Data-Driven Crowd Understanding: A Baseline for a Large-Scale Crowd Dataset

Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, *Member, IEEE*, Rong Xie, and Xiaokang Yang, *Senior Member, IEEE*

***Abstract*—Crowd understanding has drawn increasing attention from the computer vision community, and its progress is driven by the availability of public crowd datasets. In this paper, we contribute a large-scale benchmark dataset collected from the Shanghai 2010 World Expo. It includes 2630 annotated video sequences captured by 245 surveillance cameras, far larger than any public dataset. It covers a large number of different scenes and is suitable for evaluating the performance of crowd segmentation and estimation of crowd density, collectiveness, and cohesiveness, all of which are universal properties of crowd systems. In total, 53 637 crowd segments are manually annotated with the three crowd properties. This dataset is released to the public to advance research on crowd understanding. The large-scale annotated dataset enables using data-driven approaches for crowd understanding. In this paper, a data-driven approach is proposed as a baseline of crowd segmentation and estimation of crowd properties for the proposed dataset. Novel global and local crowd features are designed to retrieve similar training scenes and to match spatio-temporal crowd patches so that the labels of the training scenes can be accurately transferred to the query image. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art approaches for crowd understanding.**

***Index Terms*—Crowd features, crowd scene understanding, data-driven methods, large-scale benchmark.**

I. Introduction

CROWD understanding is an interdisciplinary topic and has been studied in physics [10], [25], [61], biology [52], [66], sociology [8], [36], [49], and psychology [6], [45] for a long time. In multimedia and computer vision community, a number of works contribute to action recognition [7], [28], event detection [47], [62] and behavior analysis [56] for the individuals or small groups in videos. Due to severe occlusion and perspective distortion, crowd understanding is a challenging topic and draws increasing attention because of the large demands on crowd video surveillance, which is especially important for metropolis security. Current works mainly focus on crowd segmentation [2], [12], crowd counting [11], [14], [15], [26], [42], crowd tracking [1], [3], [27], [53], [55], [72], and crowd behavior analysis [32], [33], [38], [41], [43], [46], [57], [58], [64], [69]–[71].

The progress of crowd understanding was mainly driven by the available public crowd datasets. Most of the above mentioned works [11], [12], [14], [15], [38], [41], [42], [64], [69] on crowd understanding are scene-specific, i.e., crowd understanding models learned from a particular scene can only be applied to the same scene. For example, the crowd counting approaches [11], [12], [14], [15], [42] require manually annotating some frames from the target scenes for training. In crowd behavior analysis [24], [33], [38], [41], [43], [46], [64], [69], behavior models trained for a target scene cannot generalize to other scenes. Therefore, the datasets proposed in [11], [15], [38], [64], [69] only contain one or two scenes.

Scientific studies [10], [13], [48] show that different crowd systems share the same underlying principles and can be characterized by a set of universal properties. Automatically understanding such general crowd properties across different scenes from videos not only has important applications, such as crowd video retrieval and crowd event detection, but also benefits scientific studies [9], [66] in other areas.

The learned crowd models are expected to generalize to new scenes not in the training set. Some research efforts [57], [70] have been made recently in this direction. Progress relies heavily on the availability of large-scale crowd datasets that include a large variety of scenes and video sequences. Existing crowd datasets do not provide enough variation. The largest one [70] (which actually combines other major crowd datasets) only contains 62 scenes. Because of security issues, only a small number of crowd videos by surveillance cameras are publicly available. The videos in [70] are mainly collected from the INTERNET. They are not of bird's-eye view and are therefore not suitable for crowd understanding. It provides only annotations of collectiveness at video-sequence level. Other multi-scene crowd datasets [2], [55] even do not provide ground-truth annotations. Benchmark datasets have become a bottleneck for the research on cross-scene crowd understanding.

Manuscript received July 27, 2015; revised February 6, 2016; accepted February 29, 2016. This work was supported in part by the NSFC under Grant 61527804, Grant 61221001, Grant 61301269, Grant 61371192, and Grant 61301269, in part by the STCSM under Grant 14XD1402100, 111 Program (B07022), in part by the General Research Fund sponsored by the Research Grants Council of Hong Kong Project CUHK 419412 and CUHK 147011, in part by the Hong Kong Innovation and Technology Support Programme Project ITS/221/13FP, in part by the Shenzhen Basic Research Program under Grant JCYJ20130402113127496, in part by the Ph.D. Programs Foundation of China under Grant 20130185120039, in part by the Sichuan Hi-tech R&D Program under Grant 2014GZX0009, and in part by the China Postdoctoral Foundation under Grant 2014M552339. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Martha Larson.

C. Zhang is with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China (e-mail: zhangcong0929@gmail.com).

K. Kang, H. Li, and X. Wang are with the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong, China (e-mail: kkang@ee.cuhk.edu.hk; hsli@ee.cuhk.edu.hk; xgwang@ee.cuhk.edu.hk).

R. Xie and X. Yang are with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xierong@sjtu.edu.cn; xkyang@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2542585

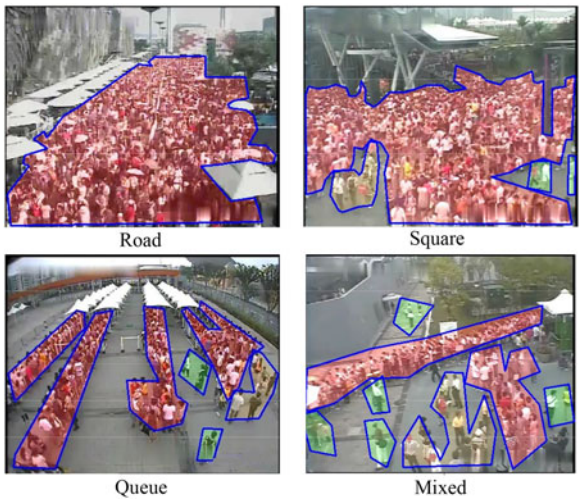


Fig. 1. Examples of four types of scenes from the WorldExpo'10 dataset. Crowd regions are manually labeled with polygons of different density levels (high = red, medium = yellow, low = green).

The first contribution of this paper is that we contribute a large-scale benchmark dataset[1] for crowd understanding. It includes 2630 annotated video sequences captured by 245 surveillance cameras, all from Shanghai 2010 WorldExpo.[2] Since all the cameras have disjoint bird's-eye views, they cover a large variety of scenes and the videos are especially suitable for cross-scene surveillance studies. Example scenes are shown in Figs. 1 and 2. We propose four challenges for this dataset: crowd segmentation, and estimation of crowd density, collectiveness and cohesiveness.

Crowd segmentation is the first step towards understanding crowds, because it answers the question of where the crowds are. Crowd counting, tracking, and behavior analysis are mainly based on the results of crowd segmentation, which itself also has important applications such as crowd trespassing detection. *Density* is a well known property of crowds and is related to other crowd properties such as collectiveness and cohesiveness [66]. Density estimation is of interest to security and traffic management, where highly dense crowds may lead to congestion or even disasters. *Collectiveness* is the degree of individuals in crowds acting with the same goal and was first studied in [70] from the perspective of computer vision. Collective behaviors widely exist in various crowd systems [8], [13], [66], [69], and have many potential applications [70]. *Cohesiveness* is another important property of crowd systems, in which some individuals move in groups and are bonded by force because of their special relationships. It measures the stability of local geometric and topological structures of crowd groups. Although it has been widely studied in crowd psychology [6], [18], it has not yet been addressed as a vision problem.

To accurately evaluate the methods that aim to solve the proposed four challenges, the WorldExpo'10 dataset is manually annotated at the region level. There are a total of 53 637 crowd segments with polygon boundaries, each of which was labeled with all the three crowd properties. Its annotations are much more comprehensive than any existing dataset. For example, the current largest one, CUHK dataset [70], only contains 413 videos with 413 annotations (because its annotations are at video level), while ours has 160 911 annotations. A detailed comparison with existing datasets is shown inTable I. Other researchers can propose new challenges and add new annotations to the WorldExpo'10 dataset. It would significantly advance the research of crowd understanding.

The second contribution is to propose a data-driven approach as a baseline of crowd segmentation and estimation of crowd properties for the proposed dataset. This is the first time to propose a unified approach that solves all the four crowd understanding challenges. As a generic solution, it has the potential of being applied to estimating other crowd properties. The large-scale annotated training data makes using data-driven approaches possible. Similar to [40], [60], no training procedure is required for our proposed data-driven method. The proposed method transfers the required information (such as density, collectiveness and cohesiveness) from the labeled training videos to the query via image matching. Our framework contains the following three steps: 1) retrieving candidate scenes similar to the query clip based on the proposed global crowd feature; 2) extracting multi-scale patches from the query and computing the proposed local crowd feature for each patch; 3) for each query patch, retrieving its nearest-neighboring patches from the candidate scenes and transferring the crowd properties from the nearest neighbors. The multiscale Markov Random Field (MRF) is utilized to enforce the smoothness of the resulting segmentation and property maps.

However, the general features widely used for scene understanding and texture description are not effective to describe crowd. Therefore, unlike other data-driven methods, new global crowd features (GCF) and local crowd features are proposed in our method. The proposed features are more effective than widely used generic features (such as GIST [51], HOG3D [31] and HOGHOF [35]) in the applications of crowd understanding. A new global crowd feature is proposed to retrieve similar crowd scenes for each input video clip. We train a series of mid-level filters as the crowd filters to generate filtering response maps for the input video clip. The global crowd feature of the input is then generated as the concatenation of the response maps. A new local crowd feature is proposed to compare similarities of spatio-temporal crowd patches. It combines eight types of features to characterize crowd appearance and motion. The combination weights of the eight features are automatically learned with relevance feedback such that the weighted features well match human perception on crowds.

The remainder of this paper is organized as follows. The existing works related to crowd understanding are reviewed and discussed in Section II. The details of our dataset, including annotation and evaluation protocols are introduced in Section III. Section IV presents our proposed data-driven approach for crowd understanding, the global crowd feature and the local crowd feature. In Section V, comprehensive experiments have been conducted to show the effectiveness of our

[1][Online]. Available: http://www.ee.cuhk.edu.hk/~xgwang/crowdexpo.html

[2]Since most exhibition pavilions have been deconstructed, and no video corresponding to those pavilions still in use is included, the data is approved to be released for academic purposes.

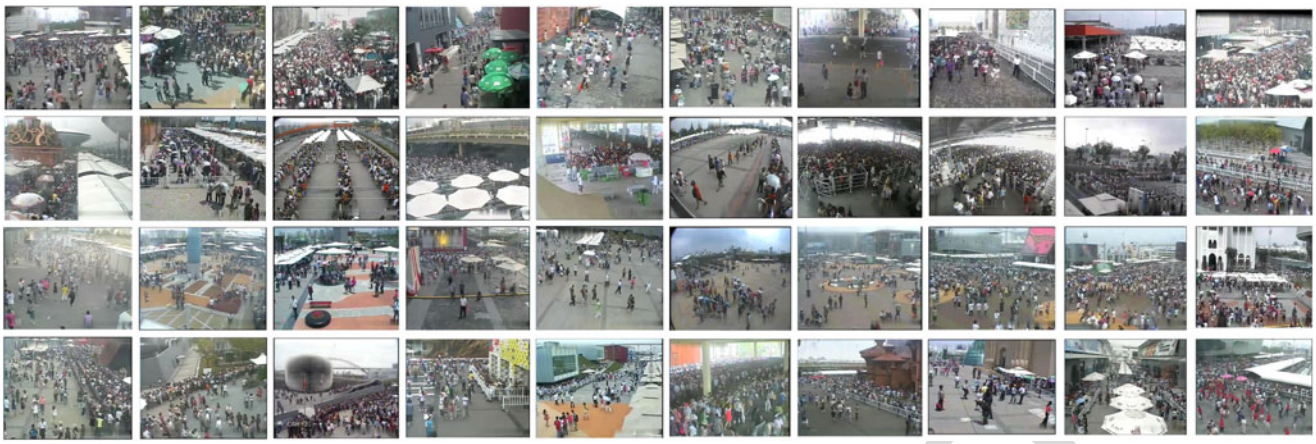


Fig. 2. Examples of four types of scenes from our dataset: road (first row), queue (second row), square (third row), and mixed (last row).

TABLE I
COMPARISON OF DIFFERENT CROWD DATASETS

	UCF [2]	Data-driven [55]	CUHK [70]	Flickr [26]	PETS [21]	WorldExpo'10
Source	internet	internet	internet	internet	surveillance	surveillance
# of scenes	38	35	62	50	8	245
# of videos	38	35	413	50 Images	40	2,630
Resolution	480×360	480×360	various	various	720×576	720×576
Annotation type	n/a	n/a	video-level	frame-level	frame-level	region-level
# of annotation	0	0	50	50	4000	160, 911
Task	segmentation	tracking	collectiveness	counting	counting	segmentation, density, collectiveness, cohesiveness

Videos or images in the CUHK [70] and Flickr [26] datasets do not have uniform resolutions.

approach and compare it with state-of-the-art methods. Finally, the future works are discussed in Section VI and conclusion is drawn in Section VII.

II. RELATED WORKS

Crowd datasets. A number of crowd datasets [2], [3], [11], [15], [21], [33], [38], [43], [55], [64], [70] have been released in recent years. They are designed for specific tasks. Since many approaches are scene-specific, most of these datasets [3], [11], [15], [33], [38], [43], [64] have one or two scenes, and cannot be used to study generic crowd understanding. Courty *et al.* [16] proposed the AGORASET dataset which contains eight three-dimensional (3-D) synthetic scenes of walking pedestrians. However, real-world surveillance videos are much more challenging and realistic for research and evaluation. Table I compares our proposed dataset to existing ones with more than five scenes. Most of them are collected from the internet. The crowd counting dataset [26] only contains 50 static images from Flickr. The PETS [21] dataset was collected by eight cameras with overlapping views on a campus. Both the above datasets annotate the total number of persons in each image/frame. The UCF [2] and data-driven [55] datasets do not provide any annotation. The CUHK dataset [70] provides collectiveness annotation for each video sequence. Since each video contains multiple groups with different collective behaviors, it is more accurate to annotate collectiveness of each crowd region as in our dataset. None of the previous datasets provide annotations on crowd segmentation, density, collectiveness and cohesiveness simultaneously.

Crowd segmentation. Crowd segmentation is an important step for crowd counting, tracking and behavior analysis. It is typically conducted through background subtraction [11], [12], [14], [15], [34], optical flow estimation [2], [38], feature point tracking [68], pedestrian detection [24], [54], [67], and SVM classifier [4]. All these approaches have major limitations in practice. For instance, some areas in the scene might be occupied by crowds for long periods and the background is invisible. Fig. 3 shows example results of crowd segmentation by some above mentioned approaches. Background subtraction does not work well when it is difficult to model and update the background [see Fig. 1(a)]. Background modeling generates a lot of false alarms due to the changes of lightings, scene clutters, and nonhuman foreground objects. Optical flow estimation and feature point tracking do not work well when the crowds are stationary or move slowly, or the video quality is low [see Fig. 1(b) and (c)]. These motion-based approaches do not utilize crowd textures that can be used to distinguish other image regions. Appearance-based pedestrian detectors perform poorly on extremely dense crowds because of heavy occlusions and small pedestrian sizes [see Fig. 1(d)]. In comparison, our data-driven approach works on crowd patches and uses both appearance and motion features.

Fig. 3. Example results by different crowd segmentation methods. (a) Result by background subtraction used in [11]. (b) Result by optical flow. (c) Result by feature point tracking with the KLT tracker [68]. (d) Result by pedestrian detection used in DPM [19]. (e) Result by our data-driven approach.

Crowd counting and density estimation. A number of works [11], [14], [15], [26], [42] directly estimate the number of pedestrians in a whole image without detecting individual persons. Most of them require annotating training frames of the target scene and rely on crowd segmentation. Meanwhile, several methods [5], [22], [37] are proposed to estimate crowd density based on pedestrian localization results. However, these methods are also scene-specific and not applicable to cross-scene crowd understanding in large-scale surveillance applications.

Crowd behavior analysis. Many approaches [32], [33], [38], [41], [43], [46], [64] for crowd behavior analysis learn motion patterns for a target scene. They are not scene-independent and do not capture universal properties of crowd behaviors. Zhou *et al.* [69] measured the collectiveness using crowd manifolds and compared it across different crowd scenes. Li *et al.* [39] surveyed some state-of-the-art techniques on crowd behavior analysis, including available features, existing models and evaluation protocols.

Data-driven approaches. Several works [40], [55], [60] were proposed to solve pixel-wise or superpixel-wise classification tasks via dense image matching. Such nonparametric and data-driven approaches are suitable for large-scale data because they do not need any training. They transfer the required information from the training images to the query via dense image matching. Liu *et al.* [40] proposed a nonparametric image parsing method by recovering dense deformation fields between the query and training images, and it can work with an arbitrary set of labels. A simpler yet more effective nonparametric approach is proposed in [60], where the label transfer is achieved by superpixel-level matching with local features. A data-driven method is also adopted for crowd tracking in [55] to search for similar behaviors among crowd motion patterns in other videos. The key of these data-driven approaches is to design effective global and local features to match query and training images, which is also the focus of our proposed approach.

III. WorldExpo'10 Crowd Dataset

We contribute a large-scale benchmark dataset for understanding crowd. All the videos are shot with actual surveillance cameras from Shanghai 2010 WorldExpo, which was the world's largest fair site ever with an area size of 5.28 square km. Over 73 million people have visited during six months and nearly 250 pavilions were built at the expo site. The abundant sources of these surveillance videos enrich the diversity and completeness of the surveillance scenes. We define four challenges and evaluation protocols on this dataset: crowd segmentation, and estimation of crowd density, collectiveness and cohesiveness. It would significantly promote the research on crowd understanding.

A. Data Collection

A huge amount of crowd videos were collected from Shanghai 2010 WorldExpo from June to October 2010. A total of 2630 video sequences from 245 cameras with disjoint views are selected. Each camera has 10–12 videos, one of which was collected at night, and at least two in each month. Each sequence lasts one minute (3000 frames), and the data size is 40 GB. Cameras were mounted on the top of buildings and had far-field views. The resolutions of videos are 720×576, which is higher than or comparable to existing datasets (Table I). The data was collected under various weather conditions: sunny, cloudy, and rainy (pedestrians held umbrellas on rainy days). All the scenes generally fall into four categories: road, square, queue at entrances, and mixture of the previous three types of scenes (e.g., the bottom-right image in Fig. 1 has both queue and crowd in square). Generally, crowds in queue or on road tend to have higher collectiveness, while crowds in queue tend to have higher cohesiveness. Examples are shown in Figs. 1 and 2.

B. Annotation

A professional labeling company was hired and 20 labelers were trained for the annotation task. Three frames were uniformly sampled from each sequence for annotation. Before labelers annotate a frame, they first browsed its surrounding frames to observe moving objects. The boundaries of crowd regions are drawn with polygons as shown in Fig. 1. Each crowd region is labeled with three properties: density, collectiveness and cohesiveness. Each crowd property is labeled as one of the three levels: low (1), medium (2), and high (3). The property of background regions is always labeled as 0.

The annotation rule for crowd segmentation is as follows. Every person has his or her own territory which is a circle with a radius of one meter.[3] If the territories of two persons overlap, the two persons are connected. A crowd region covers a connected component of multiple persons.

Crowd density is annotated with the widely used Jacobs's method [29] proposed in social science, which classifies density into three levels. It counts the average number (n) of persons

[3]The "one meter" for each person is empirically determined by the labeler as 2/3 of the person's height.

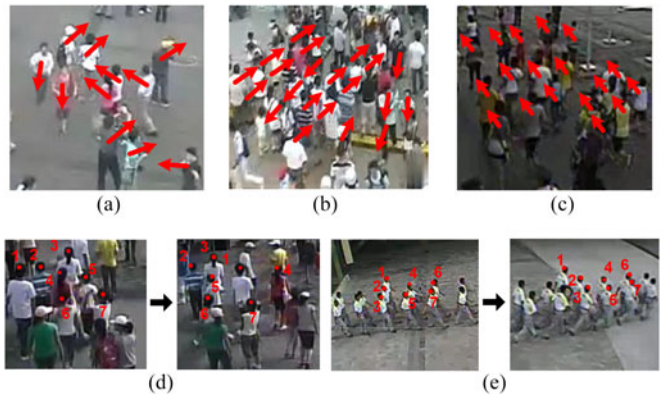


Fig. 4. Illustration of different levels of collectiveness and cohesiveness. (a) Low collectiveness. (b) Medium collectiveness. (c) High collectiveness. (d) Low cohesiveness. (e) High cohesiveness.

in every square meter. A scene is sparse if $n \leq 1$, medium if $1 < n \leq 2$, or dense if $n > 2$. Since crowd is not uniformly distributed in a scene, we empirically modified this rule to make it easier for annotation. Within a segmented crowd region, if the territory of a person includes another $0 \leq m \leq 2$ persons on average, this crowd region is annotated as sparse. Similarly, it is labeled as medium if $2 < m \leq 5$, and dense if $m > 5$. This is consistent with the Jacobs's method [29], since the area of a person's territory is around 3 square meters. Our annotation rule also implicitly considers crowd size. If a crowd region only has three persons, it is always labeled as sparse, even if all three stand tightly within one square meter, because there are no more than two persons in the territory of another person. Examples of density annotations are shown in Fig. 1.

Collectiveness and cohesiveness have been widely studied in physics [10], [25], [61] and sociology [8], [36], [49] for a long time. There is no explicit mathematical definition on crowd collectiveness and cohesiveness. Therefore, collectiveness and cohesiveness of our dataset's samples are defined in a subjective manner. For each sample, we have the same multiple human labelers to annotate its collectedness and cohesiveness (e.g., low=1, medium=2, high=3), and the average of their annotations is used as the final label. Fig. 4 shows examples of our definition on different levels. The collectiveness of Fig. 4(a) is labeled as low, since the pedestrians move in different directions without the same goal. In Fig. 4(b), a few crowd groups move in opposite directions and its collectiveness is labeled as medium. In Fig. 4(c), all the persons move in the same direction and the collectiveness is high.

Cohesiveness measures the stability of local geometrical and topological structures of crowd groups. Fig. 4(d) shows the same crowd at different frames. The topological structure of its members has changed significantly, and therefore the cohesiveness is low. Fig. 4(e) shows an example with high cohesiveness. Note that high collectiveness does not mean high cohesiveness. If a group of people move in the same direction but with very different speed, their local structures cannot remain stable.

Fig. 5 shows the histograms (on the area of crowd regions) of the three properties for the four type of scenes. According

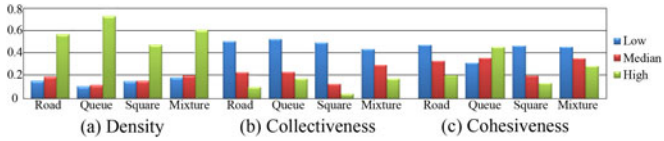


Fig. 5. Statistics of three properties in different crowd scenes (blue = low, red = medium, green = high). (a) Density. (b) Collectiveness. (c) Cohesiveness.

TABLE II
PARTITION OF TRAINING AND TEST SETS

	Road	Queue	Square	Mixture	Total
$N_{s/train}$	83	38	41	30	192
$N_{s/test}$	20	10	12	11	53
$N_{s/total}$	103	48	53	41	245
$N_{c/train}$	809	394	464	386	2053
$N_{c/test}$	207	103	147	120	577
$N_{c/total}$	1016	497	611	506	2630

N_s indicates the number of scenes and N_c indicates the number of video sequences.

to our statistics, around 75% regions are background and the remaining 25% regions are crowds. Most of the crowd regions in our dataset have high density. Generally, road and queue crowd scenes with strict man-made constraints have higher collectiveness and cohesiveness than open scenes such as square. Especially, in queue scenes, people are kept within some bounds, and most of the crowd regions have high cohesiveness.

C. Evaluation Protocols

80% of the data is partitioned for training and the other 20% for testing. The two subsets have no overlap on scenes or video sequences, In this way, the methods' capability of handling unseen scenes can be well evaluated. On the test set, we attempt to make data distribution more balanced on the four types of scenes. Detailed statistics are shown in Table II. Each crowd region in the test set was annotated by five labelers and we use the average of their scores. Since the training set is much larger, we cannot afford the cost of labeling each crowd region for multiple times. Although each crowd region is only labeled by one labeler, the whole training set is labeled by 20 labelers. The bias introduced by individual labelers can be reduced to some extent, because the learning process is based on the whole training set. Four evaluation criteria on the test set have been set for the proposed challenges.

Crowd segmentation. Every pixel in an annotated frame has a label: background (0) or crowd (1). ROC curve is used to evaluate the performance of crowd segmentation.

Crowd density estimation. Every pixel has an annotated density score ranging from 0 to 3. 0 indicates background and no crowd exists, while 3 indicates dense crowd. The estimation algorithms are expected to output continuous density scores. The Mean Square Error (MSE) is used for evaluation, and is computed as

$$\text{MSE} = \frac{1}{N_{\text{test}} N_I} \sum_{i=1}^{N_{\text{test}}} \sum_{p \in I_i} (\hat{l}_p - l_p)^2 \qquad (1)$$

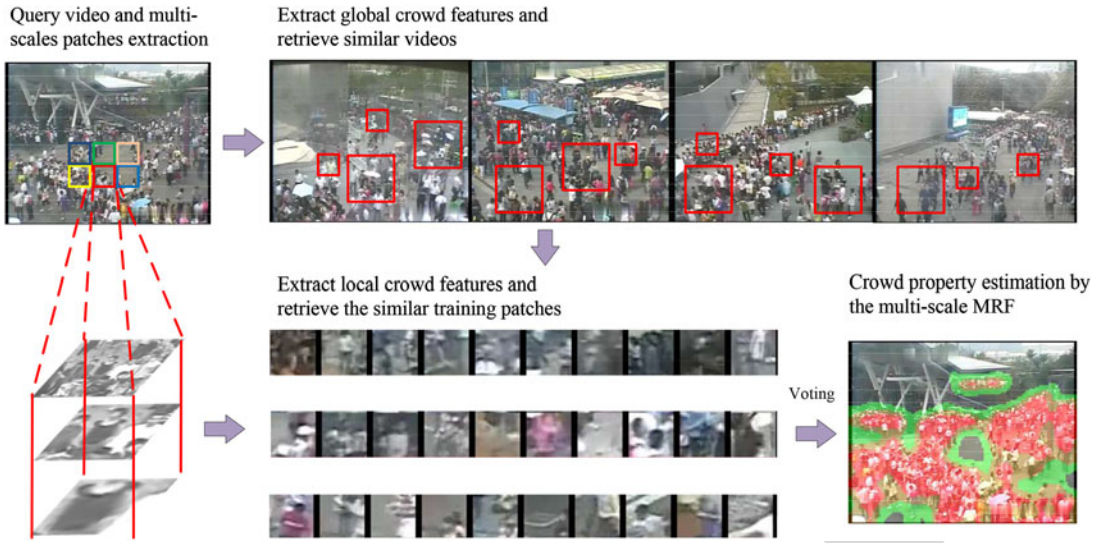


Fig. 6. Illustration of our proposed data-driven crowd understanding method.

where for pixel p in frame I_i, the ground-truth annotation is l_p, the predicted output is $\hat{l}_p$, N_{test} is the number of test samples and N_I is the number of pixels for image I_i.

Collectiveness and cohesiveness estimation. Since it is not reasonable to estimate collectiveness or cohesiveness of background, we only use manually segmented crowd regions for evaluation. The scores of both properties are in the range of 1 to 3. Similar to crowd density estimation, the MSE is used for evaluation.

IV. Data-Driven Crowd Understanding

We propose a data-driven crowd understanding approach as the baseline for our dataset. Different from most scene-specific crowd understanding methods, the data-driven method can be applied to any unseen scene without extra labeling and training. Data-driven approaches [40], [60] have achieved great success on scene understanding, which transfer the annotations of training data to test samples via dense pixel-level or superpixel-level image matching. Our large-scale annotated training set makes it possible for us to develop a data-driven approach as a baseline for our crowd understanding dataset.

A. Overview of the Proposed Method

In order to automatically annotate a query frame, the key of our data-driven method is to retrieve the most similar samples from training set and transfer their labels to the query via dense image matching. Fig. 6 illustrates the overall framework of our proposed method. In our framework, a short video clip including 30 frames surrounding the query frame is extracted as input. The training video clips are generated in the same way. To transfer labels only from training video clips that are similar to the query, the most similar scenes to the query video clip are first retrieved from the training set based on the global crowd feature as the candidate scene set. Then multi-scale crowd patches are extracted in a sliding window fashion with 50% overlap from the query video. For each patch, the most similar patches are retrieved from the candidate scene set based on local crowd features. Therefore, the key is to design effective global crowd feature to retrieve similar scenes and local crowd feature to match similar patches. Instead of using existing generic features, we learn crowd features and the optimal combination weights of different components based on training crowd videos. The crowd properties of each pixel can then be estimated by average voting. The multi-scale MRF is utilized to ensure the smoothness of the resulting crowd property map.

B. Global Crowd Feature for Candidate Scene Retrieval

For each patch in the query frame, it is costly to search among millions of crowd patches in the whole dataset for the most similar training patches. Therefore, it is more efficient to first retrieve a small set of candidate training video clips most similar to the query clip and match training patches within this subset. A global feature is needed to describe the whole crowd scene. One commonly used scene feature GIST requires convolving each image with a set of Gabor filters. However, there is no filter specifically designed to describe crowd scenes. Therefore, for our global crowd feature, we train mid-level filters to effectively describe the content of a crowd scene (see Fig. 7).

Mid-level crowd filters. Mid-level feature learning has been exploited in recent works on several vision topics, such as scene classification [59] and action recognition [30]. But existing works on mid-level feature learning did not consider the special properties of crowd understanding. The crowd property of a patch would significantly influence its appearance. Our goal is therefore to train discriminative mid-level filters that are able to distinguish patches of different appearance. We first group patches into several clusters with similar visual appearance. 16 000 spatio-temporal patches are uniformly sampled from crowd regions for clustering based on their ground truth crowd density, collectiveness and cohesiveness. In this way, the sampled patches have good diversity. The affinity propagation (AP) clustering method [23] is adopted because it does not

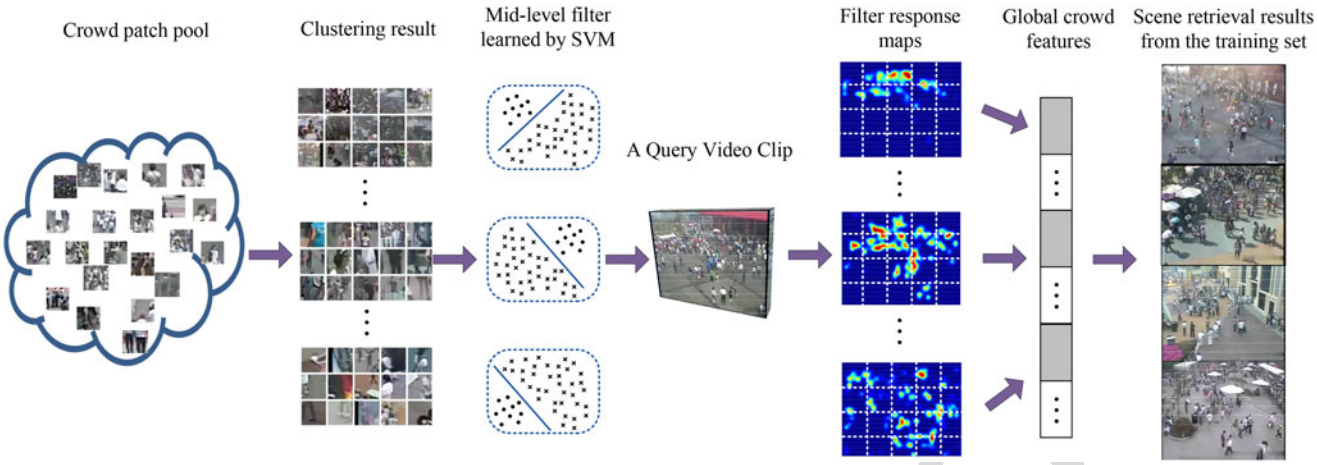


Fig. 7. Global crowd feature for scene retrieval. Red regions in the response map have high response value while blue regions are low. The retrieved scenes have similar views. Dense crowds are in areas farther to cameras.

require the number of clusters to be estimated in advance. For our dataset, $N_c = 30$ clusters are obtained for our training set by the AP algorithm. To capture distinctive appearance patterns to describe patches of different clusters, a discriminative filter is learned for each crowd cluster. For each cluster k, all the patches assigned to this cluster are regarded as positive samples, and patches from the other clusters and background are randomly sampled to form the negative samples. The number of negative samples is set 10 times as many as the positive samples. After creating the positive and negative patch sets, a linear SVM classifier $\{w_k, b_k\}_{k=1}^{N_c}$ is trained for every cluster. The SVM weights w_k and bias term b_k serve as the kth crowd mid-level filter. A response score map is obtained when crowd mid-level filters are used to convolve with a query video clip.

Global crowd feature. Global crowd feature is designed to describe the properties of the whole crowd scene for scene retrieval. Therefore, global crowd feature is extracted from the whole response maps generated by the mid-level filters for each video clip. The response maps are divided into $N_x \times N_y$ cells with no overlap. We set $N_x = 4$ and $N_y = 5$ for our proposed dataset. The average response scores of each grid is calculated. Such scores of all the filter response maps of N_c filters is concatenated as the global crowd feature to calculate its similarity between different scenes and to retrieve similar training scenes for a query (see Fig. 7). The total dimension of our global crowd feature is therefore $N_x \times N_y \times N_c = 600$.

C. Local Crowd Feature for Patch Matching

To distinguish different crowd properties, local crowd feature should describe both appearance and motion information at multiple scales. Therefore, our local crowd feature includes eight appearance and motion features extracted from each 3-D spatio-temporal crowd volume.

Multi-scale augmentation. Video surveillance data has large perspective variation, and crowds can be observed at different scales. In order to augment the training set and increase the robustness of matching with query patches, we sample both training and test patches at multiple scales and normalize them to the same size ($36 \times 36 \times 30$) as shown in Fig. 6.

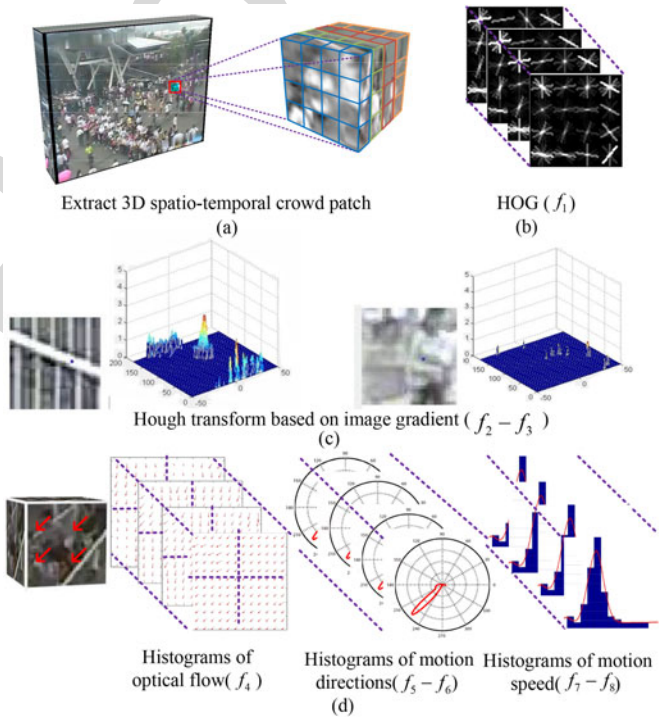


Fig. 8. Local crowd feature. (a) Uniformly sampling four frames from a 3-D crowd volume and dividing them into 3×3 cells. (b) Extracting HOG at each sampled frame. (c) Features computed from Hough transform are used to distinguish crowd patches from man-made patches with long line structures. (d) Examples of crowd patches with coherent (top) and incoherent (bottom) motions. Each sampled frame is divided into four sub-regions. Histograms of motion directions (third column) and speed (fourth column) are computed in each sub-region and the whole region.

Appearance features. The first feature f_1, HOG [17], is extracted from each sampled patch, as shown in Fig. 8(a) and (b). 4 frames are uniformly sampled from a 3-D patch, and each frame is divided into 3×3 cells with 50% overlap. The size of each cell is 18×18. Empirically, we observe that HOG cannot

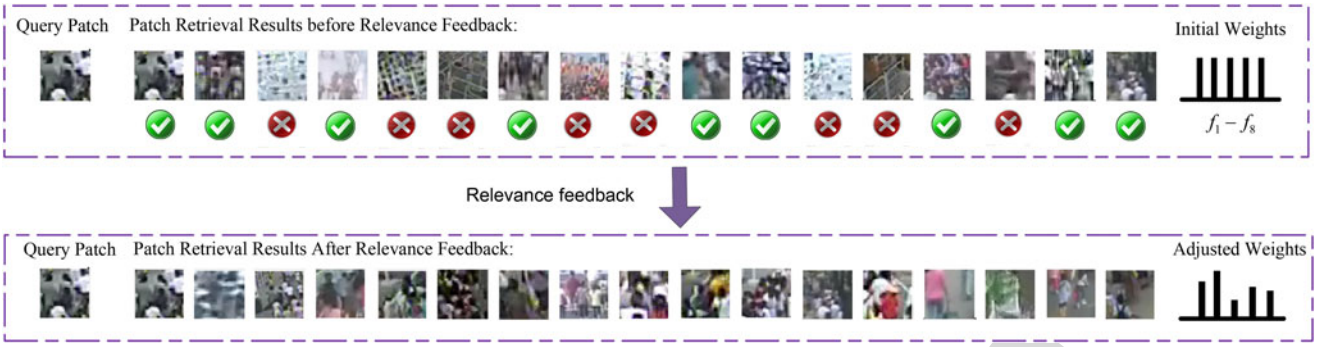


Fig. 9. Using relevance feedback to learn the optimal weights for the local crowd feature.

distinguish crowd patches with some man-made background patches with long line structures [such as fence in Fig. 8(c)], which are commonly observed in crowd scenes. We design two features (f_2 and f_3) with Hough transform to capture line structures. Traditional Hough transform is based on edge detection operators, such as Canny and Sobel. However, these edge operators ignore much texture information, especially for the surveillance video patch with relatively low resolution. Therefore, we perform Hough transform on image gradient map

$$r(\theta) = x_0 \cos\theta + y_0 \sin\theta \tag{2}$$

where θ is determined by gradient $(\Delta x, \Delta y)$ at (x_0, y_0) and is calculated as $\theta = \frac{\Delta y}{\Delta x} + \frac{\pi}{2}$. After applying a Gaussian filter, a response map $M(r, \theta)$ in the polar coordinate is obtained, as shown in Fig. 8(c). The feature $f_2 = [\mu_v, \sigma_v]$ characterizes vertical lines. μ_v and σ_v are the mean and variance of responses in the range $\theta \in [0, 10^\circ] \cup [170^\circ, 180^\circ]$. The feature $f_3 = [\mu_a, \sigma_a]$ characterizes the longest line in any direction. The mode of the highest peak is detected with the mean shift algorithm and its mean and variance are μ_a and σ_a.

Motion features. To characterize local motion of the sampled patch, the feature f_4, Histogram of Optical Flows (HOF) [35], is computed on the same sampled frames and cells. The same parameter setting is adopted as HOG. In order to further characterize whether individuals in crowd move in similar directions and keep stable local structures, the features $f_5 - f_8$ are computed based on the histograms of motion directions and speed as shown in Fig. 8(d). They are the entropy and variance of the two types of histograms. The patch at each frame is divided into four sub-regions. Histograms of the four sub-regions and of the whole region are computed. Note that besides $f_5 - f_8$, f_1 and f_4 at sampled frames are also useful for estimating collectiveness and cohesiveness, since they characterize how appearance and motion change over time.

Learning feature weights. The $f_1 - f_8$ features are concatenated as the local crowd feature. The distance between a training patch x_i and a query patch x_q is then computed as

$$d(x_i, x_q) = \sum_{k=1}^{8} \omega_k \|f_{\mathrm{ik}} - f_{\mathrm{qk}}\|_2 \tag{3}$$

where f_{ik} and f_{qk} is the kth local crowd feature of the patch x_i and x_q. It is important to assign a set of optimal weights $\{\omega_k\}$ to weight the importance of the eight features. We do not use the annotated labels in the training set to learn the weights, because it might make our crowd feature overfit to a particular task. Instead, we choose a relevance feedback approach to learn the weights that most match human perception. The weights learned in this way are more general and can be applied to various crowd understanding tasks.

It starts with uniform weights. Some examples of matching results with uniform distribution were shown Fig. 9. At each iteration t, a patch $x_q^{(t)}$ is randomly selected from the training set and is tried to match with other training patches $x_i^{(t)}$ using the current weights. Top N matches are presented to a labeler, who labels each of them as similar (1), dissimilar (-1), or uncertain (0) based on visual perception (Fig. 9). Based on the feedback, the feature weights are adjusted with adaptive SVM [65] as

$$d^{(t+1)}(x_i, x_q) = d^{(t)}(x_i, x_q) + \sum_{k=1}^{8} \Delta\omega_k^{(t)} \|f_{\mathrm{ik}} - f_{\mathrm{qk}}\|_2 \tag{4}$$

where $d^{(t)}$ represents the distance function at iteration t, and $\Delta\omega^{(t)}$ are the parameters estimated from the feedback examples at iteration t. To learn the parameter $\Delta\omega^{(t)}$, we adopted a SVM-like objective function

$$\begin{aligned} &\min_{w^{(t)}} \frac{1}{2}\left\|w^{(t)}\right\|^2 + C^{(t)} \sum_{i=1}^{N^{(t)}} \xi_i^{(t)} \\ &s.t.\ \xi_i^{(t)} \geq 0;\ C^{(t)} = \eta^{(t)}(1 - \eta^{(t)}) \\ &y_i d^{(t)}(x_i) + y_i \sum_{k=1}^{8} \Delta\omega_k^{(t)} \left\|f_{\mathrm{ik}}^{(t)} - f_{\mathrm{qk}}^{(t)}\right\|_2 \geq 1 - \xi_i^{(t)} \\ &\forall (x_i, y_i) \in D^{(t)} \end{aligned} \tag{5}$$

where $\sum_{i=1}^{N} \xi_i^{(t)}$ measures the total classification error of the tth feedback iteration. The cost factor $C^{(t)}$ represents the discriminative capability of the current iteration data to balance the contribution of previous iterations. So we define the $C^{(t)}$ as $C^{(t)} = \eta^{(t)}(1 - \eta^{(t)})$, where $\eta^{(t)}$ is the accuracy of feedback results at iteration t. $\eta = 1$ or 0 means all the feedback results are similar patches or dissimilar patches, which would not improve the retrieval results, and result in the lowest value of C. Oppositely, an equal number of positive samples and negative samples would lead to optimal weights.

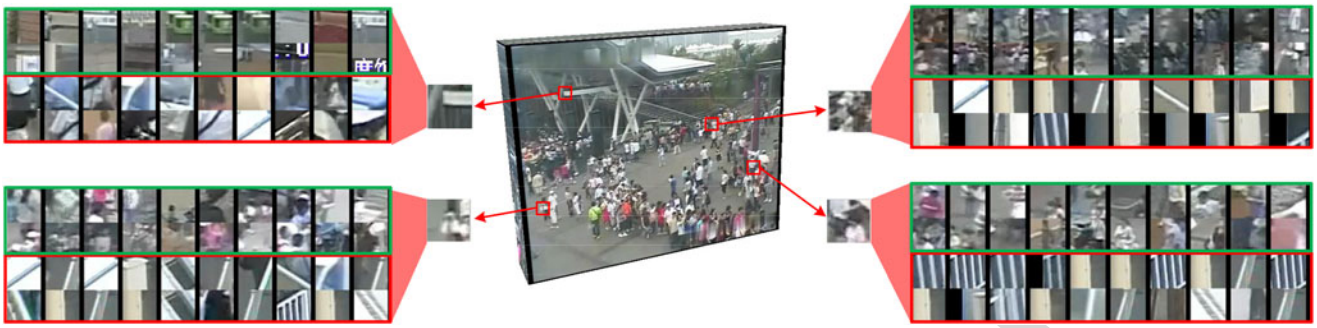


Fig. 10. Four query patches are selected from a query video clip. Their most similar (two upper rows) and dissimilar (two bottom rows) training patches based on our local crowd feature are shown in green and red rectangles, respectively.

At every iteration, a different query patch is randomly chosen. The iterations stop when the weighted features well match human perception and cannot be further improved. Some examples of matching results are shown in Fig. 10. The learned weights well distinguish background and crowds of different density levels.

D. Crowd Property Estimation

For each query video clip, we first retrieve its $M = 30$ most similar training video clips according to the global crowd feature. In addition, no more than three video clips are from the same scene to ensure their diversity. Image patches from the candidate training clips form the pool of candidate training patches. For the query video clip (30 frames), crowd patches are sampled at $S = 3$ different scales. For each patch p at scale s, its observed label score $\hat{l}_p^s$ is the averaged label score of its top $K = 20$ matched training patches from the candidate training patch pool.

The final label scores $\{l_p^s\}$ are obtained by the multi-scale MRF [26] to ensure smoothness. The graph can be represented by (V, E), where V are the pixel nodes and E are the neighbors at the same level and intermediated nodes that connect a patch to layers above and below it. The energy function with S level scales is thus given by

$$\min_{l} \sum_{s \in S} \left(\sum_{p \in V_s} D(\hat{\eta}_p^s, l_p^s) + \sum_{(p,q) \in E} V(l_p^s - l_q^s) \right) \qquad (6)$$

where l_p^s represents the estimated property of patch p at scale s, and q is the spatial neighbor of patch p. The data term is defined as $D = |\hat{\eta}_p^s - l_p^s|$, where $\hat{\eta}_p^s = \frac{1}{2}(\hat{l}_p^{s+1} + \hat{l}_p^s)$ is of the bottom two scales and $\hat{\eta}_p^s = \hat{l}_p^s$ is of the top scale. The smoothness term is defined as $V = \min(|l_p^s - l_q^s|, \varepsilon)$, which enforces the smoothness between the neighboring nodes. This multi-scale MRF model is optimized using the Max-Product Belief Propagation method on grid structure [20].

V. Experimental Evaluation

We evaluate our data-driven approach for different crowd understanding tasks, including crowd segmentation (Section V-A), crowd density estimation (Section V-B), and crowd collectiveness and cohesiveness estimation (Section V-C) on the WorldExpo'10 dataset and compare it with other methods. The evaluation metrics were explained in Section III-C. For the test set, the patches are extracted in a sliding window fashion with 50% overlap in three scales, 36×36, 72×72, and 144×144, respectively. The estimated property of each pixel is obtained by averaging all the predictions of overlapping patches. The extensive experimental results by our proposed method and the compared ones on crowd segmentation, crowd density estimation, and crowd collectiveness and cohesiveness estimation demonstrate our method's capability of handling unseen scenes.

A. Crowd Segmentation

We compare our proposed data-driven approach (*Data-driven*) with six other crowd segmentation methods. The ROC curve is used to evaluate the performance of crowd segmentation. The following approaches are compared.

1) *SVM (Codebook)*: To the best of our knowledge, the only existing method specifically designed for crowd segmentation is [4]. The proposed method modeled crowd texture with a codebook. The SIFT features are extracted from interest points in frames. The codebook of size 1000 is built through k-means clustering on the SIFT feature. Crowd-likelihood features are computed based on the codebook as in [4] and used to classify each patch with SVM with a RBF kernel.

2) *BS*: Background subtraction is used by many crowd understanding works [11], [12], [15], [42] to segment crowd. The method used in [12] is chosen for comparison.

3) *Deformable Parts Model (DPM)*: Pedestrian detection approaches might also be used for crowd segmentation. A state-of-the-art pedestrian detector with DPM [19] is applied to test frames. It is trained on the INRIA dataset [17]. A pixel is segmented as crowd if it falls into a pedestrian window. We also compare with two baselines to evaluate the effectiveness of the components of our proposed method.

4) *SVM (HOG)*: This baseline follows the same framework as [4] but utilizes the HOG as the features to describe crowd, which is a popular descriptor for pedestrians. 5) *SVM Local Crowd Feature (LCF)*. To evaluate the performance of our proposed data-driven classifier, we also create a baseline that utilizes SVM and our proposed LCF feature. For methods 1), 4) and 5), we select 192 clips from every training scenes with

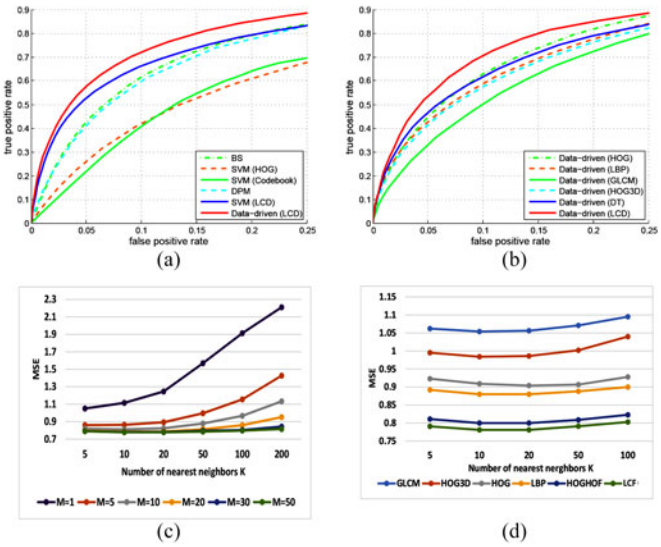


Fig. 11. (a) ROC curves of crowd segmentation results by different methods. (b) ROC curves of crowd segmentation results by using different local features in our proposed data-driven framework. (c) MSE of density estimation by our proposed framework with varying M and K. (d) MSE of density estimation by our proposed framework with varying K and different local features.

medium density distribution as the training data. The SVM is trained with approximately 230 000 patches. For fair comparison, when comparing with methods 1), 4) and 5), our data-driven method is trained with the same training set.

As shown in Fig. 11(a), our data-driven method with the proposed LCF feature outperforms the compared methods. Although BS works better than several other methods, it still performs worse than our data-driven approach. This is because background subtraction methods utilize motion information and cannot handle crowds that move slowly or are stationary. It is also affected by scene clutters. When setting the false positive ratio to 0.1, the true positive ratio of BS is 10% lower than that of ours. The pedestrian detector (DPM) does not work well neither because of severe occlusions. We also observe that when using the same classifier, i.e., SVM, our proposed local crowd feature significantly outperforms the widely used HOG and SIFT features. However, using the SVM classifier with our proposed feature is still inferior to the proposed data-driven approach, which is more effective on handling the complex distributions of crowd and background patches.

In order to further evaluate the effectiveness of our proposed local crowd feature, we compare our LCF feature to different local features by using them as the local feature in our proposed data-driven framework. The compared features include HOG [17], Local Binary Patterns (LBP) [50], Gray-Level Co-occurrence Matrix (GLCM) [44], HOG3D [31] and Dense Trajectory (DT) [63]. The general appearance features, such as LBP, GLCM and HOG, are widely used for general texture description and crowd understanding. HOG3D and DT are utilized for spatio-temporal description and achieve satisfactory performance on action recognition and crowd behavior understanding. We utilize the recommended parameters for all the compared features. Fig. 11(b) shows the ROC curves of different local features, where our proposed LCF feature outperforms other local features. LCF is more effective to describe the crowd characters. Note that DT obtains better performance than other texture features, which shows that motion information is important for the crowd segmentation task. But the general spatio-temporal features, such as DT and HOG3D, are not effective on describing crowds.

TABLE III
MSE OF CROWD DENSITY ESTIMATION BY REGRESSION-BASED METHODS (LEFT COLUMN) AND OUR PROPOSED DATA-DRIVEN METHODS WITH DIFFERENT LOCAL FEATURES (RIGHT COLUMN)

Method	MSE	Method	MSE
HOG+RR	1.10	Data-driven (HOG)	0.94
GLCM+GPR [11]	1.07	Data-driven (GLCM)	1.03
LBP+KRR [15]	0.98	Data-driven (LBP)	0.91
Lempitsky [37]	1.31	Data-driven (Ours)	**0.71**

B. Crowd Density Estimation

Our propose data-driven framework can also be utilized to estimate crowd density. The MSE (1) is used as the evaluation criterion. We compare our proposed method with some state-of-the-art regression based methods. All the major components in our methods are also evaluated. At last, we also discuss parameter selection and computational cost of our data-driven method.

Comparison with regression-based methods. We compared our proposed framework with several regression-based methods to estimate crowd density of each patch [11], [15], [37], [54]. They were originally proposed for crowd counting but can be used to estimate density in a similar way.

Gaussian Processes Regression (GPR) with GLCM feature was used for crowding counting [11]. Similarly, Kernel Ridge Regression (KRR) with LBP feature was adopted in [15]. Lempitsky [37] proposed a crowd density estimation approach that uses SIFT and regularized linear regression, which was also used in [54]. We also use the widely used HOG feature with the basic Ridge Regression (HOG+RR) as a baseline. The density estimation results of all the methods are listed in Table III. For fair comparison, we use the same training data for both regression-based methods and our data-driven method, which means that the step of candidate scene retrieval based on the global crowd feature is skipped in our method. Instead, we perform the local patch matching on all training data.

Our approach achieves the highest accuracy among all the compared methods. Most of these regression-based methods are scene-specific, and models learned from a particular scene can only be well applied to the same scene. From Table III, it is obvious that they do not show satisfactory performance in the large-scale dataset. In contrast, data-driven methods are more suitable for the large-scale and dynamic dataset. Some examples of our results are shown in Fig. 12.

Experiments are also conducted on the popular UCSD dataset [11] and MALL dataset [15], which are widely used to evaluate crowd counting and crowd density estimation. Pedestrians' positions are labeled for each scene of the two dataset. Followed by the Jacobs's method mentioned in III-B, the

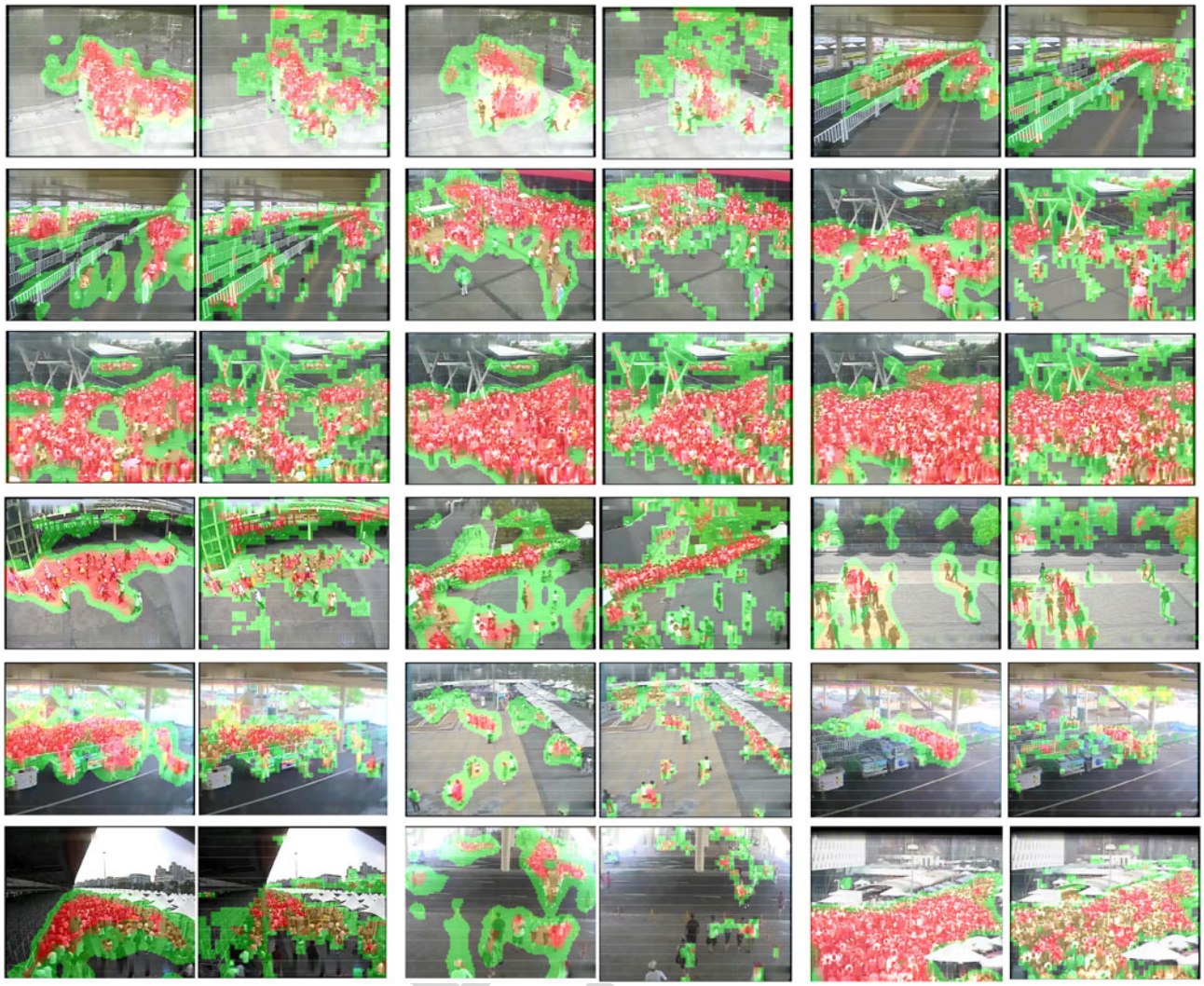


Fig. 12. Example density estimation results (red = dense, yellow = medium, green = sparse) by our data-driven framework (odd columns) and by the regression-based method LBP+KRR (even columns), which is the best regression-based method.

TABLE IV
MSE OF CROWD DENSITY ESTIMATION BY REGRESSION-BASED METHODS AND OUR PROPOSED DATA-DRIVEN METHODS ON THE UCSD DATASET AND MALL DATASET

Method	UCSD [11]	MALL [15]
Lempitsky [37]	1.07	1.16
GLCM+GPR [11]	0.73	0.91
LBP+KRR [15]	0.62	0.84
Data-driven (Ours)	**0.54**	**0.77**

TABLE V
MSE OF CROWD DENSITY ESTIMATION BY THE DATA-DRIVEN FRAMEWORK WITH DIFFERENT GLOBAL AND LOCAL CROWD FEATURES

	HOG	LBP	GLCM	HOG3D	DT	HOGHOF	LCF (UW)	LCF
GIST	1.08	1.08	1.21	1.15	1.10	0.96	0.98	0.94
GIST+MRF	0.93	0.95	1.08	1.00	0.93	0.84	0.88	0.82
GCF	1.06	1.01	1.19	1.16	1.02	0.93	0.94	0.89
GCF+MRF	0.90	0.88	1.05	0.98	0.85	0.80	0.82	**0.78**

density level annotations can be generated from the position labels. Both regression-based methods and our data-driven method are only trained from the training set of our World-Expo'10 dataset. Following, the same test partition as in [11] and [15]. Most of the regression-based methods are scene-specific, and our proposed method outperforms all the compared methods in these two datasets as shown by the results in Table IV. The results demonstrate that our proposed method is able to handle unseen target scene with our large-scale training dataset.

Evaluation of individual components. The effects of different local features are first investigated by using them in the proposed data-driven framework. Notice that spatio-temporal features, such as DT and HOGHOF [35], have better performances than texture features because of the additional motion information. Our LCF outperforms all other compared features. In addition, our LCF feature with uniform weights was compared as a baseline to demonstrate the necessity of the relevance feedback scheme as Eq. (5). We then compare our GCF for candidate scene retrieval with the GIST feature (Table V).

We observe that our GCF based on the mid-level crowd filters generate more accurate candidate scenes for label transfer with different local features. We also test the effect of the multi-scale MRF. Obviously, the MRF effectively improves the estimation accuracy because it enforces smoothness on the resulting label maps.

Parameter selection. The proposed data-driven system retrieves M most similar scenes for the query clip, and further selects K nearest neighbors for each query patch to estimate the crowd properties. We investigate the performance of our data-driven system by varying the parameters M and K.

Fig. 11(c) shows the MSE by setting different M and K. The performance improves as the number of candidate scenes (larger M) increases. The performance drops as K increases since more candidate patches may introduce noise to label transfer, especially when M is small. Although by conducting local patch matching in all training frames (equivalent to $M = 7000$), we obtain lower MSEs (as shown by the results in Table III). The computational time is proportional to M and using such large M is not practical for real-world applications. We also fix $M = 30$ and calculate the MSE of density estimation with varying K using different local features as shown in Fig. 11(d). Smaller K incorporates less information and larger K might result in more noise. We observed that $K = 20$ achieves the best performance for most types of local features. The balanced performance and computational cost are achieved when $M = 30$ and $K = 20$.

Computational cost. Our implementation is in MATLAB and is mostly parallelized. All our tests ran on a PC with a Core-i7 3.4 GHz quad core processor and 16 GB RAM. Our computational cost is mainly dominated by the extraction of global and local crowd features, which costs nearly 100 s for every query clip. But it can be easily sped up by utilizing more powerful hardware and better parallelization. Labeling one query clip with candidate scene retrieval and local patch matching takes less than 10 s. The main bottleneck of our implementation is file I/O for loading retrieval set features from hard disk. More appropriate data structure and larger RAM would improve the effectiveness of our implementation.

C. Collectivness and Cohesiveness Estimation

Our proposed method can also be extended to estimate collectiveness and cohesiveness. The MSE (1) is used as the evaluation criterion. Since it does not make sense to estimate collectiveness or cohesiveness on background, we only estimate annotated crowd regions for evaluation. We compared with the collectiveness measurement method proposed by Zhou *et al.* [70]. Since collectiveness and cohesiveness describe the motion information of crowds, we only compare our LCF feature with two spatio-temporal features, HOG3D [31] and HOGHOF [35].

Table VI reports the results of collectiveness and cohesiveness estimation by different methods. For collectiveness estimation, [70] does not work well if feature points cannot be well detected and tracked, especially when the video resolution is low. Our data-driven method performs robustly and does not rely on any detection and tracking. The experiment results show that our proposed crowd feature achieves better accuracy on collectiveness estimation than the other two spatio-temporal features. For cohesiveness estimation, there is no previous work on this topic. We only report the results by our data-driven framework with different local features. The data-driven method with our proposed crowd features also achieves the best performance.

TABLE VI
MSE OF COLLECTIVENESS AND COHESIVENESS ESTIMATION

	MSE
Zhou *et al.* [70] for collectiveness	0.71
Data-driven (HOG3D) for collectiveness	0.67
Data-driven (HOGHOF) for collectiveness	0.52
Data-driven for collectiveness (our LCF)	**0.49**
Data-driven (HOG3D) for cohesiveness	0.78
Data-driven (HOGHOF) for cohesiveness	0.66
Data-driven for cohesiveness (our LCF)	**0.64**

VI. DISCUSSION AND FUTURE WORK

The WorldExpo'10 dataset is a large-scale benchmark dataset for crowd understanding and covers a large variety of scenes with sufficient training data. Such training data would benefit learning algorithms specifically designed for big data, such as deep learning, data-driven approaches etc. Therefore, we hope the WorldExpo'10 dataset would become an important resource for more crowd video surveillance applications and can play a critical role in advancing the research on understanding crowds. We envision the following possible potential challenges:

Crowd counting. Most of existing crowd counting algorithms and datasets are scene-specific and focus on low density crowd. In comparison, the WorldExpo'10 dataset contains a large number of scenes with high variation of density. In the most crowded scenes, the number of pedestrians in a frame is close to one thousand. The crowd density also varies in a large range. Therefore, it is much more challenging and realistic to real-world surveillance applications. The baseline method of density estimation proposed in this paper would offer an important prior for crowd counting.

Abnormal event detection. Anomaly detection is an important problem in crowd understanding with extensive applications. In our high quality, diverse and large-scale WorldExpo'10 dataset, plenty of abnormal events can be observed and defined to evaluate and advance related research. The universal properties, density, collectiveness and cohesiveness, might be helpful for anomaly detection.

Crowd scene classification. We roughly summarize the crowd scenes in the WorldExpo'10 dataset into four categories. However, it can be further classified into many more categories based on different crowd behaviors, such as crowd gathering, crowd dispersing, crowd queuing, rushing, and loitering.

Deep learning has achieved great success in computer vision during recent years. However, so far little work has been done on deep learning for crowd understanding due to the lack of large-scale training data with annotation. This new dataset would significantly advance deep learning research in this area,

and more effective and discriminative crowd features and representations can be learned.

VII. Conclusion

In this paper, we contribute a large-scale annotated benchmark dataset including 245 scenes for cross-scene crowd understanding. Four challenges are proposed for this dataset based on their importance in scientific studies and crowd video surveillance applications. Benefiting from the large-scale training set, a data-driven approach with new global and local crowd features is proposed to solve crowd understanding tasks. It serves as a baseline for the proposed dataset and outperforms state-of-the-art approaches.

References

[1] I. Ali and M. N. Dailey, "Multiple human tracking in high-density crowds," *Image Vis. Comput.*, vol. 30, pp. 966–977, 2012.

[2] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–6.

[3] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 1–14.

[4] O. Arandjelovic, "Crowd detection from still images," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–10.

[5] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Interactive object counting," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 504–518.

[6] A. F. Aveni, "The not-so-lonely crowd: Friendship groups in collective behavior," *Sociometry*, vol. 40, pp. 96–99, 1977.

[7] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action representation and classification in unconstrained videos," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1234–1245, Aug. 2012.

[8] H. Blumer, "Collective behavior," *Principles of Sociology*. New York, NY, USA: Barnes & Noble, 1951.

[9] J. Buhl *et al.*, "From disorder to order in marching locusts," *Science*, vol. 312, pp. 1402–1406, 2006.

[10] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Rev. Modern Phys.*, vol. 81, pp. 591–646, 2009.

[11] A. B. Chan, Z. S. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Jun. 2008, pp. 1–7.

[12] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[13] H. Chate, F. Ginelli, G. Grgoire, and F. Raynaud, "Collective motion of self-propelled particles interacting without cohesion," *Phys. Rev.*, vol. 77, 2008, art. no. 046113.

[14] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2467–2474.

[15] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 21-1–21-11.

[16] N. Courty, P. Allain, C. Creusot, and T. Corpetti, "Using the agoraset dataset: Assessing for the quality of crowd video analysis methods," *Pattern Recog. Lett.*, vol. 44, pp. 161–170, 2014.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[18] K. L. Dion, "Group cohesion: From 'field of forces,' to multidimensional construct," *Group Dynamics Theory, Res., Practice*, vol. 4, pp. 7–26, 2000.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[20] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, pp. 41–54, 2006.

[21] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *Proc. IEEE 12th Int. Performance Eval. Tracking Surveillance*, Dec. 2009, pp. 1–6.

[22] L. Fiaschi, R. Nair, U. Koethe, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proc. 21st Int. Conf. Pattern Recog.*, Nov. 2012, pp. 2685–2688.

[23] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[24] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[25] L. F. Henderson, "The statistics of crowd fluids," *Nature*, vol. 229, pp. 381–383, 1971.

[26] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2547–2554.

[27] H. Idrees, N. Warner, and M. Shah, "Tracking in dense crowds using prominence and neighborhood motion concurrence," *Image Vis. Comput.*, vol. 32, pp. 14–26, 2014.

[28] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1031–1045, Aug. 2012.

[29] H. A. Jacobs, "To count a crowd," *Columbia J. Rev.*, vol. 6, pp. 37–40, 1967.

[30] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis, "Representing videos using mid-level discriminative patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2571–2578.

[31] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3-D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 275-1–275-10.

[32] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1446–1453.

[33] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1951–1958.

[34] D. Lan, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," in *Proc. 11th IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[35] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[36] G. Le Bon, *The Crowd: A Study of the Popular Mind*. New York, NY, USA: Macmillan, 1897.

[37] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inform. Process. Syst.*, 2010, pp. 1324–1332.

[38] J. Li, S. Gong, and T. Xiang, "Scene segmentation for behavior correlation," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 383–395.

[39] T. Li *et al.*, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[40] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[41] C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1988–1995.

[42] Z. Ma and A. Chan, "Crossing the line: Crowd counting by integer programming with local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2539–2546.

[43] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1975–1981.

[44] A. Marana, L. F. Costa, R. Lotufo, and S. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. Int. Symp. Comput. Graph., Image Process., and Vis.*, Oct. 1998, pp. 354–361.

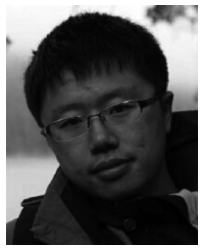
[45] C. Mcphail, *The Myth of the Madding Crowd*. Piscataway, NJ, USA: Transaction, 1991.

[46] R. Mehran, A. Oyama, and Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 935–942.

[47] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Feb. 2012.

[48] M. Moussaid, S. Garnier, G. Theraulaz, and D. Helbing, "Collective information processing and pattern formation in swarms, flocks, and crowds," *Topics Cognitive Sci.*, vol. 1, pp. 469–497, 2009.

[49] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behavior of pedestrian social groups and its impact on crowd dynamics," *PloSONE*, vol. 5, 2010, Art. ID e10047.
[50] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[51] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress Brain Res.*, vol. 155, pp. 23–36, 2006.
[52] J. K. Parrish and L. Edelstein-Keshet, "Complexity, pattern, and evolutionary trade-offs in animal aggregation," *Science*, vol. 284, pp. 99–101, 1999.
[53] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "Yoyll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 261–268.
[54] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2423–2430.
[55] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert, "Data-driven crowd analysis in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1235–1242.
[56] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, Jun. 2012.
[57] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2227–2234.
[58] S. Yi, X. Wang, C. Lu, and J. Jia, "L0 regularized stationary time estimation for crowd group analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2219–2226.
[59] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.
[60] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 352–365.
[61] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel type of phase transition in a system of self-driven particles," *Phys. Rev. Lett.*, vol. 75, pp. 1226-1229, 1995.
[62] F. Wang and C.-W. Ngo, "Summarizing rushes videos by motion, object, and event understanding," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 76–87, Feb. 2012.
[63] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis.*, Jun. 2011, pp. 3169–3176.
[64] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.
[65] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 188–197.
[66] H. P. Zhang, A. Beer, E. L. Florin, and H. L. Swinney, "Collective motion and density fluctuations in bacterial colonies," *Proc. Nat. Academy Sci. USA*, vol. 107, pp. 13626–13630, 2010.
[67] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.
[68] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 857–871.
[69] B. Zhou, X. Tang, and X. Wang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2871–2878.
[70] B. Zhou, X. Tang, and X. Wang, "Measuring crowd collectiveness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3049–3056.
[71] B. Zhou, X. Tang, H. P. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1586–1599, Aug. 2014.
[72] F. Zhu, X. Wang, and N. Yu, "Crowd tracking with dynamic evolution of group structures," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 139–154.



Cong Zhang received the B.S. degree in electronic engineering from the Shanghai Jiao Tong University, Shanghai, China, in 2009, and is currently working toward the Ph.D. degree at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University.

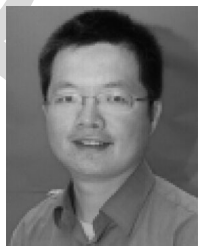
His current research interests include crowd understanding, crowd behavior detection, and machine learning.

Kai Kang received the B.S. degree in optics from the School of the Gifted Young, University of Science and Technology of China, Hefei, China, in 2009, and is currently working toward the Ph.D. degree in electronic engineering at the Chinese University of Hong Kong, Hong Kong, China.

His research interests include computer vision, video analysis, and deep learning.

Hongsheng Li received the B.S. degree in automation from the East China University of Science and Technology, Shanghai, China, in 2006, and the M.S. and Ph.D. degrees in computer science from Lehigh University, Bethlehem, PA, USA, in 2010, and 2012, respectively.

He is currently an Assistant Research Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China. His research interests include computer vision, medical image analysis, and machine learning.



Xiaogang Wang (S'03–M'10) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2001, the M.S. degree from the Chinese University of Hong Kong, Hong Kong, China, in 2004, and the Ph.D. degree in computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009.

He is currently an Assistant Professor with the Department of Electronic Engineering, Chinese University of Hong Kong. His research interests include computer vision and machine learning.

Rong Xie received the B.S. and M.S. degrees in communication engineering from the Northeast Dianli University, Jilin, China, in 1996 and 1999, respectively, and the Ph.D. degree in communication and information processing from the Zhejiang University, Hangzhou, China, in 2002.

She is currently an Associate Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China.

Xiaokang Yang (A'00–SM'04) received the B.S. degree from the Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000.

He is currently a Professor, the Vice Dean of the School of Electronic Information and Electrical Engineering, and the Deputy Director of the Institute of Image Communication and Information Processing at Shanghai Jiao Tong University. His current research interests include visual signal processing and communication, media analysis and retrieval, and pattern recognition.

Data-Driven Crowd Understanding: A Baseline for a Large-Scale Crowd Dataset

Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, *Member, IEEE*, Rong Xie, and Xiaokang Yang, *Senior Member, IEEE*

***Abstract*—Crowd understanding has drawn increasing attention from the computer vision community, and its progress is driven by the availability of public crowd datasets. In this paper, we contribute a large-scale benchmark dataset collected from the Shanghai 2010 World Expo. It includes 2630 annotated video sequences captured by 245 surveillance cameras, far larger than any public dataset. It covers a large number of different scenes and is suitable for evaluating the performance of crowd segmentation and estimation of crowd density, collectiveness, and cohesiveness, all of which are universal properties of crowd systems. In total, 53 637 crowd segments are manually annotated with the three crowd properties. This dataset is released to the public to advance research on crowd understanding. The large-scale annotated dataset enables using data-driven approaches for crowd understanding. In this paper, a data-driven approach is proposed as a baseline of crowd segmentation and estimation of crowd properties for the proposed dataset. Novel global and local crowd features are designed to retrieve similar training scenes and to match spatio-temporal crowd patches so that the labels of the training scenes can be accurately transferred to the query image. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art approaches for crowd understanding.**

***Index Terms*—Crowd features, crowd scene understanding, data-driven methods, large-scale benchmark.**

I. Introduction

CROWD understanding is an interdisciplinary topic and has been studied in physics [10], [25], [61], biology [52], [66], sociology [8], [36], [49], and psychology [6], [45] for a long time. In multimedia and computer vision community, a number of works contribute to action recognition [7], [28], event detection [47], [62] and behavior analysis [56] for the individuals or small groups in videos. Due to severe occlusion and perspective distortion, crowd understanding is a challenging topic and draws increasing attention because of the large demands on crowd video surveillance, which is especially important for metropolis security. Current works mainly focus on crowd segmentation [2], [12], crowd counting [11], [14], [15], [26], [42], crowd tracking [1], [3], [27], [53], [55], [72], and crowd behavior analysis [32], [33], [38], [41], [43], [46], [57], [58], [64], [69]–[71].

The progress of crowd understanding was mainly driven by the available public crowd datasets. Most of the above mentioned works [11], [12], [14], [15], [38], [41], [42], [64], [69] on crowd understanding are scene-specific, i.e., crowd understanding models learned from a particular scene can only be applied to the same scene. For example, the crowd counting approaches [11], [12], [14], [15], [42] require manually annotating some frames from the target scenes for training. In crowd behavior analysis [24], [33], [38], [41], [43], [46], [64], [69], behavior models trained for a target scene cannot generalize to other scenes. Therefore, the datasets proposed in [11], [15], [38], [64], [69] only contain one or two scenes.

Scientific studies [10], [13], [48] show that different crowd systems share the same underlying principles and can be characterized by a set of universal properties. Automatically understanding such general crowd properties across different scenes from videos not only has important applications, such as crowd video retrieval and crowd event detection, but also benefits scientific studies [9], [66] in other areas.

The learned crowd models are expected to generalize to new scenes not in the training set. Some research efforts [57], [70] have been made recently in this direction. Progress relies heavily on the availability of large-scale crowd datasets that include a large variety of scenes and video sequences. Existing crowd datasets do not provide enough variation. The largest one [70] (which actually combines other major crowd datasets) only contains 62 scenes. Because of security issues, only a small number of crowd videos by surveillance cameras are publicly available. The videos in [70] are mainly collected from the INTERNET. They are not of bird's-eye view and are therefore not suitable for crowd understanding. It provides only annotations of collectiveness at video-sequence level. Other multi-scene crowd datasets [2], [55] even do not provide ground-truth annotations. Benchmark datasets have become a bottleneck for the research on cross-scene crowd understanding.

Manuscript received July 27, 2015; revised February 6, 2016; accepted February 29, 2016. This work was supported in part by the NSFC under Grant 61527804, Grant 61221001, Grant 61301269, Grant 61371192, and Grant 61301269, in part by the STCSM under Grant 14XD1402100, 111 Program (B07022), in part by the General Research Fund sponsored by the Research Grants Council of Hong Kong Project CUHK 419412 and CUHK 147011, in part by the Hong Kong Innovation and Technology Support Programme Project ITS/221/13FP, in part by the Shenzhen Basic Research Program under Grant JCYJ20130402113127496, in part by the Ph.D. Programs Foundation of China under Grant 20130185120039, in part by the Sichuan Hi-tech R&D Program under Grant 2014GZX0009, and in part by the China Postdoctoral Foundation under Grant 2014M552339. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Martha Larson.

C. Zhang is with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China (e-mail: zhangcong0929@gmail.com).

K. Kang, H. Li, and X. Wang are with the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong, China (e-mail: kkang@ee.cuhk.edu.hk; hsli@ee.cuhk.edu.hk; xgwang@ee.cuhk.edu.hk).

R. Xie and X. Yang are with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xierong@sjtu.edu.cn; xkyang@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2542585

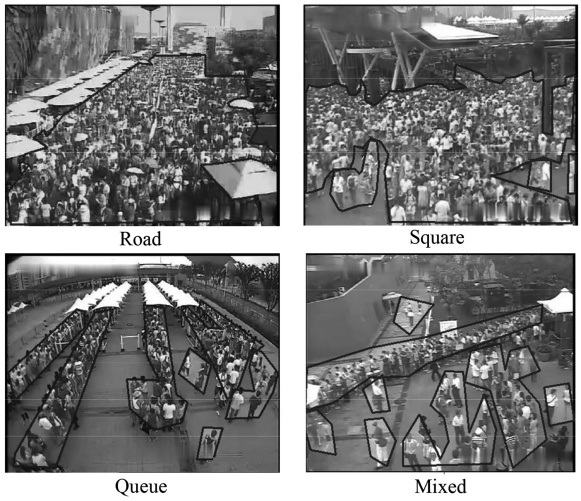


Fig. 1. Examples of four types of scenes from the WorldExpo'10 dataset. Crowd regions are manually labeled with polygons of different density levels (high = red, medium = yellow, low = green).

The first contribution of this paper is that we contribute a large-scale benchmark dataset[1] for crowd understanding. It includes 2630 annotated video sequences captured by 245 surveillance cameras, all from Shanghai 2010 WorldExpo.[2] Since all the cameras have disjoint bird's-eye views, they cover a large variety of scenes and the videos are especially suitable for cross-scene surveillance studies. Example scenes are shown in Figs. 1 and 2. We propose four challenges for this dataset: crowd segmentation, and estimation of crowd density, collectiveness and cohesiveness.

Crowd segmentation is the first step towards understanding crowds, because it answers the question of where the crowds are. Crowd counting, tracking, and behavior analysis are mainly based on the results of crowd segmentation, which itself also has important applications such as crowd trespassing detection. *Density* is a well known property of crowds and is related to other crowd properties such as collectiveness and cohesiveness [66]. Density estimation is of interest to security and traffic management, where highly dense crowds may lead to congestion or even disasters. *Collectiveness* is the degree of individuals in crowds acting with the same goal and was first studied in [70] from the perspective of computer vision. Collective behaviors widely exist in various crowd systems [8], [13], [66], [69], and have many potential applications [70]. *Cohesiveness* is another important property of crowd systems, in which some individuals move in groups and are bonded by force because of their special relationships. It measures the stability of local geometric and topological structures of crowd groups. Although it has been widely studied in crowd psychology [6], [18], it has not yet been addressed as a vision problem.

To accurately evaluate the methods that aim to solve the proposed four challenges, the WorldExpo'10 dataset is manually annotated at the region level. There are a total of 53 637 crowd segments with polygon boundaries, each of which was labeled with all the three crowd properties. Its annotations are much more comprehensive than any existing dataset. For example, the current largest one, CUHK dataset [70], only contains 413 videos with 413 annotations (because its annotations are at video level), while ours has 160 911 annotations. A detailed comparison with existing datasets is shown inTable I. Other researchers can propose new challenges and add new annotations to the WorldExpo'10 dataset. It would significantly advance the research of crowd understanding.

The second contribution is to propose a data-driven approach as a baseline of crowd segmentation and estimation of crowd properties for the proposed dataset. This is the first time to propose a unified approach that solves all the four crowd understanding challenges. As a generic solution, it has the potential of being applied to estimating other crowd properties. The large-scale annotated training data makes using data-driven approaches possible. Similar to [40], [60], no training procedure is required for our proposed data-driven method. The proposed method transfers the required information (such as density, collectiveness and cohesiveness) from the labeled training videos to the query via image matching. Our framework contains the following three steps: 1) retrieving candidate scenes similar to the query clip based on the proposed global crowd feature; 2) extracting multi-scale patches from the query and computing the proposed local crowd feature for each patch; 3) for each query patch, retrieving its nearest-neighboring patches from the candidate scenes and transferring the crowd properties from the nearest neighbors. The multiscale Markov Random Field (MRF) is utilized to enforce the smoothness of the resulting segmentation and property maps.

However, the general features widely used for scene understanding and texture description are not effective to describe crowd. Therefore, unlike other data-driven methods, new global crowd features (GCF) and local crowd features are proposed in our method. The proposed features are more effective than widely used generic features (such as GIST [51], HOG3D [31] and HOGHOF [35]) in the applications of crowd understanding. A new global crowd feature is proposed to retrieve similar crowd scenes for each input video clip. We train a series of mid-level filters as the crowd filters to generate filtering response maps for the input video clip. The global crowd feature of the input is then generated as the concatenation of the response maps. A new local crowd feature is proposed to compare similarities of spatio-temporal crowd patches. It combines eight types of features to characterize crowd appearance and motion. The combination weights of the eight features are automatically learned with relevance feedback such that the weighted features well match human perception on crowds.

The remainder of this paper is organized as follows. The existing works related to crowd understanding are reviewed and discussed in Section II. The details of our dataset, including annotation and evaluation protocols are introduced in Section III. Section IV presents our proposed data-driven approach for crowd understanding, the global crowd feature and the local crowd feature. In Section V, comprehensive experiments have been conducted to show the effectiveness of our

[1][Online]. Available: http://www.ee.cuhk.edu.hk/∼xgwang/crowdexpo.html

[2]Since most exhibition pavilions have been deconstructed, and no video corresponding to those pavilions still in use is included, the data is approved to be released for academic purposes.

Fig. 2. Examples of four types of scenes from our dataset: road (first row), queue (second row), square (third row), and mixed (last row).

TABLE I
COMPARISON OF DIFFERENT CROWD DATASETS

	UCF [2]	Data-driven [55]	CUHK [70]	Flickr [26]	PETS [21]	WorldExpo'10
Source	internet	internet	internet	internet	surveillance	surveillance
# of scenes	38	35	62	50	8	245
# of videos	38	35	413	50 Images	40	2,630
Resolution	480×360	480×360	various	various	720×576	720×576
Annotation type	n/a	n/a	video-level	frame-level	frame-level	region-level
# of annotation	0	0	50	50	4000	160, 911
Task	segmentation	tracking	collectiveness	counting	counting	segmentation, density, collectiveness, cohesiveness

Videos or images in the CUHK [70] and Flickr [26] datasets do not have uniform resolutions.

approach and compare it with state-of-the-art methods. Finally, the future works are discussed in Section VI and conclusion is drawn in Section VII.

II. RELATED WORKS

Crowd datasets. A number of crowd datasets [2], [3], [11], [15], [21], [33], [38], [43], [55], [64], [70] have been released in recent years. They are designed for specific tasks. Since many approaches are scene-specific, most of these datasets [3], [11], [15], [33], [38], [43], [64] have one or two scenes, and cannot be used to study generic crowd understanding. Courty *et al.* [16] proposed the AGORASET dataset which contains eight three-dimensional (3-D) synthetic scenes of walking pedestrians. However, real-world surveillance videos are much more challenging and realistic for research and evaluation. Table I compares our proposed dataset to existing ones with more than five scenes. Most of them are collected from the internet. The crowd counting dataset [26] only contains 50 static images from Flickr. The PETS [21] dataset was collected by eight cameras with overlapping views on a campus. Both the above datasets annotate the total number of persons in each image/frame. The UCF [2] and data-driven [55] datasets do not provide any annotation. The CUHK dataset [70] provides collectiveness annotation for each video sequence. Since each video contains multiple groups with different collective behaviors, it is more accurate to annotate collectiveness of each crowd region as in our dataset. None of the previous datasets provide annotations on crowd segmentation, density, collectiveness and cohesiveness simultaneously.

Crowd segmentation. Crowd segmentation is an important step for crowd counting, tracking and behavior analysis. It is typically conducted through background subtraction [11], [12], [14], [15], [34], optical flow estimation [2], [38], feature point tracking [68], pedestrian detection [24], [54], [67], and SVM classifier [4]. All these approaches have major limitations in practice. For instance, some areas in the scene might be occupied by crowds for long periods and the background is invisible. Fig. 3 shows example results of crowd segmentation by some above mentioned approaches. Background subtraction does not work well when it is difficult to model and update the background [see Fig. 1(a)]. Background modeling generates a lot of false alarms due to the changes of lightings, scene clutters, and nonhuman foreground objects. Optical flow estimation and feature point tracking do not work well when the crowds are stationary or move slowly, or the video quality is low [see Fig. 1(b) and (c)]. These motion-based approaches do not utilize crowd textures that can be used to distinguish other image regions. Appearance-based pedestrian detectors perform poorly on extremely dense crowds because of heavy occlusions and small pedestrian sizes [see Fig. 1(d)]. In comparison, our data-driven approach works on crowd patches and uses both appearance and motion features.

Fig. 3. Example results by different crowd segmentation methods. (a) Result by background subtraction used in [11]. (b) Result by optical flow. (c) Result by feature point tracking with the KLT tracker [68]. (d) Result by pedestrian detection used in DPM [19]. (e) Result by our data-driven approach.

Crowd counting and density estimation. A number of works [11], [14], [15], [26], [42] directly estimate the number of pedestrians in a whole image without detecting individual persons. Most of them require annotating training frames of the target scene and rely on crowd segmentation. Meanwhile, several methods [5], [22], [37] are proposed to estimate crowd density based on pedestrian localization results. However, these methods are also scene-specific and not applicable to cross-scene crowd understanding in large-scale surveillance applications.

Crowd behavior analysis. Many approaches [32], [33], [38], [41], [43], [46], [64] for crowd behavior analysis learn motion patterns for a target scene. They are not scene-independent and do not capture universal properties of crowd behaviors. Zhou *et al.* [69] measured the collectiveness using crowd manifolds and compared it across different crowd scenes. Li *et al.* [39] surveyed some state-of-the-art techniques on crowd behavior analysis, including available features, existing models and evaluation protocols.

Data-driven approaches. Several works [40], [55], [60] were proposed to solve pixel-wise or superpixel-wise classification tasks via dense image matching. Such nonparametric and data-driven approaches are suitable for large-scale data because they do not need any training. They transfer the required information from the training images to the query via dense image matching. Liu *et al.* [40] proposed a nonparametric image parsing method by recovering dense deformation fields between the query and training images, and it can work with an arbitrary set of labels. A simpler yet more effective nonparametric approach is proposed in [60], where the label transfer is achieved by superpixel-level matching with local features. A data-driven method is also adopted for crowd tracking in [55] to search for similar behaviors among crowd motion patterns in other videos. The key of these data-driven approaches is to design effective global and local features to match query and training images, which is also the focus of our proposed approach.

III. WorldExpo'10 Crowd Dataset

We contribute a large-scale benchmark dataset for understanding crowd. All the videos are shot with actual surveillance cameras from Shanghai 2010 WorldExpo, which was the world's largest fair site ever with an area size of 5.28 square km. Over 73 million people have visited during six months and nearly 250 pavilions were built at the expo site. The abundant sources of these surveillance videos enrich the diversity and completeness of the surveillance scenes. We define four challenges and evaluation protocols on this dataset: crowd segmentation, and estimation of crowd density, collectiveness and cohesiveness. It would significantly promote the research on crowd understanding.

A. Data Collection

A huge amount of crowd videos were collected from Shanghai 2010 WorldExpo from June to October 2010. A total of 2630 video sequences from 245 cameras with disjoint views are selected. Each camera has 10–12 videos, one of which was collected at night, and at least two in each month. Each sequence lasts one minute (3000 frames), and the data size is 40 GB. Cameras were mounted on the top of buildings and had far-field views. The resolutions of videos are 720×576, which is higher than or comparable to existing datasets (Table I). The data was collected under various weather conditions: sunny, cloudy, and rainy (pedestrians held umbrellas on rainy days). All the scenes generally fall into four categories: road, square, queue at entrances, and mixture of the previous three types of scenes (e.g., the bottom-right image in Fig. 1 has both queue and crowd in square). Generally, crowds in queue or on road tend to have higher collectiveness, while crowds in queue tend to have higher cohesiveness. Examples are shown in Figs. 1 and 2.

B. Annotation

A professional labeling company was hired and 20 labelers were trained for the annotation task. Three frames were uniformly sampled from each sequence for annotation. Before labelers annotate a frame, they first browsed its surrounding frames to observe moving objects. The boundaries of crowd regions are drawn with polygons as shown in Fig. 1. Each crowd region is labeled with three properties: density, collectiveness and cohesiveness. Each crowd property is labeled as one of the three levels: low (1), medium (2), and high (3). The property of background regions is always labeled as 0.

The annotation rule for crowd segmentation is as follows. Every person has his or her own territory which is a circle with a radius of one meter.[3] If the territories of two persons overlap, the two persons are connected. A crowd region covers a connected component of multiple persons.

Crowd density is annotated with the widely used Jacobs's method [29] proposed in social science, which classifies density into three levels. It counts the average number (n) of persons

[3]The "one meter" for each person is empirically determined by the labeler as 2/3 of the person's height.

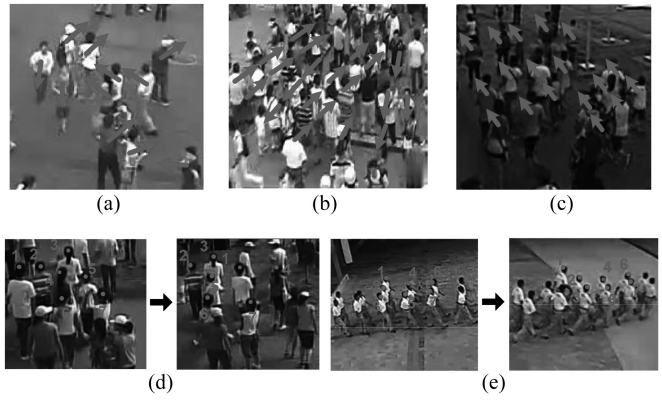


Fig. 4. Illustration of different levels of collectiveness and cohesiveness. (a) Low collectiveness. (b) Medium collectiveness. (c) High collectiveness. (d) Low cohesiveness. (e) High cohesiveness.

in every square meter. A scene is sparse if $n \leq 1$, medium if $1 < n \leq 2$, or dense if $n > 2$. Since crowd is not uniformly distributed in a scene, we empirically modified this rule to make it easier for annotation. Within a segmented crowd region, if the territory of a person includes another $0 \leq m \leq 2$ persons on average, this crowd region is annotated as sparse. Similarly, it is labeled as medium if $2 < m \leq 5$, and dense if $m > 5$. This is consistent with the Jacobs's method [29], since the area of a person's territory is around 3 square meters. Our annotation rule also implicitly considers crowd size. If a crowd region only has three persons, it is always labeled as sparse, even if all three stand tightly within one square meter, because there are no more than two persons in the territory of another person. Examples of density annotations are shown in Fig. 1.

Collectiveness and cohesiveness have been widely studied in physics [10], [25], [61] and sociology [8], [36], [49] for a long time. There is no explicit mathematical definition on crowd collectiveness and cohesiveness. Therefore, collectiveness and cohesiveness of our dataset's samples are defined in a subjective manner. For each sample, we have the same multiple human labelers to annotate its collectedness and cohesiveness (e.g., low=1, medium=2, high=3), and the average of their annotations is used as the final label. Fig. 4 shows examples of our definition on different levels. The collectiveness of Fig. 4(a) is labeled as low, since the pedestrians move in different directions without the same goal. In Fig. 4(b), a few crowd groups move in opposite directions and its collectiveness is labeled as medium. In Fig. 4(c), all the persons move in the same direction and the collectiveness is high.

Cohesiveness measures the stability of local geometrical and topological structures of crowd groups. Fig. 4(d) shows the same crowd at different frames. The topological structure of its members has changed significantly, and therefore the cohesiveness is low. Fig. 4(e) shows an example with high cohesiveness. Note that high collectiveness does not mean high cohesiveness. If a group of people move in the same direction but with very different speed, their local structures cannot remain stable.

Fig. 5 shows the histograms (on the area of crowd regions) of the three properties for the four type of scenes. According

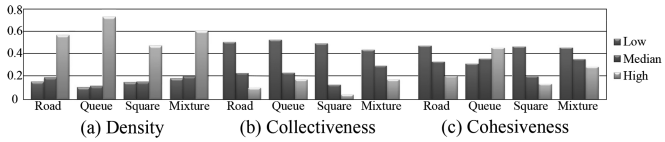


Fig. 5. Statistics of three properties in different crowd scenes (blue = low, red = medium, green = high). (a) Density. (b) Collectiveness. (c) Cohesiveness.

TABLE II
PARTITION OF TRAINING AND TEST SETS

	Road	Queue	Square	Mixture	Total
$N_{s/train}$	83	38	41	30	192
$N_{s/test}$	20	10	12	11	53
$N_{s/total}$	103	48	53	41	245
$N_{c/train}$	809	394	464	386	2053
$N_{c/test}$	207	103	147	120	577
$N_{c/total}$	1016	497	611	506	2630

N_s indicates the number of scenes and N_c indicates the number of video sequences.

to our statistics, around 75% regions are background and the remaining 25% regions are crowds. Most of the crowd regions in our dataset have high density. Generally, road and queue crowd scenes with strict man-made constraints have higher collectiveness and cohesiveness than open scenes such as square. Especially, in queue scenes, people are kept within some bounds, and most of the crowd regions have high cohesiveness.

C. Evaluation Protocols

80% of the data is partitioned for training and the other 20% for testing. The two subsets have no overlap on scenes or video sequences, In this way, the methods' capability of handling unseen scenes can be well evaluated. On the test set, we attempt to make data distribution more balanced on the four types of scenes. Detailed statistics are shown in Table II. Each crowd region in the test set was annotated by five labelers and we use the average of their scores. Since the training set is much larger, we cannot afford the cost of labeling each crowd region for multiple times. Although each crowd region is only labeled by one labeler, the whole training set is labeled by 20 labelers. The bias introduced by individual labelers can be reduced to some extent, because the learning process is based on the whole training set. Four evaluation criteria on the test set have been set for the proposed challenges.

Crowd segmentation. Every pixel in an annotated frame has a label: background (0) or crowd (1). ROC curve is used to evaluate the performance of crowd segmentation.

Crowd density estimation. Every pixel has an annotated density score ranging from 0 to 3. 0 indicates background and no crowd exists, while 3 indicates dense crowd. The estimation algorithms are expected to output continuous density scores. The Mean Square Error (MSE) is used for evaluation, and is computed as

$$\text{MSE} = \frac{1}{N_{\text{test}} N_I} \sum_{i=1}^{N_{\text{test}}} \sum_{p \in I_i} (\hat{l}_p - l_p)^2 \qquad (1)$$

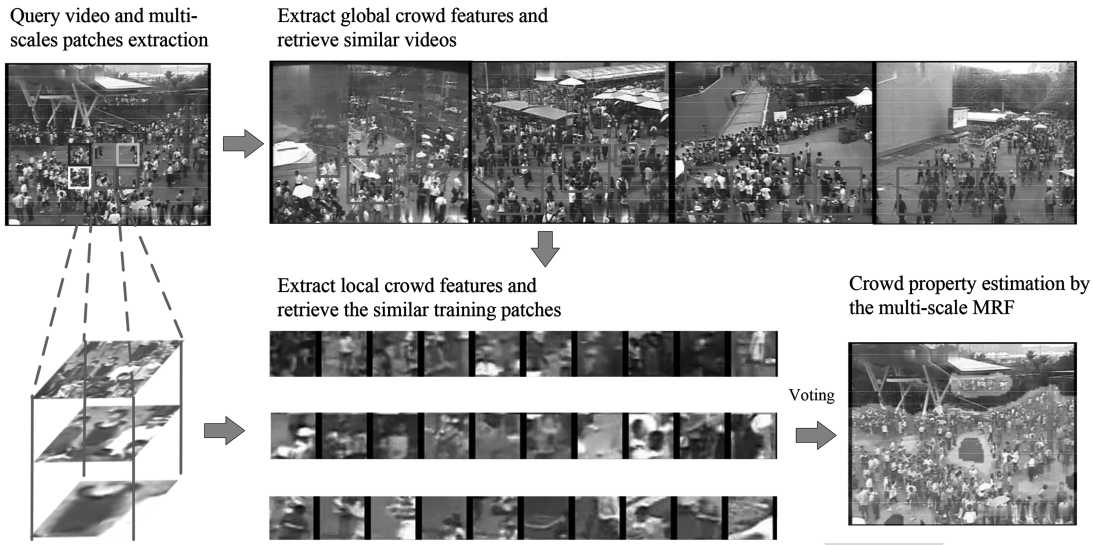


Fig. 6. Illustration of our proposed data-driven crowd understanding method.

where for pixel p in frame I_i, the ground-truth annotation is l_p, the predicted output is $\hat{l}_p$, N_{test} is the number of test samples and N_I is the number of pixels for image I_i.

Collectiveness and cohesiveness estimation. Since it is not reasonable to estimate collectiveness or cohesiveness of background, we only use manually segmented crowd regions for evaluation. The scores of both properties are in the range of 1 to 3. Similar to crowd density estimation, the MSE is used for evaluation.

IV. Data-Driven Crowd Understanding

We propose a data-driven crowd understanding approach as the baseline for our dataset. Different from most scene-specific crowd understanding methods, the data-driven method can be applied to any unseen scene without extra labeling and training. Data-driven approaches [40], [60] have achieved great success on scene understanding, which transfer the annotations of training data to test samples via dense pixel-level or superpixel-level image matching. Our large-scale annotated training set makes it possible for us to develop a data-driven approach as a baseline for our crowd understanding dataset.

A. Overview of the Proposed Method

In order to automatically annotate a query frame, the key of our data-driven method is to retrieve the most similar samples from training set and transfer their labels to the query via dense image matching. Fig. 6 illustrates the overall framework of our proposed method. In our framework, a short video clip including 30 frames surrounding the query frame is extracted as input. The training video clips are generated in the same way. To transfer labels only from training video clips that are similar to the query, the most similar scenes to the query video clip are first retrieved from the training set based on the global crowd feature as the candidate scene set. Then multi-scale crowd patches are extracted in a sliding window fashion with 50% overlap from the query video. For each patch, the most similar patches are retrieved from the candidate scene set based on local crowd features. Therefore, the key is to design effective global crowd feature to retrieve similar scenes and local crowd feature to match similar patches. Instead of using existing generic features, we learn crowd features and the optimal combination weights of different components based on training crowd videos. The crowd properties of each pixel can then be estimated by average voting. The multi-scale MRF is utilized to ensure the smoothness of the resulting crowd property map.

B. Global Crowd Feature for Candidate Scene Retrieval

For each patch in the query frame, it is costly to search among millions of crowd patches in the whole dataset for the most similar training patches. Therefore, it is more efficient to first retrieve a small set of candidate training video clips most similar to the query clip and match training patches within this subset. A global feature is needed to describe the whole crowd scene. One commonly used scene feature GIST requires convolving each image with a set of Gabor filters. However, there is no filter specifically designed to describe crowd scenes. Therefore, for our global crowd feature, we train mid-level filters to effectively describe the content of a crowd scene (see Fig. 7).

Mid-level crowd filters. Mid-level feature learning has been exploited in recent works on several vision topics, such as scene classification [59] and action recognition [30]. But existing works on mid-level feature learning did not consider the special properties of crowd understanding. The crowd property of a patch would significantly influence its appearance. Our goal is therefore to train discriminative mid-level filters that are able to distinguish patches of different appearance. We first group patches into several clusters with similar visual appearance. 16 000 spatio-temporal patches are uniformly sampled from crowd regions for clustering based on their ground truth crowd density, collectiveness and cohesiveness. In this way, the sampled patches have good diversity. The affinity propagation (AP) clustering method [23] is adopted because it does not

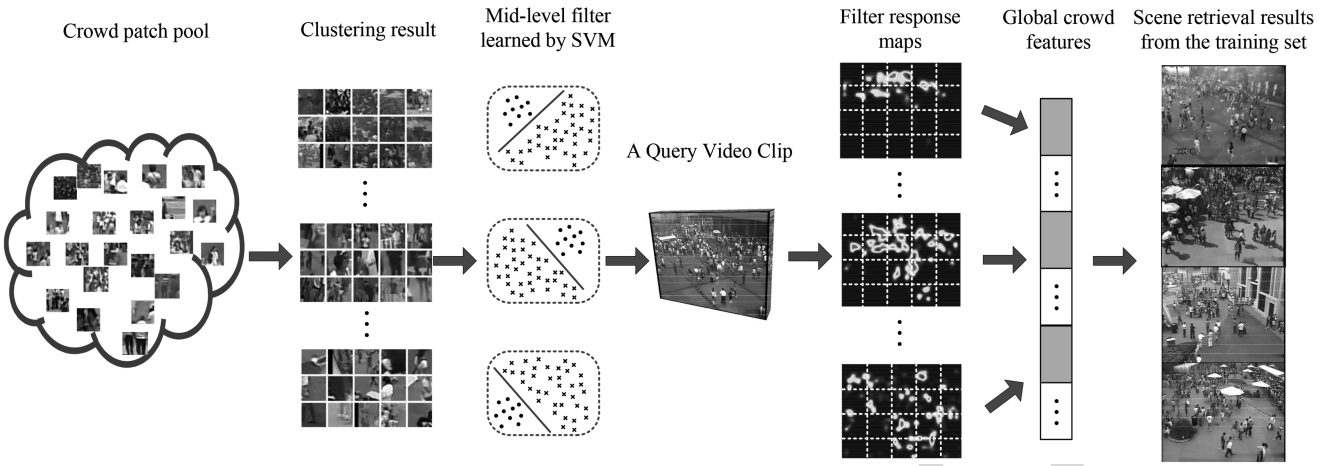


Fig. 7. Global crowd feature for scene retrieval. Red regions in the response map have high response value while blue regions are low. The retrieved scenes have similar views. Dense crowds are in areas farther to cameras.

require the number of clusters to be estimated in advance. For our dataset, $N_c = 30$ clusters are obtained for our training set by the AP algorithm. To capture distinctive appearance patterns to describe patches of different clusters, a discriminative filter is learned for each crowd cluster. For each cluster k, all the patches assigned to this cluster are regarded as positive samples, and patches from the other clusters and background are randomly sampled to form the negative samples. The number of negative samples is set 10 times as many as the positive samples. After creating the positive and negative patch sets, a linear SVM classifier $\{w_k, b_k\}_{k=1}^{N_c}$ is trained for every cluster. The SVM weights w_k and bias term b_k serve as the kth crowd mid-level filter. A response score map is obtained when crowd mid-level filters are used to convolve with a query video clip.

Global crowd feature. Global crowd feature is designed to describe the properties of the whole crowd scene for scene retrieval. Therefore, global crowd feature is extracted from the whole response maps generated by the mid-level filters for each video clip. The response maps are divided into $N_x \times N_y$ cells with no overlap. We set $N_x = 4$ and $N_y = 5$ for our proposed dataset. The average response scores of each grid is calculated. Such scores of all the filter response maps of N_c filters is concatenated as the global crowd feature to calculate its similarity between different scenes and to retrieve similar training scenes for a query (see Fig. 7). The total dimension of our global crowd feature is therefore $N_x \times N_y \times N_c = 600$.

C. Local Crowd Feature for Patch Matching

To distinguish different crowd properties, local crowd feature should describe both appearance and motion information at multiple scales. Therefore, our local crowd feature includes eight appearance and motion features extracted from each 3-D spatio-temporal crowd volume.

Multi-scale augmentation. Video surveillance data has large perspective variation, and crowds can be observed at different scales. In order to augment the training set and increase the robustness of matching with query patches, we sample both training and test patches at multiple scales and normalize them to the same size ($36 \times 36 \times 30$) as shown in Fig. 6.

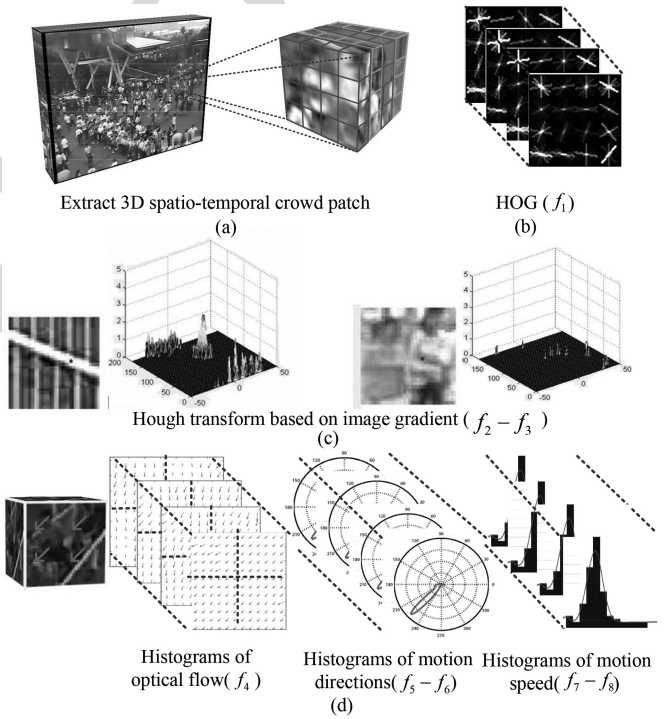


Fig. 8. Local crowd feature. (a) Uniformly sampling four frames from a 3-D crowd volume and dividing them into 3×3 cells. (b) Extracting HOG at each sampled frame. (c) Features computed from Hough transform are used to distinguish crowd patches from man-made patches with long line structures. (d) Examples of crowd patches with coherent (top) and incoherent (bottom) motions. Each sampled frame is divided into four sub-regions. Histograms of motion directions (third column) and speed (fourth column) are computed in each sub-region and the whole region.

Appearance features. The first feature f_1, HOG [17], is extracted from each sampled patch, as shown in Fig. 8(a) and (b). 4 frames are uniformly sampled from a 3-D patch, and each frame is divided into 3×3 cells with 50% overlap. The size of each cell is 18×18. Empirically, we observe that HOG cannot

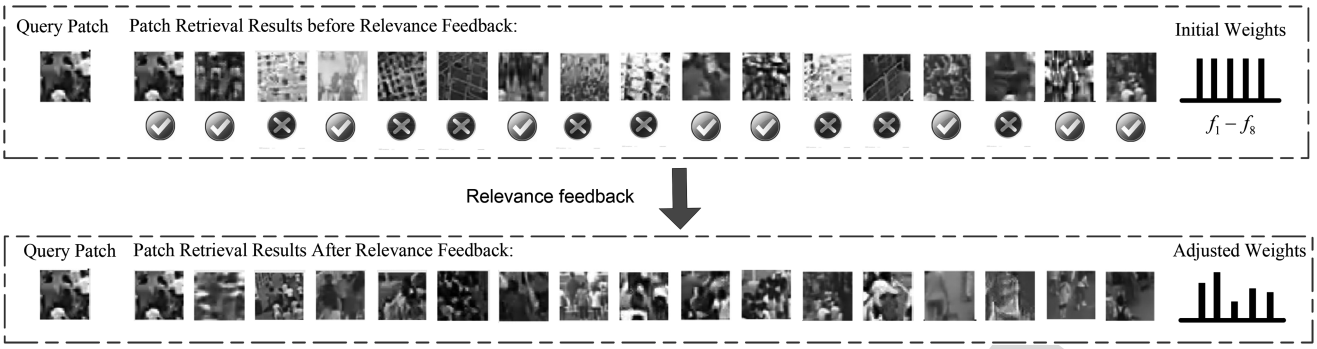


Fig. 9. Using relevance feedback to learn the optimal weights for the local crowd feature.

distinguish crowd patches with some man-made background patches with long line structures [such as fence in Fig. 8(c)], which are commonly observed in crowd scenes. We design two features (f_2 and f_3) with Hough transform to capture line structures. Traditional Hough transform is based on edge detection operators, such as Canny and Sobel. However, these edge operators ignore much texture information, especially for the surveillance video patch with relatively low resolution. Therefore, we perform Hough transform on image gradient map

$$r(\theta) = x_0 \cos\theta + y_0 \sin\theta \tag{2}$$

where θ is determined by gradient $(\Delta x, \Delta y)$ at (x_0, y_0) and is calculated as $\theta = \frac{\Delta y}{\Delta x} + \frac{\pi}{2}$. After applying a Gaussian filter, a response map $M(r, \theta)$ in the polar coordinate is obtained, as shown in Fig. 8(c). The feature $f_2 = [\mu_v, \sigma_v]$ characterizes vertical lines. μ_v and σ_v are the mean and variance of responses in the range $\theta \in [0, 10^\circ] \cup [170^\circ, 180^\circ]$. The feature $f_3 = [\mu_a, \sigma_a]$ characterizes the longest line in any direction. The mode of the highest peak is detected with the mean shift algorithm and its mean and variance are μ_a and σ_a.

Motion features. To characterize local motion of the sampled patch, the feature f_4, Histogram of Optical Flows (HOF) [35], is computed on the same sampled frames and cells. The same parameter setting is adopted as HOG. In order to further characterize whether individuals in crowd move in similar directions and keep stable local structures, the features $f_5 - f_8$ are computed based on the histograms of motion directions and speed as shown in Fig. 8(d). They are the entropy and variance of the two types of histograms. The patch at each frame is divided into four sub-regions. Histograms of the four sub-regions and of the whole region are computed. Note that besides $f_5 - f_8$, f_1 and f_4 at sampled frames are also useful for estimating collectiveness and cohesiveness, since they characterize how appearance and motion change over time.

Learning feature weights. The $f_1 - f_8$ features are concatenated as the local crowd feature. The distance between a training patch x_i and a query patch x_q is then computed as

$$d(x_i, x_q) = \sum_{k=1}^{8} \omega_k \|f_{\mathrm{ik}} - f_{\mathrm{qk}}\|_2 \tag{3}$$

where f_{ik} and f_{qk} is the kth local crowd feature of the patch x_i and x_q. It is important to assign a set of optimal weights $\{\omega_k\}$ to weight the importance of the eight features. We do not use the annotated labels in the training set to learn the weights, because it might make our crowd feature overfit to a particular task. Instead, we choose a relevance feedback approach to learn the weights that most match human perception. The weights learned in this way are more general and can be applied to various crowd understanding tasks.

It starts with uniform weights. Some examples of matching results with uniform distribution were shown Fig. 9. At each iteration t, a patch $x_q^{(t)}$ is randomly selected from the training set and is tried to match with other training patches $x_i^{(t)}$ using the current weights. Top N matches are presented to a labeler, who labels each of them as similar (1), dissimilar (-1), or uncertain (0) based on visual perception (Fig. 9). Based on the feedback, the feature weights are adjusted with adaptive SVM [65] as

$$d^{(t+1)}(x_i, x_q) = d^{(t)}(x_i, x_q) + \sum_{k=1}^{8} \Delta\omega_k^{(t)} \|f_{\mathrm{ik}} - f_{\mathrm{qk}}\|_2 \tag{4}$$

where $d^{(t)}$ represents the distance function at iteration t, and $\Delta\omega^{(t)}$ are the parameters estimated from the feedback examples at iteration t. To learn the parameter $\Delta\omega^{(t)}$, we adopted a SVM-like objective function

$$\begin{aligned}
&\min_{w^{(t)}} \frac{1}{2}\left\|w^{(t)}\right\|^2 + C^{(t)} \sum_{i=1}^{N^{(t)}} \xi_i^{(t)} \\
&s.t.\ \xi_i^{(t)} \geq 0;\ C^{(t)} = \eta^{(t)}(1 - \eta^{(t)}) \\
&y_i d^{(t)}(x_i) + y_i \sum_{k=1}^{8} \Delta\omega_k^{(t)} \left\|f_{\mathrm{ik}}^{(t)} - f_{\mathrm{qk}}^{(t)}\right\|_2 \geq 1 - \xi_i^{(t)} \\
&\forall (x_i, y_i) \in D^{(t)}
\end{aligned} \tag{5}$$

where $\sum_{i=1}^{N} \xi_i^{(t)}$ measures the total classification error of the tth feedback iteration. The cost factor $C^{(t)}$ represents the discriminative capability of the current iteration data to balance the contribution of previous iterations. So we define the $C^{(t)}$ as $C^{(t)} = \eta^{(t)}(1 - \eta^{(t)})$, where $\eta^{(t)}$ is the accuracy of feedback results at iteration t. $\eta = 1$ or 0 means all the feedback results are similar patches or dissimilar patches, which would not improve the retrieval results, and result in the lowest value of C. Oppositely, an equal number of positive samples and negative samples would lead to optimal weights.

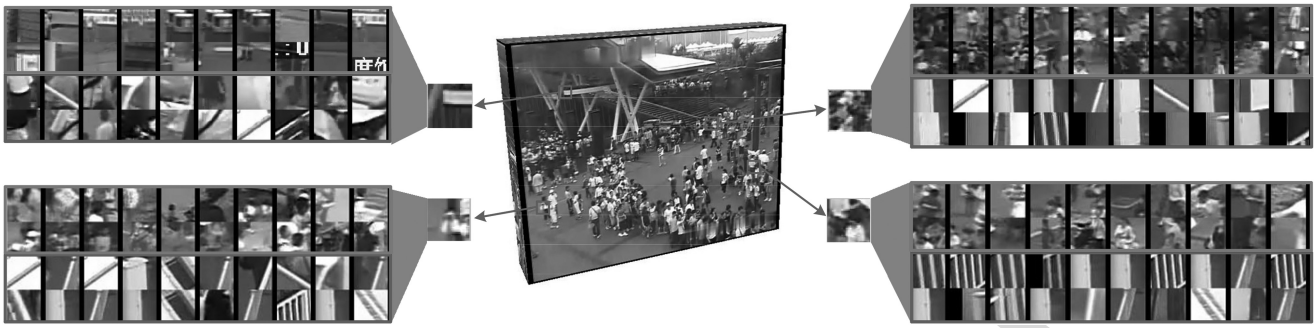


Fig. 10. Four query patches are selected from a query video clip. Their most similar (two upper rows) and dissimilar (two bottom rows) training patches based on our local crowd feature are shown in green and red rectangles, respectively.

At every iteration, a different query patch is randomly chosen. The iterations stop when the weighted features well match human perception and cannot be further improved. Some examples of matching results are shown in Fig. 10. The learned weights well distinguish background and crowds of different density levels.

D. Crowd Property Estimation

For each query video clip, we first retrieve its $M = 30$ most similar training video clips according to the global crowd feature. In addition, no more than three video clips are from the same scene to ensure their diversity. Image patches from the candidate training clips form the pool of candidate training patches. For the query video clip (30 frames), crowd patches are sampled at $S = 3$ different scales. For each patch p at scale s, its observed label score $\hat{l}_p^s$ is the averaged label score of its top $K = 20$ matched training patches from the candidate training patch pool.

The final label scores $\{l_p^s\}$ are obtained by the multi-scale MRF [26] to ensure smoothness. The graph can be represented by (V, E), where V are the pixel nodes and E are the neighbors at the same level and intermediated nodes that connect a patch to layers above and below it. The energy function with S level scales is thus given by

$$\min_{l} \sum_{s \in S} \left(\sum_{p \in V_s} D(\widehat{\eta}_p^s, l_p^s) + \sum_{(p,q) \in E} V(l_p^s - l_q^s) \right) \qquad (6)$$

where l_p^s represents the estimated property of patch p at scale s, and q is the spatial neighbor of patch p. The data term is defined as $D = |\hat{\eta}_p^s - l_p^s|$, where $\hat{\eta}_p^s = \frac{1}{2}(\hat{l}_p^{s+1} + \hat{l}_p^s)$ is of the bottom two scales and $\hat{\eta}_p^s = \hat{l}_p^s$ is of the top scale. The smoothness term is defined as $V = \min(|l_p^s - l_q^s|, \varepsilon)$, which enforces the smoothness between the neighboring nodes. This multi-scale MRF model is optimized using the Max-Product Belief Propagation method on grid structure [20].

V. Experimental Evaluation

We evaluate our data-driven approach for different crowd understanding tasks, including crowd segmentation (Section V-A), crowd density estimation (Section V-B), and crowd collectiveness and cohesiveness estimation (Section V-C) on the WorldExpo'10 dataset and compare it with other methods. The evaluation metrics were explained in Section III-C. For the test set, the patches are extracted in a sliding window fashion with 50% overlap in three scales, 36×36, 72×72, and 144×144, respectively. The estimated property of each pixel is obtained by averaging all the predictions of overlapping patches. The extensive experimental results by our proposed method and the compared ones on crowd segmentation, crowd density estimation, and crowd collectiveness and cohesiveness estimation demonstrate our method's capability of handling unseen scenes.

A. Crowd Segmentation

We compare our proposed data-driven approach (*Data-driven*) with six other crowd segmentation methods. The ROC curve is used to evaluate the performance of crowd segmentation. The following approaches are compared.

1) *SVM (Codebook)*: To the best of our knowledge, the only existing method specifically designed for crowd segmentation is [4]. The proposed method modeled crowd texture with a codebook. The SIFT features are extracted from interest points in frames. The codebook of size 1000 is built through k-means clustering on the SIFT feature. Crowd-likelihood features are computed based on the codebook as in [4] and used to classify each patch with SVM with a RBF kernel.

2) *BS*: Background subtraction is used by many crowd understanding works [11], [12], [15], [42] to segment crowd. The method used in [12] is chosen for comparison.

3) *Deformable Parts Model (DPM)*: Pedestrian detection approaches might also be used for crowd segmentation. A state-of-the-art pedestrian detector with DPM [19] is applied to test frames. It is trained on the INRIA dataset [17]. A pixel is segmented as crowd if it falls into a pedestrian window. We also compare with two baselines to evaluate the effectiveness of the components of our proposed method.

4) *SVM (HOG)*: This baseline follows the same framework as [4] but utilizes the HOG as the features to describe crowd, which is a popular descriptor for pedestrians. 5) *SVM Local Crowd Feature (LCF)*. To evaluate the performance of our proposed data-driven classifier, we also create a baseline that utilizes SVM and our proposed LCF feature. For methods 1), 4) and 5), we select 192 clips from every training scenes with

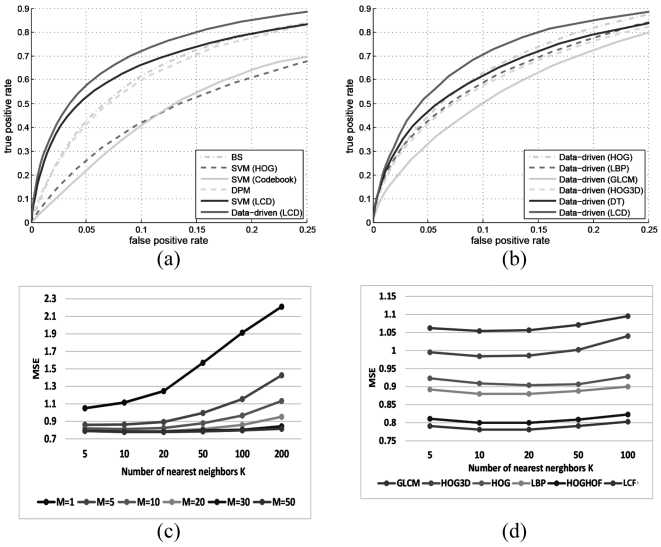


Fig. 11. (a) ROC curves of crowd segmentation results by different methods. (b) ROC curves of crowd segmentation results by using different local features in our proposed data-driven framework. (c) MSE of density estimation by our proposed framework with varying M and K. (d) MSE of density estimation by our proposed framework with varying K and different local features.

medium density distribution as the training data. The SVM is trained with approximately $230\,000$ patches. For fair comparison, when comparing with methods 1), 4) and 5), our data-driven method is trained with the same training set.

As shown in Fig. 11(a), our data-driven method with the proposed LCF feature outperforms the compared methods. Although BS works better than several other methods, it still performs worse than our data-driven approach. This is because background subtraction methods utilize motion information and cannot handle crowds that move slowly or are stationary. It is also affected by scene clutters. When setting the false positive ratio to 0.1, the true positive ratio of BS is 10% lower than that of ours. The pedestrian detector (DPM) does not work well neither because of severe occlusions. We also observe that when using the same classifier, i.e., SVM, our proposed local crowd feature significantly outperforms the widely used HOG and SIFT features. However, using the SVM classifier with our proposed feature is still inferior to the proposed data-driven approach, which is more effective on handling the complex distributions of crowd and background patches.

In order to further evaluate the effectiveness of our proposed local crowd feature, we compare our LCF feature to different local features by using them as the local feature in our proposed data-driven framework. The compared features include HOG [17], Local Binary Patterns (LBP) [50], Gray-Level Co-occurrence Matrix (GLCM) [44], HOG3D [31] and Dense Trajectory (DT) [63]. The general appearance features, such as LBP, GLCM and HOG, are widely used for general texture description and crowd understanding. HOG3D and DT are utilized for spatio-temporal description and achieve satisfactory performance on action recognition and crowd behavior understanding. We utilize the recommended parameters for all the compared features. Fig. 11(b) shows the ROC curves of different local features, where our proposed LCF feature outperforms other local features. LCF is more effective to describe the crowd characters. Note that DT obtains better performance than other texture features, which shows that motion information is important for the crowd segmentation task. But the general spatio-temporal features, such as DT and HOG3D, are not effective on describing crowds.

TABLE III
MSE OF CROWD DENSITY ESTIMATION BY REGRESSION-BASED METHODS (LEFT COLUMN) AND OUR PROPOSED DATA-DRIVEN METHODS WITH DIFFERENT LOCAL FEATURES (RIGHT COLUMN)

Method	MSE	Method	MSE
HOG+RR	1.10	Data-driven (HOG)	0.94
GLCM+GPR [11]	1.07	Data-driven (GLCM)	1.03
LBP+KRR [15]	0.98	Data-driven (LBP)	0.91
Lempitsky [37]	1.31	Data-driven (Ours)	**0.71**

B. Crowd Density Estimation

Our propose data-driven framework can also be utilized to estimate crowd density. The MSE (1) is used as the evaluation criterion. We compare our proposed method with some state-of-the-art regression based methods. All the major components in our methods are also evaluated. At last, we also discuss parameter selection and computational cost of our data-driven method.

Comparison with regression-based methods. We compared our proposed framework with several regression-based methods to estimate crowd density of each patch [11], [15], [37], [54]. They were originally proposed for crowd counting but can be used to estimate density in a similar way.

Gaussian Processes Regression (GPR) with GLCM feature was used for crowding counting [11]. Similarly, Kernel Ridge Regression (KRR) with LBP feature was adopted in [15]. Lempitsky [37] proposed a crowd density estimation approach that uses SIFT and regularized linear regression, which was also used in [54]. We also use the widely used HOG feature with the basic Ridge Regression (HOG+RR) as a baseline. The density estimation results of all the methods are listed in Table III. For fair comparison, we use the same training data for both regression-based methods and our data-driven method, which means that the step of candidate scene retrieval based on the global crowd feature is skipped in our method. Instead, we perform the local patch matching on all training data.

Our approach achieves the highest accuracy among all the compared methods. Most of these regression-based methods are scene-specific, and models learned from a particular scene can only be well applied to the same scene. From Table III, it is obvious that they do not show satisfactory performance in the large-scale dataset. In contrast, data-driven methods are more suitable for the large-scale and dynamic dataset. Some examples of our results are shown in Fig. 12.

Experiments are also conducted on the popular UCSD dataset [11] and MALL dataset [15], which are widely used to evaluate crowd counting and crowd density estimation. Pedestrians' positions are labeled for each scene of the two dataset. Followed by the Jacobs's method mentioned in III-B, the

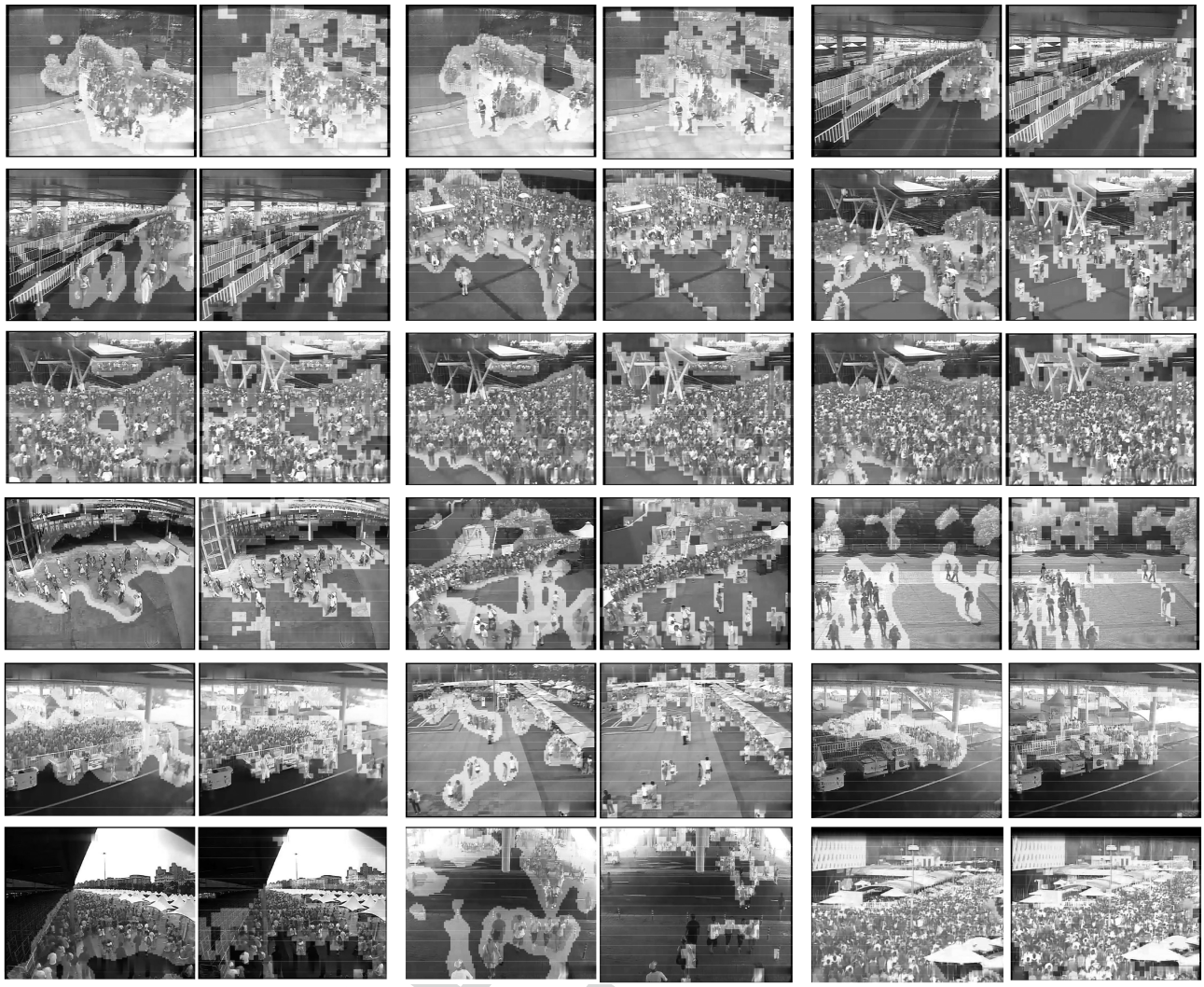


Fig. 12. Example density estimation results (red = dense, yellow = medium, green = sparse) by our data-driven framework (odd columns) and by the regression-based method LBP+KRR (even columns), which is the best regression-based method.

TABLE IV
MSE OF CROWD DENSITY ESTIMATION BY REGRESSION-BASED METHODS AND OUR PROPOSED DATA-DRIVEN METHODS ON THE UCSD DATASET AND MALL DATASET

Method	UCSD [11]	MALL [15]
Lempitsky [37]	1.07	1.16
GLCM+GPR [11]	0.73	0.91
LBP+KRR [15]	0.62	0.84
Data-driven (Ours)	**0.54**	**0.77**

TABLE V
MSE OF CROWD DENSITY ESTIMATION BY THE DATA-DRIVEN FRAMEWORK WITH DIFFERENT GLOBAL AND LOCAL CROWD FEATURES

	HOG	LBP	GLCM	HOG3D	DT	HOGHOF	LCF (UW)	LCF
GIST	1.08	1.08	1.21	1.15	1.10	0.96	0.98	0.94
GIST+MRF	0.93	0.95	1.08	1.00	0.93	0.84	0.88	0.82
GCF	1.06	1.01	1.19	1.16	1.02	0.93	0.94	0.89
GCF+MRF	0.90	0.88	1.05	0.98	0.85	0.80	0.82	**0.78**

density level annotations can be generated from the position labels. Both regression-based methods and our data-driven method are only trained from the training set of our World-Expo'10 dataset. Following, the same test partition as in [11] and [15]. Most of the regression-based methods are scene-specific, and our proposed method outperforms all the compared methods in these two datasets as shown by the results in Table IV. The results demonstrate that our proposed method is able to handle unseen target scene with our large-scale training dataset.

Evaluation of individual components. The effects of different local features are first investigated by using them in the proposed data-driven framework. Notice that spatio-temporal features, such as DT and HOGHOF [35], have better performances than texture features because of the additional motion information. Our LCF outperforms all other compared features. In addition, our LCF feature with uniform weights was compared as a baseline to demonstrate the necessity of the relevance feedback scheme as Eq. (5). We then compare our GCF for candidate scene retrieval with the GIST feature (Table V).

We observe that our GCF based on the mid-level crowd filters generate more accurate candidate scenes for label transfer with different local features. We also test the effect of the multi-scale MRF. Obviously, the MRF effectively improves the estimation accuracy because it enforces smoothness on the resulting label maps.

Parameter selection. The proposed data-driven system retrieves M most similar scenes for the query clip, and further selects K nearest neighbors for each query patch to estimate the crowd properties. We investigate the performance of our data-driven system by varying the parameters M and K.

Fig. 11(c) shows the MSE by setting different M and K. The performance improves as the number of candidate scenes (larger M) increases. The performance drops as K increases since more candidate patches may introduce noise to label transfer, especially when M is small. Although by conducting local patch matching in all training frames (equivalent to $M = 7000$), we obtain lower MSEs (as shown by the results in Table III). The computational time is proportional to M and using such large M is not practical for real-world applications. We also fix $M = 30$ and calculate the MSE of density estimation with varying K using different local features as shown in Fig. 11(d). Smaller K incorporates less information and larger K might result in more noise. We observed that $K = 20$ achieves the best performance for most types of local features. The balanced performance and computational cost are achieved when $M = 30$ and $K = 20$.

Computational cost. Our implementation is in MATLAB and is mostly parallelized. All our tests ran on a PC with a Core-i7 3.4 GHz quad core processor and 16 GB RAM. Our computational cost is mainly dominated by the extraction of global and local crowd features, which costs nearly 100 s for every query clip. But it can be easily sped up by utilizing more powerful hardware and better parallelization. Labeling one query clip with candidate scene retrieval and local patch matching takes less than 10 s. The main bottleneck of our implementation is file I/O for loading retrieval set features from hard disk. More appropriate data structure and larger RAM would improve the effectiveness of our implementation.

C. Collectivness and Cohesiveness Estimation

Our proposed method can also be extended to estimate collectiveness and cohesiveness. The MSE (1) is used as the evaluation criterion. Since it does not make sense to estimate collectiveness or cohesiveness on background, we only estimate annotated crowd regions for evaluation. We compared with the collectiveness measurement method proposed by Zhou *et al.* [70]. Since collectiveness and cohesiveness describe the motion information of crowds, we only compare our LCF feature with two spatio-temporal features, HOG3D [31] and HOGHOF [35].

Table VI reports the results of collectiveness and cohesiveness estimation by different methods. For collectiveness estimation, [70] does not work well if feature points cannot be well detected and tracked, especially when the video resolution is low. Our data-driven method performs robustly and does not rely on any detection and tracking. The experiment results show that our proposed crowd feature achieves better accuracy on collectiveness estimation than the other two spatio-temporal features. For cohesiveness estimation, there is no previous work on this topic. We only report the results by our data-driven framework with different local features. The data-driven method with our proposed crowd features also achieves the best performance.

TABLE VI
MSE OF COLLECTIVENESS AND COHESIVENESS ESTIMATION

	MSE
Zhou *et al.* [70] for collectiveness	0.71
Data-driven (HOG3D) for collectiveness	0.67
Data-driven (HOGHOF) for collectiveness	0.52
Data-driven for collectiveness (our LCF)	**0.49**
Data-driven (HOG3D) for cohesiveness	0.78
Data-driven (HOGHOF) for cohesiveness	0.66
Data-driven for cohesiveness (our LCF)	**0.64**

VI. DISCUSSION AND FUTURE WORK

The WorldExpo'10 dataset is a large-scale benchmark dataset for crowd understanding and covers a large variety of scenes with sufficient training data. Such training data would benefit learning algorithms specifically designed for big data, such as deep learning, data-driven approaches etc. Therefore, we hope the WorldExpo'10 dataset would become an important resource for more crowd video surveillance applications and can play a critical role in advancing the research on understanding crowds. We envision the following possible potential challenges:

Crowd counting. Most of existing crowd counting algorithms and datasets are scene-specific and focus on low density crowd. In comparison, the WorldExpo'10 dataset contains a large number of scenes with high variation of density. In the most crowded scenes, the number of pedestrians in a frame is close to one thousand. The crowd density also varies in a large range. Therefore, it is much more challenging and realistic to real-world surveillance applications. The baseline method of density estimation proposed in this paper would offer an important prior for crowd counting.

Abnormal event detection. Anomaly detection is an important problem in crowd understanding with extensive applications. In our high quality, diverse and large-scale WorldExpo'10 dataset, plenty of abnormal events can be observed and defined to evaluate and advance related research. The universal properties, density, collectiveness and cohesiveness, might be helpful for anomaly detection.

Crowd scene classification. We roughly summarize the crowd scenes in the WorldExpo'10 dataset into four categories. However, it can be further classified into many more categories based on different crowd behaviors, such as crowd gathering, crowd dispersing, crowd queuing, rushing, and loitering.

Deep learning has achieved great success in computer vision during recent years. However, so far little work has been done on deep learning for crowd understanding due to the lack of large-scale training data with annotation. This new dataset would significantly advance deep learning research in this area,

and more effective and discriminative crowd features and representations can be learned.

VII. Conclusion

In this paper, we contribute a large-scale annotated benchmark dataset including 245 scenes for cross-scene crowd understanding. Four challenges are proposed for this dataset based on their importance in scientific studies and crowd video surveillance applications. Benefiting from the large-scale training set, a data-driven approach with new global and local crowd features is proposed to solve crowd understanding tasks. It serves as a baseline for the proposed dataset and outperforms state-of-the-art approaches.

References

[1] I. Ali and M. N. Dailey, "Multiple human tracking in high-density crowds," *Image Vis. Comput.*, vol. 30, pp. 966–977, 2012.

[2] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–6.

[3] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 1–14.

[4] O. Arandjelovic, "Crowd detection from still images," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–10.

[5] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Interactive object counting," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 504–518.

[6] A. F. Aveni, "The not-so-lonely crowd: Friendship groups in collective behavior," *Sociometry*, vol. 40, pp. 96–99, 1977.

[7] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action representation and classification in unconstrained videos," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1234–1245, Aug. 2012.

[8] H. Blumer, "Collective behavior," *Principles of Sociology*. New York, NY, USA: Barnes & Noble, 1951.

[9] J. Buhl *et al.*, "From disorder to order in marching locusts," *Science*, vol. 312, pp. 1402–1406, 2006.

[10] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Rev. Modern Phys.*, vol. 81, pp. 591–646, 2009.

[11] A. B. Chan, Z. S. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Jun. 2008, pp. 1–7.

[12] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[13] H. Chate, F. Ginelli, G. Grgoire, and F. Raynaud, "Collective motion of self-propelled particles interacting without cohesion," *Phys. Rev.*, vol. 77, 2008, art. no. 046113.

[14] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2467–2474.

[15] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 21-1–21-11.

[16] N. Courty, P. Allain, C. Creusot, and T. Corpetti, "Using the agoraset dataset: Assessing for the quality of crowd video analysis methods," *Pattern Recog. Lett.*, vol. 44, pp. 161–170, 2014.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[18] K. L. Dion, "Group cohesion: From 'field of forces,' to multidimensional construct," *Group Dynamics Theory, Res., Practice*, vol. 4, pp. 7–26, 2000.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[20] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, pp. 41–54, 2006.

[21] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *Proc. IEEE 12th Int. Performance Eval. Tracking Surveillance*, Dec. 2009, pp. 1–6.

[22] L. Fiaschi, R. Nair, U. Koethe, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proc. 21st Int. Conf. Pattern Recog.*, Nov. 2012, pp. 2685–2688.

[23] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[24] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[25] L. F. Henderson, "The statistics of crowd fluids," *Nature*, vol. 229, pp. 381–383, 1971.

[26] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2547–2554.

[27] H. Idrees, N. Warner, and M. Shah, "Tracking in dense crowds using prominence and neighborhood motion concurrence," *Image Vis. Comput.*, vol. 32, pp. 14–26, 2014.

[28] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1031–1045, Aug. 2012.

[29] H. A. Jacobs, "To count a crowd," *Columbia J. Rev.*, vol. 6, pp. 37–40, 1967.

[30] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis, "Representing videos using mid-level discriminative patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2571–2578.

[31] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3-D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 275-1–275-10.

[32] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1446–1453.

[33] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1951–1958.

[34] D. Lan, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," in *Proc. 11th IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[35] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[36] G. Le Bon, *The Crowd: A Study of the Popular Mind*. New York, NY, USA: Macmillan, 1897.

[37] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inform. Process. Syst.*, 2010, pp. 1324–1332.

[38] J. Li, S. Gong, and T. Xiang, "Scene segmentation for behavior correlation," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 383–395.

[39] T. Li *et al.*, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[40] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[41] C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1988–1995.

[42] Z. Ma and A. Chan, "Crossing the line: Crowd counting by integer programming with local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2539–2546.

[43] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 1975–1981.

[44] A. Marana, L. F. Costa, R. Lotufo, and S. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. Int. Symp. Comput. Graph., Image Process., and Vis.*, Oct. 1998, pp. 354–361.

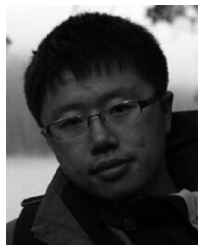
[45] C. Mcphail, *The Myth of the Madding Crowd*. Piscataway, NJ, USA: Transaction, 1991.

[46] R. Mehran, A. Oyama, and Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 935–942.

[47] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Feb. 2012.

[48] M. Moussaid, S. Garnier, G. Theraulaz, and D. Helbing, "Collective information processing and pattern formation in swarms, flocks, and crowds," *Topics Cognitive Sci.*, vol. 1, pp. 469–497, 2009.

[49] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behavior of pedestrian social groups and its impact on crowd dynamics," *PloSONE*, vol. 5, 2010, Art. ID e10047.
[50] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[51] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress Brain Res.*, vol. 155, pp. 23–36, 2006.
[52] J. K. Parrish and L. Edelstein-Keshet, "Complexity, pattern, and evolutionary trade-offs in animal aggregation," *Science*, vol. 284, pp. 99–101, 1999.
[53] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "Yoyll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 261–268.
[54] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2423–2430.
[55] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert, "Data-driven crowd analysis in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1235–1242.
[56] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, Jun. 2012.
[57] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2227–2234.
[58] S. Yi, X. Wang, C. Lu, and J. Jia, "L0 regularized stationary time estimation for crowd group analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2219–2226.
[59] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.
[60] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 352–365.
[61] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel type of phase transition in a system of self-driven particles," *Phys. Rev. Lett.*, vol. 75, pp. 1226-1229, 1995.
[62] F. Wang and C.-W. Ngo, "Summarizing rushes videos by motion, object, and event understanding," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 76–87, Feb. 2012.
[63] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis.*, Jun. 2011, pp. 3169–3176.
[64] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.
[65] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 188–197.
[66] H. P. Zhang, A. Beer, E. L. Florin, and H. L. Swinney, "Collective motion and density fluctuations in bacterial colonies," *Proc. Nat. Academy Sci. USA*, vol. 107, pp. 13626–13630, 2010.
[67] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.
[68] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 857–871.
[69] B. Zhou, X. Tang, and X. Wang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2871–2878.
[70] B. Zhou, X. Tang, and X. Wang, "Measuring crowd collectiveness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3049–3056.
[71] B. Zhou, X. Tang, H. P. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1586–1599, Aug. 2014.
[72] F. Zhu, X. Wang, and N. Yu, "Crowd tracking with dynamic evolution of group structures," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 139–154.



Cong Zhang received the B.S. degree in electronic engineering from the Shanghai Jiao Tong University, Shanghai, China, in 2009, and is currently working toward the Ph.D. degree at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University.

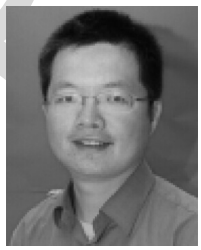
His current research interests include crowd understanding, crowd behavior detection, and machine learning.

Kai Kang received the B.S. degree in optics from the School of the Gifted Young, University of Science and Technology of China, Hefei, China, in 2009, and is currently working toward the Ph.D. degree in electronic engineering at the Chinese University of Hong Kong, Hong Kong, China.

His research interests include computer vision, video analysis, and deep learning.

Hongsheng Li received the B.S. degree in automation from the East China University of Science and Technology, Shanghai, China, in 2006, and the M.S. and Ph.D. degrees in computer science from Lehigh University, Bethlehem, PA, USA, in 2010, and 2012, respectively.

He is currently an Assistant Research Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China. His research interests include computer vision, medical image analysis, and machine learning.



Xiaogang Wang (S'03–M'10) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2001, the M.S. degree from the Chinese University of Hong Kong, Hong Kong, China, in 2004, and the Ph.D. degree in computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009.

He is currently an Assistant Professor with the Department of Electronic Engineering, Chinese University of Hong Kong. His research interests include computer vision and machine learning.

Rong Xie received the B.S. and M.S. degrees in communication engineering from the Northeast Dianli University, Jilin, China, in 1996 and 1999, respectively, and the Ph.D. degree in communication and information processing from the Zhejiang University, Hangzhou, China, in 2002.

She is currently an Associate Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China.

Xiaokang Yang (A'00–SM'04) received the B.S. degree from the Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000.

He is currently a Professor, the Vice Dean of the School of Electronic Information and Electrical Engineering, and the Deputy Director of the Institute of Image Communication and Information Processing at Shanghai Jiao Tong University. His current research interests include visual signal processing and communication, media analysis and retrieval, and pattern recognition.