Deep Learning Approach for Crowd Segmentation in Complex Videos

Adwan Alownie Alanazi[a], Sultan Daud Khan[b]

[a] College of Computer Science & Software Engineering, University of Hail, Saudi Arabia
[b] Department of Computer Science, National University of Technology, Pakistan

Abstract

Crowd analysis has numerous applications in crowd safety and security. In order to automate the process of crowd analysis, crowd segmentation is the pre-processing step. In this paper, we propose crowd segmentation framework that extract crowd regions from the background. We can extract crowd regions by employing background modeling and motion segmentation techniques. Since these techniques use motion cues, therefore accumulate false positives in the scenes where the crowd is stationary. In order to avoid using motion cues, we propose a fast and robust crowd segmentation framework that exploits appearance and structure cues to distinguish between the crowd region and background. We train appearance and structure based models separately and then jointly optimized the pre-trained models. To evaluate the performance of our proposed framework, we collect a data set that includes images from different complex scenes. From the experiment results, we observe that our proposed framework achieve superior performance compared to other state-of-the-art methods.

1. Introduction

Crowd safety and security is extremely important, particularly in populous urban areas. Recognizing the importance of crowd safety, research community from different domains are developing methods and techniques to ensure crowd safety. Conventionally, crowd analysis is performed through manual analysis of the scene, where analyst inside the surveillance room looks over large number of displays over a long duration to detect malicious activities. However, this manual analysis of crowd is tedious job and usually prone to errors. An alternative solution is to use automated analysis of the crowd that can efficiently and effectively analyze the crowd scene. The design of such intelligent system become the focus of computer vision's scientists. Several strides have been made toward the design of intelligent crowd management system. However, automated crowd analysis is still an open issue. Automated crowd analysis is challenging due to the following factors: (1) In high density crowded scenes, significantly large number of people gather in a limited area, that cause severe clutter and occlusions. Furthermore, extremely small size of head (few pixels) make the problem even worse. Therefore, alternatively, researcher develop crowd simulation models to understand the crowd dynamics. Unfortunately, these models failed to mimic the real time situations.

For analyzing crowd dynamics, detection of pedestrians and tracking are the main pre-processing steps. However, these two techniques failed to achieve desired results in high density crowds. An intuitive solution is to employ motion segmentation and background modeling methods to extract crowd from the scene. However, these models only extract crowds in motion and can not segment stationary crowd. Furthermore, motion segmentation also detects motion of foreign objects (other than crowd) that results in low in precision and recall rate. In real time surveillance application, for example, anomaly detection, behavior understanding, etc, crowd segmentation serves as important pre-processing step. The performance of these methods rely on the performance of crowd segmentation algorithm. However, crowd segmentation is a challenging task due to the following reasons, (1) Severe occlusion: In high density crowds, pedestrians usually stand very close to each other that cause severe occlusion. (2) Usually in high density crowds, sever clutter in the scenes confuse detector to distinguish crowd from the background.

Convolutional neural network (CNN) has achieved tremendous success in object detection, classification and segmentation tasks. However, to the best of our knowledge, CNN has not be explored for static crowd segmentation. We argue that CNN can learn hierarchical appearance features that can increase the precision and recall rates of crowd segmentation. In this paper, we proposed a CNN based framework named as , Crowd Segmentation Network (CS-CNN) for crowd segmentation in complex scenes. Compare to other existing methods, our proposed method has following contributions:

- Our proposed method does not use motion cues and appearance features for detection and tracking of individuals.
- Our approach reduce the computational cost by detecting crowd from a single image instead of using whole video sequence.
- Our approach do not rely on background modelling for crowd segmentation.

Manuscript received June 6, 2020
Manuscript revised June 20, 2020

- Our approach learn hierarchical feature from the scene, therefore, can be applicable in both low and high density crowds.
- Our approach improve the performance of other crowd applications, i.e., crowd counting, behaviour understand and anomaly detection by precisely segmenting crowd regions.
- We evaluate our method on different scenes. The experiments results shows that our proposed method can precisely localize the crowd.

2. Related Work

Crowd segmentation is an emerging topic and limited amount of work is reported in literature. Most of existing methods generally focus on counting, density estimation, crowd tracking, and crowd behaviour understanding. Crowd behaviour understanding has a lot of application in anomaly detection [30, 31, 19, 36, 33, 34] and congestion detection [13] in crowded scenes. Other studies focus on crowd counting in dense crowds [25, 10, 4, 3, 27, 11, 21]. Crowd motion analysis and flow segmentation methods [15, 29, 28, 35, 1, 32, 37, 38, 20] have been studies extensively. Other methods focus on detecting social groups in crowd
scene [5, 17, 16, 2, 14, 23, 12, 40] . Generally there has been a growing interest in crowd counting and density estimation, however, crowd segmentation is not fairly discussed and very few papers reported in literature discussing crowd segmentation problem. Most of existing methods use background subtraction and motion flow segmentation [6, 8, 22, 42] to segment crowd. Other approaches rely on detection and tracking methods [18, 41, 39] to segment crowd.

However, these methods work fine in low density situations, but suffers significant set back in high density situations. Incorporating multiple visual cues has also been explored in [8, 26]to segment crowd. Most of these methods are not applicable in real world scenes, since there methods use same data for training and testing. Deep neural networks have achieved tremendous success in object classification, object detection and semantic segmentation tasks. Traditional semantic segmentation methods [9, 7, 24]employ patch-by-patch scanning strategy and require input of fixed size

3. Proposed Methodology

In this section, we discuss our propose crowd segmentation network that learn hierarchical appearance features from the input images. Our segmentation network consists of fully covolutional layers in the last layer that predict the probability of each pixel in the output segmentation mask. Our segmentation network takes an arbitrary size image and outputs corresponding segmentation mask, where high values represent the presence of crowd and lower values indicate the background. The main advantage of our crowd segmentation network is the network is translation invariant, as it only uses convolutional and pooling layers. The network also incorporate contextual information by predicting segmentation mask for a small region surrounding a pixel. Furthermore, incorporating six convolutional and two pooling layers increase the receptive filed size in the input image and much contextual information is captured. This enable our network to predict segmentation mask with high accuracy. Moreover, our network is independent of the size of input image and normalization of image size is not required. The overall pipeline of our proposed crowd segmentation framework is shown in Figure 1.

Our crowd segmentation network fuses appearance and structure cues for crowd segmentation. Intuitively, we can combine these cues as separate channel of input image, however, these cues have different roles to play in crowd segmentation. For instance, for a given patch, if we achieve high confidence using appearance, we will assign label ”crowd” to that patch. We do not rely on motion cues, the reason is the motion may be caused by other moving objects. We train appearance and structural filters separately and then jointly optimized them via fine-tuning.

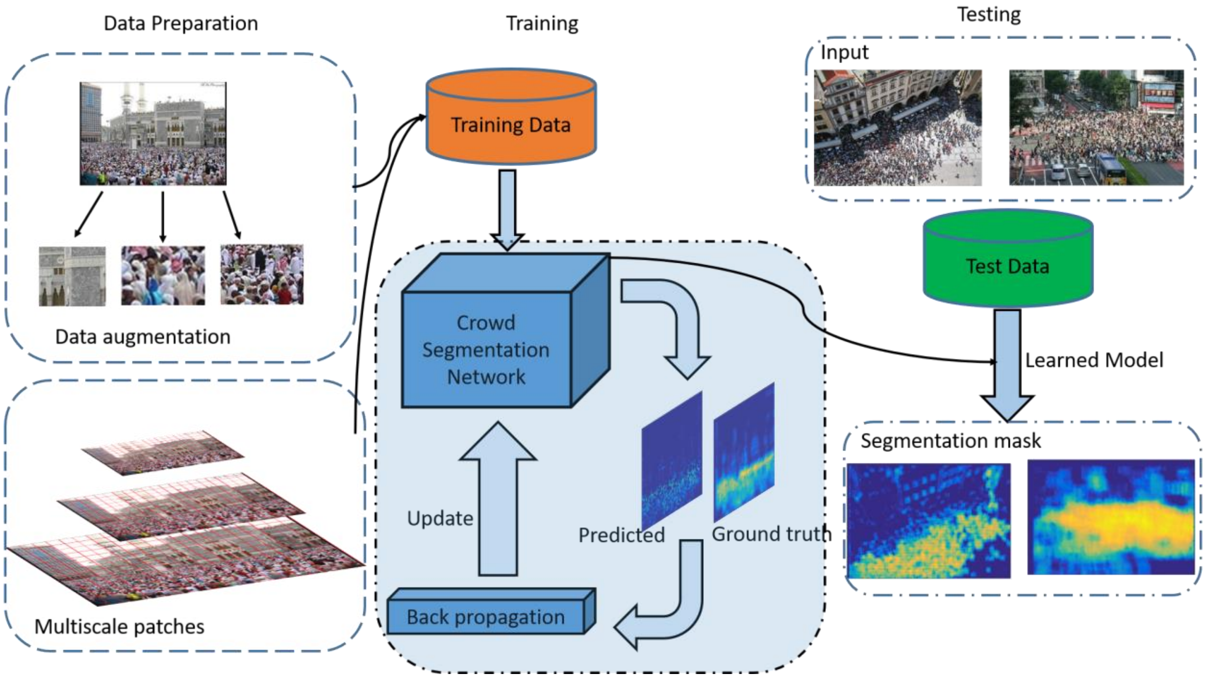


Fig. 1 Pipeline of proposed framework both during testing and training phase. Input image is divided into block and the into cells. Features are then extracted from each cell and train SVM classifier.

3.1 Appearance Cues

Appearance cues has been exploited by CNNs and achieved tremendous success in object classification, detection and segmentation tasks. Even with large scale classification and segmentation tasks, appearance cues achieved significant success. The main advantage of using CNNs that these models learn hierarchical appearance cues automatically by adjusting weights and back propagating the loss. This is due to the fact that we also employed appearance of crowd that has an obvious difference from background and other objects in the scene.

3.2 Temporal Cues

Learning motion cues from videos play an important role in crowd segmentation. For precise crowd segmentation, it is imperative to capture long term motion information. However, most of the existing methods are based on motion descriptor that can not capture long term motion. For capturing motion information, optical flow has been widely used in variational approaches for flow segmentation. These approaches computes the displacement of each pixel by matching its pixel value in the next frame. With the advent of covolutional neural networks, for example, DeepMatching [31] and DeepFlow [50] compute the optical flow by learning hierarchical features. However, these methods depend on several parameters that need to be set manually.

In order to address above challenges, several methods are reported in literature to learn motion information from videos. Ji et al. [14] proposed 3D-CNN that accepts multiple channel

input and perform 3D convolution to capture spatial-temporal information. However, this CNN based model performed lower than hand-craft feature based statistical model [49]. Simonyan et al. [37] achieved state-of-the-art performance by proposing two-stream model that incorporates spatial and temporal features by employing two parallel CNN networks.

To model long duration motion information, our temporal network takes multiple optical flow fields computed between consecutive frames from a fixed temporal window. In contrast to spatial CNNs that learn hierarchical features from input image, temporal network takes the stack of optical flow fields and learn long term motion information from the temporal window. The stack of optical flow fields generally captures the motion in the temporal segment defined by a window and makes crowd segmentation trivial. With this architecture,

our temporal network does not need to learn the optical flow explicitly

For optical flow computation, we employ [7] that precisely computes optical flow for every pixel of the image by using gradient and smoothness consistency constraint. Let i represents a feature point in the image at time t of a temporal segment S. Let its displacement vector contains the spatial location and Vi,t represents velocity. Li,t = [x, y]t represents spatial horizontal and vertical coordinates, while Vi,t = [u, v]t represents the displacement in horizontal and vertical directions. We compute dense flow Dt by computing displacement vector for each pixel of the image. We observed that Dt accumulates noise due to sensitivity of optical flow towards illumination changes, therefore, we refine Dt by applying threshold value Ω. The resultant Dt0 contains precise information about the displacement of every pixel in the image. Once we compute optical flow for whole temporal windows, we then prepare input volume Ψ by stacking optical flow fields as Ψ(u, v, S − 1) = Dt(u, v) .

3.3 Structure Cues

From empirical studies, we observe that structure of background, dimensions of scene and perspective distortions also play positive role in detecting crowds. For extracting structure from the scene, we employ edge detection algorithm [7] and feed the resultant image to the network. The edge detection models provide more information about the kind of structure and can easily distinguish between the crowd and background.

Edges are the salient features that contain structural information of the scene. Edge detection has been widely used in many applications, for example, object detection, image segmentation, scene segmentation and scene classification. Traditional methods extract low level cues like pixel intensity, color, texture, and gradient features to classify pixels into edge and non-edge pixels. Although, the performance of these hand-crafted feature model is promising, yet suffer from several limitations. For example, it is non-trivial to use low level cues to extract high-level semantic information. Since CNN achieved tremendous success in automatically learning high level representation from raw images, researchers are employing different CNN models to learn edges in natural images, for example, DeepEdge [5], DeepContour [36], and holistically nested edge detection (HED) [51].

Our structural network follows the backbone of VGG16 [38] network that consists of thirteen convolutional layers and three fully connected layers at the top. The convolutional layers are divided into five stages, where pooling layer is applied after each stage. This shallow layers of the network with small receptive field sizes capture information about the small objects. while top layers capture meaningful semantic knowledge about large objects. The

details about architecture and function of VGG16 is provided in [38]. We exploit hierarchical features extracted from the last convolutional layer to hypothesized the presence of edge. The details of our structural network are as follows.

1. In order to make network to take images of arbitrary size, we remove fully connected layers and replace those layers by 1 x 1 convolutional layers. We also add pool5 layer to the network to increase the stride by 2 to obtain better generalization of the edges.
2. We keep kernel size of each convolutional layer is 1 × 1 and channel depth 31. The resulting feature map from all stages are fused and accumulated using eltwise layer to obtain fused features.
3. An 1 x 1 conv layer is applied after each eltwise operation. We use deconvolution layer to upsample the feature maps.
4. After upsampling the feature maps, a cross entropy layer is connected for loss calculation.
5. Since the receptive fields of propose structure network are different that enables the network to learn and accumulates multi-scale features from all convolutional layers. These multiscale features provide support in precisely detecting the edges and capture structure of the scene.

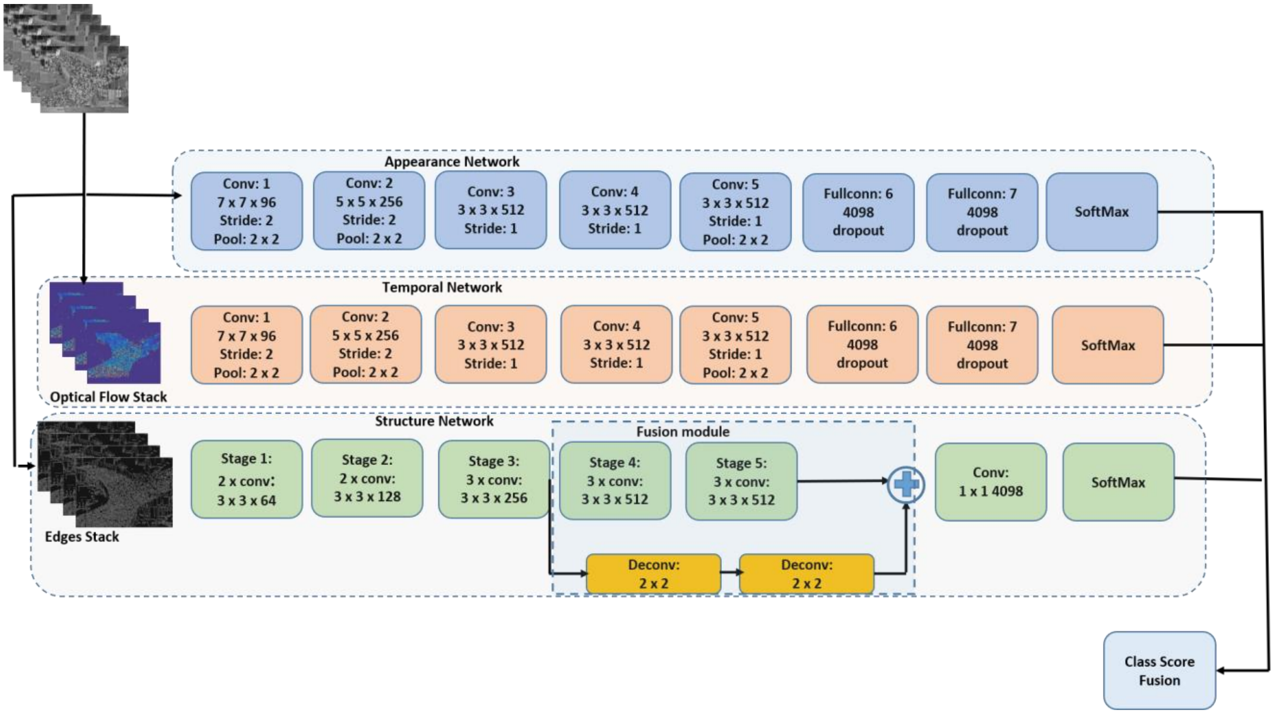

Appearance Network
Conv: 1 7 x 7 x 96 Stride: 2 Pool: 2 x 2
Conv: 2 5 x 5 x 256 Stride: 2 Pool: 2 x 2
Conv: 3 3 x 3 x 512 Stride: 1
Conv: 4 3 x 3 x 512 Stride: 1
Conv: 5 3 x 3 x 512 Stride: 1 Pool: 2 x 2
Fullconn: 6 4098 dropout
Fullconn: 7 4098 dropout
SoftMax
Temporal Network
Optical Flow Stack
Conv: 1 7 x 7 x 96 Stride: 2 Pool: 2 x 2
Conv: 2 5 x 5 x 256 Stride: 2 Pool: 2 x 2
Conv: 3 3 x 3 x 512 Stride: 1
Conv: 4 3 x 3 x 512 Stride: 1
Conv: 5 3 x 3 x 512 Stride: 1 Pool: 2 x 2
Fullconn: 6 4098 dropout
Fullconn: 7 4098 dropout
SoftMax
Structure Network
Edges Stack
Stage 1: 2 x conv: 3 x 3 x 64
Stage 2: 2 x conv: 3 x 3 x 128
Stage 3: 3 x conv: 3 x 3 x 256
Fusion module
Stage 4: 3 x conv: 3 x 3 x 512
Stage 5: 3 x conv: 3 x 3 x 512
Deconv: 2 x 2
Deconv: 2 x 2
Conv: 1 x 1 4098
SoftMax
Class Score Fusion

Fig. 2 Three stream network for crowd segmentation

The first convolutional neural network learns hierarchical features from raw images. The second is temporal network that takes stack of optical flow fields as input. The third network is structure network that takes stack of edges. The softmax scores of these networks are combined to classify crowd from non crowd regions.

3.4 Fusion Schemes

In order to combine appearance and structural cues, we explore three different fusion schemes: early fusion, middle fusion and final map fusion. In early fusion scheme, input appearance and structural maps concatenated directly before feeding to the network. The middle fusion combine features maps of different convolutional layers while in final map fusion, the output segmentation masks generated using appearance and structural cues are combined.

In joint fusion model, we fuse the feature maps from appearance, motion and structure model by first applying pooling operation to respective score vector (obtained after softmax layer) and obtain a unified score vector. Since it is desirable for all score vector to be in the same size before applying unification step. Typically, score vector of structure network is of high dimension since it contains information about the edges, therefore, in order to bring it back to the size equal to the size of score vector of appearance and temporal model, we add extra linear layer. The joint fusion model is trained using stochastic gradient descent (SGD) by minimizing the negative likelihood [11].

In joint feature fusion, feature map from three network are in the same feature space. It is to be noted that join fusion model treats score vector from each network as different and exploit no relationship among the score vectors. An auxiliary strategy is adopted to enforce and exploit the similarities among score vectors of all three networks

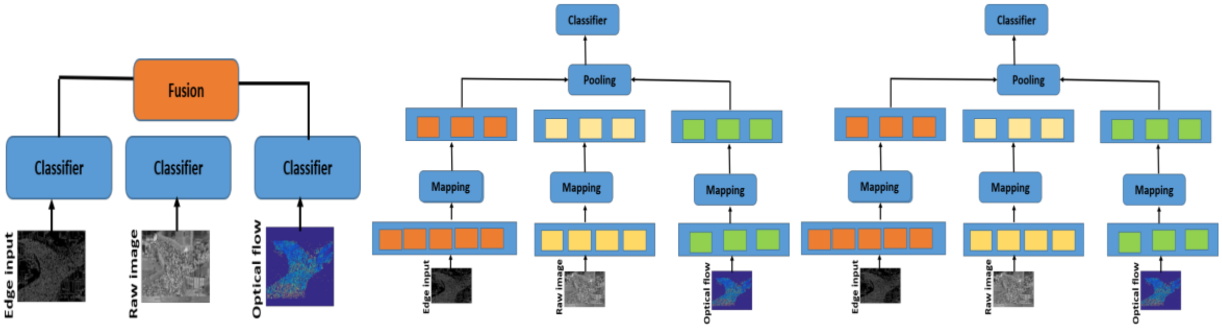
We train the first scheme independently. For training other schemes, we use the pre-trained model of first scheme, keep the parameters of previous layers and fine tune new layers. This training strategy has following advantages: (1) We reduce the computational complexity by keeping the parameters of the previous layers and only train the last layers. (2) Enables a network to learn complementary information that improve the performance of the network.

The overall framework of our proposed crowd segmentation is shown in Figure 1. As obvious from Figure, the input to our framework is arbitrary size image and corresponding ground truth segmentation mask.

The output segmentation map represents the confidence of being crowd. With proper configuration, all convolutional and pooling layers keep similar dimensions of the input image. In order to keep the same dimension, ground truth segmentation maps are fed to two average pooling layers. We use cross entropy loss function to minimize the loss between ground truth segmentation mask and the predicted map. We observe that cross entropy loss defined in Equation 1 is useful for crowd segmentation problem.

$$L = -\frac{1}{M} \sum_{k=1}^{M} T_k \log s_k + (1 - T_k) \log(1 - s_k) \qquad 1$$

where L represents the loss function, M is the total number of samples, Tk is the ground truth segmentation mask for sample k, sk is the predicted segmentation mask. Our proposed crowd segmentation model is trained in a global fashion by utilizing all images with corresponding ground truth segmentation masks.



(a) Late fusion (b) Early fusion (c) Joint fusion

Fig, 3 Different fusion strategies.

The first figure illustrates pipeline of late fusion, where features from classification layers are fused. The second figure shows the early fusion strategy, where features are combined before feeding to classification layer. The last figure shows join fusion layer, where feature maps are first mapped to a common space and then fused together.

4. Experiment Results

In this section, we evaluate our proposed method in both qualitative and quantitative way. Training and testing the proposed segmentation model require considerable amount of labeled data. Therefore, we keep the following requirements for data: (1) the images should be acquired from distinct camera view points; (2) the size of the annotated data should be considerably large; (3) the test data should also contain pixel-level segmentation ground truth masks. In order to acquire this kind of data, to the best of our knowledge, no data set is publicly available that satisfy these requirements, therefore we collect images that stratify the requirements.

Shanghai World Expo dataset is first proposed by [52]. The dataset is collected from 235 cameras installed in different places of shanghai World Expo in 2010. For training, we select

185 camera views and the rest of 51 camera views are left for testing. For ground truth annotation, we select one frame from each video and annotate the crowd with polygons. The

polygons that cover the forground region (where pedestrian are standing) are regarded as positive while the rest of the scene is considered as background. For the pilot experiment, we select 10 frames from each video sequence and label the frame at pixel level for precise evaluation.

City dataset is first proposed by [15]. This dataset contains 11 different scenes captured from different cameras installed at different public places. These places include parks, squares,

railway stations, subways, bus stops, streets, etc. We use this dataset for cross-scene validation and non of its sample was used for training. We annotate the dataset at pixel-level same as the first dataset

We evaluate and compare the performance of different baseline methods. The first baseline method is Gaussian Mixture model (GMM) for background subtraction. The second baseline model is Histogram of Oriented Gradients (HOG). In this baseline method, we first divide the image into patches and then from each patch, we extract HOG features. Then a linear SVM classifier is trained. During testing phase, we randomly select images to test the trained model.

In addition to baseline methods, we use different variants of our proposed framework. The first variant use appearance features named as CS-CNN(appearance) and other use structural feature for crowd segmentation and termed as CSCNN(structure). The third variant CNN(temporal) that accepts stack of optical flow fields and capture motion information.

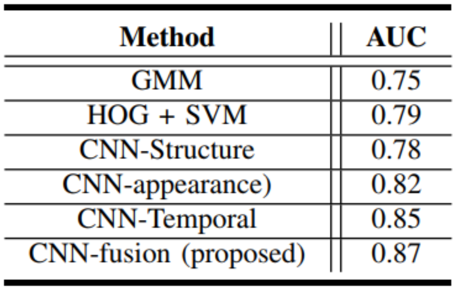
The fourth variant is the fusion of appearance, structure, and temporal features and termed as CS-CNN(fusion).

We also employ data augmentation to further increase the amount of data for training. For this purpose, we take the input image and randomly cropped patches each of size 259 x 259 pixels with horizontal and vertical flipping up to 50%. For ground truth segmentation, the segmentation maps are cropped and flipped in the same way. The same augmentation strategy is also carried out for optical flow and edge samples. Since the loss function of proposed framework is very sensitive to the initialization of weights. To solve this, we adopt an incremental strategy. We first train the two convolution and pooling layers with last fusion layer. We then add covolutional layer one by one before the fusion layer and re-train the network. After training all the layers, we then fine-tuned the network on datasets.

We evaluate and compare proposed approach with other reference methods using these two datasets. We follow the same convention used in other research work for segmentation task. For segmentation model, Area Under Curve (AUC) is widely adopted as performance measure. We report the performance of different methods in Table I using Shanghai Expo dataset. From the Table,it is obvious that Gaussian Mixture Model (GMM) performed relatively lower than other reference methods. We observed that GMM could not precisely segment the stationary crowd. The low performance of GMM attributes to limited number of sample used for training GMM. HOG+SVM method achieved comparable performance to extract the crowded regions from background. Although there is no overlap between training and testing data, there is overlap in local image patches due to similarity in appearance between the training and testing samples. From the Table, it is obvious that appearance and temporal network achieve similar performance, however, temporal model slightly outperforms appearance model. This is due to reason that input to the temporal model includes both appearance and motion information, since we are applying the background subtraction without setting up a threshold. Structure model takes stack of edge images as input, which is similar to HOG model, however, structure model performs slightly better than HOG+SVM method. Fusion model which is blend of three models, i.e., appearance network, structure network and temporal network achieves state-of-the-art performance. The fusion model learns hierarchical feature using appearance network. The integration of temporal and structural information further improves the results.

In Table II, we report the performance of different methods using City dataset. From the Table, it is obvious that GMM performance better on City dataset compare to Shanghai Expo dataset. This is due to fact that City dataset contains video sequences, where most of people are in motion, therefore, GMM model can easily detect the background. Moreover, compare to Shanghai Expo dataset, City dataset contains video sequences of longer duration, that means enough data is available for training GMM. HOG+SVM model performs lower on City dataset. This is due to reason that HOG features could not generalize most of video sequences. Similar to Shanghai Expo dataset, the appearance and temporal network perform higher compared to structure network. Fusion model, on the other hand beats all reference methods by an obvious margin.

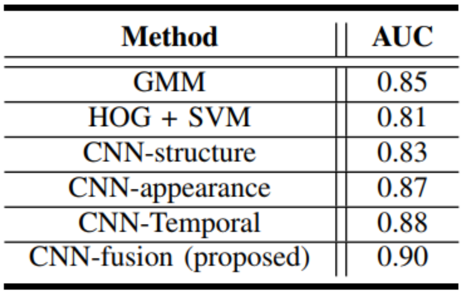
Table 1: Comparisons on Shanghai Expo Dataset



Method	AUC
GMM	0.75
HOG + SVM	0.79
CNN-Structure	0.78
CNN-appearance)	0.82
CNN-Temporal	0.85
CNN-fusion (proposed)	0.87

The evaluation results reported in Table I, II are obtained with a single threshold value (0.5). This threshold value is used to classify the pixel into two classes, i.e., crowd/noncrowd. From the empirical evidence, we observed that using a single threshold value can not alone justifies the performance of model. Although a single threshold divide the given data into obvious classes and it may be applicable in some of applications. As far as, crowd segmentation is concerned, a single threshold value can not generalize the performance of a model. Therefore, in order to get better insight into the performance of different methods, we use Receiver operating characteristic (ROC) curve.

Table 2: Comparisons on City Dataset



Method	AUC
GMM	0.85
HOG + SVM	0.81
CNN-structure	0.83
CNN-appearance	0.87
CNN-Temporal	0.88
CNN-fusion (proposed)	0.90

ROC curve plot the values between True Positive Rate TPR and False Positive Rate FPR with different threshold values. TPR is calculated as $\frac{TP}{TP+FN}$and FPR is computed As $\frac{FP}{FP+TN}$ True positive (TP) value shows the data is correctly classified, while False Negative that data is not correctly classified. We compute ROC curves for all methods using both dataset and the results are reported in Figure 4. Figure 4 illustrates that proposed method beats other reference methods by a significant margin in both datasets.

From the Tables I and II, it is obvious that appearance features perform better than structure and other background modeling techniques. In Figure 5, we show the qualitative results of our proposed method. From the Figure, it is obvious that our proposed method (CS-CNN(fused) precisely identify crowded areas in different scenes.

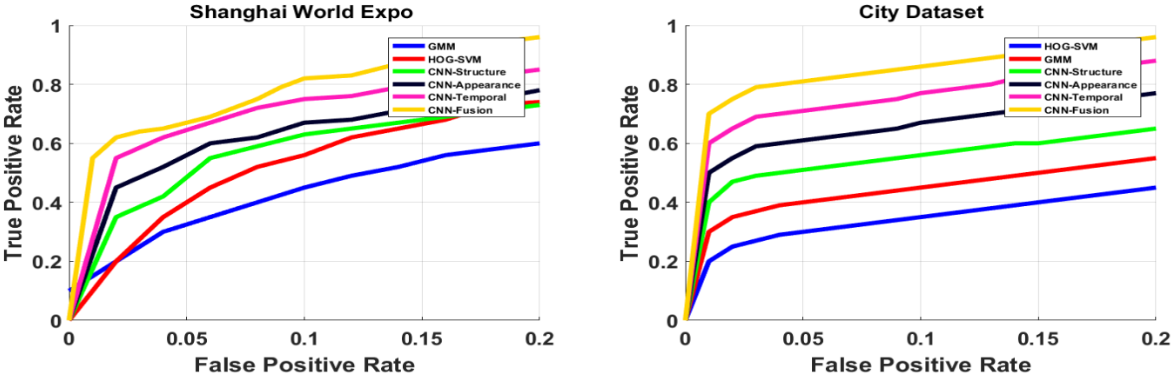


Fig. 4 Performance of different methods using ROC curves.

The left figure shows the performance of different methods using Shanghai Expo dataset. The figure on the right shows the performance of different methods on City dataset.

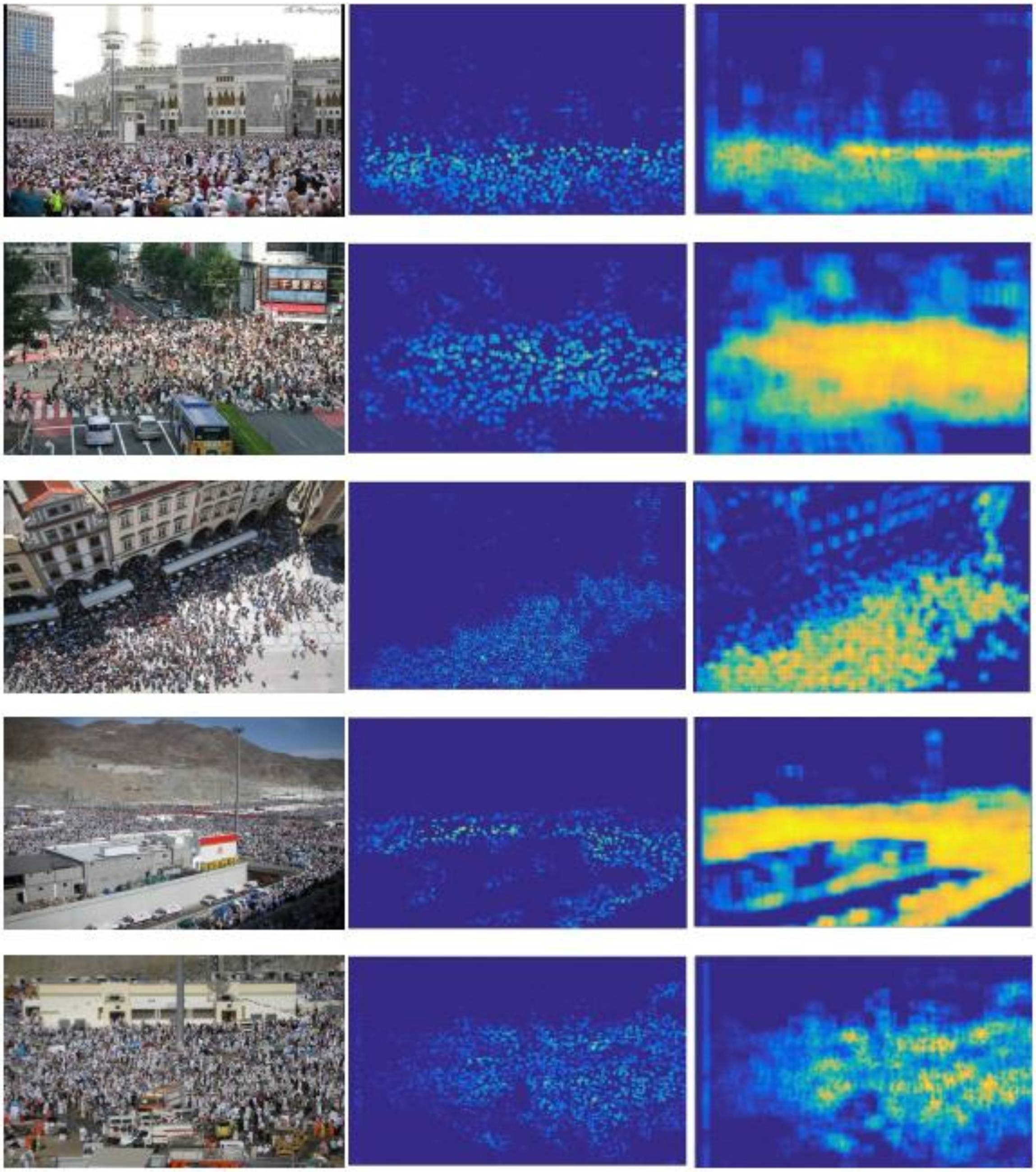


Figure 2: Visualization of segmentation maps. The first row show the sample images, the second row show segmentation maps obtained via CS-CNN(appearance) method while the third row depicts the segmentation results obtained via CS-CNN(fused) method. Blue color shows the background and yellow color represents the crowd regions

5. Conclusion

In this paper, we propose a novel crowd segmentation network, that exploits both structure and appearance features to precisely identify crowded regions in the image. We evaluate our framework on images acquired from different source. The images show different scenes, with significant variations in camera view points, densities and illumination. The proposed framework precisely identify crowd regions in complex scenes. We show that our proposed framework outperforms other state-of-the-art methods in both quantitative and qualitative way.

We believe that proposed framework will serve as a pre-processing unit in applications, like crowd counting, crowd behavior understanding and anomaly detection. In future, we will integrate the propose framework with aforementioned applications and will also evaluate the significance of its integration on their performance.

Acknowledgements

This work is supported by Deanship of Research, University of Ha'il under the grant no: BA-1910. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research

References

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–6. IEEE, 2007.

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In European conference on computer vision, pages 1–14. Springer, 2008.

[3] M. Arif, S. Daud, and S. Basalamah. People counting in extremely dense crowd using blob size optimization. Life Science Journal 9(3):1663–1673, 2012.

[4] M. Arif, S. Daud, and S. Basalamah. Counting of people in the extremely dense crowd using genetic algorithm and blobs counting. IAES International Journal of Artificial Intelligence, 2(2):51, 2013.

[5] G. Bertasius, J. Shi, and L. Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4380–4389, 2015.

[6] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 1, pages 594–601. IEEE, 2006.

[7] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In European conference on computer vision, pages 25–36. Springer, 2004.

[8] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE transactions on pattern analysis and machine intelligence, 30(5):909–926, 2008.

[9] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In Advances in neural information processing systems, pages 2843–2851, 2012.

[10] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami. Fast crowd segmentation using shape indexing. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE, 2007.

[11] C. T. Duong, R. Lebret, and K. Aberer. Multimodal classification for analysing social media. arXiv preprint arXiv:1708.02099, 2017. [12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence, 35(8):1915–1929, 2012.

[12] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multiscale counting in extremely dense crowd images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2547–2554, 2013.

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1):221–231, 2012.

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

[15] S. Khan, G. Vizzari, S. Bandini, and S. Basalamah. Detecting dominant motion flows and people counting in high density crowds. 2014.

[16] S. D. Khan. Estimating speeds and directions of pedestrians in realtime videos: A solution to road-safety problem. In CEUR Workshop Proceedings, page 1122, 2014.

[17] S. D. Khan. Congestion detection in pedestrian crowds using oscillation in motion trajectories. Engineering Applications of Artificial Intelligence, 85:429–443, 2019.

[18] S. D. Khan, F. Porta, G. Vizzari, and S. Bandini. Estimating speeds of pedestrians in real-world using computer vision. In International Conference on Cellular Automata, pages 526–535. Springer, 2014.

[19] S. D. Khan, G. Vizzari, and S. Bandini. Identifying sources and sinks and detecting dominant motion patterns in crowds. Transportation Research Procedia, 2:195–200, 2014.

[20] S. D. Khan, G. Vizzari, S. Bandini, and S. Basalamah. Detection of social groups in pedestrian crowds using computer vision. In International Conference on Advanced Concepts for Intelligent Vision Systems, pages 249–260. Springer, 2015.

[21] L. Kratz and K. Nishino. Tracking with local spatio-temporal motion patterns in extremely crowded scenes. 2010.

[22] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 878–885. IEEE, 2005.

[23] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 1975–1981. IEEE, 2010.

[24] H. Mansour, C. Dicle, D. Tian, M. Benosman, and A. Vetro. Method and system for segmenting pedestrian flows in videos, Feb. 13 2018. US Patent 9,892,520.

[25] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In Advanced Video and Signal

Based Surveillance (AVSS), 2017 14th IEEE International Conference on, pages 1–7. IEEE, 2017.

[26] A. Mumtaz, W. Zhang, and A. B. Chan. Joint motion segmentation and background estimation in dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 368–375, 2014.

[27] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In Computer Vision, 2009 IEEE 12th International Conference on, pages 261–268. IEEE, 2009.

[28] P. H. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In 31st International Conference on Machine Learning (ICML), number CONF, 2014.

[29] V. Rabaud and S. Belongie. Counting crowded moving objects. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 1, pages 705–711. IEEE, 2006.

[30] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Deepmatching: Hierarchical deformable dense matching. International Journal of Computer Vision, 120(3):300–323, 2016.

[31] J. Rittscher, P. H. Tu, and N. Krahnstoever. Simultaneous estimation of segmentation and shape. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 486–493. IEEE, 2005.

[32] M. Saqib, S. D. Khan, and M. Blumenstein. Texture-based feature mining for crowd density estimation: A study. In Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on, pages 1–6. IEEE, 2016.

[33] M. Saqib, S. D. Khan, and M. Blumenstein. Detecting dominant motion patterns in crowds of pedestrians. In Eighth International Conference on Graphic and Image Processing (ICGIP 2016), volume 10225, page 102251L. International Society for Optics and Photonics, 2017.

[34] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein. Extracting descriptive motion information from crowd scenes. In 2017 International Conference on Image and Vision Computing New Zealand (IVCNZ), pages 1–6. IEEE, 2017.

[35] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3982–3991, 2015.

[36] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pages 568–576, 2014.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[38] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah. Anomalous entities detection and localization in pedestrian flows. Neurocomputing, 290:74–86, 2018.

[39] H. Ullah and N. Conci. Crowd motion segmentation and anomaly detection via multi-label optimization. In ICPR workshop on Pattern Recognition and Crowd Analysis, 2012.

[40] H. Ullah and N. Conci. Structured learning for crowd motion segmentation. In 2013 IEEE International Conference on Image Processing, pages 824–828. IEEE, 2013.

[41] H. Ullah, L. Tenuti, and N. Conci. Gaussian mixtures for anomaly detection in crowded scenes. In Video Surveillance and Transportation Imaging Applications, volume 8663, page 866303. International Society for Optics and Photonics, 2013.

[42] H. Ullah, M. Ullah, H. Afridi, N. Conci, and F. G. De Natale. Traffic accident detection through a hydrodynamic lens. In Image Processing (ICIP), 2015 IEEE International Conference on, pages 2470–2474. IEEE, 2015.

[43] H. Ullah, M. Ullah, and N. Conci. Dominant motion analysis in regular and irregular crowd scenes. In International Workshop on Human Behavior Understanding, pages 62–72. Springer, 2014.

[44] H. Ullah, M. Ullah, and N. Conci. Real-time anomaly detection in dense crowded scenes. In Video Surveillance and Transportation Imaging Applications 2014, volume 9026, page 902608. International Society for Optics and Photonics, 2014.

[45] H. Ullah, M. Ullah, and M. Uzair. A hybrid social influence model for pedestrian motion segmentation. Neural Computing and Applications, pages 1–17, 2018.

[46] H. Ullah, M. Uzair, M. Ullah, A. Khan, A. Ahmad, and W. Khan. Density independent hydrodynamics model for crowd coherency detection. Neurocomputing, 242:28–39, 2017.

[47] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision, 63(2):153–161, 2005.

[48] H. Wang and C. Schmid. Action recognition with improved trajectories. In Proceedings of the IEEE international conference on computer vision, pages 3551–3558, 2013.

[49] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In Proceedings of the IEEE international conference on computer vision, pages 1385– 1392, 2013.

[50] S. Xie and Z. Tu. Holistically-nested edge detection. In Proceedings of the IEEE international conference on computer vision, pages 1395– 1403, 2015.

[51] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 589–597, 2016.

[52] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In null, pages 406–413. IEEE, 2004.

[53] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. IEEE transactions on pattern analysis and machine intelligence, 30(7):1198–1211, 2008.

[54] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., volume 2, pages 28–31. IEEE, 2004