# 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
jshao@ee.cuhk.edu.hk, kkang@ee.cuhk.edu.hk, ccloy@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

## Abstract

*Crowded scene understanding is a fundamental problem in computer vision. In this study, we develop a multi-task deep model to jointly learn and combine appearance and motion features for crowd understanding. We propose crowd motion channels as the input of the deep model and the channel design is inspired by generic properties of crowd systems. To well demonstrate our deep model, we construct a new large-scale WWW Crowd dataset with* $10,000$ *videos from* $8,257$ *crowded scenes, and build an attribute set with* $94$ *attributes on WWW. We further measure user study performance on WWW and compare this with the proposed deep models. Extensive experiments show that our deep models display significant performance improvements in cross-scene attribute recognition compared to strong crowd-related feature-based baselines, and the deeply learned features behave a superior performance in multi-task learning.*

## 1. Introduction

During the last decade, the field of crowd analysis had a remarkable evolution from crowded scene understanding, including crowd behavior analysis [38, 24, 26, 33, 3, 48, 46, 45, 15, 27, 41, 40, 44, 47], crowd tracking [2, 32, 49], and crowd segmentation [1, 7, 16, 42]. 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 [7, 38, 3, 48, 24, 27, 15] 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 [11, 20, 4], faces [19, 25], actions [13, 23, 39], and scenes [31, 28, 12, 30] 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 [5, 8] 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 [33, 45] have made efforts on crowd attribute profiling. But the number of at-


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.

tributes in their work is limited (only four in [33, 45] ), as well as the dataset is also small in terms of scene diversity.

In this paper, we introduce a new large-scale crowd video dataset designed to understand crowded scenes named as the *Who do What at someWhere* (WWW) Crowd Dataset[1]. It contains $10,000$ videos from $8,257$ crowded scenes. To our best knowledge, the WWW Crowd Dataset is the largest crowd dataset to date. 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.

From the modeling perspective, we are interested in exploring whether deeply learned crowd features can exceed traditional hand-craft features. Since videos possess motion information in addition to appearance, we examine deeply learned crowd features from both the appearance and motion aspects. Compared with the method that directly inputs a single frame and multiple frames to the deep neural network, we propose the motion feature channels as the input of the deep model. 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) is correlated with the results of computational models.

Our contributions are listed as follows:

1) *The largest crowd dataset with crowd attributes annotations* - We establish a large-scale crowd dataset with $10,000$ videos from $8,257$ scenes. $94$ crowd-related attributes are designed and annotated to describe each video in the dataset. It is the first time such a large set of attributes on crowd understanding is defined.

2) *Deeply learned features for crowd scene understanding* - We develop a multi-task learning deep model to jointly learn appearance and motion features and effectively combine them. Instead of directly inputting multiple frames to a deep model to learn motion features as most existing works [17] did for video analysis, we specially design crowd motion channels as the input of the deep model. The motion channels are inspired by generic properties of crowd systems, which have been well studied in biology and physics. With multi-task learning, the correlations among attributes are well captured when learning deep features.

3) *Extensive experiments evaluation and user study to explore the WWW dataset* - They provide valuable insights on how static appearance cues and motion cues behave differently and complementarily on the three types of attributes: "where", "who" and "why". It also shows that the features specifically learned for human crowds are more effective than state-of-the-art handcrafted features.

## 2. WWW Crowd Dataset Construction

Most of the existing public crowd datasets [6, 9, 22, 38, 48] contain only one or two specific scenes, and even the largest one [33] merely provides $474$ videos from $215$ crowded scenes. On the contrary, our proposed WWW dataset provides $10,000$ videos[2] with over $8$ million frames from $8,257$ diverse scenes, therefore offering a superiorly comprehensive dataset for the area of crowd understanding. The abundant sources of these videos also enrich the diversity and completeness. We compare our WWW dataset with the other publicly available crowd datasets in Table 1. Over all the comparison items listed in the table, our dataset surpasses the rest both in scale and diversity.

### 2.1. Crowd Video Construction

**Collecting Keywords.** In order to obtain a large scaled and comprehensive crowd dataset, we selected a set of keywords related to common crowd scenarios (*e.g.* street, stadium, and rink) and crowd events (*e.g.* marching, chorus, and graduation) for the sake of searching efficiency and effectiveness.

For the purpose of generalization, we did not include keywords referring to specific places, but used general keywords that describe the functionalities of places instead. For instance, we chose "landmark" rather than names of specific places like "Time Square" and "Grand Central Station". It is common sense that "landmark" attracts crowds of tourists. Besides keywords about functional places like "station", "restaurant", and "conference center", we also included several specific types of places, such as "escalator" and "stage". Although these can be seen as objects, they are known to have high correlation with crowd.

**Collecting Crowd Videos[3].** The gathered keywords were used to search for videos from several public video search engines including Getty Images[4], Pond5[5], and YouTube[6].

[1] http://www.ee.cuhk.edu.hk/~jshao/WWWCrowdDataset.html

[2] The average length of all videos is around 23 seconds, and its std. is around 26 seconds.

[3] Our collection covers major existing crowd video datasets such as [33, 45, 14].

[4] http://www.gettyimages.com/

[5] http://www.pond5.com/

[6] http://www.youtube.com/

| | CUHK [33] | Collectiveness [45] | Violence [14] | Data-driven [32] | UCF [1] | **WWW** |
|---|---|---|---|---|---|---|
| # video | 474 | 413 | 246 | 212 | 46 | **10,000** |
| # scene | 215 | 62 | 246 | 212 | 46 | **8,257** |
| # frame | 60,384 | 40,796 | 22,074 | 121,626 | 18,196 | **$>$ 8 million** |
| resolution | multiple | $670 \times 1000$ | $320 \times 240$ | $640 \times 360$ | multiple | **$640 \times 360$** |
| source | Getty Images, Pond5, surveillance | web, surveillance | YouTube | web | BBC Motion Gallery, Getty Images | **Getty Images, Pond5, YouTube, surveillance, movies** |

Table 1. Comparison of WWW and other existing datasets. WWW offers the largest number of videos, scenes, and frames.

To increase the chance of retrieving crowd videos, we added "crowd" or "group" in most of the keywords, except keywords that explicitly describe the crowd (*e.g.* "chorus" and "marathon"). Besides these three sources, we further collected 469 videos from 23 movies[7]. To control the video quality, we removed videos with blurred motion, synthetic crowd, and extremely short length. In addition, all the duplicated videos were filtered.

## 2.2. Crowd Attribute Annotation

Given a video collection of many different crowded scenes, there is a enormous number of possible attributes describing different scenarios, subjects, and events. The capacity to infer attributes allows us to describe a crowded scene by answering "Where is the crowd?", "Who is in the crowd?", and "Why is crowd here?". Importantly, when faced with a new crowded scene, we can still describe it with these three types of cues (*e.g.* newly-wed couple [Who] are in the wedding [Why] at beach [Where]).

Furthermore, there is a large number of possible interactions among these attributes. Some attributes are likely to co-occur with each other, whilst some seem exclusive. For example, the scenario "street" attribute is likely to co-occur with subject "pedestrian" when the subject is "walking", and also likely to co-occur with subject "mob" when the subject is "fighting", but not related to subject "swimmer" because the subject cannot "swim" on "street". In addition, there exist attributes that are grouped hierarchically, *e.g.* "outdoor/indoor" contains almost all the other attributes of location. Some attributes, like "stadium" and "stage" belong to both "outdoor" and "indoor".

**Collecting Crowd Attributes from Web Tags.** To build the attribute taxonomy, we first collected tags from Pond5 and Getty Images[8] as shown in Fig. 2 as a form of wordle[9]. By carefully examining the contents of the retrieved tags with a total number of 7000+, we found that it is laborious and non-trivial to define attributes from these raw tags since the majority of them are not relevant to the problem we are interested in, even not related to crowd, such as video qualities and environment condition. In addition, tags with the highest frequency (*e.g.* people, adult, time, and ethnicity) are also likely to be discarded. We spent efforts to clean raw tags and finally constructed an attribute set with 94 crowd-related attributes which is shown in Fig. 1. It includes 3 types of attributes: (1) Where (*e.g.* street, temple, and classroom), (2) Who (*e.g.* star, protester, and skater), and (3) Why (*e.g.* walk, board, and ceremony). A complete list can be found in our project website.


Figure 2. Raw tag wordle (partial tag set). Bigger font size of a word suggests higher frequency of a tag appears in dataset.

**Crowd Attribute Annotation.** We hired 16 annotators to label attributes in the WWW dataset, and another 3 annotators to refine labeling for all videos. All the attributes in our dataset are commonly seen and experienced in daily life, so the annotation does not require special background knowledge. Each annotator was provided with a sample set containing both positive and negative samples of each attribute. They were asked to select possible attributes for each test video in a label tool containing three attribute lists of where, who, and why, respectively. In every round, each annotator was shown a 10-second video clip and was required to label at least one attribute from each attribute list without time constraint.

Fig. 3 shows several examples in the WWW Crowd dataset. As shown in the first row of Fig. 3, not all the video clips can tell a complete story in the form of "somebody" "do something" "at somewhere". Therefore, before labeling, we add "ambiguous" options in each attribute list. Totally, the annotators labeled 2855 "ambiguous" among all the marked $980,000$ labels, taking $0.1\%$, $0.2\%$, and $0.4\%$

[7]The list can be found from our project website.

[8]Other websites, *i.e.* YouTube, movies, and existing datasets do not have tag information.

[9]http://www.wordle.net/


Figure 3. Several video examples in the WWW dataset. Both two videos in the first row have ambiguous attributes. While the other two videos in the second row have multiple attributes in where, who, and why.

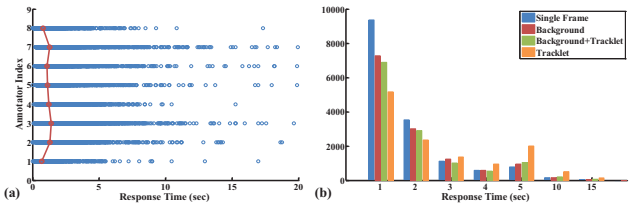


Figure 4. Visualize users' response time. The blue circles in (a) plot the response time of all annotators on labeling tasks, and the red line marks the average response time of each annotator. (b) shows the histograms of response time of different cues.

| | Single Frame | Background | Background + Tracklet | Tracklet | Average |
|---|---|---|---|---|---|
| Accuracy | 0.82 | 0.71 | 0.74 | 0.41 | 0.67 |

Table 2. User accuracy with four types of cues.

in Where, Who and Why, respectively. Two videos shown in the second row of Fig. 3 demonstrate that a video might have quite a number of attributes, i.e., multiple subjects doing different tasks at different locations in a single video.

## 3. User Study on Crowd Attribute

Appearance and motion cues play different roles in crowd scene understanding. In this section, we conduct a user study on the WWW crowd dataset to investigate human performance if only one type of cues is shown. This also serves as a reference for comparison with our empirical experiments in Section 5 and to explore the correlation between human perception and the computational models.

We distributed 8 users with four types of data, including single frame image, background[10], tracklets, and background with tracklets. The compared ground truth is the set of annotations in Section 3 from whole videos. To avoid bias, every user is provided with all the four types of data and randomly selected $10 \sim 15$ attributes. Before starting labeling, we provide each annotator $5 \sim 10$ positive as well as negative samples to help them get familiar with the attributes. Users were informed that their response time would be recorded.

(1) *Response time*: The average response time of all the users is 1.1094 seconds, as shown in Fig. 4(a). Fig. 4(b) shows that labeling with only tracklets is more laborious, and it is not easy for human to recognize crowd attributes simply from motions without seeing images.

(2) *Accuracy*: Table 2 shows that with single frames users can achieve much higher accuracy than with only tracklets or background. It means that the appearance of moving people and their poses are useful, but they are blurred on the background image. It is found that the background cue and tracklet cue are complementary. Figure 6(a) shows how many samples were wrongly labeled only with the background cue and how many of them were corrected after users also seeing the tracklet cue. Very few failure cases in the first 17 attributes are corrected by the tracklet cues, because these attributes belong to "where". Tracklets are more effective on the last 23 attributes belonging to "who" and "why". Figure 6(b) shows that the tracklets perform poorly on recognizing attributes belonging to "where".

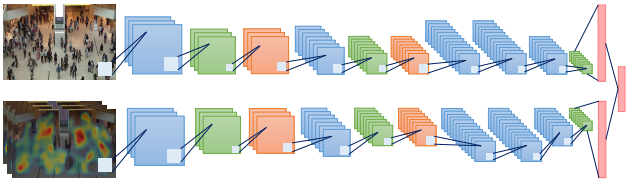

Figure 5. Deep model. The appearance and motion channels are input in two separate branches with the same deep architecture. Both branches consist of multiple layers of convolution (blue), max pooling (green), normalization (orange), and one fully-connected (red). The two branches then fuse together to one fully-connected layers (red).

## 4. Method

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.

### 4.1. Deep Network Structure and Model Setting

Fig. 5 shows the network structure of our deep model. The network contains two branches with the same archi-

[10]The average image of all frames of each video.

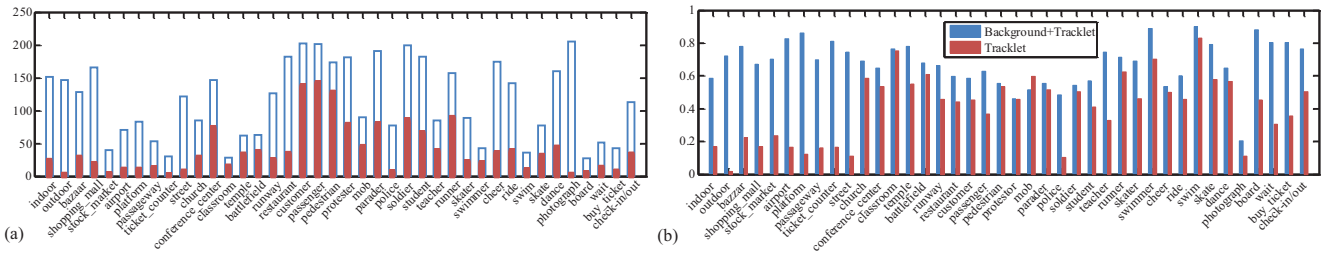


Figure 6. (a) The number of wrongly labeled samples with the background cue (indicated by blue bars) and how many of them can be corrected after adding the track let cue. (b) Accuracy comparison between the tracklet cue and tracklet + background. All the results are obtained from the user study described in Sec. 3.

tecture. We use simple notations to represent parameters in the networks: (1) *Conv*(*N,K,S*) for convolutional layers with $N$ outputs, kernel size $K$ and stride size $S$, (2) *Pool*(*T,K,S*) for pooling layers with type $T$, kernel size $K$ and stride size $S$, (3) *Norm*($K$) for local response normalization layers with local size $K$, and (4) *FC*($N$) for fully-connected layers with $N$ outputs, (5) The activation functions in each layer are represented by *ReLU* for rectified linear unit and *Sig* for sigmoid function. Then the two branches have parameters: *Conv*(96,7,2)-*ReLU*-*Pool*(3,2)-*Norm*(5)-*Conv*(256,5,2)-*ReLU*-*Pool*(3,2)-*Norm*(5)-*Conv*(384,3,1)-*ReLU*-*Conv*(384,3,1)-*ReLU*-*Conv*(256,3,1)-*ReLU*-*Pool*(3,2)-*FC*(4096). The output fully-connected layers of two branches are concatenated to be *FC*(8192). Finally, we have *FC*(8192)-*FC*(94)-*Sig* producing 94 attribute probability predictions. The loss function of the network is cross entropy as in Equation (1). The network parameters of Appearance branch are initialized using a pre-trained model for ImageNet detection task [29].

$$E = -\frac{1}{N}\sum_{n=1}^{N} t_n \log o_n + (1 - t_n)\log(1 - o_n) \qquad (1)$$

where the $N = 94$ denotes the number of output neurons, $t_n$ $(n = 1, \ldots, N)$ are the target labels and $o_n$ $(n = 1, \ldots, N)$ are the output probability predictions.

### 4.2. Motion Channels

The traditional input of deep model is a map of single frame (RGB channels) or multiple frames [17]. In this paper, we propose three scene-independent motion channels as the complement of the appearance channels. Some well-known motion features like optical flow cannot well characterize motion patterns in crowded scenes, especially across different scenes. Scientific studies have shown that different crowd systems share similar principles that can be characterized by some generic properties. Inspired by [33] that introduced several scene-independent properties (*e.g.* collectiveness, stability, and the conflict) for groups in crowd, we find that these properties also exist in the whole scene space and can be quantified from scene-level. After our reformulation, the collectiveness indicates the degree of individuals in the whole scene acting as a union in collective motion, and the stability characterizes whether the whole scene can keep its topological structures, and conflict measures the interaction/friction between each pair of nearest neighbors of interest points. Examples shown in the Fig. 7 illustrate each property intuitively.

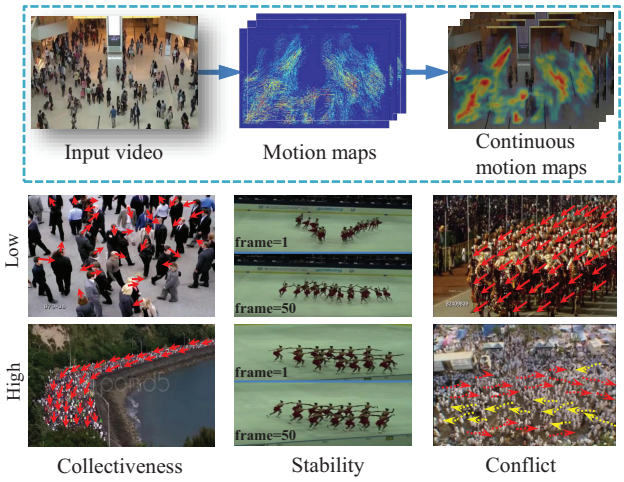


Figure 7. Motion channels. The first row gives an example to briefly illustrate three motion channels construction procedure. For each channel, two examples are shown in the second and third rows. Individuals in crowd moving randomly indicates low collectiveness, while the coherent motion of crowd reveals high collectiveness. Individuals have low stability if their topological structure changes a lot, whereas high stability if topological structure changes a little. Conflict occurs when individuals move towards different directions.

All the descriptors are defined upon tracklets detected by the KLT feature point tracker, and each of them is computed on 75 frames of each video in the WWW dataset. We first define a $K$-NN ($K = 10$) graph for the whole tracklet point set. Since we do not detect groups in advance, the descriptor proposed in [45] is more suitable to extract collectiveness for each tracklet point in the whole scene. Following the similar idea in [33], we design the descriptor for stability

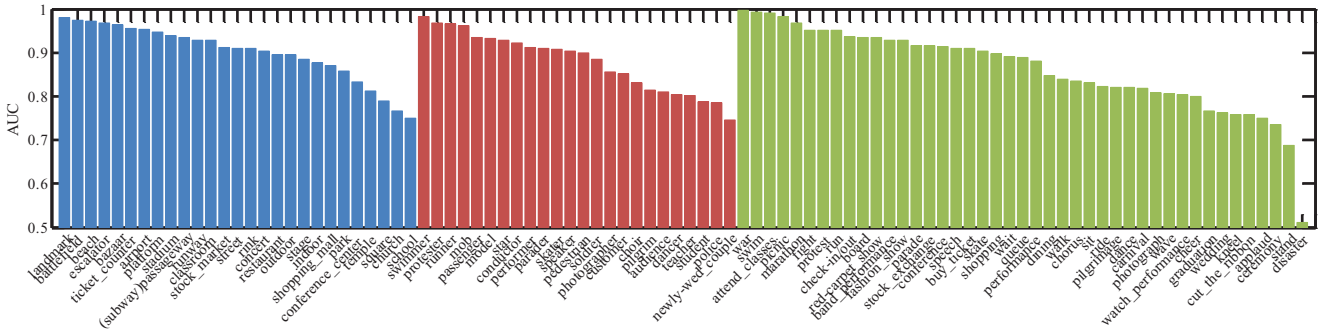


Figure 8. AUC of each attribute obtained with DLSF+DLMF. Blue, red, and green indicate attributes of where, who and why respectively.

by counting and averaging the number of invariant neighbors of each point in the $K$-NN graph. It reveals the fact that the stable crowd needs to maintain a similar set of nearest neighbors. The conflict descriptor defined in [33] is based on the group-based transition prior, thus is not suitable in our case. Instead, we generalize this descriptor by computing the velocity correlation between each nearby tracklet points within the $K$-NN graph. We average the per-frame descriptor map for each motion feature across the temporal domain to output three motion maps, which act as the input of the deep model. Although a single frame owns tens or hundreds of tracklets, the total tracklet points are still sparse. We then interpolate these sparse points to output a complete and continuous feature map. The brief channel construction procedure is shown in the first row in Fig. 7.

As shown in the Section 5.4 , these motion channels can facilitate appearance to improve the performance on attribute recognition.

# 5. Experimental Results

## 5.1. Settings

We split all the WWW dataset randomly into training, validation, and test sets with a ratio of $7:1:2$. Note that all the three sets are guaranteed to have positives and negatives of the 94 attributes, and they do not have overlap on scenes to guarantee that the attributes are learned scene-independently. In all the experiments, we employ the area under ROC curve (AUC) as the evaluation criteria.

## 5.2. Evaluation on Deeply Learned Static Features

To evaluate our deeply learned static features (DLSF) from the appearance channels only[11], we select a set of state-of-the-art static features that have been widely used in scene classification for comparison. Literature shows that Dense SIFT [21] and GIST [28] have good performance good on describing general image content, while HOG [10] has been widely used in pedestrian detection. They all have the potential of being applied to crowd scene understanding. To capture global information, we add a color histogram in the HSV color space and the self-similarity (SSIM) [34] descriptor. In addition, we also employ local binary patterns (LBP) [43] to quantify texture in crowded scenes.

We extract the six types of features from the first frame of each video and construct the static feature histogram (SFH) following a standard bag-of-words pipeline with K-means clustering and locality-linear coding [37]. Linear SVM is used to train independent classifiers with SFH on each attribute. As shown in the second row of Table 3, our DLSF method outperforms the SFH baseline. The mean AUC is improved by 6%. Out of the total 94 attributes, it has higher AUC on 64 attributes (shown in the last column).

## 5.3. Evaluation on Deeply Learned Motion Features

We also report the performance of the deeply learned motion features in Table 3, compared with two baselines. One is the histogram of our proposed motion descriptor (MDH) in Sec. 4.2. And another is dense trajectory [36] showed state-of-the-art result in action recognition. Both baseline features are trained with independent classifiers via linear SVM similar to the SFH baseline.

According to the results shown in the third row of Table 3, DLMF outperforms the other two baselines by 10% and 5% on mean AUC respectively. Over 77% attributes, DLMF achieves higher AUC than the baselines. On the other hand, DLMF has a nearly 20% drop compared with DLSF. This is consistent with our observation from the user study in Table 2 that the motion cue is less effective on recognizing attributes compared with the appearance cue in general.

## 5.4. Evaluation on Combined Deep Model

The deep model combining DLSF and DLMF is shown in Fig. 5. It is compared with five baselines. The first two baselines are the combination of the static feature (SFH) with two motion features (MDH and dense trajectory). We add a baseline [18] that extracts spatio-temporal motion patterns (STMP) by modeling the input video as the assembly

[11] The first row in Fig. 5 with the last fully-connected layer is substituted by three fully-connected layers.

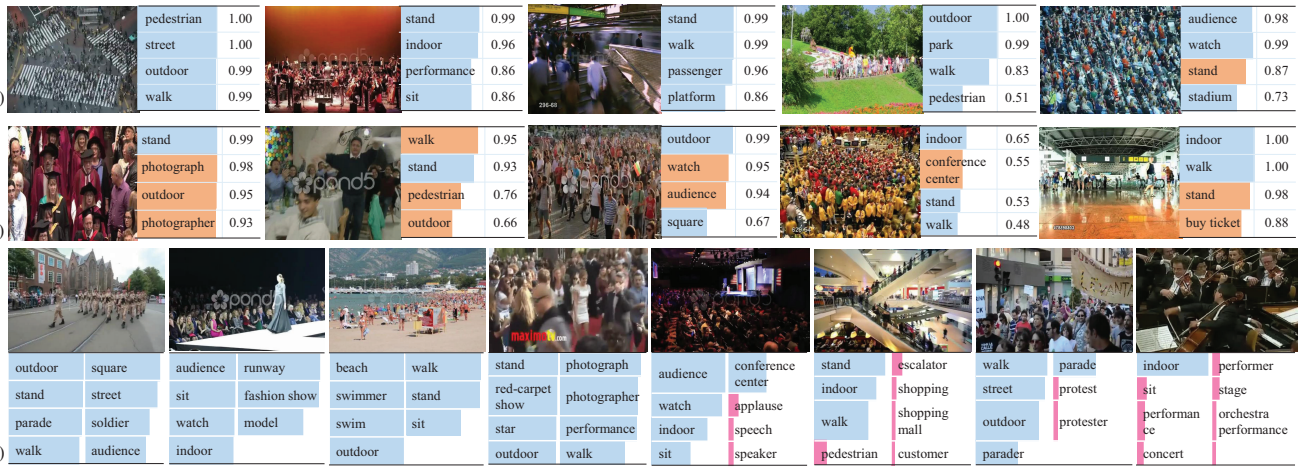


Figure 9. Good and bad attribute prediction examples are shown in (a) and (b). For each image, its top four attributes with the highest prediction scores with our DLSF + DLMF are shown. The heights represent prediction scores. Blue indicates correctly predicted attributes and orange indicates wrong prediction (false alarms). In (c) all the user annotated ground truth attributes are shown for each image example. If a prediction score on an attribute is lower than 0.5, is represented by red, which means miss detection. Otherwise it is blue.

| Our Methods | mean AUC | Baselines | mean AUC | # wins |
|---|---|---|---|---|
| DLSF | **0.87** | SFH | 0.81 | 67/94 |
| DLMF | **0.68** | MDH | 0.58 | 85/94 |
| | | DenseTrack [36] | 0.63 | 72/94 |
| DLSF + DLMF | **0.88** | SFH+MDH | 0.80 | 78/94 |
| | | SFH+DenseTrack | 0.82 | 72/94 |
| | | STMP [18] | 0.72 | 89/94 |
| | | Slow Fusion [17] | 0.81 | 74/94 |
| | | Two-stream [35] | 0.76 | 89/94 |

Table 3. Compare deeply learned features with baselines. The last column shows the number of attributes (out of the total number of 94) on which our proposed deep features have higher AUC than baselines.

of spatio-temporal cuboids. It combine both appearance and motion cues. The fourth baseline is the slow fusion scheme with multi-frames as input in deep model proposed in [17] recently. It is a state-of-the-art deep learning method for video analysis, and it has achieved the best performance in [17] for sports classification. It is interesting to investigate whether this deep learning framework can learn crowd features well. And the last baseline is the two-stream convolutional networks for action recognition [35]. We substitute our motion channels with optical flow maps (*i.e.* 2 maps for each frame, and 5 frames for each video) and keep the appearance channels unchanged. According to the last row in Table 3, our combined deep features DLSF+DLMF outperform all the baselines and STMP is the worst. Slow Fusion [17] does not outperform handcrafted features. This reason might be its way of inputting multiple frames to the deep model in order to capture motion information. It leads to a much larger net structure with many more parameters, and therefore requires larger scale training data. Similarly, the two-stream structure [35] also involves more parameters caused by ten motion channels, and optical flow itself cannot characterize common features across different scenes. Instead the input of our deep model is three motion channels, which well summarize motion information and reduce the network size. Summarizing all the resulting in Table 3, we achieve the conclusion that the motion cue alone cannot get good result on crowd attribute recognition. By adding deeply learned motion features (DLMF) to deep learned static features (DLSF), the mean AUC has been improved by $1\%$. A detailed investigation shows the AUC of $41$ attributes gets improved by adding DLMF. Most of these attributes belong to "Who" and "Why". The averaged improvement of AUC is $5\%$.

**Quantitative Evaluation.** The AUC for each attribute with DLSF+DLMF is shown in Fig. 8. Different colors represent "Where", "Who", and "Why" from left to right, and the results are sorted in a descending order. The attribute "war" achieves the highest AUC score whereas "disaster" is the lowest. The lowest score may result from too few positive samples in the training set. Some attributes such as "battlefield", "mob", and "war" have strong correlation, although they belong to 'where", "who" and "why" respectively. They all have high AUC.

**Qualitative Evaluation.** Some good and bad examples on attribute prediction are shown in Fig. 9(a) and (b). Noted that the last example in Fig. 9(a) shows an attribute "stand" with a high prediction probability of $0.87$, while the groundtruth recommends "sit". It is actually quite challenging to determine the action as "stand" or "sit" in this example even from human perception.The third example shown in Fig. 9(b) has two of the top four attributes mistakenly predicted. The fourth is actually "stock market" but wrongly recognized as "conference center". This is because

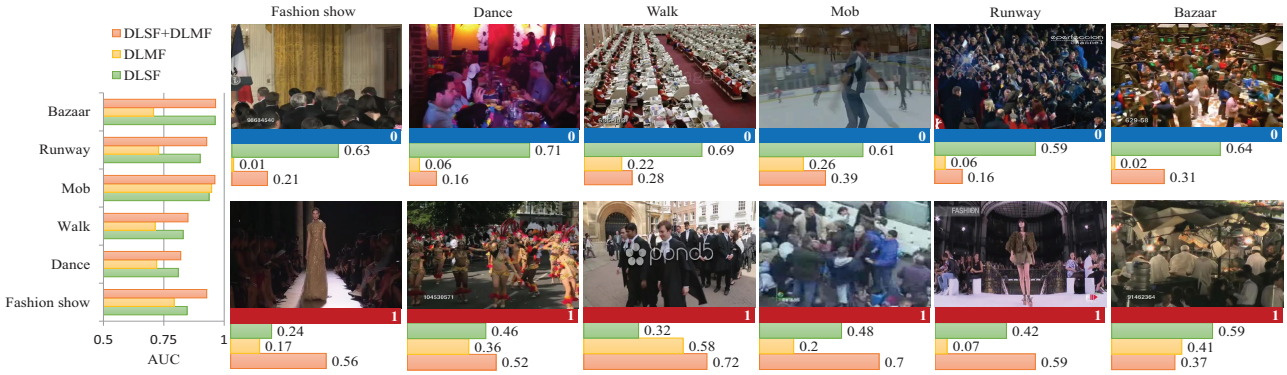


Figure 10. Six attributes predicted by DLSF, DLMF, and DLSF + DLMF. The first row shows negative examples with blue bar, and the second row shows positive examples with red bar. The prediction scores with DLSF, DLMF and DLSF + DLMF are represented by green, yellow, and orange.

people in this example move coherently, behaving like "audience" or "watching performance", and its appearance cue cannot distinguish it as "conference center" or "stock market". Fig. 9(c) shows some examples whose attributes are miss detected. The proposed deep model can well recognize attributes with distinctive motion and appearance patterns, such as the attributes related to the beach scene. But it may be poor for attributes owning complex and diverse appearance and motions, such as attributes related to the shopping mall scene.

**Combined Deep Features vs. Separate Deep Features.** To further verify that our combined deep features outperform both DLSF and DLMF, we show 6 attributes with their quantitative (AUC scores) and qualitative results in Fig. 10. The first row shows negative examples, and the higher prediction probability indicates higher error. On the contrary, the higher prediction probability in the second row indicates higher accuracy. Generally, DLSF extracts static appearance features and thus works poorly at several attributes specified with motion patterns, *e.g.* "fashion show" and "walk". But only motion features cannot effectively explore the difference between attributes with similar motion patterns. Likewise, the negative example in the fourth column is actually "skate", but the given frame shows a short cut image that is similar to "mob" or "fight". Combinational model fusing the appearance and motion channels and complement the missing cues in DLSF or DLMF, therefore reveals superior performances over all the sample attributes.

### 5.5. Multi-task learning

Deep models are ideal for multi-task learning. We compare the result of training three different deep models for the three sets of attributes "where", "who" and "why" separately. This is called single-task learning. In comparison, the deep model discussed above is called multi-task learning. Since there exist correlations between different types of attributes, joint training of the three sets of attributes implicitly emphasizes the common features that shared by the correlated attributes. For instance, the "swimmer" should be at "beach" or "pedestrian" walks on "street". Table 4 reports the average AUC of each set of attributes by single-task and multi-task learning in the first two columns. The last column shows the numbers of attributes where multi-task learning outperforms single-task learning. It is obvious that the multi-task learning improves the overall AUC from 0.81 to 0.87. The accuracies on most attributes get improved.

| | Multi-task | Single-task | # wins |
|---|---|---|---|
| Where | 0.89 | 0.84 | 22/27 |
| Who | 0.86 | 0.79 | 18/24 |
| Why | 0.86 | 0.79 | 36/43 |
| Mean | 0.87 | 0.81 | 76/94 |

Table 4. Compare average AUC with single-task learning and multi-task learning. The last column is the number of attributes where multi-task learning outperforms single-task learning.

## 6. Conclusion

In this paper, we build a large-scale crowd dataset with $10,000$ videos from $8,257$ scenes, and propose $94$ crowd-related attributes. This is a significant contribution to the field of crowd scene understanding. Both appearance features and motion features are learned by our designed deep models. Instead of inputting multiple frames to deep models as existing works [17] did for video analysis, we design motion channels motivated by generic properties of crowd systems. Crowd features are learned with multi-task learning, such that the correlations among crowd attributes are well captured. The learned crowd features and crowd attribute predictors have many potential applications in the future work, such as crowd video retrieval and crowd event detection.

## Acknowledgment

This work is partially supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK 419412, CUHK 417011, CUHK 14206114, and CUHK 14207814), Hong Kong Innovation and Technology Support Programme (Project reference ITS/221/13FP), Shenzhen Basic Research Program (JCYJ20130402113127496), and a hardware donation from NVIDIA Corporation.

## References

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007. 1, 3

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*. 2008. 1

[3] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR*, 2006. 1

[4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*. 2010. 1

[5] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009. 1

[6] A. B. Chan, Z.-S. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008. 2

[7] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *TPAMI*, 30(5):909–926, 2008. 1

[8] H. Chaté, F. Ginelli, G. Grégoire, and F. Raynaud. Collective motion of self-propelled particles interacting without cohesion. *Physical Review E*, 77(4):046113, 2008. 1

[9] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012. 2

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1

[12] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10, 2007. 1

[13] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*. 2012. 1

[14] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR*, 2012. 2, 3

[15] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *CVPR*, 2009. 1

[16] K. Kang and X. Wang. Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*, 2014. 1

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2, 5, 7, 8

[18] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009. 6, 7

[19] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1

[20] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 6

[22] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *ECCV*. 2008. 2

[23] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 1

[24] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009. 1

[25] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013. 1

[26] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010. 1

[27] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 1

[28] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 6

[29] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv preprint arXiv:1409.3505*, 2014. 5

[30] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 1

[31] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 1

[32] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011. 1, 3

[33] J. Shao, C. C. Loy, and X. Wang. Scene-independent group profiling in crowd. In *CVPR*, 2014. 1, 2, 3, 5, 6

[34] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 6

[35] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 7

[36] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 6, 7

[37] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 6

[38] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using

hierarchical bayesian models. *TPAMI*, 31(3):539–555, 2009. 1, 2

[39] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 1

[40] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *CVPR*, 2015. 1

[41] S. Yi, X. Wang, C. Lu, and J. Jia. L0 regularized stationary time estimation for crowd group analysis. In *CVPR*, 2014. 1

[42] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015. 1

[43] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen. Rotation-invariant image and video description with local binary pattern features. *TIP*, 21(4):1465–1477, 2012. 6

[44] B. Zhou, X. Tang, and X. Wang. Coherent filtering: detecting coherent motions from crowd clutters. In *ECCV*. 2012. 1

[45] B. Zhou, X. Tang, and X. Wang. Measuring crowd collectiveness. In *CVPR*, 2013. 1, 2, 3, 5

[46] B. Zhou, X. Tang, and X. Wang. Learning collective crowd behaviors with dynamic pedestrian-agents. *IJCV*, 111(1):50–68, 2015. 1

[47] B. Zhou, X. Tang, H. Zhang, and X. Wang. Measuring crowd collectiveness. *TPAMI*, 36(8):1586–1599, 2014. 1

[48] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *CVPR*, 2012. 1, 2

[49] F. Zhu, X. Wang, and N. Yu. Crowd tracking with dynamic evolution of group structures. In *ECCV*. 2014. 1