Crowded Scene Understanding by Deeply Learned Attributes*  

Jing Shao1  Kai Kang1  Chen Change Loy2  Xiaogang Wang1  
1Department of Electronic Engineering, The Chinese University of Hong Kong  
2Department of Information Engineering, The Chinese University of Hong Kong  
jshao@ee.cuhk.edu.hk, kkang@ee.cuhk.edu.hk, ccloy@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

1. Introduction  
During the last decade, the field of crowd analysis had a remarkable evolution from crowded scene understanding, including crowd behavior analysis [13, 6, 7, 10, 8, 14, 16], crowd tracking [1, 9, 17], and crowd segmentation [2, 3, 15]. 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 as they 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”. Recently, some works [10, 16] have made efforts on crowd attribute profiling. But the number of attributes in their work is limited, as well as the dataset is also small in terms of scene diversity.  

2. Methodology and Experiment  
In this paper, we introduce a new large-scale crowd video dataset with crowd attribute annotation designed to under-stand crowded scenes. We exploit deep models to learn the features for each attribute from the appearance and motion information of each video, and apply the learned models for recognizing attributes in unseen crowd videos.  

The largest crowd dataset with crowd attribute annotation. To our best knowledge, the Who do What at some-Where (WWW) Crowd Dataset¹ is the largest crowd dataset to date. It contains 10,000 videos from 8,257 crowded scenes. The videos in the WWW crowd 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 navigated by tag information of the crowd videos from Internet. They cover the common crowded places, subjects, actions, and events.  

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 the WWW dataset.

1http://www.ee.cuhk.edu.hk/~jshao/WWWcrowd.html
Extensive experiment evaluation. Since videos possess motion information in addition to appearance, we examine deeply learned crowd features from both the appearance and motion aspects. Composed with the method that directly inputs a single frame and multiple frames to the deep neural network, we propose the motion feature channels inspired from [10] as the input of the deep model and develop a multi-task deep model to jointly learn and combine appearance and the proposed motion features for crowded scene understanding. The network is shown in Fig. 2. In all the experiments, we employ the area under ROC curve (AUC) as the evaluation criteria.

1) Deeply learned static features (DLSF). To evaluate our DLSF from the appearance channels only, we select a set of state-of-the-art hand-craft static features (i.e., SIFT, GIST, HOG, SSIM, and LBP) that have been widely used in scene classification for comparison, named as SFH in Table 1.  
2) Deeply learned motion features. We also report the performance of the deeply learned motion features in Table 1, compared with two baselines. One is the histogram of our proposed motion descriptor (MDH), and another is dense trajectory (DenseTrack) [12].

3) Deeply learned motion features. The deep model combining the DLSF and DLMF is compared with five baselines. It includes two combinations of appearance and motion (i.e., SFH+MDH and SFH+DenseTrack), a hand-craft feature extracting spatio-temporal motion patterns (STMP) [5], and two state-of-the-art deep models (i.e., Slow Fusion [4] and Two-stream [11]).

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 and crowd video retrieval.

User study on the WWW dataset. Appearance and motion cues play different roles in crowded scene understanding. We further conduct 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.

### References


