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

Web images, which capture events that occur in our personal lives or social activities are booming thanks to the online photo sharing services. These images are a precious source for analyzing human life and convey useful information for research and industry. Despite the sheer amount of study on event recognition, most existing methods rely on videos and are not directly applicable to this task. For still images, although a few attempts have been made to analyze simple events like human actions, we rarely see effort in understanding the complex event in web images. The major obstacle standing in our way is the large gap between high-level event semantics and traditional low-level visual features. Generally, complex events 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. The difficulty of event recognition can be seen from Fig. 1. Two images capturing the same kind of events can be vastly different in their visual structures.

In this paper, we aim to develop an effective method for recognizing events from static images. Inspired by the recent success of deep learning, we formulate a multi-layer framework to tackle this problem. One key step towards the understanding is to identify the entities within these images. The natural way of representing them is using detection bounding boxes. 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. We propose a novel way to incorporate them with a deep network. Instead of directly using the bounding boxes' coordinates, 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.

Overall, this framework integrates two channels of convolutional networks and forms a directed acyclic graph deep model. The first channel is devised to capture the visual appearance, while the second channel is devised to capture the interactions among humans and objects. The second channel 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 takes into account both visual appearance and the interactions among humans and objects, and combines them via semantic fusion.

Datasets are an important force in driving the advancement in a research area. A large dataset appears absent for supporting research on still image event recognition. 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 a diverse set 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. To construct this dataset, we use different approaches to stem event classes and query raw images from search Engines. Automatic filtering and manual check are adopted for screening the query results to make the databset clean and precise. As a result, images in this dataset demonstrate substantial variations in visual patterns among the images within each category. The dateset is made public along with the paper to foster future research on this topic.

We conducted experiments on the WIDER dataset to evaluate the proposed method and compare it with representative methods on image classification. The proposed method achieved the top-1 accuracy of 42.4%. This substantial improvement over the state-of-the-art raises the accuracy of event recognition by over 10%. The implementation of the model is detailed in the paper and will be public together with the dataset.

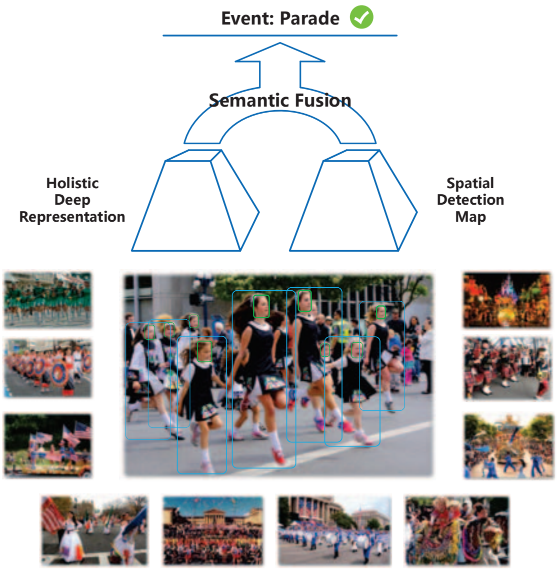

Figure 1: Event recognition is highly challenging due to the large semantic gap. Even in the same event class, *Parade*, the images can look very different.

As a conclusion, we demonstrated the potential of deep learning techniques in understanding the complex events from static images. The effectiveness of multi-scale spatial map is well examined by experiments. 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.

Method	Top-1 Accuracy	Top-5 Accuracy
Gist [3]	13.8%	34.6%
SPM [2]	26.8%	47.2%
RCNNBank	37.7%	62.5%
CNN [1]	38.5%	65.5%
FCNN+H	42.1%	67.3%
FCNN+H+O	**42.4%**	**67.5%**

Table 1: Class averaged recognition accuracy.

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

[2] Svetlana Lazebnik, Cordelia Schmid, and Jean 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.

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.