

In the context of video analysis, the most popular benchmark challenge is Video Retrieval Evaluation (TRECVID), which organizes benchmarking activities focusing on different application domains. Multimedia Event Detection (MED) has also been part of the challenge since 2010. Each year, the

Table 5. Summary of some relevant works in event recognition in videos: event types, dataset, modality (Single, multi-modal) of the information and a brief description of the method.

Refs.	Events	Dataset	Mod.	Method
[159]	Daily life	TRECVID 13 and TRECVID 2014 [109]	S	Extracts the general features from pool-5 layer of an existing model as vectors of latent concept descriptors, followed by an encoding method to generate the video representation.
[167]	Daily life	TRECVID MED 2014	S	Relies on existing pre-trained models with several spatial and temporal pooling and feature normalization techniques. SVMs classifier are trained on the extracted features, and individual classification scores are then combined in a late fusion method
[102]	Daily life	TRECVID 2013 and 2015	M	Relies on features extracted through two different pre-trained models on complete set of ImageNet dataset (21,814 classes and more than 14 million images) instead of relying on conventional 1,000 classes, followed by averaging the representations of the frames over each video. Subsequently, SVMs are used for the final classification of the videos.
[162]	Daily life	Self-collected	S	A CNN model is trained on 95, 321 videos over the 500 events, and the model is then used to extract deep learning feature from video content. With the learned deep learning feature, 4, 490 binary SVM classifiers are trained as the event-specific concept library.
[56]	Daily life	TRECVID 2014	S	Generate a spatial-temporal saliency map by back passing through DevNet, which then used to find the key frames which are most indicative to the event, as well as to localize the specific spatial position, usually an object, in the frame of the highly indicative area.
[117]	Natural disasters	Self-collected dataset	S	Features are extracted through several deep models, and separate SVMs classifiers are trained on the features extracted through individual models. The classification scores are then fused in a late fusion method by assigning different weights learned through a deep architecture to each model.
[169]	Daily life	TRECVID MED 2011	S	Relies on a deep architecture to combine multiple semantic cues. Initially, each type of features is modeled by feeding it into a corresponding multi-layer feature abstraction pathway, followed by connection of all of these features through a fusion layer. The interaction/correlation among the semantic cues is modeled an auto-encoder.
[118]	Daily life	TRECVID	S	Proposes a deep learning framework to overcome the issues associated with imbalance data and over-fitting due to a single model by combining the classification scores of different SVMs used in the layer of the ensemble architecture on the top of the feature extracted through 3 different individual models.
[164]	Daily life	TRECVID MED14	S	Relies on CNNs and a RNN for frame-level visual features and temporal information, respectively. Both feature vectors are then aggregated followed by an activation layer with labels to obtain the final event detection model.
[165]	Daily life	TRECVID MED14	M	Proposes a hybrid fusion techniques with multimodal information including visual, audio and text. Moreover, semantic features based on low-level features are used along with deep features.

organizers provide a dataset covering specific events for the evaluation of the approaches proposed by the participants. Over the years the challenge has evolved in terms of type and number of events. Table 6 summarizes the main properties of the TRECVID datasets used for event recognition in videos.

6 EVENT RECOGNITION IN AUDIO RECORDINGS

Another facet of event analysis is to explore the information provided by audio recordings. Typical applications of Acoustic/Audio Event Detection (AED) includes multimedia indexing and retrieval [168], surveillance [65] and robotics [37]. AED frameworks are typically composed of feature extraction and inference/classification phases, and aim to recognize a distinct sound pattern/event in a continuous acoustic signal [18].

Table 6. Summary of the datasets for event recognition in videos where "B" indicates whether the dataset has been part of a benchmark competition or not.

Refs.	Name	Features	B	Notes
[48]	TRECVID MED-2010	3492 video clips with audio recordings	Y	Collected for the evaluations of TRECVID MED 2010 task. Mainly covers three events, namely "Making a cake", "Batting a run", and "Assembling a shelter". Separate development and evaluation sets
[30]	TRECVID MED-2011	15 events; a total of 370 hours of video clips	Y	Collected for the evaluations of TRECVID MED 2011 task. One of the initially widely used dataset for event detection in videos
[108]	TRECVID MED-212	20 events, and a collection of over 4000 hours of multimedia clips	Y	Combined collection for MED and MER tasks; cover two different types of events, namely pre-specified events and Ad Hoc events
[85]	TRECVID MED-2013	20 events, video clips are provided in MPEG-4	Y	Covers the development sets from TRECVID MED-12 and 11 with additional events and data; training exemplar conditions are 100, 10 and 0 exemplars
[109]	TRECVID MED-2014	20 events and a total of 7580 hours video clips (almost double of previous years)	Y	Video clips were provided in MPEG-4 formatted and encoded to the H.264 standard. The audio was encoded using MPEG-4's Advanced Audio Coding (AAC) standard
[109]	SMED	Total of 45 hours video clips	Y	First dataset for surveillance event detection in videos taken from the Imagery Library for Intelligent Detection Systems (iLIDS)

Ballan et al. [20] proposed a probabilistic neural network for audio event detection in soccer recordings. In [58], a Deep Neural Network (DNN) is used for the detection and classification of isolated acoustic events, such as motorbiking, baby crying and rain events. After pre-processing, features from the left and right channels are extracted to feed the layers of the network. A classifier is then trained on the extracted features. In [28], the same authors extended their framework by treating the task as a multi-label learning problem with no bound on the number of simultaneous events. Initially, spectral domain features are used to represent the audio signals, and DNNs are then used to learn a mapping between extracted features and underlying events. In [133], a framework relying on a deep network with 9 layers is proposed, which allows the network to directly model entire audio events. The proposed learning model is inspired from VGGNet [127], where larger convolutional kernels are replaced with a stack of 3x3 kernels without pooling layers. Moreover, in order to train the network, a novel data augmentation technique is proposed by producing more samples through a random mixing of two sounds of a class at randomly selected time slots. Lee et al. [86] propose an ensemble of CNNs, each processing a different length of analysis window for multiple input scaling. The models analyzing signals at different scales complement each other in both sound event detection and localization. Similarly, in [88], a CNN-based framework is proposed where a CNN with 1-D convolution is combined with a RNN with long short term memory units (LSTM). In both networks, log-amplitude mel-spectrogram is used as an input feature set. The 1-D ConvNet is used on the time-frequency frame to convert the spectral feature followed by RNN-LSTM to incorporate the temporal dependency of the extracted features. Adavanne et al. [2] rely on CNNs with 3-D convolution for sound event detection in audio recordings. In details, a stacked Convolutional and Recurrent Neural Network (CRNN) is proposed with a 3-D convolution operation in the first layer to learn the inter and intra channel features in a multi-channel input signal. In order to deal with memory requirement issues of CNNs, Meyer et al. [103] rely on structural optimization of CNNs. With the proposed strategy, the authors report a significant reduction in the memory requirement as well as an improvement in the performance. Takahashi et al. [134] proposed a deep architecture, namely AENet, for audio event recognition in videos. In order to incorporate long-time frequency structure of audio events, in the proposed framework a CNN operating on a large temporal input video is used. Kons et al. [81] proposed a DNN-based approach for audio event detection/classification in outdoor environments. Results of different classification techniques along with the combined score of a late fusion method are reported. Choi et al. [36] use DNNs for AED with a noise reduction strategy. An exemplar-based noise reduction scheme is used for enhancing mel-band energy features extracted from audio signals. A multi-label DNN classifier is then trained on the extracted features to model mel-band energy to the underlying events. Hou et al. [72] proposed a multi-model framework composed of a DNN, and five models based on Bi-directional Gated Recurrent Units Recurrent Neural Networks (BGRU-RNN) for different types of events. The deep model is developed for the detection of sound events related to car while the BGRU-RNNs based models are meant for other types of sound events, such as brakes squeaking, children and large vehicles noises.

Acoustic scene/event classification is also proposed in [17], where a deep model called SoundNet is developed for learning audio representation in unlabelled videos, by leveraging the synchronization between vision and sound. The underlying insight of the proposed work is transferring discriminative knowledge from visual recognition networks into sound networks. Visual data is used in the training phase only, and the network has no dependence on vision during classification/inference of audio events. In the classification phase, an SVM is trained on audio features extracted through SoundNet with a one-versus-all strategy to deal with multi-class classification. In [111], a bi-directional long short term memory (BLSTM) framework is proposed for AED in daily life audio recordings, where the model is trained to map audio features of a mixture signals consisting of sounds from multiple classes to a two class indicator for each class. Hayashi et al. [66] proposed a bi-directional Long Short-Term Memory (BLSTM) combined with Hidden Markov Model (BLSTM-HMM), where a hybrid model of neural network and HMM is extended to a multi-label classification problem. In [67], the same authors proposed a temporal structure modeling technique with a hidden Markov model (HMM) in combination with a bidirectional long short-term memory (BLSTM) for sound event detection. In [1], a RNN-based approach to AED is presented. The framework combines spatial features, extracted through a long short term memory (LSTM) network with harmonic features. In details, three different characteristics, namely log mel-band energies, pitch frequency, periodicity, and time difference of arrival (TDOA) in sub-bands, are considered. These features are extracted at a hop length of 20 ms to ensure consistency across them. Moreover, the RNN-LSTM model is composed of two hidden layers each with 32 LSTM units.

Wang et al. [153] rely on RNN acoustic features for event recognition in multimedia contents, where a deep RNN along with temporal information is used for both representation and classification purposes. In another work from the same authors [152], a sequence-to-sequence model namely Connectionist Temporal Classification (CTC) is proposed to overcome the limitation of RNNs due to their dependence on frame by frame prediction. CTC provides ordered sequences of audio events without exact starting and ending times. In [114], another CNN-based approach with comparatively less number of layers has been proposed for AED. In total, the network is composed of a convolutional, a pooling and a softmax layer. Another distinguishing characteristic of the model is its varying-size of the convolutional filters at the convolutional layer, and 1-max pooling at the pooling layer. The authors in [40] propose a DNN model for audio event analysis. Hidden layers are embedded with a number of functionalities, such as batch normalization, dropout, L2 regularization and the rectified linear unit (ReLU). In [140], a DNN relying on Kullback-Leibler (KL) divergence is used to combine several classifiers/models, namely GMM, a DNN, and LSTM for final classification of the underlying audio event. Cakir et al. [29] proposed a DNN to combine human perception and learning capabilities of deep architectures. In the first layer of the proposed deep architecture, instead of random initialization, weights and biases are initialized with the coefficients of a filter bank, which are then updated during the training process to provide better discrimination over the target audio events. The filter-bank is designed on the basis of human perception, so the sound intervals consist of equal perceptual pitch increments. Wang et al. [151] approached audio-based multimedia event detection with recurrent SVMs by combining kernel mapping and large-margin optimization criterion of SVMs along with the processing capabilities of RNNs for variable length sequences of audio signals. Several experiments were conducted to compare the performances of recurrent SVMs against individual SVMs and RNNs. Mesaros et al. [101] provided a detailed discussion on segment and event-based definitions of metrics used for the evaluation of AED approaches along with a toolbox containing implementations of these metrics. Table 7 summarizes some properties of the approaches presented in this section.

6.1 Datasets

Audio event recognition has been part of benchmarking challenges. In this regards, Detection and Classification of Acoustic Scenes and Events (DCASE)[5] is one of the most popular benchmarking activities on audio-based analysis of multimedia. Sound event detection [84] has been part of the activity for the last three years. In the first year, two different datasets have been provided for the two tasks, namely (i) Sound Event Detection in Synthetic Audio (SEDSA), and (ii) Sound event detection in real life audio (SEDRA).

The dataset provided for SEDSA covers 11 different events, where 20 samples were provided for each sound event for training set along with a development set consisting of 18 minutes of synthetic mixture material in 2 minute-long audio files as additional materials. The audio files, recorded in a calm environment, are sampled at 44.1kHz. The dataset consists of two types of acoustic events, namely home events (indoor events such as, rustling, cutlery, dishes drawer etc.) and residential area (outdoor events, such as banging, car passing by and people talking etc.). The provided audio files, represent common environments of interest in applications for safety and surveillance as well as human activity monitoring or home surveillance. TUT Sound Events 2017 [99] has been introduced for the DCASE 2017 challenge on event recognition in audio recordings focusing on human activities and hazardous situations. This dataset is a subset of TUT Acoustic scenes 2017 [99], and covers outdoor acoustic scenes with various levels of traffic and other outdoor activities. The dataset consists of 7 different events, namely brakes squeaking, car noise, children talking, large vehicles, people speaking and walking in the streets.

Apart from the datasets proposed for DCASE challenge on audio event detection, there are several other datasets proposed for the evaluations of AED frameworks [70, 83, 116, 142]. For instance, Foggia et al. [50] proposed a dataset for event detection in road surveillance applications. The same authors also proposed another dataset [49] that focus on events, such as glass breaking, gun shots and screams. Salamon et al. [123] as well as Piczak et al. [115] proposed a dataset for urban events detection. Other datasets are also proposed for audio event recognition in isolated environments [38, 59]. In Table 8, we provide a summary of different properties of some of the datasets discussed in this section.

7 DISCUSSION AND CONCLUSIONS

In this survey paper, we conducted a comprehensive analysis of deep learning-based frameworks for event recognition. We have discussed several event detection and recognition approaches in four different sub-domains, namely event recognition in single images, personal photo-collections, videos, and in audio recordings. We also discussed the challenges associated with event recognition

[5] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/index

Table 7. Summary of some relevant works in event recognition in audio recordings: event types, dataset, modality (Single, multi-modal) of the information and a brief description of the method.

Refs.	Events	Dataset	Mod.	Method
[58]	Daily life events	Acoustic dataset [100]	S	Relies of deep neural networks with an early fusion of audio features extracted from current frames and the neighbouring frames (left and right frames) to obtain a larger feature set by covering overlapping frames
[28]	Daily life events	Acoustic dataset [100]	S	treats the task as a multi-label learning problem with no bound on the number of simultaneous events. Initially, spectral domain features are used to represent the audio signals, and DNNs are then used to learn a mapping between extracted features and underlying events.
[133]	daily life events	Acoustic events [51]	M	A deeper network with 9 layers and a larger input filed allowing the network to directly model entire audio events and be able to be trained completely.
[88]	Rare sound events	DCASE 2017 Benchmarking	S	A CNNs-based framework, where a CNN with 1-D convilution is combined with a RNN with long shortterm memory units (LSTM). In both networks, a log-amplitude mel-spectrogram is used as an input feature.
[2]	Street events	TUT-SED 2017 [99]	M	Relies on a stacked Convolutional and a Recurent Neural Network (CRNN) with a 3-D convolution operation in the first layer to learn the inter and intra channel features in a multi-channel input signal.
[81]	Outdoor events	FreeSound [51]	S	Proposes a DNNs based approach for audio event detection/classification of outdoor environments. In addition, results of different classification techniques along with combined score of a late fusion method are reported.
[36]	Daily life events	DCSE 2016 [84]	S	Relies on a DNN for AED with a noise reduction strategy. An exemplar-based noise reduction scheme is used for enhancing mel-band energy feature extracted from audio signals. A multi-label DNN classifier is then trained on the extracted features to model mel-band energy feature to underlying events in audio recordings.
[72]	Daily life events	DCSE 2017	M	A multi-model framework composed of a DNN and five models based on Bi-directional Gated Recurrent Units Recurrent Neural Networks (BGRU-RNN). The deep model is developed for the detection of sound events related to car while the BGRU-RNNs based models are meant for other types of sound events, such as brakes squeaking, children and large vehicles noises.
[111]	Sports and daily life events	Self-collected dataset	S	A bi-directional long short term memory (BLSTM) recurrent neural networks (RNNs) based architecture trained to map audio features of a mixture signal consisting of sounds from multiple classes to a two class indicator for each class.
[67]	Daily life events	DCASE 2016	S	A temporal structure modeling technique with a hidden Markov model (HMM) in combination a bidirectional long short-term memory (BLSTM) recurrent neural network (RNN).
[40]	Daily life events	DCASE 2016	S	Provides a novel DNN composed of an input, several hidden layers and an output layer. Hidden layers are embedded with a number of functionalities, such as batch normalization, dropout, L2 regularization and the rectified linear unit (ReLU).
[17]	Miscellaneous events	Acoustic Scene Classification DCASE	M	The underlying insight of the proposed work is transferring discriminative knowledge from visual recognition networks into sound networks. Visual data is used in the training phase only, and the network has no dependence on vision during classification/inference of audio events..

in these sub-domains and presented the most common benchmarking datasets available for the evaluation.

We observed a trend towards the use of existing pre-trained models as feature descriptors or fine-tuning them on event-related images for event recognition in single images. To this aim, most of the models pre-trained on ImageNet are exploited, showing a superiority of object-level

Table 8. Summary of the datasets for event recognition in audio recordings where "B" indicates whether the dataset has been part of a benchmark competition or not.

Refs.	Name	Features	B	Notes
[40]	SEDSA	11 events and 20 samples per event along with additional materials of 18 minutes	Y	Recorded through shotgun microphone AT8035 connected to a ZOOM H4n recorder, and are sampled at 44.1kHz; Other parameters include the EBR with different values.
[99]	TUT Sound Events 2017	7 events with 3-5 mins audio files	Y	Focuses on human activities and hazard situations recorded through a binaural Soundman OKM II Klassik/studio A3 electret in-ear microphone and a Roland Edirol R-09 wave recorder.
[50]	MIVIA road audio events	2 events and 400 files	N	Road surveillance files; Recorded with an Axis P8221Audio Module and an Axis T83 omni-directional microphone; sampled at 32000 Hz and quantized at 16 bits per PCM sample.
[123]	UrbanSound	10 events; 27 hours of audio files	N	Targets sudio events in urban areas; data is collected from Freesound[6], an online free repository containing over 160,000 user-uploaded recordings under a creative commons license
[115]	ESC	3 subsets with different number of files and classes.	N	Downloaded for public recordings; 5-second-long clips, 44.1 kHz, single channel, Ogg Vorbis compressed @ 192 kbit/s
[49]	Mivia AED	6000 events, 4200 for training and 1800 in the test set	N	Provides each audio event at 6 different values of signal-to-noise ratio and overimposed to different combinations of environmental sounds in order to simulate their occurrence in different ambiences.

information. However, a joint use of object and scene-level information has also been employed in several works showing a clear advantage of the fusion over the individual models. The literature shows that most of the efforts are spent on late fusion aiming to combine the classification scores of classifiers trained on features extracted through several models.

We also observed that most of the proposed approaches for event recognition in personal photo-collections aim to deal with ambiguous training samples/irrelevant images. To this aim, a number of frameworks mostly relying on semi-supervised learning techniques have been proposed.

The literature on event-based analysis of videos shows a clear trend towards the use of both CNN features extracted from static-frames, and RNNs for motion-based information. A number of fusion techniques have been proposed, where a joint use of both types of features through early fusion techniques have shown significant improvement over the individual modality.

Before concluding, we would like to also provide the reader with an insight about the state-of-the-art performances in quantitative terms, related to the research areas mentioned above. Although the figures are expected to be growing over time, our aim is to set a marker line according to the existing literature, to be used as a reference for future works. In terms of single event images from social media, and considering the most widespread datasets, WIDER and Cultural events dataset, methods like [9, 112, 147] and [124] have demonstrated the ability to achieve an average classification accuracy of 56% and 87%, respectively, highlighting the complexity of WIDER compared to other datasets. As far the natural disaster analysis in single images is concerned, the average accuracy reached by the most relevant methods [11, 12, 22, 107] is in the range of 95% and 85% for social media and satellite imagery, respectively. Considering this is a recent research trend, the higher performances are probably mostly linked to the available datasets, rather than to the scientific problem itself. Similar performances have also been reached in event recognition in personal photo collections, showing an average accuracy of about 85% to 87% on PEC dataset (see [8, 63]). As probably expected, the performances for event recognition in videos are generally lower, due to the high variability of a visual content over time. State-of-the-art methods, such as [159, 164], achieve a mean average precision of 38% to 44% on widely used datasets, such as TRECVID MED-2013 and TRECVID MED-2014. In the case of audio event recognition, the methods proposed in [36, 67, 86?] can achieve an average F-score of 80% and 58.9% on the widespread datasets DCASE-2016 and DCASE-2017, respectively, demonstrating a significantly higher complexity of the latter.

REFERENCES

[1] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, and Tuomas Virtanen. 2017. Sound event detection in multichannel audio using spatial and harmonic features. *arXiv preprint arXiv:1706.02293* (2017).

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. 2018. Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features. *arXiv preprint arXiv:1801.09522* (2018).

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

[4] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco GB De Natale. 2017. Event recognition in personal photo collections via multiple instance learning-based classification of multiple images. *Journal of Electronic Imaging* 26, 6 (2017), 060502.

[5] Kashif Ahmad, Nicola Conci, and FGB De Natale. 2018. A saliency-based approach to event recognition. *Signal Processing: Image Communication* 60 (2018), 42–51.

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

[7] Kashif Ahmad, ML Mekhalfi, Nicola Conci, Giliua Boato, F Melgani, and FGB De Natale. 2017. A pool of deep models for event recognition. In *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2886–2890.

[8] Kashif Ahmad, Mohamed Lamine Mekhalfi, and Nicola Conci. [n. d.]. Event Recognition in Personal Photo Collections: An Active Learning Approach. ([n. d.]).

[9] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Farid Melgani, and Francesco De Natale. 2018. Ensemble of Deep Models for Event Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2 (2018), 51.

[10] Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Nicola Conci, and Pål Halvorsen. 2018. Social media and satellites. *Multimedia Tools and Applications* (2018), 1–39.

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

[12] Kashif Ahmad, Amir Sohail, Nicola Conci, and Francesco De Natale. 2018. A Comparative study of Global and Deep Features for the analysis of user-generated natural disaster related images. In *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 1–5.

[13] Sheharyar Ahmad, Kashif Ahmad, Nasir Ahmad, and Nicola Conci. 2017. Convolutional neural networks for disaster images retrieval. In *Proceedings of the MediaEval 2017 Workshop (Sept. 13–15, 2017). Dublin, Ireland*.

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

[15] Nazia Attari, Ferda Ofli, Mohammad Awad, Ji Lucas, and Sanjay Chawla. 2016. Nazr-CNN: Fine-Grained Classification of UAV Imagery for Damage Assessment. *arXiv preprint arXiv:1611.06474* (2016).

[16] Konstantinos Avgerinakis, Anastasia Moumtzidou, Stelios Andreadis, Emmanouil Michail, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2017. Visual and textual analysis of social media and satellite images for flood detection@ multimedia satellite task MediaEval 2017. In *Proceedings of the Working Notes Proceeding MediaEval Workshop, Dublin, Ireland*. 13–15.

[17] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. 892–900.

[18] Elham Babaee, Nor Badrul Anuar, Ainuddin Wahid Abdul Wahab, Shahaboddin Shamshirband, and Anthony T Chronopoulos. 2018. An Overview of Audio Event Detection Methods from Feature Extraction to Classification. *Applied Artificial Intelligence* (2018), 1–54.

[19] Siham Bacha, Mohand Said Allili, and Nadjia Benblidia. 2016. Event recognition in photo albums using probabilistic graphical models and feature relevance. *Journal of Visual Communication and Image Representation* 40 (2016), 546–558.

[20] Lamberto Ballan, Alessio Bazzica, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. 2009. Deep networks for audio event classification in soccer videos. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*.

IEEE, 474–477.

[21] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer, 404–417.

[22] Benjamin Bischke, Prakriti Bhardwaj, Aman Gautam, Patrick Helber, D Borth, and A Dengel. 2017. Detection of flooding events in social multimedia and satellite imagery using deep neural networks. In *Working Notes Proceedings MediaEval Workshop*. 2.

[23] Benjamin Bischke, Damian Borth, Christian Schulze, and Andreas Dengel. 2016. Contextual enrichment of remote-sensed events with social media streams. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1077–1081.

[24] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth. 2017. The Multimedia Satellite Task at MediaEval 2017: Emergence Response for Flooding Events. In *Proceedings of the MediaEval 2017 Workshop (Sept. 13-15, 2017). Dublin, Ireland*.

[25] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 401–408.

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

[27] Markus Brenner and Ebroul Izquierdo. 2011. MediaEval Benchmark: Social Event Detection in collaborative photo collections.. In *MediaEval*.

[28] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. 2015. Polyphonic sound event detection using multi label deep neural networks. In *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 1–7.

[29] Emre Cakir, Ezgi Can Ozan, and Tuomas Virtanen. 2016. Filterbank learning for deep neural network based polyphonic sound event detection. In *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 3399–3406.

[30] Liangliang Cao, Shih-Fu Chang, Noel Codella, Courtenay Cotton, Dan Ellis, Leiguang Gong, Matthew Hill, Gang Hua, John Kender, Michele Merler, et al. 2011. Ibm research and columbia university trecvid-2011 multimedia event detection (med) system. In *NIST TRECVID Workshop*, Vol. 28.

[31] Xiaojun Chang, Yao-Liang Yu, Yi Yang, and Eric P Xing. 2017. Semantic pooling for complex event analysis in untrimmed videos. *IEEE transactions on pattern analysis and machine intelligence* 39, 8 (2017), 1617–1632.

[32] S Chatzichristofis, Y Boutalis, and Mathias Lux. 2009. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Proceedings of the 6th IASTED International Conference*, Vol. 134643. 064.

[33] Savvas A Chatzichristofis and Yiannis S Boutalis. 2008. CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *International Conference on Computer Vision Systems*. Springer, 312–322.

[34] Ming-yu Chen and Alexander Hauptmann. 2009. Mosift: Recognizing human actions in surveillance videos. (2009).

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

[36] Inkyu Choi, Kisoo Kwon, Soo Hyun Bae, and Nam Soo Kim. 2016. DNN-based sound event detection with exemplar-based approach for noise reduction. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. 16–19.

[37] Selina Chu, Shrikanth Narayanan, C-C Jay Kuo, and Maja J Mataric. 2006. Where am I? Scene recognition for mobile robots using audio features. In *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 885–888.

[38] Courtenay V Cotton and Daniel PW Ellis. 2011. Spectral vs. spectro-temporal features for acoustic event detection. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 69–72.

[39] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. AutoAugment: Learning Augmentation Policies from Data. *arXiv preprint arXiv:1805.09501* (2018).

[40] Juncheng Li Dai Wei, Phuong Pham, Samarjit Das, Shuhui Qu, and Florian Metze. 2016. Sound event detection for real life audio DCASE challenge. In *Proceedings Workshop Detection and Classification of Acoustic Scenes and Events*.

[41] Minh-Son Dao, Duc-Tien Dang-Nguyen, and Francesco GB De Natale. 2014. Robust event discovery from photo collections using Signature Image Bases (SIBs). *Multimedia Tools and Applications* 70, 1 (2014), 25–53.

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

[43] Terrance DeVries and Graham W Taylor. 2017. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538* (2017).

[44] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48, 10 (2015), 2993–3003.

[45] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. 2015. 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*. 1–9.

[46] Lijie Fan, Wenbing Huang, Stefano Ermon Chuang Gan, Boqing Gong, and Junzhou Huang. 2018. End-to-End Learning of Motion Representation for Video Understanding. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition. 6016–6025.

[47] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. 2016. Deep representation for abnormal event detection in crowded scenes. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 591–595.

[48] Jonathan G Fiscus. 2010. TRECVID Multimedia Event Detection 2010 Evaluation. (2010).

[49] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* 65 (2015), 22–28.

[50] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2016. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems* 17, 1 (2016), 279–288.

[51] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 411–412.

[52] Alexandre RJ Francois, Ram Nevatia, Jerry Hobbs, Robert C Bolles, and John R Smith. 2005. VERL: an ontology framework for representing and annotating video events. *IEEE multimedia* 12, 4 (2005), 76–86.

[53] Steve Frolking, Jianjun Qiu, Stephen Boles, Xiangming Xiao, Jiyuan Liu, Yahui Zhuang, Changsheng Li, and Xiaoguang Qin. 2002. Combining remote sensing and ground census data to develop new maps of the distribution of rice agriculture in China. *Global Biogeochemical Cycles* 16, 4 (2002).

[54] Jianlong Fu, Yue Wu, Tao Mei, Jinqiao Wang, Hanqing Lu, and Yong Rui. 2015. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *Proceedings of the IEEE international conference on computer vision*. 1985–1993.

[55] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. 2016. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *European Conference on Computer Vision*. Springer, 849–866.

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

[57] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. 2016. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 923–932.

[58] Oguzhan Gencoglu, Tuomas Virtanen, and Heikki Huttunen. 2014. Recognition of acoustic events using deep neural networks. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 506–510.

[59] D Giannoulis, E Benetos, D Stowell, M Rossignol, M Lagrange, and M Plumbley. 2013. IEEE AASP challenge: Detection and classification of acoustic scenes and events. *Queen Mary University of London: London, UK* (2013).

[60] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[61] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

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

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

[64] Cong Guo, Xinmei Tian, and Tao Mei. 2017. Multi-granular Event Recognition of Personal Photo Albums. *IEEE Transactions on Multimedia* (2017).

[65] Aki Harma, Martin F McKinney, and Janto Skowronek. 2005. Automatic surveillance of the acoustic activity in our living environment. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 4–pp.

[66] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Takaaki Hori, Jonathan Le Roux, and Kazuya Takeda. 2016. Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. 35–39.

[67] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Takaaki Hori, Jonathan Le Roux, and Kazuya Takeda. 2017. BLSTM-HMM hybrid system combined with sound activity detection network for polyphonic sound event detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 766–770.

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

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

[70] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Moncef Gabbouj. 2013. Supervised model training for overlapping sound events based on unsupervised source separation.. In *ICASSP*. 8677–8681.

[71] Somboon Hongeng, Ram Nevatia, and Francois Bremond. 2004. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding* 96, 2 (2004), 129–162.

[72] Yuanbo Hou and Shengchen Li. 2017. *Sound event detection in real life audio using multimodel system.* Technical Report. DCASE2017 Challenge, Tech. Rep.

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

[74] I-Hong Jhuo and DT Lee. 2014. Video event detection via multi-modality deep learning. In *Pattern Recognition (ICPR), 2014 22nd International Conference on.* IEEE, 666–671.

[75] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia.* ACM, 675–678.

[76] Lu Jiang, Alexander G Hauptmann, and Guang Xiang. 2012. Leveraging high-level and low-level features for multimedia event detection. In *Proceedings of the 20th ACM international conference on Multimedia.* ACM, 449–458.

[77] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 2018. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE transactions on pattern analysis and machine intelligence* 40, 2 (2018), 352–364.

[78] Brendan Jou and Shih-Fu Chang. 2016. Deep cross residual learning for multitask visual recognition. In *Proceedings of the ACM Conference on Multimedia.* ACM, 998–1007.

[79] Andreas Kamilaris and Francesc X Prenafeta-Boldú. 2018. Disaster monitoring using unmanned aerial vehicles and deep learning. *arXiv preprint arXiv:1807.11805* (2018).

[80] Nogueira Keiller, Fadel Samuel, Dourado Ícaro, Werneck Rafael, Muñoz Javier, Penatti Otávio, and Calumby Rodrigo. [n. d.]. Data-Driven Flood Detection using Neural Networks. In *Proceedings of the MediaEval 2017 Workshop* (Sept. 13-15, 2017). Dublin, Ireland.

[81] Zvi Kons and Orith Toledo-Ronen. 2013. Audio event classification using deep neural networks.. In *Interspeech.* 1482–1486.

[82] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* 1097–1105.

[83] Julian Kürby, Rene Grzeszick, Axel Plinge, and Gernot A Fink. 2016. Bag-of-features acoustic event detection for sensor networks. In *Proceedings Workshop Detect. Classification Acoust. Scenes Events.* 55–59.

[84] Ying-Hui Lai, Chun-Hao Wang, Shi-Yan Hou, Bang-Yin Chen, Yu Tsao, and Yi-Wen Liu. 2016. DCASE report for task 3: Sound event detection in real life audio. *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events* (2016).

[85] Zhen-Zhong Lan, Lu Jiang, Shoou-I Yu, Shourabh Rawat, Yang Cai, Chenqiang Gao, Shicheng Xu, Haoquan Shen, Xuanchong Li, Yipei Wang, et al. 2013. Cmu-informedia at trecvid 2013 multimedia event detection. In *TRECVID 2013 Workshop*, Vol. 1. 5.

[86] Donmoon Lee, Subin Lee, Yoonchang Han, and Kyogu Lee. 2017. *Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input.* Technical Report. Tech. Rep., DCASE2017 Challenge.

[87] Li-Jia Li and Li Fei-Fei. 2007. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* IEEE, 1–8.

[88] Hyungui Lim, Jeongsoo Park, Kyogu Lee, and Yoonchang Han. [n. d.]. Rare Sound Event Detection Using 1D Convolutional Recurrent Neural Networks. ([n. d.]).

[89] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).

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

[91] Xueliang Liu and Benoit Huet. 2013. Heterogeneous features and model selection for event-based media classification. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval.* ACM, 151–158.

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

[93] Xiang Long, Chuang Gan, Gerard de Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. 2018. Multimodal keyless attention fusion for video classification. AAAI.

[94] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. 2018. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7834–7843.

[95] Laura Lopez-Fuentes, Joost van de Weijer, Marc Bolanos, and Harald Skinnemoen. 2017. Multi-modal deep learning approach for flood detection. In *Proc. of the MediaEval 2017 Workshop (Sept. 13–15, 2017). Dublin, Ireland.*

[96] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[97] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proc. of the 7th International Conference on Multimedia Systems*. ACM, 30.

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

[99] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

[100] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *Signal Processing Conference, 2010 18th European*. IEEE, 1267–1271.

[101] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. Metrics for polyphonic sound event detection. *Applied Sciences* 6, 6 (2016), 162.

[102] Pascal Mettes, Dennis C Koelma, and Cees GM Snoek. 2016. The imagenet shuffle: Reorganized pre-training for video event detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 175–182.

[103] Matthias Meyer, Lukas Cavigelli, and Lothar Thiele. 2017. Efficient Convolutional Neural Network For Audio Event Detection. *arXiv preprint arXiv:1709.09888* (2017).

[104] Dao Minh-Son, Pham Quang-Nhat-Minh, and Dang-Nguyen Duc-Tien. [n. d.]. A Domain-based Late-Fusion for Disaster Image Retrieval from Social Media. In *Proc. of the MediaEval Workshop* (Sept. 13-15, 2017). Dublin, Ireland.

[105] Hanif Muhammad, Atif Muhammad, Khan Mahrukh, and Rafi Mohammad. [n. d.]. Flood detection using Social Media Data and Spectral Regression based Kernel Discriminant Analysis. In *Proceedings of the MediaEval 2017 Workshop* (Sept. 13-15, 2017). Dublin, Ireland.

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

[107] Keiller Nogueira, Samuel G Fadel, Ícaro C Dourado, Rafael de O Werneck, Javier AV Muñoz, Otávio AB Penatti, Rodrigo T Calumby, Lin Tzy Li, Jefersson A dos Santos, and Ricardo da S Torres. 2017. Exploiting ConvNet Diversity for Flooding Identification. *arXiv preprint arXiv:1711.03564* (2017).

[108] Dan Oneata, Matthijs Douze, Jérôme Revaud, Schwenninger Jochen, Danila Potapov, Heng Wang, Zaid Harchaoui, Jakob Verbeek, Cordelia Schmid, Robin Aly, et al. 2012. Axes at trecvid 2012: Kis, ins, and med. In *TRECVID workshop*.

[109] Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. 2014. Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*. 52.

[110] Symeon Papadopoulos, Raphael Troncy, Vasileios Mezaris, Benoit Huet, and Ioannis Kompatsiaris. 2011. Social Event Detection at MediaEval 2011: Challenges, dataset and evaluation.. In *MediaEval*.

[111] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. 2016. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 6440–6444.

[112] Sungheon Park and Nojun Kwak. 2015. Cultural event recognition by subregion classification with convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 45–50.

[113] Georgios Petkos, Symeon Papadopoulos, Vasileios Mezaris, Raphael Troncy, Philipp Cimiano, Timo Reuter, and Yiannis Kompatsiaris. 2014. Social event detection at MediaEval: a three-year retrospect of tasks and results. In *Proc. of the International Conference on Multimedia Retrieval Workshop on Social Events in Web Multimedia (SEWM)*.

[114] Huy Phan, Lars Hertel, Marco Maass, and Alfred Mertins. 2016. Robust audio event recognition with 1-max pooling convolutional neural networks. *arXiv preprint arXiv:1604.06338* (2016).

[115] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 1015–1018.

[116] Axel Plinge, Rene Grzeszick, and Gernot A Fink. 2014. A bag-of-features approach to acoustic event detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 3704–3708.

[117] Samira Pouyanfar and Shu-Ching Chen. 2016. Semantic event detection using ensemble deep learning. In *Multimedia (ISM), 2016 IEEE International Symposium on*. IEEE, 203–208.

[118] Samira Pouyanfar and Shu-Ching Chen. 2017. Automatic video event detection for imbalance data using enhanced ensemble deep learning. *International Journal of Semantic Computing* 11, 01 (2017), 85–109.

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

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

[121] Jinyoung Rhee, Jungho Im, and Gregory J Carbone. 2010. Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data. *Remote Sensing of Environment* 114, 12 (2010), 2875–2887.

[122] Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. 2015. Sports videos in the wild (SVW): A video dataset for sports analysis. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1. IEEE, 1–7.

[123] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1041–1044.

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

[125] Emmanouil Schinas, Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2012. CERTH@ MediaEval 2012 Social Event Detection Task.. In *MediaEval*. Citeseer.

[126] Yuhui Shi and Russell C Eberhart. 1999. Empirical study of particle swarm optimization. In *Evolutionary computation, 1999. CEC 99. Proceedings of the 1999 congress on*, Vol. 3. IEEE, 1945–1950.

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

[128] Bharat Singh, Xintong Han, Zhe Wu, Vlad I Morariu, and Larry S Davis. 2015. Selecting relevant web trained concepts for automated event retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 4561–4569.

[129] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world Anomaly Detection in Surveillance Videos. *Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)* (2018).

[130] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. 2018. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1390–1399.

[131] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[132] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[133] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool. 2016. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv preprint arXiv:1604.07160* (2016).

[134] Naoya Takahashi, Michael Gygli, and Luc Van Gool. 2018. Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia* 20, 3 (2018), 513–524.

[135] Planet Team. 2016. Planet Application Program Interface: In Space for Life on Earth. San Francisco, CA. (2016).

[136] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[137] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. of the IEEE international conference on computer vision*. 4489–4497.

[138] Shen-Fu Tsai, Thomas S Huang, and Feng Tang. 2011. Album-based object-centric event recognition. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 1–6.

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

[140] Dmitrii Ubskii and Alexei Pugachev. 2016. Sound event detection in real-life audio. *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events* (2016).

[141] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision* 104, 2 (2013), 154–171.

[142] MWW Van Grootel, Tjeerd C Andringa, and JD Krijnders. 2009. DARES-G1: Database of annotated real-world everyday sounds. In *Proceedings of the NAG/DAGA International Conference on Acoustics*.

[143] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 3169–3176.

[144] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 3551–3558.

[145] Jun Wang and Jean-Daniel Zucker. 2000. Solving multiple-instance problem: A lazy learning approach. (2000).

[146] Limin Wang, Zhe Wang, Sheng Guo, and Yu Qiao. 2015. Better exploiting os-cnns for better event recognition in images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 45–52.

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

[148] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. [n. d.]. Non-local neural networks. ([n. d.]).

[149] Xiaolong Wang and Abhinav Gupta. 2018. Videos as Space-Time Region Graphs. *arXiv preprint arXiv:1806.01810* (2018).

[150] Xiaoyang Wang and Qiang Ji. 2015. Video event recognition with deep hierarchical context model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4418–4427.

[151] Yun Wang and Florian Metze. 2016. Recurrent Support Vector Machines for Audio-Based Multimedia Event Detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 265–269.

[152] Yun Wang and Florian Metze. 2017. A first attempt at polyphonic sound event detection using connectionist temporal classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2986–2990.

[153] Yun Wang, Leonardo Neves, and Florian Metze. 2016. Audio-based multimedia event detection using deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2742–2746.

[154] Xiu-Shen Wei, Bin-Bin Gao, and Jianxin Wu. 2015. Deep spatial pyramid ensemble for cultural event recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 38–44.

[155] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. 2016. Understanding data augmentation for classification: when to warp? *arXiv preprint arXiv:1609.08764* (2016).

[156] Zifeng Wu, Yongzhen Huang, and Liang Wang. 2015. Learning representative deep features for image set analysis. *IEEE Transactions on Multimedia* 17, 11 (2015), 1960–1968.

[157] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. 2015. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1600–1609.

[158] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553* (2015).

[159] Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2015. A discriminative CNN video representation for event detection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 1798–1807.

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

[161] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. 4507–4515.

[162] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 471–480.

[163] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. 2018. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision* 126, 2-4 (2018), 375–389.

[164] Litao Yu, Xiaoshuai Sun, and Zi Huang. 2016. Robust spatial-temporal deep model for multimedia event detection. *Neurocomputing* 213 (2016), 48–53.

[165] Shoou-I Yu, Lu Jiang, Zexi Mao, Xiaojun Chang, Xingzhong Du, Chuang Gan, Zhenzhong Lan, Zhongwen Xu, Xuanchong Li, Yang Cai, et al. 2014. Informedia@ trecvid 2014 med and mer. In *NIST TRECVID Video Retrieval Evaluation Workshop*, Vol. 24.

[166] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. 2015. Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 53–61.

[167] Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, and Ruslan Salakhutdinov. 2015. Exploiting image-trained CNN architectures for unconstrained video classification. *arXiv preprint arXiv:1503.04144* (2015).

[168] Dongqing Zhang and Dan Ellis. 2001. Detecting sound events in basketball video archive. *Dept. Electronic Eng., Columbia Univ., New York* (2001).

[169] Xishan Zhang, Hanwang Zhang, Yongdong Zhang, Yang Yang, Meng Wang, Huanbo Luan, Jintao Li, and Tat-Seng Chua. 2016. Deep fusion of multiple semantic cues for complex event recognition. *IEEE Transactions on Image Processing* 25, 3 (2016), 1033–1046.

[170] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[171] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision* 124, 3 (2017), 409–421.