Recognize Complex Events from Static Images by Fusing Deep Channels

Yuanjun Xiong[1] Kai Zhu[1] Dahua Lin[1] Xiaoou Tang[1,2]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Shenzhen key lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China
xy012@ie.cuhk.edu.hk zk013@ie.cuhk.edu.hk dhlin@ie.cuhk.edu.hk xtang@ie.cuhk.edu.hk

Abstract

A considerable portion of web images capture events that occur in our personal lives or social activities. In this paper, we aim to develop an effective method for recognizing events from such images. Despite the sheer amount of study on event recognition, most existing methods rely on videos and are not directly applicable to this task. Generally, events are complex phenomena that involve interactions among people and objects, and therefore analysis of event photos requires techniques that can go beyond recognizing individual objects and carry out joint reasoning based on evidences of multiple aspects. Inspired by the recent success of deep learning, we formulate a multi-layer framework to tackle this problem, which takes into account both visual appearance and the interactions among humans and objects, and combines them via semantic fusion. An important issue arising here is that humans and objects discovered by detectors are in the form of bounding boxes, and there is no straightforward way to represent their interactions and incorporate them with a deep network. We address this using a novel strategy that projects the detected instances onto multi-scale spatial maps. On a large dataset with $60,000$ *images, the proposed method achieved substantial improvement over the state-of-the-art, raising the accuracy of event recognition by over* 10%.

1. Introduction

The explosive growth of web images, driven primarily by the thriving of online photo sharing services such as Flickr and Instagram, has been gradually and profoundly transforming our lives and the way we communicate. Many of these images are event photos, namely the ones that capture human activities in either private or social contexts. Such images not only provide valuable records of our lives and our world, but also convey useful information that one can exploit to analyze consumer preferences or study socioeconomic trends. The primary goal of this paper is to develop an effective method for recognizing events from images.

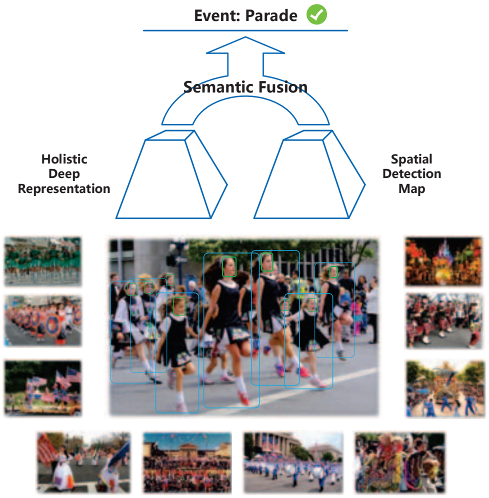


Figure 1: Event recognition is highly challenging due to the large semantic gap. Even in the same event class, *Parade*, the images can look very different. This calls for methods that are capable of reasoning about high-level semantics by fusing evidences of multiple aspects.

Event recognition is not a new story in computer vision. However, most existing efforts [1, 5, 32] are devoted to recognizing events from *videos*. This is not surprising, as it is a common conception that dynamics play a critical role in defining an event. *Do we really need videos for event recognition?* Our experience seems to suggest otherwise – people can effortlessly identify events from photos most of the time. This motivates us to explore a new approach, one that is able to recognize events from static images.

This is a challenging problem. A major obstacle standing in our way is the large gap between high-level event semantics and low-level visual features. Event images are complex as compared to object images. They usually involve multiple objects interacting with each other. As we can see in Figure 1, two images capturing the same kind of events can be vastly different in their visual structures. Traditional methods that rely mainly on shallow analysis of visual appearance would be faced with substantial difficulties when applied to this task.

Recently, the use of convolutional neural networks (CNN) has led to remarkable progress in several important vision tasks, including image classification [12], object detection [7], and face verification [31]. This line of work clearly demonstrates the superior capability of deep models in capturing complex variations and the critical role of intermediate layers in bridging the semantic gap. Following the lead of these efforts, we explore the use of deep learning in this work, with an aim to bring its success to the next level – from recognizing individual objects to understanding complex images as a whole.

Events, by nature, are defined by the interactions among key *entities*, including *humans* and *objects*. Therefore, identifying such entities in an image is a key step towards event understanding. While a convolutional network formulated upon entire images is very powerful in modelling visual appearance, we found empirically that it is not as effective as an dedicated detector, especially in detecting humans. Our idea to tackle this problem is very simple – use dedicated detectors to locate relevant entities and incorporate them with the convolutional network to predict the event class.

However, bounding boxes of detected objects and visual appearance features are very different by nature, and can not be combined using conventional feature combination methods. In this paper, we propose a novel way to address this. Instead of directly using the bounding boxes, we project them onto multi-scale spatial maps, bring the resultant maps together, and thereon construct a convolutional network to derive a higher-level representation. This construction not only provides a way to express detecting results that is suited for higher-level analysis, but also makes it possible to exploit the spatial co-occurrences of different objects, which are important cues of their interactions. With two convolutional networks, one upon the image and the other upon the detection maps, we integrate them via *semantic fusion* and obtain a fused representation that captures key semantic elements of the event image.

The major contributions of this work are summarized here: (1) We explore a new approach to event recognition, which, unlike most previous methods, rely solely on static images. (2) Recognizing that interactions among people and objects are essential for event understanding, we propose using dedicated detectors to locate key entities, and develop a novel strategy, namely multi-scale spatial maps, to uniformly represent the detected results. (3) We propose a new framework that combines evidences from multiple channels via semantic fusion. (4) To facilitate this study and to promote future efforts towards image-based event recognition, we construct a large dataset comprised of nearly $60,000$ images annotated with event classes. The dataset can be found in the project website listed in the supplementary materials.

The rest of the paper is organized as follows. Section 2 provides a brief review of related work. Section 3 introduces a new dataset for image-based event recognition, called WIDER. Section 4 discusses the proposed framework in detail. Section 5 presents the experimental results. Finally, we conclude this paper in Section 6.

2. Related Work

Event recognition is a very active area in computer vision [39]. Most existing methods rely on videos to recognize events [4], with emphasis placed on the use of dynamics and temporal relations [1, 32]. These methods generally fall in three categories: *feature-based* [4, 27, 34], *concept-based* [35], and *model-based* [9, 37]. Recently, Duan *et al*. proposed a new method [5] that utilizes web images to help video-based event recognition. Despite the technical differences among these methodologies, they all rely heavily on using the dynamics extracted from videos and therefore can not be directly applied to static images.

Understanding of still images is an active field of research. Efforts on holistic scene understanding [26] are also related to this work, as they both target high-level interpretations of given images. Yet, essential differences exist. Prior work on scene understanding mainly considers visual patterns, with relatively less attention to human activities, which, however, are a key factor in event analysis. In this paper, we take into account this factor through a dedicated channel and derive a novel way, namely multi-scale maps, to incorporate it.

Analyzing human actions [10] with the help of human poses [21, 38] and human-object interactions [2, 40] also provides significant cues in recognizing certain categories of events. However it is worth to note that the events we investigate here are usually characterized by combinations of multiple aspects, including background appearance, spatial patterns of people, and their interactions. Hence this work is related but different from recognizing actions of individuals.

Among previous work on image understanding, a CRF-based method proposed by Li and Fei-fei [15] that jointly infers the classes of event, scene, and objects is perhaps the most related. This method couples two LDA models formulated directly upon low-level features, and therefore lacks the capability of capturing complex variations and bridging the semantic gap. It is also worth noting that many previous methods [16] require training sets with detailed annotations (*e.g*. object bounding boxes), which are often very costly to obtain. On the contrary, our method only needs training images labeled with event classes, making it particularly ap-

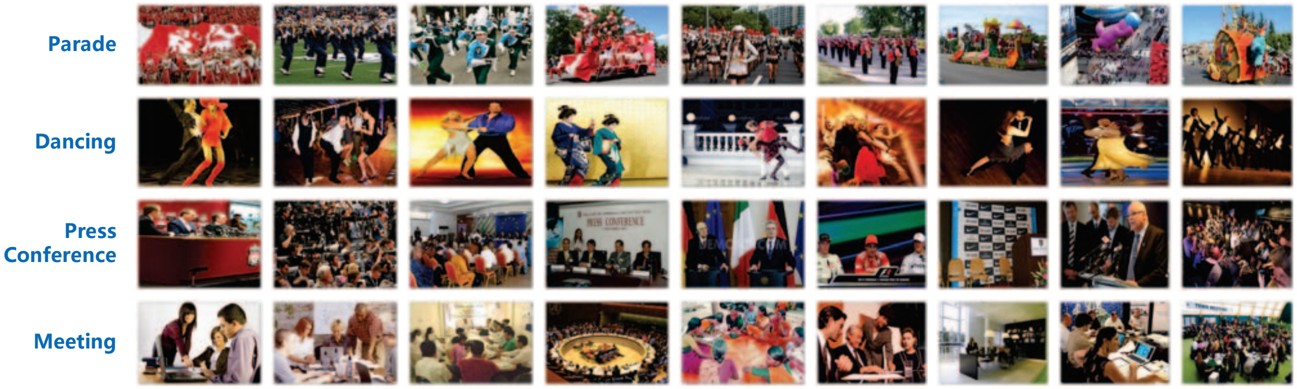


Figure 2: Examples of several categories in the WIDER dataset, which exhibit diverse visual patterns.

pealing to large-scale applications.

A key strategy adopted in this work is to combine information from multiple channels. This strategy has been widely used in previous work. In conventional frameworks, the fusion of channels is usually accomplished by combining features [18] or optimization objectives [19]. A limitation of these approaches is that they are not able to exploit the relations among the constituent elements of different channels. Following the recent success of deep models [6, 12, 20], attempts [23, 30] have been made to connect multiple modalities through deep networks. In recent work, auxiliary channels, such as depth [8] and optical flow [29], are captured using additional networks. It is worth emphasizing that depth maps or optical flows are both spatial maps by nature and thus it is relatively easy to construct CNNs thereon. However, incorporating external detectors that produce bounding boxes is not as straightforward. In this work, we develop a novel method, namely the multi-scale maps, which provides a principled solution to this problem. This method enables us to directly draw on state-of-the-art detectors [3, 7, 14] for improving the overall recognition performance.

3. WIDER: A New Dataset

Datasets are an important force in driving the advancement in a research area. Whereas there have been plenty of datasets for object recognition [28], scene understanding [36], and video-based event recognition [25]. A dataset to support the research on image-based recognition remains needed. Along with this work, we constructed a large dataset from web images, called *Web Image Dataset for Event Recognition (WIDER)*. This dataset contains $60,000$ images of 60 event classes, where the numbers of images in different classes are balanced. All images have been carefully annotated with event labels, which can be used for model training and performance evaluation. Figure 1 show some examples of the data. We can see that the dataset comprises a diverse set of event categories and there exist substantial variations in visual patterns among the images within each category. We will make the dataset available to the public following the publication of this paper in order to foster future research on this topic.

Construction of this dataset took a lot of efforts. This course is comprised of three stages:

Selecting event categories. A majority of the event categories are from the *Large Scale Ontology for Multimedia (LSCOM)* [22], which provides a list of around $1,000$ concepts relevant to video event analysis. Many of these concepts are the names of objects or low-level actions. Hence, we manually go through the list, picking those representing event classes while filtering out the others. We also noticed that the concepts in *LSCOM* are primarily from TV news, and consequently events in personal lives were not thoroughly covered. To enrich the dataset, we invited a group of students to propose activities related to their daily lives and find a number of new categories therefrom, *e.g. car-driving*. Altogether, we obtained 60 event classes.

Collecting images. We resorted to search engines like *Google* and *Bing* to collect images. Specifically, we retrieve 1000 to 3000 images for each category using the class name as the input query. We found that many images resulted from this process are simply irrelevant. To obtain more qualified images, we adopt the *query expansion* strategy. In particular, we acquire additional queries for each event class by finding highly frequent phrases from a variety of sources, such as WordNet, Wikipedia, and the text snippets that come with the retrieved images. Using these phrases as queries to expand the search substantially enrich the pool of candidate images for building the dataset.

Screening data. The collection process above results in hundreds of thousands of candidate images. In this pool, lots of samples are cartoons or cliparts while many others are irrelevant to the events of interest. To clean the data, we first filter out cartoons, cliparts, and blank images using bilateral filtering[1]. Then we asked human annotators to identify irrelevant images in the remaining set. To expedite this process, we developed a GUI tool, where the images are grouped into pages and hence the annotator can inspect 80 images at the same time. In this way, we can process a large quantity of images very quickly and reliably. The screening retained about 60, 000 images in the dataset.

4. Framework: Fusing Multiple Channels

Generally, an event can be considered as an activity taking place in a certain environment. Hence, it can be reasoned from two aspects: (1) Environment: *e.g. is it by the seashore or in a forest? is there a large crowd of people?* (2) Activity: *e.g. is the man running? are the people in the scene sitting together?* Event recognition, in essense, is a process to answer such questions and arrive at a prediction by bringing the answers together.

Following this consideration, we develop a multi-layer framework as shown in Figure 3. This framework is comprised of two major channels, one is to model the observed visual patterns, which are important for reasoning about the environment; while the other is to capture the interactions among humans and objects, which are significant cues of the activity taking place. Particularly, to ensure the reliability of detection, the latter channel employs state-of-the-art detectors to locate the entities of interest (*i.e.* humans and objects), and subsequently uses spatial maps to express the distribution of the detected results. This enables the use of deep models to capture the variations in their spatial configurations. These two channels are combined through a semantic fusing layer, resulting in a fused representation that captures the key semantic elements of the image. In what follows, we will introduce these components in detail.

4.1. Model Visual Appearance with CNN

We use a deep convolutional neural network (CNN) to model the visual appearance of event images. In previous work [12], CNNs have demonstrated excellent capability of capturing complex variations in visual patterns. Here, we are interested in studying how well they perform in higher-level tasks, *e.g.* event recognition. Particularly, we adopt the architecture of AlexNet presented in [12].

This network comprises eight layers, five convolutional and three fully-connected, and takes as input a 3-channel color image of size 224 × 224. The 1st, 2nd, and 5th convolutional layers are each followed by a max-pooling layer to compress the inputs. Each fully connected layer has 4096 neurons. The last layer is linked to a multi-way softmax classifier with dense connections. The settings of these layers follow [12]. The detailed model specification will be provided in the supplemental materials.

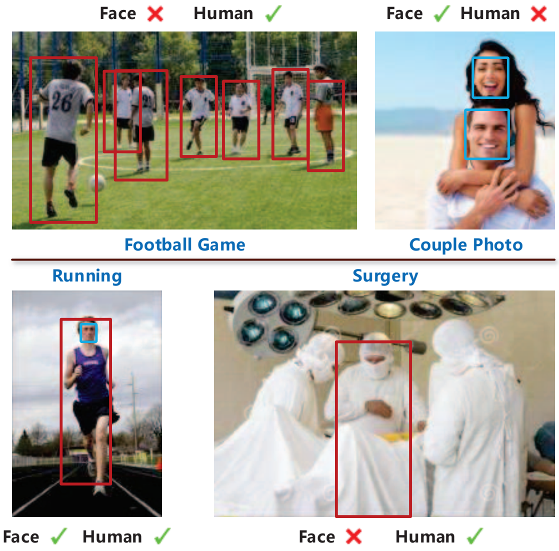


Figure 4: The face and human detectors are complementary. In case one detector fails, the other tends to find out the missed humans in image.

4.2. Find Humans with Complementary Detectors

We found empirically that humans appear in a majority of images in our dataset. This is not surprising. The interactions among humans are often a key factor in defining an event. However, locating humans from event images is very challenging. In such images, people are often occluded by one another, and their facial appearance can be seriously blurred when they are far away from the camera. There are also cases where faces of some people are completely invisible, as they are facing towards the opposite side. To tackle this problem, we combine two complementary techniques: *face detection* and *human detection*. As Figure 4 illustrates, this strategy can substantially increase the chance of successful detection even under adverse circumstances – when one technique fails, the other can come to rescue.

Specifically, we use the SURF cascade presented in [14] for face detection. This method uses multi-dimensional SURF features for describing local patches together with an improved weak classifier for boosting, thus significantly increasing the run-time efficiency without compromising the accuracy. For human detection, we employ the ACF detector developed in [3], which uses a feature pyramid for multi-scale detection with an approximation to speed up the computation. Both detectors are highly efficient and thus are suited for large-scale applications.

[1] The overall response of a bilateral filter can be used to test whether an image has enough textures to be qualified as a real-world photo.

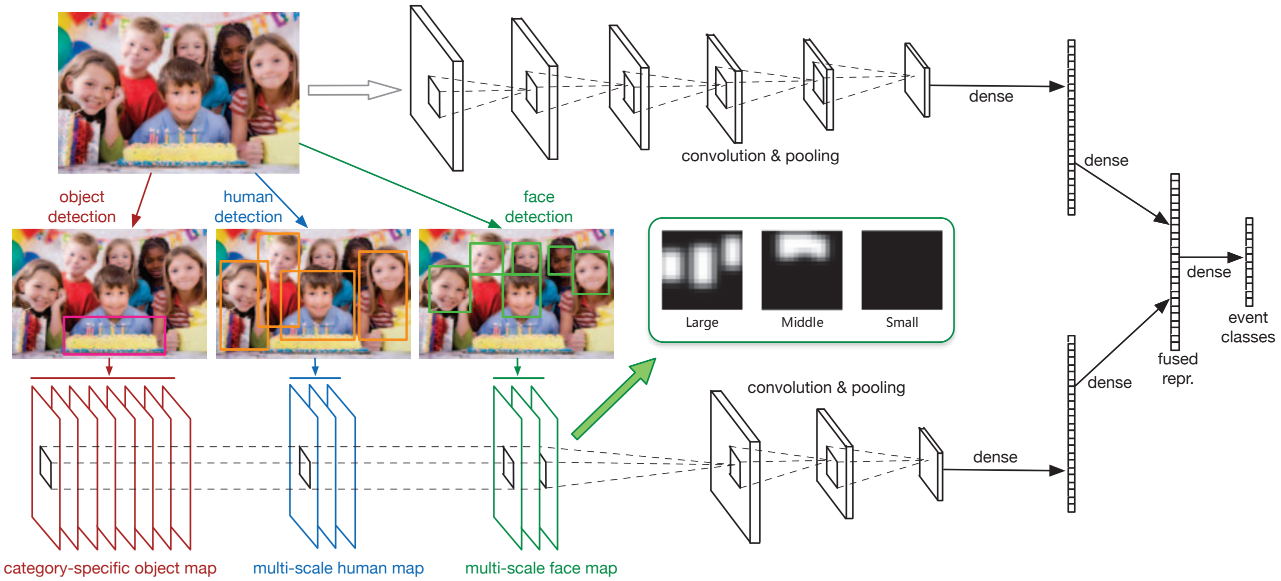


Figure 3: Overall, this framework integrates two channels. The upper channel, devised to capture the visual appearance, is formulated directly upon the input images; while the lower channel, devised to capture the interactions among humans and objects, takes as input the results of three detectors, respectively for faces, humans, and objects. In this channel, the bounding boxes obtained by the detectors are projected onto multi-scale spatial maps, which are then modeled by another CNN. On top of both CNNs, a fused representation is introduced, which is linked to the top representations of both networks, respectively via a fully-connected layer.

4.3. Multi-scale Spatial Maps

Detectors output bounding boxes. Each bounding box is represented by a 4-tuple comprised of the corner coordinates. These boxes contain rich information about the event, which, for example, include the spatial distribution of entities and their geometric relations *i.e*. relative location and size. However, a question arises here: *how can CNNs understand the bounding boxes?* This problem is not as trivial as it seems to be. Simply concatenating the coordinates of all bounding boxes does not yield a sensible representation.

Our idea to tackle this problem is simple. Since the primary message conveyed by these bounding boxes is the spatial configuration of the entities, to get this message, we can project the boxes onto a spatial map. Here, a *spatial map* is a binary image with the elements covered by detected objects set to one. However, there is an issue with this approach. Consider the two images in Figure 5, one containing a group of people, while the other containing two larger faces that cover a similar region. While these images represent very different events, one cannot distinguish them by inspecting their spatial maps.

Here, we propose a solution – *multi-scale spatial maps*. Instead of using a single channel to capture all detected entities, we expand the map into multiple channels, each for a scale level, so that entities of different scales will be reflected by different channels. In particular, we use a multi-scale spatial map comprised of three scale channels to represent the detected faces, where each channel is a binary map of size 18×18. We use two scale thresholds τ_1 and τ_2 with $\tau_1 < \tau_2$ to determine the choice of channels. Given a bounding box, we normalize its coordinates w.r.t. the 18×18 frame and compute its area a. If $a < \tau_1$, we project it to the 1st channel, setting all the covered elements of this channel to one. Otherwise, we project it to the 2nd or 3rd channel, depending on whether $a < \tau_2$ holds.

Likewise, we can apply this multi-scale representation to express the results obtained from the human detector. Altogether, we have a spatial map with 6 channels, 3 for faces and the other 3 for human bodies. This method not only provides a uniform representation that can be readily handled by higher level models, *e.g*. CNN, but also makes it possible to differentiate the spatial configurations at different scales, *e.g*. crowded gathering *vs.* private conversation.

4.4. Detect and Characterize Objects

Besides humans, the presence of objects of certain categories is often a strong indicator of some event classes. Figure 6 uses several examples to illustrate this relation. In this paper, we use R-CNN [7], a state-of-the-art technique in object detection, to locate objects of interest.

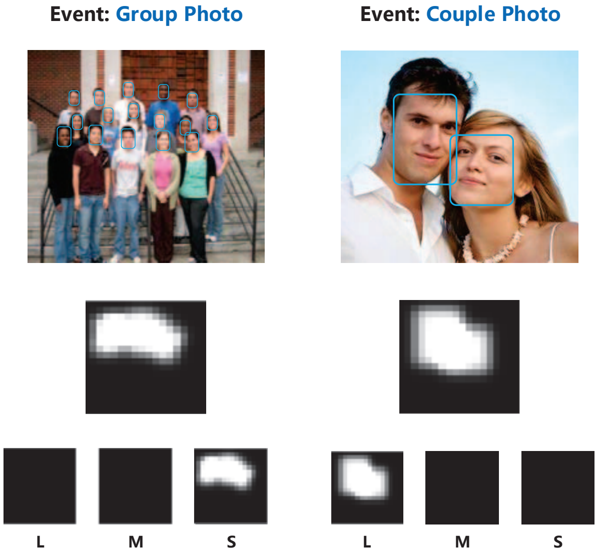


Figure 5: Here is an illustration of multi-scale spatial maps. Over these two images, the face detector produces bounding boxes of different sizes. Spatial maps resulted from the projection of these boxes are difficult to be distinguished from each other. However, when boxes of different sizes are projected onto different channels (L, M, and S), the distinction between these maps becomes much more obvious.

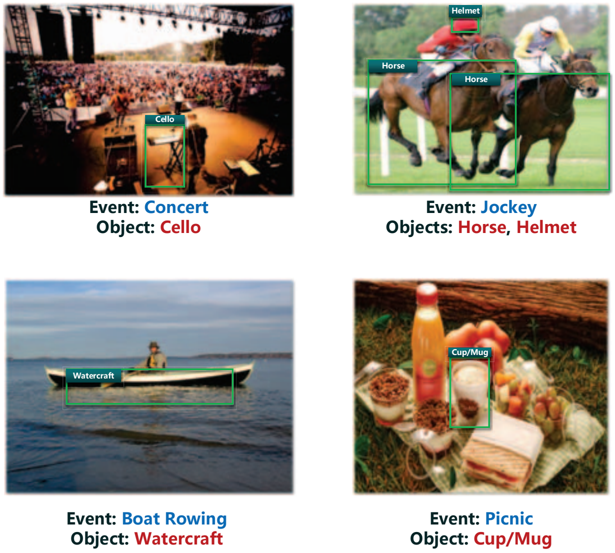


Figure 6: Existence of significant objects indicates the event categories. For example, the presence of horses and helmets is a strong indicator to the class *Jockey*.

The R-CNN method consists of three steps. First, it requires object candidates to be generated. For this purpose, we use a latest technique, called *Edge boxes* [41], which is much more efficient than the standard selective search algorithm [33]. On average, *Edge boxes* takes 0.25 second to process an image, while selective search takes about 10 seconds. Subsequently, a 4096-dimension CNN feature is derived for each candidate, which is then fed to SVMs to predict whether it belongs to specific object classes or not. Finally, a greedy non-maximum suppression procedure is applied to filter out redundant candidates.

We observed objects of thousands of different classes in our dataset. Many of them, however, are irrelevant to event understanding. To choose the ones that are truly pertinent to our task, we run a large collection of object detectors over a subset of event images, and select the 30 most frequently occurring classes[2].

Again, we use spatial maps to express detected objects. Unlike humans, we have a number of object classes but the presence of a specific object class is generally quite sparse. Hence, we use *class-specific maps* instead of *multi-scale maps* for general objects (except humans). In particular, we construct a spatial map with 30 channels, each for an object class. When an object is detected, the bounding box will be projected onto the corresponding channel. This representation enables one to exploit the interactions among objects, *e.g.* co-occurrences of objects of different categories.

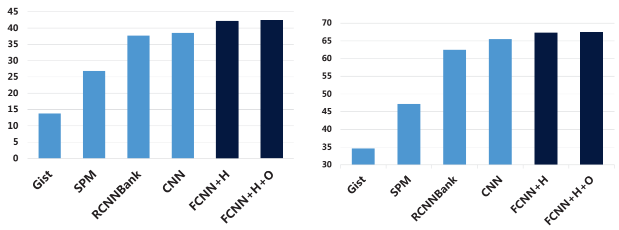
4.5. Channel Fusion

Stacking the spatial maps for faces, humans, and objects, we obtain an integrated spatial map with 36 channels, each of size 18×18. We construct a convolutional network thereon to derive a higher-level representation. Through a series of empirical experiments, we obtain an architecture suitable for modelling such spatial maps. This architecture comprises two convolutional layers. The first layer filters the inputs with 64 kernels, each of size $3 \times 3 \times 36$, producing an output of size $18 \times 18 \times 64$. This is followed by a max-pooling layer that compresses the result into an array of size $6 \times 6 \times 64$. The second convolutional layer, with 128 kernels of size $1 \times 1 \times 64$ is then applied, yielding an output of size $6 \times 6 \times 128$. Here, the first convolutional layer is to exploit the spatial interactions among neighboring parts and the co-occurrence patterns of different entities, while the second layer is mainly to adjust the relative contribution of different channels. The output of the second layer is then linked to a representation layer via a fully-connected network, resulting in a 4096-dimensional vector to capture the information derived from the detectors. Note that the detection channel needs less layers compared to the network for visual appearance. This is partly due to the reason that the detectors perform a series of visual analysis internally, which already narrows the semantic gap to some extent.

[2] The number of object classes was determined using cross validation. We found the risk of overfitting to be higher as we use more object classes.

Method	Top-1 Accuracy	Top-5 Accuracy
Gist [24]	13.8%	34.6%
SPM [13]	26.8%	47.2%
RCNNBank	37.7%	62.5%
CNN [12]	38.5%	65.5%
FCNN+H	42.1%	67.3%
FCNN+H+O	**42.4%**	**67.5%**

Table 1: Class averaged recognition accuracy.



(a) Top-1 Accuracy

(b) Top-5 Accuracy

Figure 7: Average recognition accuracy by percentages.

The visual appearance channel and the detection channel respectively yield a 4096-dimensional representation at the top. Through the computation across multiple layers, these representations are abstracted away from the low-level variations and thus are more consistent in expressing the semantics. To integrate both aspects, we further introduce a *semantic fusion layer*, which is linked to the top layers of both channels via dense connections, and thereon derive a 4096-dimensional *fused representation*. Like in other discriminative networks, this fused representation will be linked directly to the event classes via a softmax layer.

4.6. Training Algorithms

At the training stage, the CNN of the first channel was pre-trained on ImageNet [12]. This relives the over-fitting problem of deep models in the sense that natural images share similar low-level features. For the CNN of the second channel, the weights were randomly initialized from a zero-mean normal distribution. After initialization, the entire framework is jointly trained using stochastic gradient descent. Training strategy like data augmentation, weight-decay, and dropout are also used to alleviate over-fitting. The learning rate is initialized at 0.001 for the pre-trained CNN, while the learning rates of the two convolutional layers for the second channel are set to 5 and 2 times the base rate.

5. Experimental Results

We conducted experiments on the WIDER dataset (described in Section 3) to evaluate the proposed method and compare it with representative methods on image classification. The entire dataset, which contains $60,000$ images in 60 classes, is randomly divided into two disjoint halves, one for training and the other for testing.

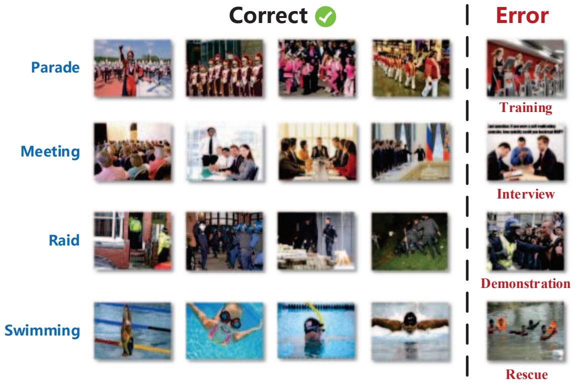


Figure 8: Successful and failed prediction examples on the testing set. Misclassified samples are shown with their ground-truth categories.

We tested our method under two settings: *"FCNN+H"* and *"FCNN+H+O"*. The former is a simplified version where the detection channel only uses the results of face & human detection, while the latter is the full version with both humans and objects taken into account. We also compared it with *"Gist"* [24], *"Spatial Pyramid Matching (SPM)"* [13], *"ObjectBank"* [17], and *"CNN"* [12]. These methods have been widely adopted in practical systems. Note that when we implemented *ObjectBank*, we made an important improvement, using the responses of R-CNN instead of the original SVM detectors. This change, which we call *"RCNNBank"*, leads to much better performance.

For all these methods, we learned the model on the training set and assessed them on the testing set. The performance was evaluated in terms of *top-1* and *top-5* accuracies. Specifically, each method was used to predict a ranked list of class labels for each testing image based on classification scores, which is then compared with the ground-truth. If the ground-truth is within top k positions of the list, we call the prediction *top-k accurate*. Then, *top-k accuracy* is defined to be the fraction of top-k accurate predictions.

Comparison of results. The performance is compared in Table 1 and Figure 7. The results show that methods using deep learning techniques outperform all others, *i.e*. Gist and SPM, by a large margin. This, again, demonstrates the superior capability of deep models in capturing complex visual variations as compared to traditional techniques. More importantly, our framework, with the detection channel incorporated, takes this capability to a next level, significantly improving the classification accuracies. Compared to CNN, the top-1 accuracy increases from 0.385 to 0.424 – the gain is over 10%. This result corroborates with our intuition that

	Para.	Hand.	Demo	Riot	Daci.	Acci.	Fune.	Chee.	Elec.	Pres.	Marc.	Meet.	Grou.	Inte.	Traf.	Stoc.	awar.	Cere.	Conc.	Coup.
CNN	.389	.270	**.641**	.117	.217	.584	.242	**.252**	.042	**.579**	.169	.406	**.274**	.134	.494	**.474**	.200	.254	.207	.193
FCNN+H+O	**.425**	**.279**	.617	**.182**	**.257**	**.663**	**.315**	.240	**.124**	.522	**.328**	**.454**	.258	**.187**	**.529**	.471	**.306**	**.336**	**.316**	**.386**

	Fami.	Fest.	Picn.	Shop.	Firi.	Patr.	Dril.	Spa	Fans	Stud.	Surg.	Wait.	Labor	Runn.	Base.	Bask.	Foot.	Socc.	Tenn.	IceS.
CNN	.176	.246	.576	.306	**.426**	.385	**.186**	.588	.120	**.132**	.416	**.584**	.239	**.582**	**.576**	.556	.488	.533	.611	.493
FCNN+H+O	**.318**	**.254**	**.593**	**.336**	.306	**.492**	.175	**.603**	**.183**	.132	**.462**	.543	**.260**	.480	.553	**.592**	**.511**	**.570**	**.636**	**.584**

	Gymn.	Swim.	Race.	Rowi.	Aero.	Ball.	Jock.	Bull.	Para.	Gree.	Cele.	Wear.	Phot.	Raid	Resc.	Trai.	Voti.	Fish.	Hock.	Driv.
CNN	**.667**	**.652**	.764	**.680**	**.420**	.293	.195	.712	.372	.141	.319	**.624**	.246	.219	.237	**.092**	**.158**	**.409**	.527	**.339**
FCNN+H+O	.641	.652	**.767**	.663	.389	**.319**	**.214**	**.761**	**.417**	**.160**	**.370**	.593	**.290**	**.253**	**.300**	.069	.124	.393	**.580**	.321

Table 2: Comparison of per class recognition accuracy. To save space, we only show abbreviations of category names here. We compare the accuracy of FCNN with the original fine-tuned CNN on these categories. With the help of spatial detection maps, accuracies on 40 out of 60 categories have been improved.

Method	CNN	FCNN+H	FCNN+H+O
Top-1 Accuracy	45.56%	48.9%	49.6%
Top-5 Accuracy	71.4%	73.6%	75.3%

Table 3: Performance comparison on the "with-object" set.

the detection channel conveys complementary information and that the multi-scale maps provide an effective means to utilize such information. Also, the use of face and human detectors makes up for the weakness of appearance-based CNN in object localization.

Table 2 offers class-specific comparisons. For 40 out of 60 classes, our method outperforms CNN [12]. For those classes where humans play a crucial role, the gain is remarkable. For example, the top-1 accuracies are nearly doubled for classes like *"Marching (Marc.)"* and *"Couple Photo(Coup.)"*. Figure 8 presents some successful and failed predictions of our model. Taking a closer look here, we can see that this model is able to identify images relevant to the same event in spite of the large variations in their visual appearance. On the other hand, many of the examples that are incorrectly classified tend to be easily confused, as the "true" classes and the predicted classes of these examples often look very similarly.

Contribution of object detection. Compared to the significant improvement due to the use of face and human detection, the performance gain brought by the object channels doesn't seem to be as notable. When investigating this issue, we found that non-human objects are only detected in about one-fourth of the images. Particularly, out of all the testing images, about 7100 contain detected non-human objects. We specifically evaluated the performance on this subset, and observed greater performance gain due to the object channels, as shown in Table 3. We note that the effectiveness of the object channels hinges largely on the performance of the object detectors. While the R-CNN detectors [7] already represent the state-of-the-art, the overall performance remains quite limited (with AP at 31.4%). However, the computer vision community is making steady progress in object detection [28]. It is reasonable to believe that with better detectors, we can see even greater improvement with the use of object channels.

Run-time performance. We implemented the framework based on Caffe [11], a popular programming platform for deep learning. The training phase involves preprocessing (detecting humans and objects) and parameter learning. A majority of the computation is performed on GPU. With a GTX Titan, it takes about 3 seconds on average to preprocess an image, and 3 hours to train the deep networks over the entire training set with about $30,000$ images. Given a new image, it also takes about 3 seconds to preprocess. Compared to preprocessing, the time needed to make the prediction is negligible (about 2.4 milliseconds per image).

6. Conclusions and Future Work

We presented a new framework for recognizing complex events from static images. This framework integrates evidences from a visual appearance channel and a detection channel, both via deep convolutional networks, to predict the event class for a given image. It is particularly worth noting that we use multi-scale spatial maps in expressing the results obtained from dedicated detectors, thus enabling the use of higher-level models, *e.g.* CNN, to capture the spatial configurations of objects and their variations.

The experiments over a large dataset clearly demonstrated the effectiveness of the proposed method. In particular, our method achieves notable improvements over state-of-the-art visual recognition techniques, increasing the accuracy by over 10%. Event recognition is a challenging task. While we have taken one step forward here, there remains much room for further improvement. We plan to explore new aspects in our future work, which include attributes of individuals, detailed characterization of interactions, and even the context. We wish that this work along with the WIDER dataset can promote the research on this topic.

References

[1] W. Brendel, A. Fern, and S. Todorovic. Probabilistic event logic for interval-based event recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3329–3336, June 2011.

[2] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *Computer Vision–ECCV 2012*, pages 158–172. Springer, 2012.

[3] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1532–1545, Aug 2014.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[5] L. Duan, D. Xu, I.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1667–1680, Sept 2012.

[6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587, June 2014.

[8] S. Gupta, R. Girshick, P. Arbelez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision ECCV 2014*, volume 8695 of *Lecture Notes in Computer Science*, pages 345–360. Springer International Publishing, 2014.

[9] C.-L. Huang, H.-C. Shih, and C.-Y. Chao. Semantic analysis of soccer video using dynamic bayesian network. *Multimedia, IEEE Transactions on*, 8(4):749–760, 2006.

[10] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[14] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3468–3475. IEEE, 2013.

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

[16] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043, June 2009.

[17] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.

[18] M. Li, X.-B. Xue, and Z.-H. Zhou. Exploiting multi-modal interactions: A unified framework. In *IJCAI*, pages 1120–1125, 2009.

[19] W. Li, L. Niu, and D. Xu. Exploiting privileged information from web data for image categorization. In *Computer Vision ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 437–452. Springer International Publishing, 2014.

[20] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2648–2655. IEEE, 2013.

[21] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3177–3184. IEEE, 2011.

[22] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *MultiMedia, IEEE*, 13(3):86–91, 2006.

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

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

[25] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[26] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1307–1314, Nov 2011.

[27] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Computer Vision–ECCV 2010*, pages 577–590. Springer, 2010.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[29] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[30] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems 25*, pages 2222–2230. Curran Associates, Inc., 2012.

[31] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *CoRR*, abs/1406.4773, 2014.

[32] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1250–1257, June 2012.

[33] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[34] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, 2009.

[35] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2665–2672. IEEE, 2014.

[36] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3485–3492, June 2010.

[37] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[38] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2030–2037. IEEE, 2010.

[39] Y. Yang and M. Shah. Complex events detection using data-driven concepts. In *Computer Vision ECCV 2012*, volume 7574, pages 722–735. Springer Berlin Heidelberg, 2012.

[40] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1691–1703, 2012.

[41] C. Zitnick and P. Dollr. Edge boxes: Locating object proposals from edges. In *Computer Vision ECCV 2014*, volume 8693, pages 391–405. Springer International Publishing, 2014.