# Recognize Complex Events from Static Images by Fusing Deep Channels Supplementary Materials

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

## 1. Model Specification

Our model is implemented on Caffe [1]. It is a common programming framework for deep learning. The detailed specification of the model can be downloaded from the project website http://personal.ie.cuhk.edu.hk/~xy012/event_recog. It can be read by Caffe and edited with any text editor.

## 2. WIDER Dataset

The dataset can be downloaded from the project website http://personal.ie.cuhk.edu.hk/~xy012/event_recog/WIDER.

## 3. Parameters of Compared Methods

We compared our method with three different methods. The detailed settings of their parameters are described below.

**Gist** We use the implementation described in [4]. Images are resized to $128 \times 128$. The orientation scales are $(8, 8, 8, 8)$.

**Spatial Pyramid Matching (SPM)** The implementation of spatial pyramid matching algorithm is based on [2]. We use pyramids of three levels. The low-level visual features are characterized by SIFT descriptors. These features are encoded with a codebook of 1500 visual terms.

**RCNNBank** The RCNNBank is implemented based on ObjectBank [3], except that the detector responses are replaced by the activation features obtained using CNN. For each image, we first obtain 500 bounding boxes with highest proposal scores and apply the CNN to derive action features. This results in 500 activation feature vectors of 4096 dimensions. Then we splice the image into evenly spaced regions of size $3 \times 3$. The activation features are accumulated to the corresponding regions according to the positions of the bounding boxs. Finally, we get a representation of $4096 \times 9 = 36864$ dimensions.

## 4. Dataset Examples

Figure 1, 2, 3, 4, 5, and 6 present sample images of the WIDER dataset, five for each class. Note we resized all images to squares for consistent layout.

## References

[1] 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. 1

[2] 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. 1

[3] 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. 1

[4] 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. 1



978-1-4673-6964-0/15/$31.00 ©2015 IEEE


Figure 1: Sample images in the dataset.


Figure 2: Sample images in the dataset.


Figure 3: Sample images in the dataset.

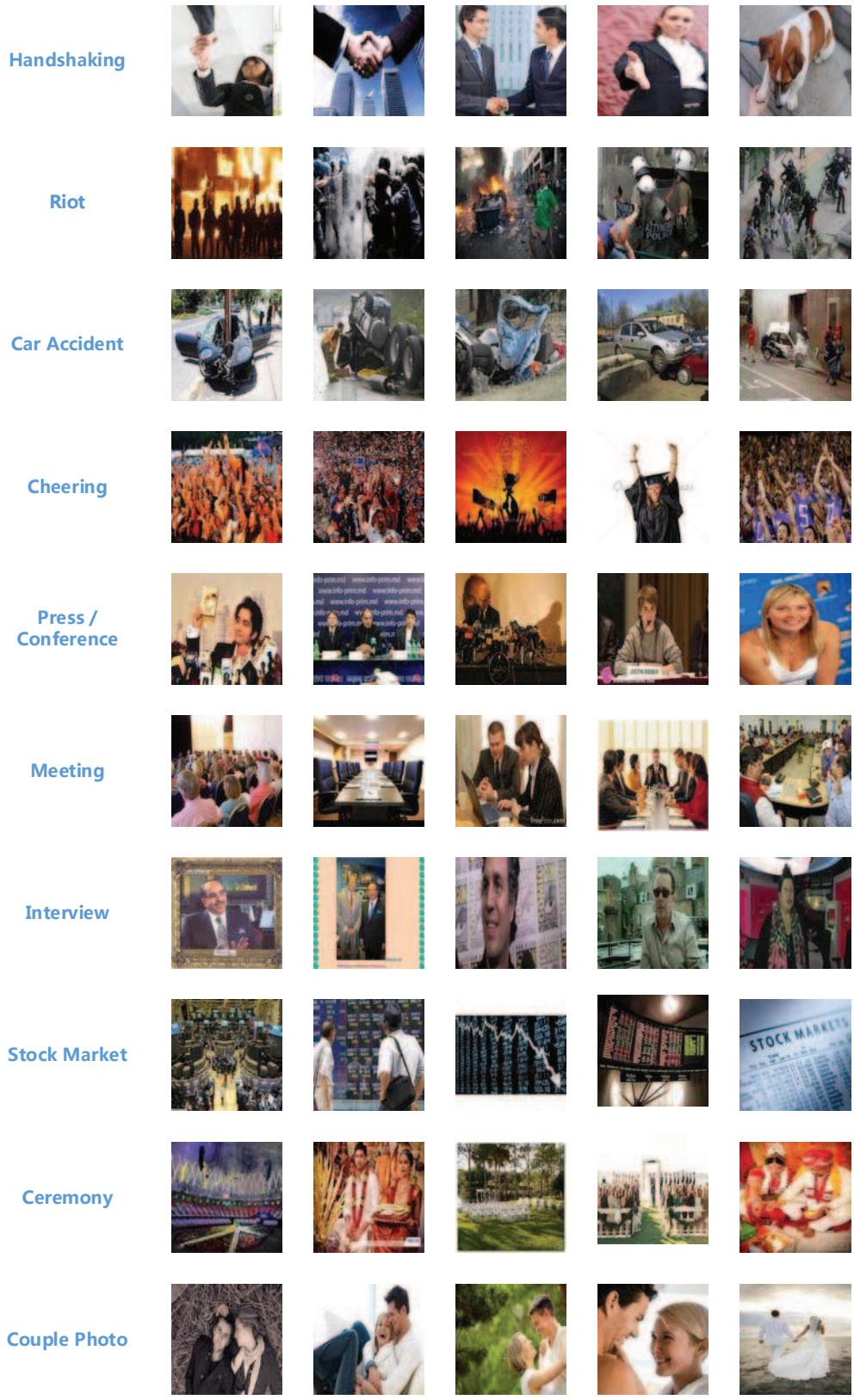


Figure 4: Sample images in the dataset.

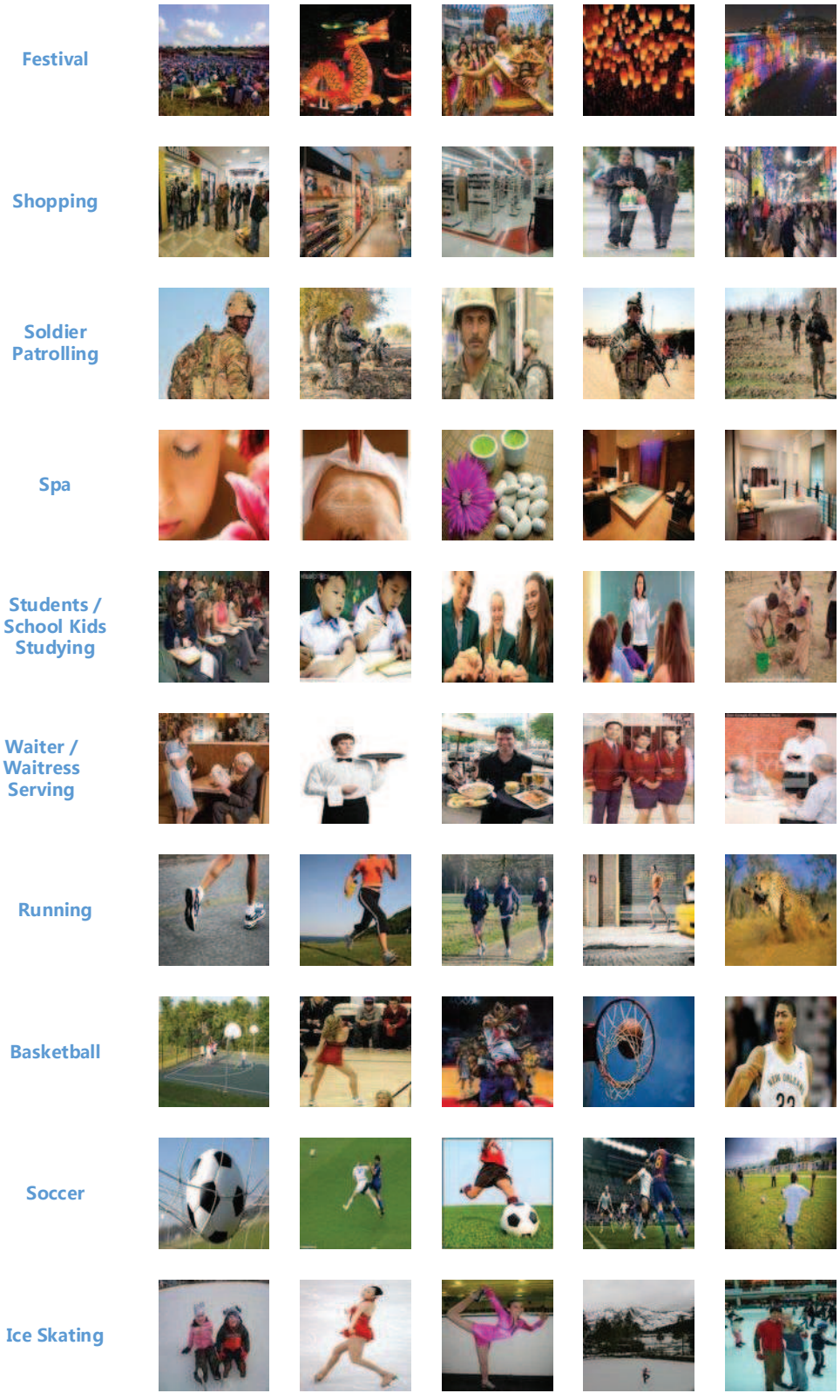


Figure 5: Sample images in the dataset.

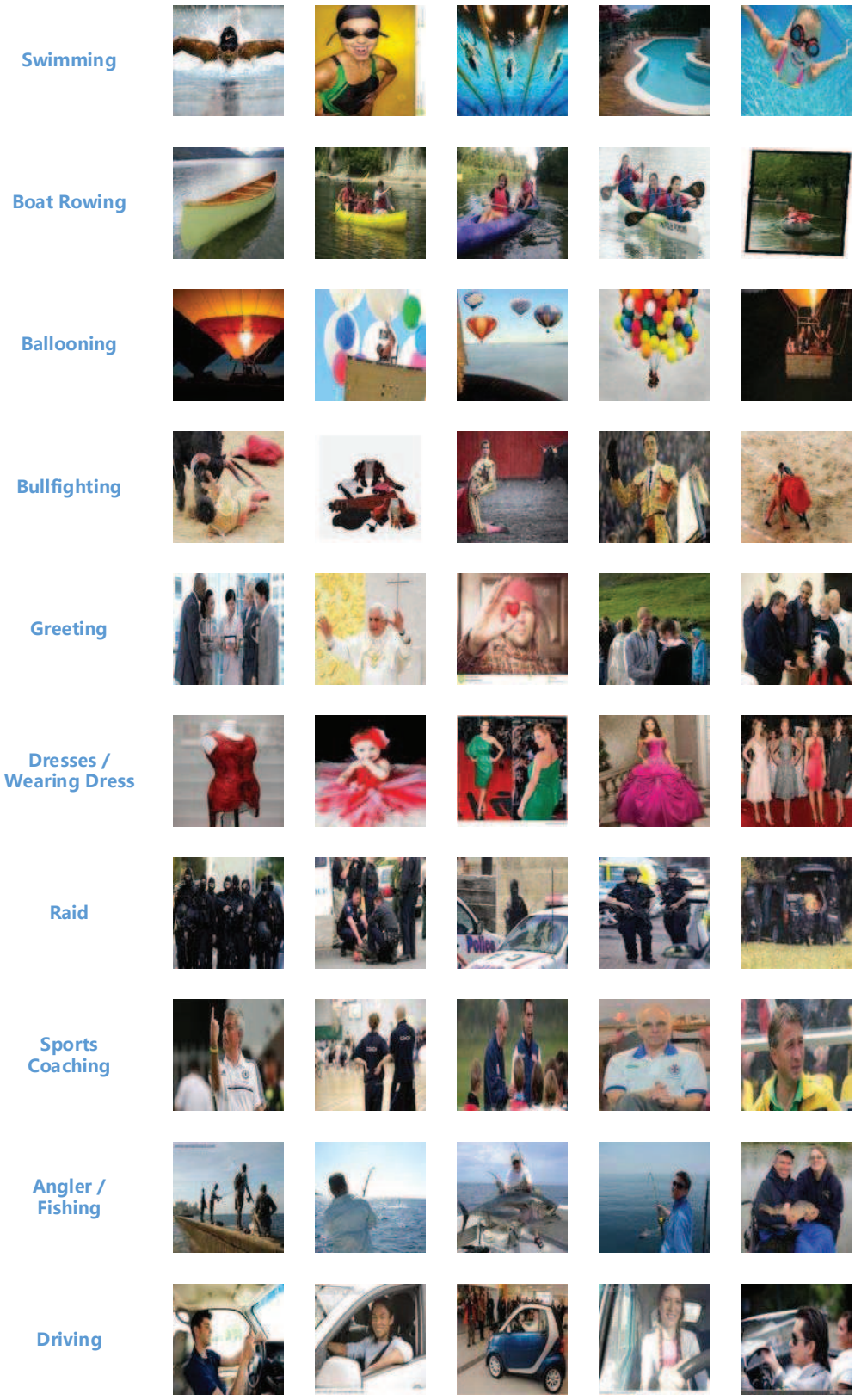


Figure 6: Sample images in the dataset.