Deeply Learned Attributes for Crowded Scene Understanding

Jing Shao[1], Kai Kang[1], Chen Change Loy[2], Xiaogang Wang[1]
[1]Department of Electronic Engineering, The Chinese University of Hong Kong.
[2]Department of Information Engineering, The Chinese University of Hong Kong.

During the last decade, the field of crowd analysis had a remarkable evolution from crowded scene understanding, including crowd behavior analysis [6, 8, 11], crowd tracking [2, 7], and crowd segmentation [1, 3]. Much of this progress was sparked by the creation of crowd datasets as well as the new and robust features and models for profiling crowd intrinsic properties. Most of the above studies on crowd understanding are scene-specific, that is, the crowd model is learned from a specific scene and thus poor in generalization to describe other scenes. Attributes are particularly effective on characterizing generic properties across scenes.

In the recent years, studies in attribute-based representations of objects, faces, actions, and scenes have drawn a large attention as an alternative or complement to categorical representations. These studies characterize the target subject by several attributes rather than discriminative assignment into a single specific category, which is too restrictive to describe the nature of the target subject. Furthermore, scientific studies have shown that different crowd systems share similar principles that can be characterized by some common properties or attributes. Indeed, attributes can express more information in a crowd video as they can describe a video by answering "Who is in the crowd?", "Where is the crowd?", and "Why is crowd here?", but not merely define a categorical scene label or event label to it. For instance, an attribute-based representation might describe a crowd video as the "conductor" and "choir" perform on the "stage" with "audience" "applauding", in contrast to a categorical label like "chorus".

In this paper, we introduce a new large-scale crowd video dataset designed to understand crowded scenes. It is named as the *Who do What at some-Where* (WWW) Crowd Dataset. It contains 10,000 videos from 8,257 crowded scenes. To our best knowledge, WWW Crowd Dataset is the largest crowd dataset to date. The videos in WWW dataset are all from real-world, collected from various sources, and captured by diverse kinds of cameras. We further define 94 meaningful attributes as high-level crowd scene representations, shown in Fig. 1. These attributes are determined with the helps from tag information of the crowd videos from Internet. They cover the common crowded places, subjects, actions, and events.

From the modeling perspective, we are interested to explore whether deeply learned crowd features can exceed traditional hand-craft features [5, 10]. Since videos possess motion information in addition to appearance, we examine deeply learned crowd features from both the appearance and motion aspects. Compared with two state-of-the-art deep learning methods for sport video classification [4] and action recognition [9], we develop a multi-task deep model to jointly learn and combine appearance and the proposed motion features [8] for crowded scene understanding. The network is shown in Fig. 2. From the experimental results with the proposed deep model, we show that our attribute-centric crowd dataset allows us to do a better job in the traditional crowded scene understanding and provides potential abilities in cross-scene event detection, crowd video retrieval, crowd video classification. We further design a user study to measure how accurately humans can recognize crowd attributes, and with which type of data that users can achieve the highest accuracy. This study is necessary and essential to provide a reference evaluation to our empirical experiments. Specifically, it is interesting to see how human perception (when given different data types) correlated with the results of computational models.

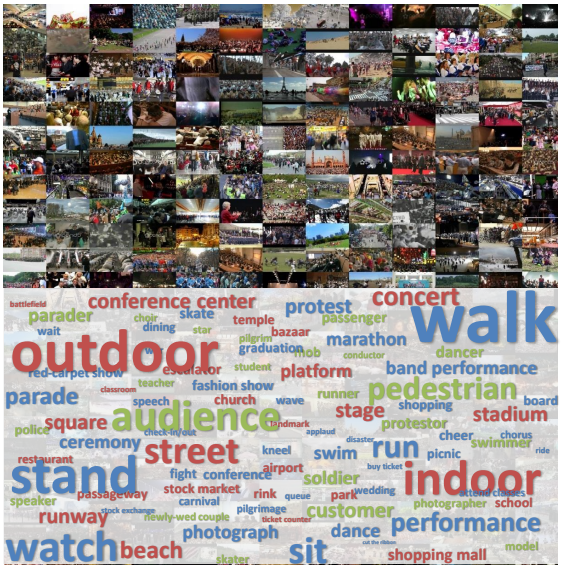

Figure 1: A quick glance of WWW Crowd Dataset with its attributes. Red represents the location (Where), green represents the subject (Who), and blue refers to event/action (Why). The area of each word is proportional to the frequency of that attribute in WWW dataset.

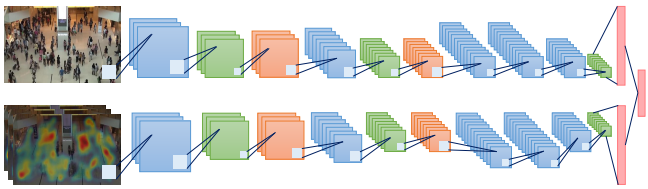

Figure 2: Deep model. The appearance and motion channels are input in two separate branches with the same deep architecture. Both branches consist of multiple layers of convolution (blue), max pooling (green), normalization (orange), and one fully-connected (red). The two branches then fuse together to one fully-connected layers (red).

[1] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007.

[2] Saad Ali and Mubarak Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*. 2008.

[3] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *TPAMI*, 30(5):909–926, 2008.

[4] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[5] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009.

[6] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009.

[7] Mikel Rodriguez, Josef Sivic, Ivan Laptev, and J-Y Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011.

[8] Jing Shao, Chen Change Loy, and Xiaogang Wang. Scene-independent group profiling in crowd. In *CVPR*, 2014.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[10] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[11] Xiaogang Wang, Xiaoxu Ma, and W Eric L Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *TPAMI*, 31(3):539–555, 2009.

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