IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 27, 2020, accepted May 14, 2020, date of publication May 19, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995764

Robust Head Detection in Complex Videos Using Two-Stage Deep Convolution Framework

SULTAN DAUD KHAN[1], YASIR ALI[2], (Member, IEEE), BASIM ZAFAR[2], AND ABDULFATTAH NOORWALI[3], (Member, IEEE)

[1]Department of Computer Science, National University of Technology, Islamabad 42000, Pakistan
[2]Expert Vision Consulting, Makkah 24381, Saudi Arabia
[3]Department of Electrical Engineering, Umm Al-Qura University, Makkah 24381, Saudi Arabia

Corresponding author: Sultan Daud Khan (sultandaud@nutech.edu.pk)

The work was funded by grant number 14-INF1015-10 from the National Science, Technology, and Innovation Plan (MAARIFAH), the King Abdul-Aziz City for Science and Technology (KACST), Kingdom of Saudi Arabia. We thank the Science and Technology Unit at Umm Al-Qura University for their continued logistics support.

ABSTRACT Pedestrian head detection plays an important role in identifying and localizing individuals in real world visual data. Head detection is a nontrivial problem due to considerable variance in camera viewpoints, scales, human poses, and appearances in the scene. Thanks to the translation invariance property of convolutional neural networks (CNNs) which enables large capacity CNNs to handle the problem of appearance and pose variations in the scene. However, the problem of scale invariance is still an open issue. To address this problem, this paper presents a two-stage head detection framework that utilizes fully convolutional network (FCN) to generate scale-aware proposals followed by CNN that classifies each proposal into two classes, i.e. head and background. Experiments results show that using scale-aware proposals obtained by FCN, the object recall rate and mean average precision (mAP) are improved. Additionaly, we demonstrate that our framework achieved state-of-the-art results on four challenging benchmark datasets, i.e. HollywoodHeads, Casablanca, SHOCK, and WIDERFACE.

INDEX TERMS Convolutional neural networks, non-maximal suppression, head detection, crowd counting, motion analysis.

I. INTRODUCTION

For many vision based applications, pedestrian and human face detection is a pre-processing step. These applications include person identification [53], [56], action recognition [14], [37], tracking [40], autonomous driving, behaviors understanding [15], [16]. While these algorithms have gained maturity in recent years [28], [46], the problem of detecting pedestrians in natural images and videos is still challenging. Face detector can not extract facial feature for a person whose face is not visible. On the other hand, person detection is challenging job. This is due to reason that large portion of human body is not visible due to occlusion and clutter in the scene. This is due to reason that face and pedestrian detection methods are not applicable in natural scenes. Therefore, to find people in unconstrained images and videos, head is an indispensable choice.

The goal of head detector is to precisely detect and localize human heads in naturalistic conditions. Precise head detection is an important element and used as a pre-processing step in many video surveillance applications, for example, tracking [3], [12], person authentication [25] and density estimation [36]. During the recent years, few strides have been made towards head detection in crowds [7], [21], [39] in complex scenes, however, head detection is still a challenging task. Significant variations in poses, scales, and appearances of human heads, make the head detection problem even more challenging.

A reliable head detection system should be invariant to scales, appearances and poses. Figure 1, highlights these problems, where three human heads are marked in red, green and yellow colors. From the Figure, it is obvious, that heads have different scales (sizes), poses and appearances. Convolutional neural networks (CNN) are inherently transnational invariant. Due to this property, large capacity CNN can handle variation in pose and appearance. However, CNNs are not inherently scale invariant and still have room for improvement.

Generally, most of the existing methods deal the head detection as a special case of generic detection problem.

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

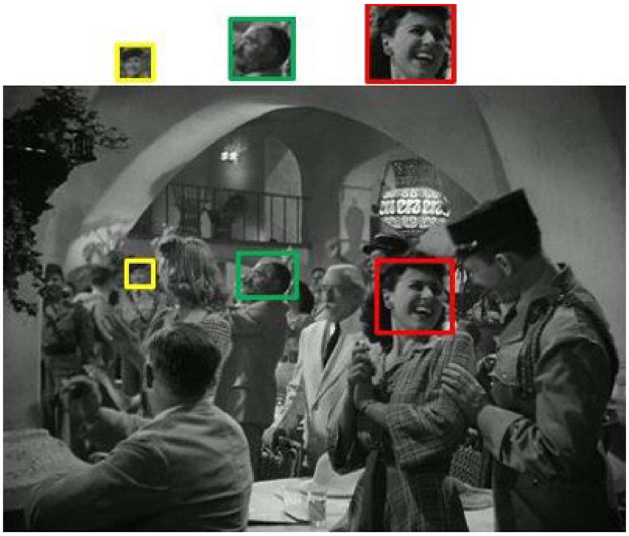


FIGURE 1. Illustrates scale, pose and appearance variations of human heads highlighted in red, green and yellow colors. Perspective distortions cause drastic changes in scale of human head due to due to which the size of head (in red) appears large compared to heads with green and yellow bounding boxes.

The detection pipeline of these detectors consist of two stages, (1) object proposal generation, (2) classification of object proposals. Therefore, for two-stage object detectors, object proposal generation is an important pre-processing step. Generally, in contrast to exhaustive search for object in image, object proposals guide the search for objects. Generating object proposals is the preferred choice for object detection over sliding windows approaches due to the following reasons: (1) save computation time by passing small number of proposal to the detector, and (2) improves the precision and recall rate. Acknowledging the importance of object proposals in object detection tasks, several methods for generating object proposls have been reported in literature during the recent years.

Recent object proposal generators exploit saliency, gradient and edge information [5], [59] to hypothesize the location of objects in images. Later on, DeepBox [20] move a step forward and refined the proposal generated by EdgeBox [59]. DeepProposal [9] utilize initial and final layers of the network in inverse cascade fashion to generate object proposals. Multi-Box [23] employs regression to extract object regions.

Usually, detector face challenges to detect head in natural scenes, since human heads have significant variations in object scales, appearances, and poses as mentioned before. Therefore, current existing two-stage methods usually achieve low precision and recall rates when tested in natural scenes. To address the problem of scale, we propose a novel strategy to detect human heads in complex scenes.

Precisely, we propose head detector to detect heads with multiple scales in various complex scenes and follows the following sequential pipeline:

1) The first part is multi-scale object proposal generation network, that captures the distribution of scales in the input image by generating scale-specific object proposals. Concisely, a binary classifier is trained by employing [24] using patches belonging to human heads. The input to network is arbitrary size image and output is a dense heat map. Dense heat map represents the confidence whether a specific region contains a human head or background for every pixel. In order to generate multi-scale object proposals, we re-size the input image into multi-scales (image pyramid), pass image pyramid to the network and obtain the multiple heat maps corresponding to levels of pyramid. Non-maximal suppression technique is then employed to reduce the redundant and obtain refine proposals.
2) The second stage is the classification, where each proposal is classified into two classes (head/background). We use different state-of-the-art network architectures, like AlexNet [19], VGGS [4], VGG-verydeep-16 [38] and ZF [54]for classification.

The contribution of this paper lies in the first part of the proposed framework. Compared with the existing methods, our framework has the following contributions:

1) We propose a novel framework to handle the scale problem by generating scale-aware proposals using Fully Covolutional Network that generate pixel-wise head scores and square shape bounding boxes of the head instances through various scales and location of the input image
2) With the adoption of anchor-free scale-specific region proposal network, our framework has significantly reduced the time cost as compared to feed-forwarding single object proposal through the CNN.
3) Compared to dense networks, e.g. GoogleNet, we train a shallow network for head/background classification. This model can be adapted to dense prediction of human heads in images with arbitrary sizes.
4) The proposed framework shows superior performance using challenging datasets.

Comparison and Difference: The proposed proposal generation network is superficially similar to typical Region proposal network (RPN) adopted by FRCNN. However, it differs in many aspects, for example, FRCNN uses a large receptive field to detect generic objects in images. Usually, these objects are large and occupy large portion of the image. These objects can easily be detected by FRCNN, however, FRCNN faces difficulties in detecting small objects, where the size of objects is less than 16 pixels. This is due to the reason, that ROI pooling layers of FRCNN use feature maps from highest convolutional layer. These feature maps have reduced resolutions and lost most of the important information related to small objects. Therefore, FRCNN can not precisely classify and predict the location of small objects. Another flip side of FRCNN is that it uses anchor boxes with predefined sizes and scales. To achieve high precision and recall, anchor boxes should be of different sizes and scales to cover size and shape variations of generic objects in image. As in crowded scenes, the size and shape of heads change significantly as compared to generic large objects, it requires much more complex design of anchor boxes to capture wide range of scales. Therefore, anchor boxes based methods are inefficient in such cases. Our proposed framework is different from FRCNN in following ways. (1) The most important difference is that the proposed RPN is anchor-free and class specific proposal generator in contrast to anchor based generic proposal generator.(2) We trained a head descriptor that can detect head in extreme scales by incorporating features from multi-scales using image pyramid.

The rest of the paper organized as follows. In Section II, we discuss related works. Section III discusses proposed methodology. Experiments results on different data sets are reported in Section IV. Conclusion is presented in Section V.

II. RELATED WORKS

Since our framework has two sequential parts, i.e,. object proposal generation and head detection, therefore we discuss related work in separate subsections.

A. GENERATING OBJECT PROPOSALS

We categorize object proposal methods into two categories: *1. Segment based methods* and *2. Window scoring methods*. In addition to these methods, we also discuss CNN based approaches.

The goal of segment based methods is to generate multiple segments from the image that may contain objects. These methods typically start with initial over-segmentation followed by different merging strategies to cluster similar segments based on color, texture, and shape into object proposals. For example, Selective Search (SS) [41] generates object proposals by greedily merging super-pixels without learning. Randomized Prim [26] utilizes connectivity graph to learn randomized merging strategy. Graph cut is used in [29] to merge super-pixels to generate proposals. Multiscale Combinatorial Grouping (MCC) [2] expolits mutli-scale hierarchical segmentation to obtain object proposals. Reference [18] measures geodesic distance transform between multiple segments, where distance transform represents object proposals. The above mentioned methods achieve high recall rates, however, these methods are computationally expensive since proposals are obtained by multiple segmentation in multiple scales and color spaces.

On the other hand window scoring methods show the likelihood of a window to contain an object of interest and therefore are computationally efficient as compared to segmentation based methods. Generally, these methods first generate candidate object proposals (bounding boxes) in multiple locations and scales. Then high confidence boxes are selected as object proposals. Objectness [1] selects the high rank proposal on the basis of low-level cues, such as, edge, size, location and color. BING [5] trained a linear Support Vector Machine and applied it in a sliding window fashion on gradient map. Similarly, Edge Boxes [59] also follows sliding window fashion and associates score to the windows base on edge map. In contrast to segment based methods, window scoring approaches are fast. However, these methods, due to sampling of proposal at discrete levels, results in poor detection accuracy.

Due to the popularity of deep learning models, CNNs are also explored for object proposal generation task. Overfeat [34] trained a deep model that operates in sliding windows fashion and simultaneously predict bounding box and score for each object. MultiBox [8] also trained a CNN that generates fixed number of proposals without adopting sliding window strategy. DeepBox [20] on the other hand, does not output the proposals by itself but re-ranks the proposals generated by other methods.

Our proposed object proposal approach falls into the category of *Window scoring method*. However, our methods adopts different scheme by exploiting fully convolutional network that outputs heat maps. Our method is similar to Region Proposal Network (RPN) [32], which is employed in [32] for object proposal generation. RPN generates fixed number of proposals based on pre-defined anchor boxes. Compare to RPN, our method neither generates fixed number of proposals nor based on pre-defined anchor boxes. Unlike other methods Overfeat [34], MultiBox [8], and RPN [32], our method does not regress the bounding boxes. Instead we adopt a mapping scheme, where each pixel in heatmap corresponds to the a window in the input image. The integration of localization information with scale-specific strategy achieves better performance and achieves high recall rates than bounding box regression methods.

B. HEAD DETECTION

Most of related works deals head detection problem as special case of object detection. Traditional head detection methods learn hand-craft features by a non-linear classifier. For example, the classical method proposed by Viola and Jones [43] extracts Haar-like features from the image and employed cascade booting classier for classification. In [33], authors move a step forward and refine the results of Viola and Jones by exploiting spatial and temporal information using Conditional Random Field (CRF). Deformable part model (DPM) [47] utilized Histogram of oriented gradients (HOG) features and was widely adopted model in object detection tasks. However, these traditional methods receive performance setback and cause high computational cost in real world scenes.

Convolutional Neural Networks achieve enjoyed tremendous success in classification, and segmentation task. Following the success of CNN, deep neural networks becomes the first choice for object detection task. The most efficient step in this direction is taken by Region based convolutional neural network (RCNN) [10]. RCNN is a two-stage framework, where the first step involves generation of object proposal (around 2000) by employing Selective Search (SS) method. The proposals generated by SS method are then feed to feed-forward network. The network then extract hierarchical features from the (5th layer of AlexNet). A linear

SVM classifier is then learned using the hierarchical features extracted from the last convolutional layer. Although R-CNN achieved state-of-the-art results, however, it also suffers from computational complexity. A more refined version, Faster R-CNN is proposed that replaced traditional Selective Search strategy by RPN. You only look once (YOLO) [30] generates bounding boxes using regression and classify each bounding box by assigning class scores to the bounding boxes. YOLO beast Faster-RCNN in terms of inference speed on most of existing object detection datasets, however, at the cost of accuracy. Single shot detector (SSD) [23] generate fixed number of bounding boxes by utilizing fully convolutional network.

Although the above existing models achieve considerable performance in classifying multiple objects in image, however, they face challenges in detecting small objects. It is due to the fact that most models utilize features from the last convolutional layer for object detection. However, last convolutional layers contain inadequate information regarding small objects. Since, in head detection problem, where the size of target (head) is usually small (upto 10-20 pixels), therefore, current existing methods in the current form are not applicable for detecting small objects.

III. PROPOSED METHODOLOGY

A. NETWORK DESIGN FOR OBJECT PROPOSALS

In this section, we discuss the proposed architecture for generating scale-aware proposals. Fully Convolutional Networks (FCNs) become dominant in image segmentation tasks that take an arbitrary size input and predict dense output of the same size. The output of FCN may also be used in dense prediction tasks (e.g., image restoration, depth estimation and semantic segmentation). Our multi-scale object proposals generation framework is based on FCN which takes whole image as input and produces a high level semantic heat map. All pixels in the output heat map represent to what extent different regions in the input image contain human heads. In short, we train a binary classifier (head/background) using patch wise training strategy with annotated heads. The framework slides over the image with a network stride and feed-forward each sampled window to a binary classifier. The output is heat map, where each pixel represents the confidence value of one of the window (corresponds to patch) in the input image as shown in Figure 3. Generally, fully convolutional networks are more efficient compare to existing sliding window methods as they share the computation among overlapping windows. Moreover, FCNs are translation invariant and take arbitrary size image as input.

For input image, the size of image patch (window size) corresponds to pixel in the heat map is called the *Scale* (receptive field size) of the network. Several parameters affect the scale of the network, for example, depth of the network, sizes of convolution and pooling layers and stride settings. Lets assume R_i represents the receptive field of network layer i, where $i = \{1, 2 \ldots, n\}$ and n represents the total number of layers in the network. Then the scale of the network is R_1 and we can compute the receptive field of any layer i of the network using the recursive formulation 2.

$$R_{i-1} = r_i(R_i - 1) + k_i \tag{1}$$

where r_i represents the convolution or pooling stride and k_i shows the size of kernel of the i^{th} convolution/pooling layer. R_i and R_{i-1} represents the receptive field of $i - 1^{th}$ and i^{th} layer respectively. To precisely map any pixel in the heat map to the corresponding window region in an image, we need to compute receptive field (scale) and stride (or network stride), N_s. One inherent issue with FCN is that N_s is computed by the network itself and is equal to the product of strides of all network layers.

With the known receptive field size R_1 and network stride N_s, we can compute window region in the input image which corresponds to the pixel in the heat map. Let (x_o, y_o) represents a pixel in the heat map. We can compute its corresponding window $W = \{x_{min}, x_{max}, y_{min}, y_{max}\}$ in the input image as follows, $x_{min} = x_o N_s$, $x_{max} = x_{min} + R_1$, $y_{min} = y_o N_s$, $y_{max} = y_{min} + R_1$.

We train the network from scratch and the details of the proposed network architecture is shown in the Table 1. Our architecture follows the geometry of the AlexNet [19] for first five convolutional layers. In our design, we convert the 6^{th} of AlexNet to full convolution layer with the kernel size of 6×6. The last 1×1 convolutional layers follows Network in Network (NN) [22]. Each convolutional layer of the network is followed by a ReLU layer. We use softmax layer on the top of the network that predicts the confidence score within the rage of 0 and 1 by optimizing cross entropy loss 2.

$$L = t_g \log(p_k) + (1 - t_g) \log(1 - p_k) \tag{2}$$

where t_g represents k^{th} ground truth value and p_k denotes k^{th} prediction value.

For training, in contrast to feeding whole image, we adopt patch wise training strategy. For generating the training data, we crop positive patches with annotated heads and re-size them to the input size of the network (224 in our case). We also crop several patches around the human heads with Intersection over Union (IoU) ≥ 0.5 and are treated as positive samples. This step is performed to increase the amount of positive patches and to balance the data (number of positive and negative patches). For the negative samples, we sparsely sampled patches from the background with IoU < 0.5. IoU is computed as the intersection of a candidate box and ground truth box divided by area of their union.

For all layers of the network, we use zero-mean Gaussian distribution to initialize the weights. We keep standard deviation to 0.01 and the biases with 0. We adopt stochastic gradient descent (SGD) during the training process and learning rate to 0.01. We reduced the learning rate 10 after every 40 epochs. We set the batch size to 256.

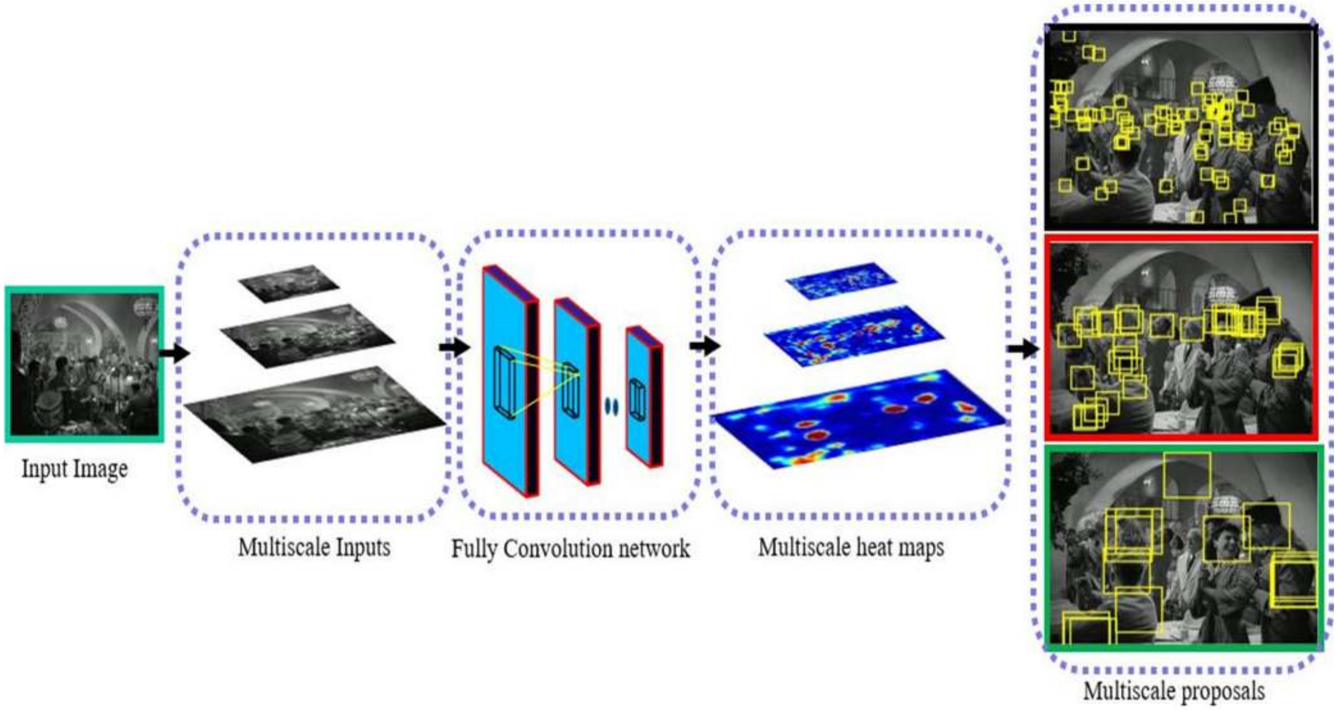


FIGURE 2. Pipeline of framework for generating scale-specific proposals.

TABLE 1. Fully convolutional network based scale-aware proposal architecture.

Layers	1	2	3	4	5	6	7	8	9	10
Type	conv + pooling	conv + pooling	conv	conv	conv + pooling	conv	conv	conv	conv	conv
#Channels	96	256	384	384	256	6144	6144	2048	1024	2
Conv. size	11 x 11	5 x 5	3 x 3	3 x 3	3 x 3	6 x 6	1 x 1	1 x 1	1 x 1	1 x 1
Conv. stride	4 x 4	1 x1	1 x 1	1 x 1	1 x 1	1 x 1	1 x 1	1 x 1	1 x 1	1 x 1
Pooling size	3 x 3	3 x 3	-	-	3 x 3	-	-	-	-	-
Pooling stride	2 x 2	2 x 2	-	-	2 x 2	-	-	-	-	-
Zero padding	-	2 x 2	-	1 x 1	1 x 1	-	-	-	-	-

B. MULTISCALE OBJECT PROPOSALS

In this section, we discuss our strategy of generating multiscale proposals. For fully convolution network discussed above, pixels in the heatmap cover windows of fixed size R_1 in an image. Therefore, FCN can only detect heads with size R_1 in the original image. However, the size of heads varies significantly due to perspective distortions. Therefore, to generate object proposals that captures different sizes of the human heads, we re-size the original input into multiple sizes and generate an image pyramid.

After generating image pyramid, we then feed each re-sized image of the pyramid to the network and predict the corresponding heatmap. The heatmaps generated by different layers of the pyramid will have different receptive fields. Figure 4 shows the input original image which is re-sized to different sizes, i.e. 28 × 28, 56 × 56 and 112 × 112 and then feed to the network one by one. We predict the corresponding heatmaps as shown in the Figure 4. From the Figure, we infer that heatmap corresponding to the smaller scale (28 × 28), the network gives higher response on smaller heads while low response on bigger heads. In the same way, the network characterizes bigger heads in large scale, i.e. 112 × 112. With this motivation, we propose multiscale strategy to generate scale-aware proposals that captures different sizes of heads in the image. The proposed pipeline for generating multiscale proposals is shown in Figure 2.

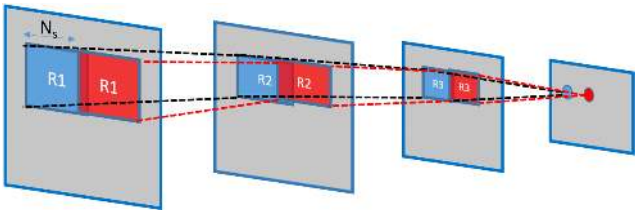


FIGURE 3. Illustration of the effect of receptive field. Two pixels red and blue heat map show the classification confidence values of the red and blue windows R_1 and will not be affected by each other. N_S is the stride of the network.

Acknowledging the effectiveness of multi-scale strategy, we now find the set of scales required to precisely detect all

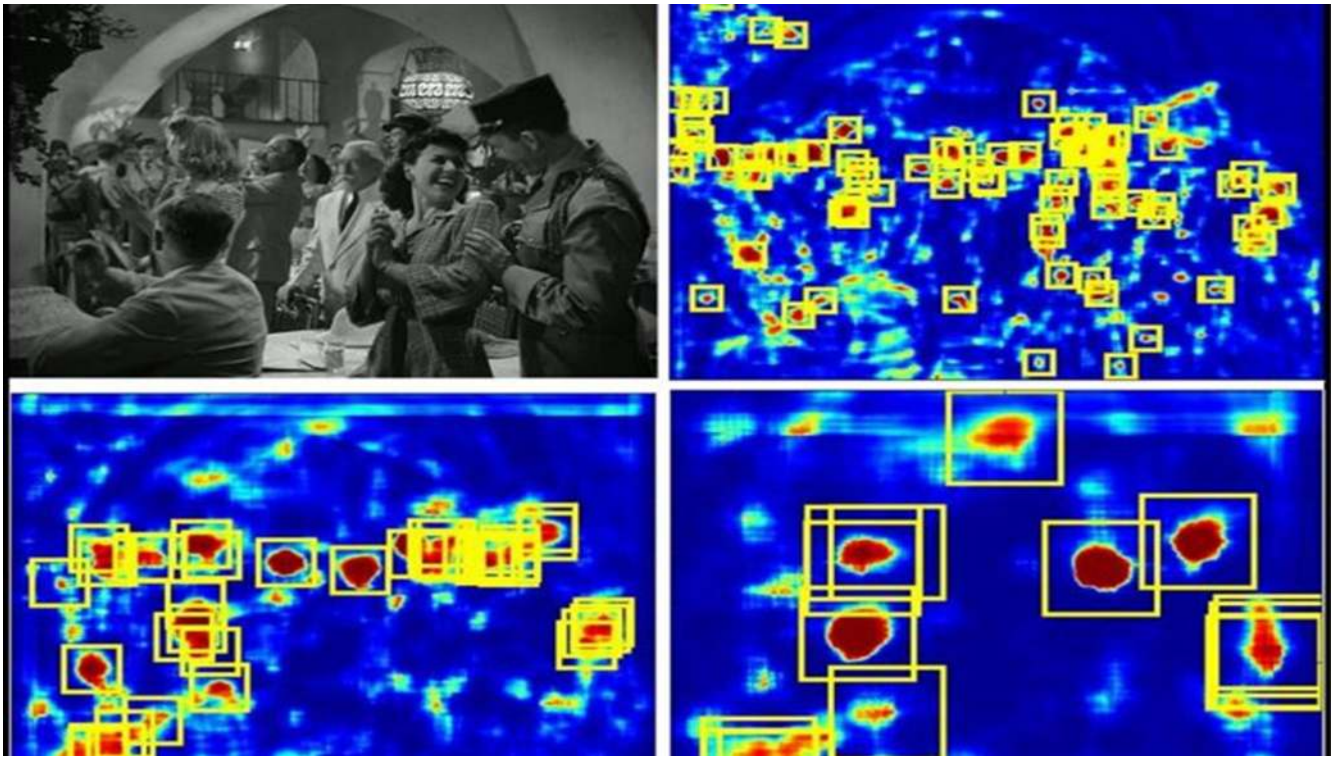


FIGURE 4. Depiction of original image (top left), heatmap of scale 28 × 28 (top right), heatmap of scale 56 × 56 (bottom left) and heatmap of scale 112 × 112 (bottom right). All heatmaps are re-sized to original image size. Pixels in the heatmap represent the score of corresponding square window contain human head.

human heads in the given image. Generally, large set of scales results in large number of proposals concentrated around the regions containing head. However, this setting produces large number of false bounding boxes (not likely to contain head) which may lower the recall. On the other hand, small set of scales usually missed the objects in the image and results in lower precision. This issue rises a trade off in selecting the parameter for multiscale settings.

We use the values of the scale, ranges from a minimum bounding box size of 28 × 28 (784 pixels area) to the full resolution of an image. For the head detection, we keep the aspect ratios as $\Re \in [\frac{2}{3}, \frac{3}{2}]$ for all bounding boxes. The exact values of the scale S can computed as follows,

$$S = \sqrt{784}(\sqrt{\frac{1}{\alpha}})^{r} \tag{3}$$

where r is the index and takes value from range 0 to $[\log(\frac{I}{\sqrt{784}})/\log(\sqrt{\frac{1}{\alpha}})]$, where I is the image size and we define α as the step size of the scale and representing IoU for neighboring boxes [59]. In all our experiments, we fix the value of α to 0.65 as it is ideal for most of cases [59]. After obtaining multiscale proposals using different heatmaps, we then remove bounding boxes with score lower than 0.3. This step will significantly minimize the number of proposals. In the next step, we sort all the remaining proposal in descending order and apply non-maximal suppression (NMS). In all our experiments, we fixed the threshold value to 0.8.

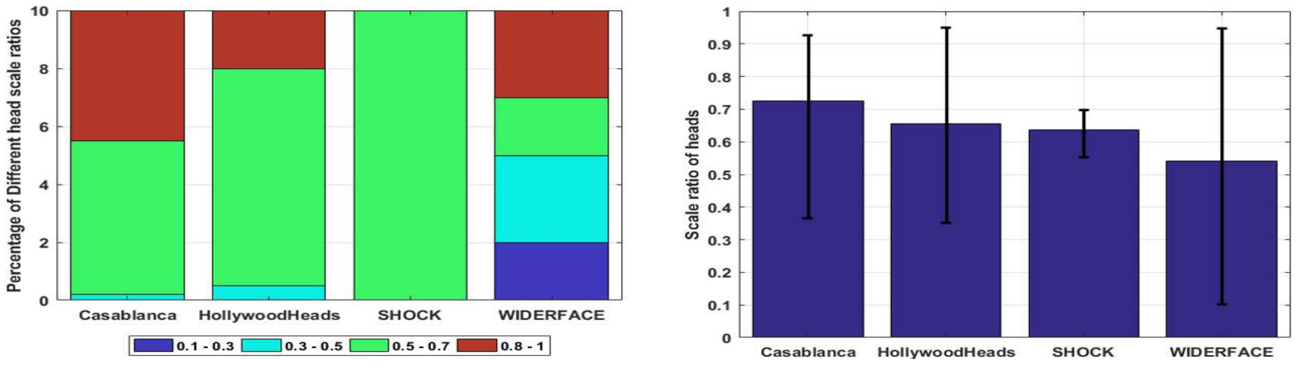
C. HEAD DETECTION

After obtaining multiscale proposals by using the above mentioned multiscale strategy, we then classify each proposal into two classes, i.e. head and background. Our head detection framework follows classical R-CNN [10] approach and instead of selective search [42], we use proposals generated by our multiscale strategy. Before feeding the proposal to a network, for each proposal, we process each proposal in following way. 1) Extend the bounding box by a small scalar value. 2) Crop patch corresponding to each proposal from image. 3) Re-size image patch to make it fit to the input layer of the CNN. For the classification, we use different architectures, AlexNet [19], VGGS [4], VGG-verydeep-16 [38], and ZF [54].

IV. EXPERIMENT RESULTS

In this section, we evaluate the performance of proposed framework using four publicly available datasets, i.e. SHOCK [6], WIDERFACE [51], HollywoodHeads [45] and Casablanca [33].

SHOCK dataset is proposed by Conigliaro *et al.* [6]. The dataset captures 100,000 spectators from all over the world to watch an ice hockey match held in Trento, Italy. The datasets contains 75 video sequences captured from five different



(a) Summary of the distribution of number of scales belong to four groups.

(b) Standard deviation of scales

FIGURE 5. Comparison of datasets in terms of head scale ratios.

cameras and covers four ice hockey matches on different days. Two different types of cameras were used to record the video sequences. To capture panoramic and ice rink view, full HD camera with resolution of 1920 × 1080, focal length 4 mm and with frame rate of 30 fps is used. To cover different locations of spectators crowd, three cameras with resolution of 1280 × 1024, focal length 12 mm, with frame rate of 30 fps were mounted at different locations of the stadium. The video sequences are annotated in different ways to evaluate different crowd analysis methods, for example, face detection, pose estimation, action recognition, and posture detection.

WIDERFACE dataset is proposed by Yang *et al.* [51]. This dataset is used to evaluate face detection methods. The dataset is composed of 32,203 images and 300,000 face annotations (bounding boxes). The dataset is 10 times larger than existing face detection datasets. The images collected from different sources with varying view points, resolutions, scales, poses and densities. This data set has unique properties. The faces are divided into groups based on scales, occlusion, pose and events. The dataset has unique property of arranging the faces into three groups, i.e., small, medium and large based on face size. The small group covers faces of size 10-50 pixels, medium (50-300 pixels), and large contains human faces of size greater than 300 pixels. In the same way, to evaluate detector performance on handling occlusion, faces are divided into three categories, high occlusion, no occlusion, and medium level occlusion. We use three groups of scales, small, medium and large to evaluate and compare the performance of proposed framework and other reference methods.

HollywoodHeads dataset is first proposed by Vu *et al.* [45]. The dataset is collected from 21 Hollywood movies scenes and contain 224,740 images. This dataset contains 369,846 annotations. The human heads were annotated in different key frames and remaining frames are annotated by using linear interpolated. These annotation are then verified by multiple coders. The dataset is divided into 216,719 training frames from 15 movies, 6,719 frames for validation sampled from other 3 movies and 1,302 frames are sample from remaining 3 movies. We followed the same convention in our evaluation of proposed framework.

Casablanca dataset is first proposed by Ren [33]. The dataset is collected from old movie named "Casablanca". The dataset contains 147600 frame of resolution 464 × 640. Casablanca dataset contains the annotations that mostly cover the frontal heads which have different scales and aspect ratios.

1) COMPLEXITY OF DATASETS

In this section, we discuss and compare the complexity of datasets. As discussed above, scale problem is caused by perspective distortions in the image that is induced by camera view point. Due to perspective distortions, the size of human heads near to camera appear large, while the size of human heads become smaller as with distance from the camera increases. Objects appears at various scales in natural images that may compromise the detector's performance. Therefore, scale problem lies in the heart of every object detector [57] and good object detector should overcome the scale problem. To demonstrate the complexity of dataset in terms of scale variations, we plot the distribution of the entire scale space of heads/face for all datasets as shown in Figure 5 (a). *Scale* is computed as ratio of size of head/face, H to image size I. Size of head/face H is the maximum of width and height of bounding box and I is the maximum of width and height of image. We compute scale of all heads/faces in all datsets.We divide entire scale space into four groups, i.e., very small (0.1 - 0.3), small(0.3 - 0.5), medium(0.5 - 0.7), large(> 0.75). We count the number of scales belonging to four groups and generate histogram for all dataset as shown in Figure 5. From the Figure, it is obvious that *casablanca* and *HollywoodHeads* datasets contain human heads belonging to medium and large groups. *SHOCK* dataset contains heads belonging to the medium group while *WIDERFACE* dataset is diverse and contains heads from all four groups. We further illustrate the complexities of the datasets by plotting standard deviation of scales in Figure 5 (b). From the Figure, it is clear that *SHOCK* dataset produces low standard deviation compare to other datasets. The small standard deviation shows small scale variance that can be easily capture by a single scale detector. On the other hand, standard deviation of *WIDERFACE* is high that shows that the existence of large scale variance and

TABLE 2. Summary of datasets.

Dataset	# of frames	Type	Resolution	Color	Location	# of Annotations
HollywoodHeads	224,740	Video sequences		Grey	Outdoor/Indoor	369,486
Casablanca	147,600	Video sequences	976 x 720	Grey	Indoor	3700
WIDERFACE	32,203	Images	Various	RGB	Outdoor/Indoor	300,000
SHOCK	69,750	Video sequences	1920 x 1080 1280 x 1024	RGB	Indoor	$\geq$ 60,000

requires multi-scale detector. We also report the summary of datasets in Table 2.

We compare the performance of proposed framework with other state-of-the-art methods using following performance metrics: object recall, and detection mean average precision (mAP).

For two stage detectors, it is important that object proposal generator should cover all object of interest. Objects missed during the object proposal stage will never be classified during the classification stage. This will reduce the object recall rate for the classifier. Generally, the performance of detector depends on the performance of object proposal generator. Therefore, it is important to evaluate the performance of object proposal stage. Generally, object recall rate is used as an evaluation metric to evaluate the performance of object proposal stage. We compare our proposed object proposal generation framework with the other reference methods, including Bing [5], Region proposal network (RPN) [32], MultiBox [23], EdgeBox [59] and SelectiveSearch [10].

For computing the object recall, we find the matching by computing intersection over union (IoU) between the object proposal and the ground truth. Figure 6 (a) shows the object recall of different methods at fixed IoU threshold (0.6) with the increasing number of proposals. From the Figure 6 (a), it is obvious that our approach out performs other state-of-the-art methods for both small and large number of proposal at fixed threshold (0.6). It can also be noticed from the Figure 6(a) that even for small number of proposal (1000), our approach performs comparatively better.

We next evaluate the performance of different methods by computing object recall for fixed number of proposals (2000) and change IoU values within the range of [0.5, 1] as shown in Figure 6(b). It can be seen that our approach beats other state-of-the-art methods by a considerable margin with IoU changes from 0.5 to 1. The superior performance of our approach attributes to the fact that we utilize multi-scale prediction strategy. This strategy has the ability to capture scale variations and results in high object recall rates.

A. COMPARISON WITH GENERIC DETECTORS

We now evaluate and compare the performance of our proposed framework with other generic object detectors. Generally, we categorize generic object detectors into two groups: (1) two stage frameworks and (2) single stage frameworks. Two stage detection frameworks incorporate generation of region proposals as a pre-processing step while one stage detection frameworks are free from region proposals.

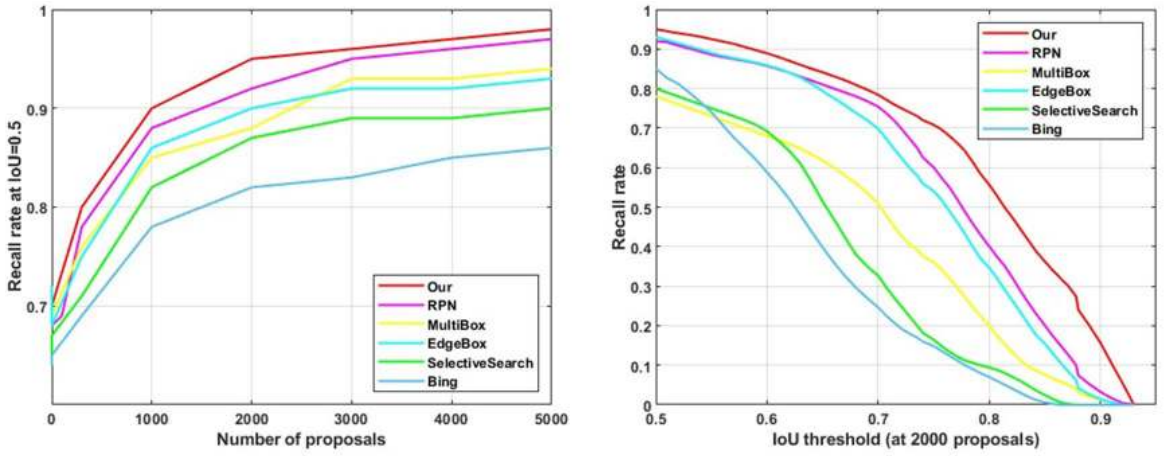
We utilize different region proposal methods with different backbone CNN architectures. Faster-RCNN [32] uses a fully convolution network named as Region Proposal Network (RPN) for generating region proposals. The features map generated from the last convolution layer is used to generate regions proposals of different sizes and aspect ratios. We combine R-CNN with MultiBox and Selective Search which utilize low-level image features for generating object proposals. We also compare our results with Cascade Rejection classifier (SDP+CRC) [49] which utilizes EdgeBoxes for object proposals.

It is important to note that for generating object proposals, we fine-tuned the pre-trained models of Selective Search, MultiBox and EdgeBoxes on HollywoodHeads dataset according to the original splits. We also use single stage detection frameworks and directly employ the publicly available pre-trained models of You Only Look Once (Yolo) [31] and Single Shot Detector (SSD) [23] during testing phase. We use average precision (AP) with a threshold of 0.5 IoU as a performance measure based on precision-recall curves. The results are summarized in Table 3. From the Table, it is obvious that the proposed framework outperforms all state-of-the-art detectors. While Faster-RCNN produce comparable results. The performance of Yolo and SSD are relatively lower than rest of detectors. We attribute their inferior performance to the following two reasons: (1) show poor generalization capability when applied to new datasets. (2) Both Yolo and SSD suffer from the problem of detecting small objects compared to region based object detection methods. It may be the reason that both these methods are using feature maps of low resolution due to which small objects features become too small to be detected.

B. COMPARISON WITH SPECIFIC HEAD DETECTORS

In this section, we evaluate and compare proposed framework with specific detectors. We divide specific detectors into two groups: (1) *Face detectors* and (2) *Head detector*. For comparison on SHOCK, HollywoodHeads and Casablanca dataset, we use one set of face detectors that includes VJ-LBP [43], VJ-HOG [43], TinyFace [13], Faceness [50] and FaceHunter [27]. For WIDERFACE dataset, we use different group of face detectors that includes CMS-RCNN [58], Multitask-CNN [55], ACF [48], and TinyFace [13]. While head detector group includes, DPM-Head [52], DHD [35], FCHD [44] and DISAM [17].

For fair comparisons, we first train each detector on ImageNet dataset and then finetune on each of the analyzed



(a) Recall vs number of proposals at IoU = 0.6

(b) Recall vs varing IoUs using fixed number of proposals (2000)

FIGURE 6. Comparison of object recall of different object proposal methods.

TABLE 3. Performance of different detectors using different region proposal methods on HollywoodHead, Casablanca, SHOCK, and WIDERFACE datasets.

Detector	Region proposal method	Backbone CNN	mAP @ IoU = 0.5			
			HollywoodHeads	Casablanca	SHOCK	WIDERFACE
Faster-RCNN [32]	RPN	VGG [4]	0.78	0.51	0.67	0.48
		VGG16 [38]	0.79	0.55	0.69	0.51
		ZF [54]	0.76	0.48	0.62	0.46
RCNN [10]	MultiBox [8]	VGG [4]	0.72	0.52	0.65	0.47
		ZF [54]	0.71	0.54	0.64	0.45
		VGG16 [38]	0.74	0.48	0.69	0.53
RCNN [10]	Selective Search [41]	VGG16 [38]	0.67	0.46	0.56	0.37
		ZF [54]	0.69	0.53	0.58	0.39
		AlexNet [19]	0.62	0.43	0.54	0.32
SDP + CRC [49]	EdgeBoxes [59]	VGG16 [38]	0.76	0.55	0.68	0.42
Yolo [31]	Free	13-layer	0.39	0.31	0.43	0.27
SSD [23]	Free	VGG16 [38]	0.43	0.38	0.52	0.38
Mask-RCNN [11]	RPN	VGG16 [38]	0.81	0.56	0.70	0.52
Proposed	MultiScale (Proposed)	VGG [4]	0.82	0.57	0.81	0.79
		VGG16 [38]	0.85	0.65	0.85	0.83
		ZF [54]	0.84	0.61	0.75	0.77

dataset. We observed from the experiments that the performance of detectors improve after fine tuning.

We evaluate the performance of all detectors using precision-recall curve with varying threshold values. We report precision-recall curves of all the methods on all datasets in Figure 7. We also report precision, recall and F-score of all detectors in Table 5, 6, 7 and 8 for SHOCK, WIDERFACE, HollywoodHeads and Casablanca datasets, repectively.

From experiment results, we observe that two variants of Viola-Jones, i.e, VJ-HOG and VJ-LBP showed lower performance on all data sets as compared to to other specific detectors. This is due to the reason that Viola-Jones is affected by orientation of heads and faces. Furthermore, it is sensitive to illumination and accumulates many bounding boxes on face location due to sliding window approach that lowers precision-recall rate. We further observed that DPM-head also achieved lower performance on all datasets. The lower performance attributes to the small size of the human head. Due to small size of head, DPM detector could not detect the heads with the size less than 23 × 23 pixels. DISAM [17], on the other hand achieved comparable results by tackling scale problem to some extent, however the method suffers from the following limitations: (1) The models follows the traditional pipeline of R-CNN which uses scale-aware strategy for object proposal generation. The strategy typically requires human efforts to generate a scale map. (2) The inference speed of the model is very slow since the model extracts samples proposals in a sliding window fashion and feed forward each proposal in a single pass.

However, proposed framework efficiently address all above problems. The superior performance of proposed

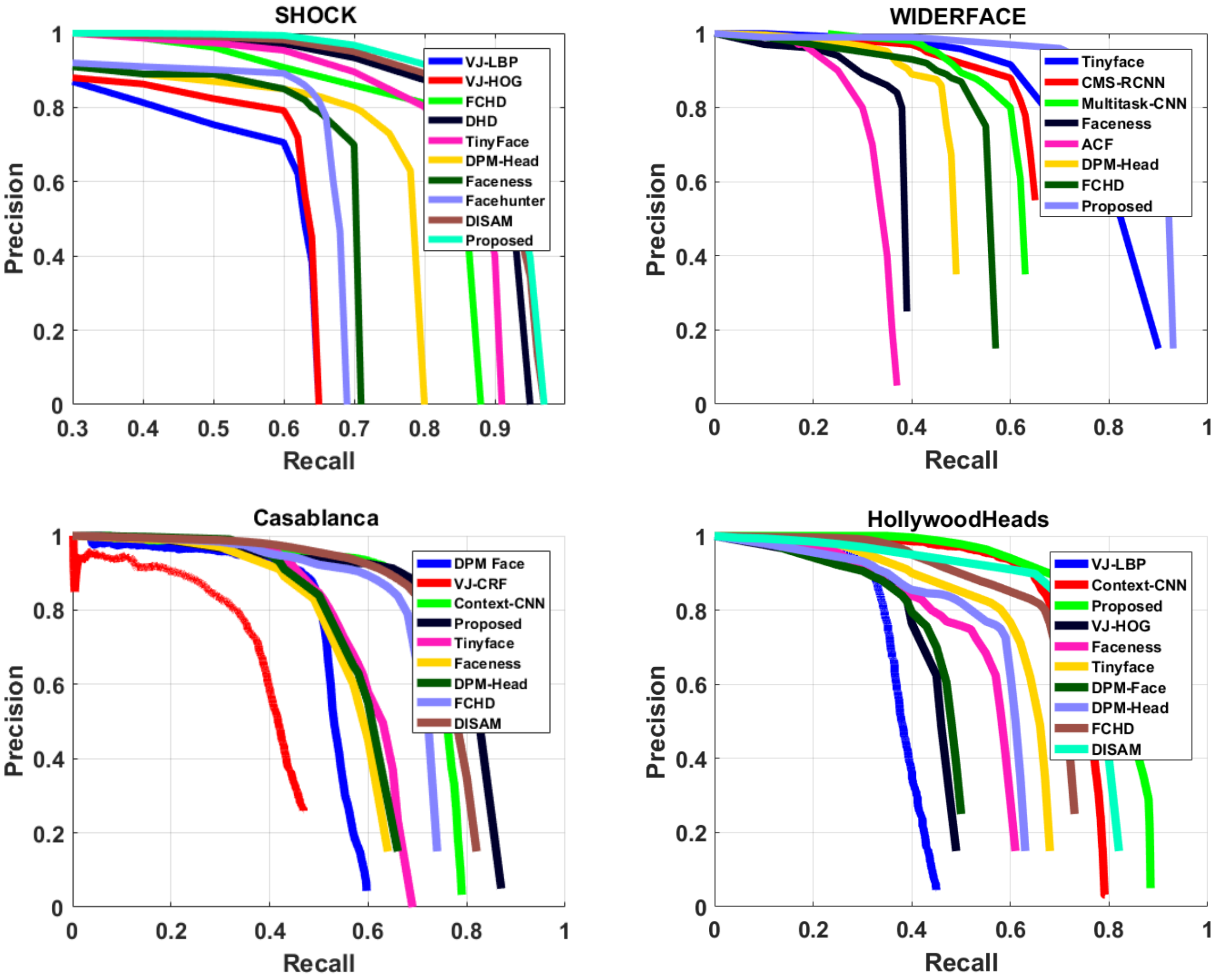


FIGURE 7. Precision Recall curves of different specific detectors on different datasets.

method attribute to the adoption of scale-aware strategy that covers large range of scales of heads. To demonstrate the effectiveness of proposed approach, we present qualitative results in Figure 8 for samples from all data sets.

We further summarize the experiment results in two points:

1) The performance of specific detectors (head/face) is comparatively higher on SHOCK dataset than other datasets. In SHOCK dataset, people are sitting in front of the camera, where most of body part of human are visible. Furthermore, human head/faces lie in limited range of scale and the variance in scale is not significant as also obvious from Figure 5 (b). Due to these properties, specific detectors perform well on this dataset compare to other datasets.
2) The performance of head detectors is higher than face detectors. Face detectors rely on facial features for detection, however, in crowded scenes, facial features are not visible due to occlusions, lighting conditions, and camera view point. For example, face detector can not detect face of a person who turns his back to the camera. Due to these limitations, face detectors perform comparatively low than head detectors.

C. EVALUATION ON EXTREME SCALES

To evaluate the performance of different methods on detecting different sizes of heads, we divide human heads into three categories, i.e., small, medium and large based on sizes of heads. The size of head corresponds to the height of bounding box overlaid on head. Bounding boxes with sizes of 8-60 pixels belong to small category, medium category contains bounding boxes of sizes 60-160 pixels, and bounding boxes greater than 160 pixels fall into large category. Since WIDERFACE dataset contains heads in wide range of scales and sizes, therefore we use WIDERFACE dataset for evaluation purpose. We evaluate the performance of all detectors in terms of mean Average Precision (mAP) and results are summarized in Table 4. From the Table, it is obvious that all detectors perform well on both medium and large groups. However, the performance of these detectors degrades on small group. It is obvious that these detectors face difficulty in detecting small objects. This is due to the reason that network uses single deep convolutional neural network with a fixed receptive field size. For example, Faster-RCNN and SSD showed inferior performance compared to other detectors. Faster-RCNN uses deep layers for ROI-pooling which

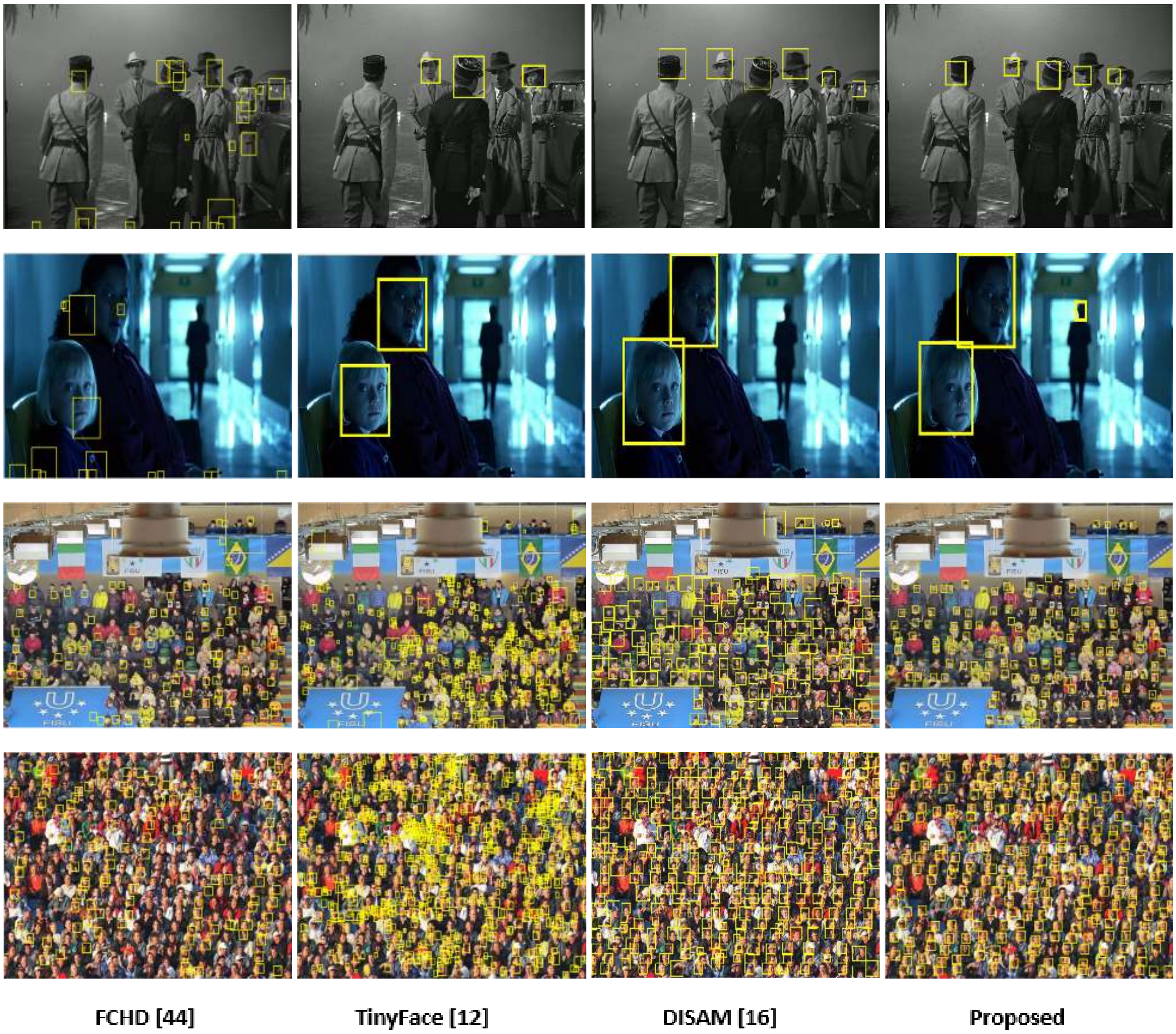


FIGURE 8. Visualization of head detection results using different methods. The first rows shows the performance of different methods on casablanca dataset. The second shows the results on HollywoodHeads dataset, third and fourth rows show results on SHOCK and WIDERFACE dataset respectively. (Best view zoom in).

TABLE 4. Performance of different detectors on WIDERFACE dataset with small, medium and large categories.

Methods	Small	Medium	Large
TinyFace [13]	65.29%	72.64%	83.29%
Faster-RCNN [32]	25.23%	55.76%	71.65%
SSD [23]	20.76%	54.39%	68.82%
FCHD [44]	42.64%	62.29%	73.48%
Multitask-CNN [55]	47.83%	69.34%	75.62%
CMS-RCNN [58]	52.73%	69.76%	78.10%
Proposed	72.34%	82.41%	83.94%

misses critical information about the small objects. On the other hand, single shallow network can not capture contextual information. SSD uses shallow layers to capture better representation of small objects. However, shallow layers can not capture contextual information and have discriminating power. Our proposed framework addresses the above issue and achieves significant improvement by using multi-scale feature that proves helpful in finding wide range of scales.

TABLE 5. Performance of different head and face detectors on SHOCK dataset.

Category	Methods	Precision	Recall	F-Score
Face Detectors	VJ-LBP [43]	0.71	0.57	0.63
	VJ-HOG [43]	0.76	0.60	0.67
	TinyFace [13]	0.82	0.83	0.82
	Faceness [50]	0.76	0.69	0.72
	FaceHunter [27]	0.75	0.65	0.69
Head Detectors	DPM-Head [52]	0.82	0.74	0.77
	DHD [35]	0.88	0.82	0.85
	FCHD [44]	0.85	0.75	0.80
	DISAM [17]	0.89	0.85	0.86
	Proposed	0.88	0.87	0.87

TABLE 6. Performance of different head and face detectors on WIDERFACE dataset.

Category	Methods	Precision	Recall	F-Score
Face Detectors	TinyFace [13]	0.84	0.78	0.80
	CMS-RCNN [58]	0.70	0.59	0.64
	Multitask-CNN [55]	0.65	0.56	0.60
	Faceness [50]	0.34	0.29	0.31
	ACF [48]	0.33	0.27	0.29
Head Detectors	DPM-Head [52]	0.54	0.47	0.50
	FCHD [44]	0.66	0.52	0.59
	DISAM [17]	0.86	0.78	0.81
	Proposed	0.87	0.80	0.83

TABLE 7. Performance of different head and face detectors on HollywoodHeads dataset.

Category	Methods	Precision	Recall	F-Score
Face Detectors	DPM-Face	0.52	0.48	0.49
	VJ-LBP [43]	0.39	0.42	0.40
	VJ-HOG [43]	0.49	0.46	0.47
	TinyFace [13]	0.65	0.63	0.64
	Faceness [50]	0.55	0.50	0.52
Head Detectors	DPM-Head [52]	0.62	0.59	0.60
	Context-CNN [45]	0.70	0.74	0.72
	FCHD [44]	0.68	0.67	0.67
	DISAM [17]	0.78	0.72	0.74
	Proposed	0.83	0.79	0.81

TABLE 8. Performance of different head and face detectors on Casablanca dataset.

Category	Methods	Precision	Recall	F-Score
Face Detectors	DPM-Face	0.54	0.55	0.54
	VJ-CRF [33]	0.47	0.38	0.42
	TinyFace [13]	0.71	0.65	0.67
	Faceness [50]	0.67	0.62	0.64
Head Detectors	DPM-Head [52]	0.64	0.62	0.62
	Context-CNN [45]	0.81	0.74	0.77
	FCHD [44]	0.76	0.68	0.71
	DISAM [17]	0.82	0.76	0.79
	Proposed	0.85	0.81	0.82

TABLE 9. Inference time(in seconds) of different methods per image.

Object proposal methods	Time complexity
RPN [32]	1.8s
EdgeBoxes [59]	0.3s
MultiBox [8]	0.79s
Selective Search [41]	10s
Proposed	1.3s

D. TIME COMPLEXITY

We also compare time complexity of our method with other state-of-the-art object proposal methods. The detailed time complexity of our proposed object proposal method as well as other state-of-the-art methods is reported in Table 9.

For SelectiveSearch method, we use its fast version while for other methods, we directly employed their codes. For testing, we use images from Casablanca dataset. From the Table, it is obvious that our method is not the fastest method but still running comparatively faster than most of the state-of-the-art methods.

V. CONCLUSION

In this paper, we exploit fully convolutional network (FCN) to handle the problem of scale variance in images by generating scale-aware proposals. The heatmap produced by FCN helps to identify whether a patch contains head or not. We observed from experiments, that proposals produced are more stable towards image perturbation compared to other object proposal methods. From the experiments, we also showed that our object proposal generation strategy results in high object recall and mean average precision. We believe that our proposed framework can also be extended to dynamic video sequences. Therefore, in future, we will extend the current framework to incorporate motion information to gain stronger power in identifying and localizing human behaviors and emotion recognition.

REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[2] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 328–335.

[3] S. Breuers, L. Beyer, U. Rafi, and B. Leibe, "Detection-tracking for efficient person analysis: The DetTA pipeline," 2018, *arXiv:1804.10134*. [Online]. Available: http://arxiv.org/abs/1804.10134

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

[5] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3286–3293.

[6] D. Conigliaro, P. Rota, F. Setti, C. Bassetti, N. Conci, N. Sebe, and M. Cristani, "The S-HOCK dataset: Analyzing crowds at the stadium," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2039–2047.

[7] D. Ballotta, G. Borghi, R. Vezzani, and R. Cucchiara, "Fully convolutional network for head detection with depth images," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 752–757.

[8] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2147–2154.

[9] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "DeepProposal: Hunting objects by cascading deep convolutional layers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2578–2586.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 2961–2969.

[12] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1509–1518.

[13] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 951–959.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1725–1732.

[15] S. Khan, G. Vizzari, S. Bandini, and S. Basalamah, "Detecting dominant motion flows and people counting in high density crowds," *J. WSCG*, vol. 22, no. 1, pp. 21–30, 2014. [Online]. Available: http://www.unimib.it

[16] S. D. Khan, S. Bandini, S. Basalamah, and G. Vizzari, "Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows," *Neurocomputing*, vol. 177, pp. 543–563, Feb. 2016.

[17] S. D. Khan, H. Ullah, M. Uzair, M. Ullah, R. Ullah, and F. A. Cheikh, "Disam: Density independent and scale aware model for crowd counting and localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4474–4478.

[18] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 725–739.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[20] W. Kuo, B. Hariharan, and J. Malik, "DeepBox: Learning objectness with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2479–2487.

[21] W. Li, H. Li, Q. Wu, F. Meng, L. Xu, and K. N. Ngan, "HeadNet: An end-to-end adaptive relational network for head detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 482–494, Feb. 2020.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] A. Malhotra, R. Singh, M. Vatsa, and V. M. Patel, "Person authentication using head images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 409–417.

[26] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized Prim's algorithm," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2536–2543.

[27] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 720–735.

[28] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[29] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2417–2424.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[31] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[33] X. Ren, "Finding people in archive films through tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*. [Online]. Available: http://arxiv.org/abs/1312.6229

[35] M. Shaban, A. Mahmood, S. A. Al-Maadeed, and N. Rajpoot, "An information fusion framework for person localization via body pose in spectator crowds," *Inf. Fusion*, vol. 51, pp. 178–188, Nov. 2019.

[36] M. B. Shami, S. Maqbool, H. Sajid, Y. Ayaz, and S.-C.-S. Cheung, "People counting in dense crowd images using sparse head detections," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2627–2636, Sep. 2019.

[37] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[39] Z. Sun, D. Peng, Z. Cai, Z. Chen, and L. Jin, "Scale mapping and dynamic re-detecting in dense head detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1902–1906.

[40] Y. Tian, A. Dehghan, and M. Shah, "On detection, data association and segmentation for multi-target tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2146–2160, Sep. 2019.

[41] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[42] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1879–1886.

[43] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[44] A. Vora and V. Chilaka, "FCHD: Fast and accurate head detection in crowded scenes," 2018, *arXiv:1809.08766*. [Online]. Available: http://arxiv.org/abs/1809.08766

[45] T.-H. Vu, A. Osokin, and I. Laptev, "Context-aware CNNs for person head detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2893–2901.

[46] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7774–7783.

[47] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2497–2504.

[48] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–8.

[49] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2129–2137.

[50] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3676–3684.

[51] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[52] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[53] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[54] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 818–833.

[55] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[56] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[57] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 528–537.

[58] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*. Springer, 2017, pp. 57–79.

[59] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 391–405.

SULTAN DAUD KHAN received the B.Sc. degree (Hons.) in computer engineering from the University of Engineering and Technology, in 2005, the M.Sc. degree (Hons.) in electronics and communication engineering from Hanyang University, South Korea, in 2010, and the Ph.D. degree in computer science from the University of Milano-Bicocca, in 2016. He is currently an Associate Professor with the Department of Computer Science, National University of Technology, Pakistan. He has published several papers in conferences and journals, such as AVSS, IVCNZ, ICGIP, *Neurocomputing*, the *Journal of Cellular Automata*, and IEEE ACCESS. His research interests include crowd analysis, action recognition and localization, object detection, visual tracking, multi-camera, and airborne surveillance using deep learning techniques. He received the Best Reviewer Award from *Pattern Recognition*, in 2017. He is also an active Reviewer of prestigious journals, *Neurocomputing*, *Pattern Recognition*, the IEEE IET SIGNAL PROCESSING, ACM Multimedia, IEEE ACCESS, and *ACM TOMM*.

YASIR ALI (Member, IEEE) received the Ph.D. degree from the Universiti Teknologi PETRONAS, Malaysia, in 2015. He worked as a Researcher with Universiti Teknologi PETRONAS, from 2011 to 2013. He also worked as a Visiting Researcher with Qatar University. He was an Assistant Professor with the Electrical Engineering Department, Umm Al-Qura University. He is currently working as a Senior consultant and the PMO Manager with Experts Vision Consulting Company, Saudi Arabia, focusing on delivering crowd management and smart city solutions for Hajj and Umrah sectors. His research interests include deep learning for crowd understanding, real-time crowd monitoring analytics, crowd management, and smart city.

BASIM ZAFAR received the Ph.D. degree in electrical engineering from Colorado State University. He is currently an expert in digital transformation, smart city, and crowd management using artificial intelligence. He served as an Assistant Professor with Electrical Engineering Department, Umm Al-Qura University and as a Director of the Center for Consulting Research and the Vice Dean of the Hajj Research Institute. He was the CIO for Makkah and Madina Development Commission. He is currently the Founder and CEO of Experts Vision Consulting Company.

ABDULFATTAH NOORWALI (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Western Ontario, London, ON, Canada, in 2017. The title of his thesis was Modeling and Analysis of Smart Grids for Critical Data Communication. He is currently the Chairman of the Electrical and Computer Engineering Department, Faculty of Engineering and Islamic Architecture, Umm Al-Qura University, where he is also an Assistance Professor. He is also a Senior Consultant with Umm Al-Qura Consultancy Oasis, Institute of Consulting Research and Studies (ICRS), Umm Al-Qura University, where he is also the Owner of Vision office of consultancy. He has authored many technical articles in journals and international conferences. His research interests includes smart grid communications, cooperative communications, wireless networks, the Internet of Things, crowd management applications, and smart city solutions.