Deep Neural Network Based Multi-Resolution Face Detection for Smart Cities.

Gary Storey
Northumbria University
United Kingdom
gary.storey@northumbria.ac.uk

Richard Jiang
Northumbria University
United Kingdom
richard.jiang@northumbria.ac.uk

Ahmed Bouridane
Northumbria University
United Kingdom
ahmed.bouridane@northumbria.ac.uk

Ranjith Dinakaran
Northumbria University
United Kingdom
ranjith.dinakaran@northumbria.ac.uk

Chang-Tsung Li
School of Computing and Mathematics
Charles Sturt University, Australia
chli@csu.edu.au

ABSTRACT

Face detection from unconstrained "in the wild" images such as those obtained from CCTV and other image capture devices used within urban environments can provide a rich source of information about citizens within the urban environments benefitting tasks such as pedestrians counting and biometric security. In recent years Deep Convolutional Neural Networks have revolutionized the state-of-the-art for face detection tasks, for utilization within smart cities through leveraging existing CCTV networks, some challenges still exist such as the scale and resolution of the faces within an image. We present a single multi-resolution deep neural network and trained on publicly available image databases that splits the face detection task into small and large face detection at a feature level. We show how our proposed network outperforms single task face detection Faster R-CNN architectures across three challenging test sets (AFW, AFLW and Wider Face).

KEYWORDS

Face Detection, Urban Computing, Biometric-as-a-service, Deep Neural Networks.

INTRODUCTION

Face detection from unconstrained "in the wild" images such as those obtained from CCTV and other image capture devices (as shown in Figure 1) used within urban environments can provide a rich source of information about citizens for the advancement of smart cities [1]. Applications for leveraging the information gained from face detection within a smart cities environment include areas such as urban computing where face detection can provide information on number of pedestrians and meta data such as gender and age which can be applied to managing areas such as air pollution and traffic flow [2]. Face detection can also be applied to security applications that use facial biometrics like facial recognition, more specifically deep neural networks can be used through cloud services like the Google Cloud Platform to provide biometric-as-a-service for urban security applications [3].

Figure 1: Example of deep neural network multi-resolution face detection, successfully identifying large and small faces.

While face detection is a computer vision problem which has seen excellent research progress over the previous decades, most recently research has focused in on unconstrained "in the wild" images where the wide range of face pose in terms of pitch, yaw and roll and occlusion still provides challenges. The traditional approach for face detection are scanning window classifiers [4] which have a proven to be both computationally efficient and accurate when dealing with full frontal posed face images. Deformable Part Model (DPM) [5] based techniques have also proved popular and accurate, especially when deployed in multi-model approaches where multiple faces models for the variant yaw angles of facial pose are defined [6]. Computational overhead is a major issue with these DPM type approaches due to the multiple models and scales they operate in. Deep Convolutional Neural Networks (CNN) have most recently been applied to the face detection problem and have surpassed previous methods producing state-of-the-art results [7][8].

A further challenge is the detection of the multiple scales and resolutions of faces within an image. This is especially prevalent when attempting to reduce the computational overhead of face detection. Two large scale facial image databases have been used in previous CNN based face detection methods for the purpose of model training, the Annotated Facial Landmarks in the Wild (AFLW) [9] and the Wider Face databases [10]. While both contain tens of thousands of face examples the AFLW generally contains more higher resolution single face images, while Wider Faces has many more examples of images with small faces and multiple faces. This can lead to issues of poor model generalization when models built with a single dataset are presented with faces of a higher or lower resolution than those used within model training. This challenge directly impacts the capability of face detection systems to work well in smart cities where the use of CCTV and other image collecting equipment can be at varying distances from pedestrians resulting in large scale difference in the faces captured.

We propose a deep neural network based multi-resolution face detection model that aims to address the challenge of multiple face resolution that can be found in images acquired in urban environments. Through exploiting multi-tasks [7] [6] using a single deep neural network we propose a model which generalizes well on unseen faces at multi-resolutions. We also highlight the issues that can arise in model generalization when training models with a single source of facial data.

Our paper consists of a short discussion on the applications within smart cities where face detection can be of huge benefits. We provide a short overview of some relevant research in the areas of object and face detection followed by an overview of the architecture applied in our model. Finally, we display our experimental results on three challenging datasets along with benchmarks followed by a conclusion.

APPLICATION DOMAINS

As briefly highlighted within the introduction there are a wealth of applications within an urban smart city environment where accurate face detection systems can be employed. In this section we will further expand on these applications and how they can be beneficial for smart cities.

Pedestrian Counting: One such key application is in the task of pedestrian counting which can provide the urban authorities with information about how many, where and when people are flowing through the urban environment. Recent market research [11] highlights large growth in this area across the globe. Modelling the flow of pedestrian across the urban environment can be used for a multitude of scenarios. An accurate people metering system provides an insight to changes in people's behavior within the city and allows those in charge of city planning to effectively manage the flow of people and vehicles around the area. As issues such as street level air pollution are present in today's urban environments having the capacity to manage pedestrian flow away from high pollution areas is a valuable tool. Accurate footfall from pedestrian counting could help to attract new business which are interested in such metrics.

Facial Biometrics: Face detection techniques are a fundamental foundation for security applications such as those that apply face recognition, which aims to identify an individual based upon facial biometrics. Facial biometrics can be applied to many scenarios within the smart city environment like restricting access to sensitive locations to only relevant people this may be access to building or street access for vehicles. While facial biometrics are not 100% foolproof they are often employed in a multi-modal approach with other biometrics or ID cards to provide additional layer of security. Cloud based Biometrics-as-a-Service [3] such as [12] have recently became a new technology for delivering biometrics services to clients, all of these technologies are first build on robust facial detection.

Privacy Concerns: As with all mass data collection techniques both ethical and privacy concerns can be raised. Pedestrian counting through face detection methods such as the proposed method in this paper can alleviate these, firstly face detection is only concerned with finding all faces within an image or frame of video data, unlike face recognition there is no specific identification of an individual. Secondly there is no specific requirement to store the actual image data, metrics can instead be stored such as a total number associated with data such as location and time. When the discussion moves to the idea of facial recognition there are greater privacy concerns and security concerns about the storage of biometric data. While there are no definite answers to these issues and relevant data legislation are likely to play a key part in how these technologies are applied in the future.

PREVIOUS RESEARCH

A brief overview of general object detection and the more specific task of face detection are covered within this section. Finally, multi-task learning methods previously used within the literature are discussed.

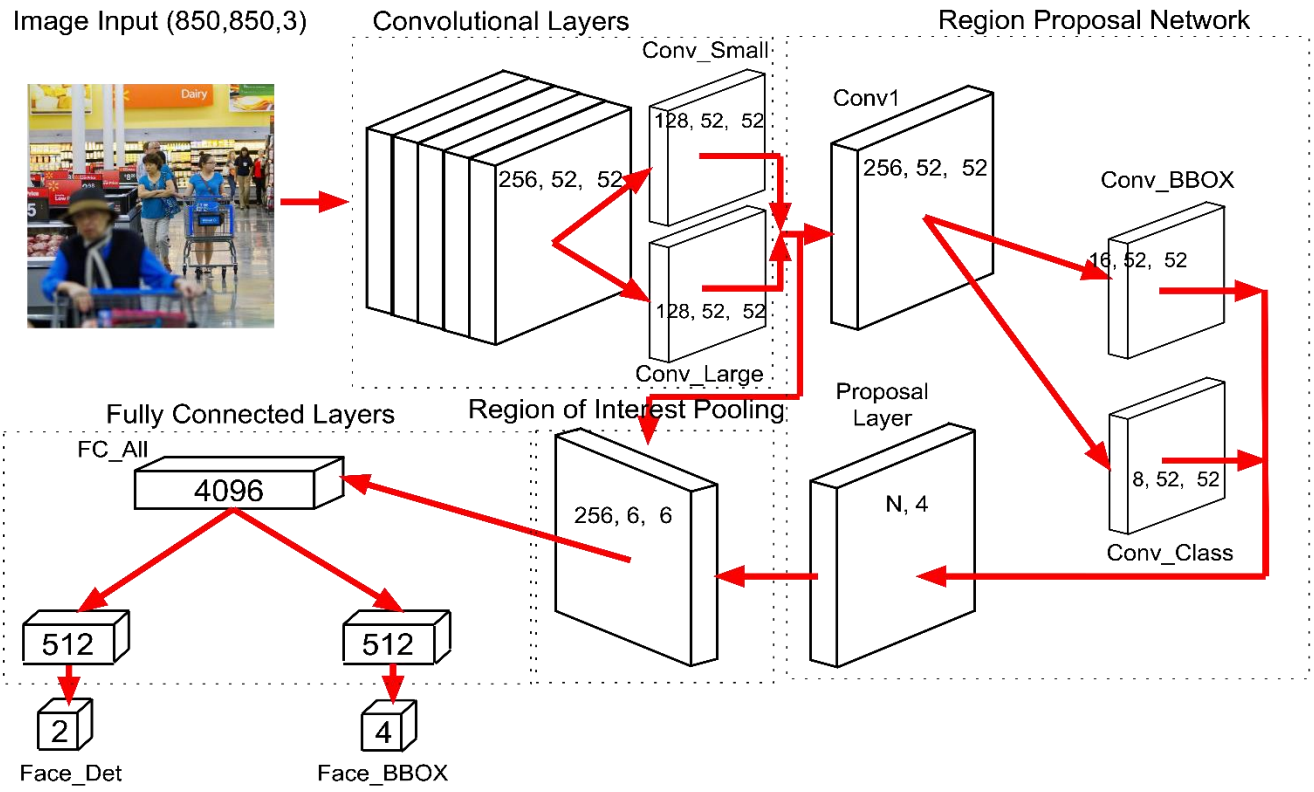

Figure 2: Overview of the deep neural network multi-resolution face detection framework.

Object Proposal: Object proposal methods were traditionally used as external modules acting independently to the object detectors. These methods provide a selection of regions from within an image that potentially contain a desired object. Popular techniques include grouping super-pixels methods such as Selective Search [13] and CPMC [14] and sliding windows such as EdgeBoxes [15]. Most recently CNNs have been proposed for this task including techniques such as the MultiBox method [16] and the Region Proposal Network (RPN) used in Faster R-CNN [17]. The MultiBox method [16] generates region proposals from a network whose last fully-connected layer simultaneously predicts multiple class-agnostic boxes. The Faster R-CNN [17] incorporates the RPN layer in a unified network, providing the network with learnt regions of interest within an image for the object detector to look.

Face Detection: Face detection the process for finding faces within images has a rich history of research in the field of computer vision which discriminatively-trained scanning window classifiers [18]–[20] have proven to be both computationally efficient and accurate. The Viola Jones detector [4]is a real-time face detector though accuracy suffers when faces are not full, frontal, and well lit, it is specifically well known due to the implementation in a number of computer vision software libraries. Deformable Parts Model (DPM) [5] based face detection methods have also been proposed in the literature where a face is essentially defined as a collection of parts [6], to overcome pose variation detection issues multiple models are applied, while this increases accuracy the computational overhead is increased significantly. In recent research deep CNN's have been applied which have shown state-of-the-art results on challenging image databases in this area specifically when dealing with non-frontal faces [21], [22]. For a full review of face detection methods we refer the reader to the following survey papers [23]–[25].

Multi-task learning: An early method of learning of multiple tasks in a joined manner was presented in [6]which used a mixture of trees models for face detection, pose estimation and landmarks localization tasks. Each tree is composed of a number of parts from a shared pool, where each part is a feature representation of a facial

landmark. Multi-task learning using CNNs was recently studied in [26] the method shows an improvement in landmark localization by also learning gender attributes. Person pose estimation and action detection is research in [27] where a CNN is trained applying only the features the last layer. Finally [28] trains a CNN model for the task of face detection, landmark localization, gender recognition and pose estimation. They apply a feature fusion method using features from different convolutional layers to provide the final learnt feature.

METHOD

We propose a single deep CNN for multi-resolution face detection which could be used with smart cities which is heavily influenced by the Faster R-CNN network architecture [17] consisting of three layers as shown in figure 2. The first layer consists of a convolutional network using a modified AlexNet architecture with four shared convolutional layers which learn the features associated with the facial detection. To change the architecture from single task of face detection to multi-task of small and large face detection we propose altering the fifth convolution layer, this is split into two smaller convolutional layers, one which is trained on larger faces and the other on smaller faces, the layers are then concatenated. The second is the region proposal network (RPN) layer which learns n regions of interests of probable face locations within an image. Finally, we have region of interest (ROI) pooling and fully connected layers which produce the final output of the network. The input to the proposed network is an n by m by 3 image, where $n = 850$ and $m = 850$ respectively. Whole images are used as the network input rather than image sections containing faces as used in other work such as [7]. Images are scaled and padded prior entering the network if they their native size is not an n by m.

Multi-resolution Loss: Our network applies a multi-task loss function to train both the RPN and face detection tasks in a single CNN network, the total loss of our network is described as in equation 1.

$$\mathrm{loss}_{\mathrm{total}} = \sum_{\mathrm{i}=\mathrm{n}} loss_i \lambda_i \quad (1)$$

Each individual loss corresponding to the ith task is defined as $\mathrm{loss}_{\mathrm{i}}$. A weighting parameter λ_{i} is applied to balance the learning priorities.

RPN loss is learnt via regions based upon a set of potential anchor points of an image that most likely contain the face for detection. Given a set of k anchors a binary class label based is assigned upon the Intersection-over-Union (IoU). An anchor that has an IoU overlap higher than 0.7 is assigned a positive detection label while those anchors registering an IoU of less than 0.3 are labelled as negative. Other anchors which IoU value between 0.7 and 0.3 are not used for training.

$$\mathrm{loss}_{\mathrm{cls}} = \frac{1}{\mathrm{N}_{\mathrm{cls}}} \sum_{\mathrm{i}=\mathrm{n}} -(1 - p_i^*) \cdot log(1 - p_i) - p_i^* \cdot log(p_i) \quad (2)$$

The SoftMax loss function given in equation 2 is used for learning an object ($\mathrm{p}_{\mathrm{i}} = 1$) and a non-object ($p_i = 0$) classification, where $\mathrm{p}_{\mathrm{i}}^*$ is the ground truth class label and $\mathrm{p}_{\mathrm{i}}$the predicted class for the *i*th anchor respectively. This loss function is normalized by the mini-batch size $\mathrm{N}_{\mathrm{cls}}$.

$$\mathrm{loss}_{\mathrm{reg}} = \frac{1}{\mathrm{N}_{\mathrm{reg}}} \sum_{\mathrm{i}=\mathrm{n}} \mathrm{p}_{\mathrm{i}}^* \mathrm{smooth}_{\mathrm{L1}}(\mathrm{t}_{\mathrm{i}} - \mathrm{t}_{\mathrm{i}}^*) \quad (3)$$

Bounding box regression is defined in equation 3, where for the *i*th anchor the $L1$ loss between the ground-truth box $\mathrm{t}_{\mathrm{i}}^*$ and the predicted bounding box t_{i} is calculated. Both $\mathrm{t}_{\mathrm{i}}^*$ and t_{i} are vectors representing the 4 parameterised coordinates of the predicted bounding box. Only positive anchors affect the loss as described by the term $\mathrm{p}_{\mathrm{i}}^* \mathrm{smooth}_{\mathrm{L1}}$. $\mathrm{N}_{\mathrm{reg}}$ represents the total number of anchors which normalizes the loss function. For a full technical overview of the RPN architecture we refer the reader to [17].

Our proposed network redefines the object detection task of [17] for the purpose of face detection. Face detection is a binary classification problem of a face or not within a proposed region, and the regression of a bounding box for the location of the face within an image. Face detection loss applies similar loss functions to the RPN loss as defined previously. Principally the difference between RPN loss and face detection loss is that RPN loss is concerned with finding a subset of anchors which best describe objects, face detection loss learns whether these anchors contain a face or not.

RESULTS

We present our experimental results within this section. To test the capability of our network two large scale image databases where used in several configurations for training the models. We used the Annotated Facial Landmarks in the Wild (AFLW) and the Wider Face [10] databases. The AFLW [9] database contains around 25,000 images of

annotated faces in real-world images capturing multiple viewpoints, different expressions and illumination conditions. We use 20,077 images for training and 900 images for testing from the AFLW. The Wider Face training set consists of 12,878 with around 200,000 faces, while the validation set contains 3,221 images. Wider Face database has images with a high degree of variability in scale, pose and occlusion as depicted in the sample images. The annotated faces in-the-wild (AFW) database [6] was applied as test database. The database consists of 205 images where each image contains at least a single face. In total there are 468 faces located within the database. We adopt the PASCAL VOC precision-recall protocol for object detection requiring 50% overlap respectively for positive face detection.

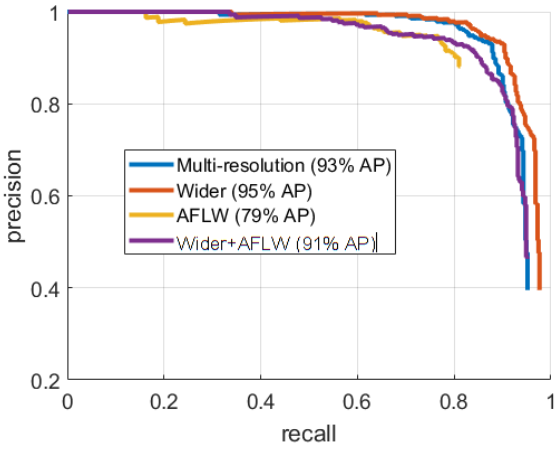

Figure 3: Precision-recall curve and average precision results for the AFW database.

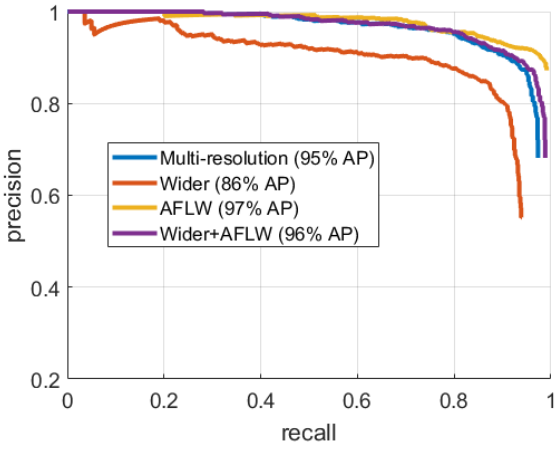

Figure 4: Precision-recall curve and average precision results for the AFLW database.

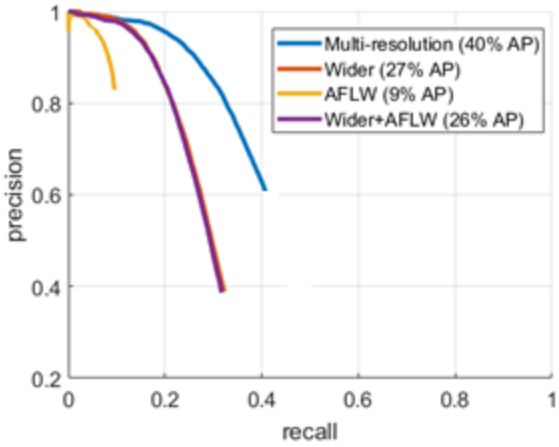

Figure 5: Precision-recall curve and average precision results for the Wider Face database.

Four models where trained for benchmarking, one using the AFLW training set, one with the Wider Face training set, one with a merged AFLW-Wider Face training set using a standard Faster R-CNN network and finally one which applies our method as shown in figure 2. Training for our multi-resolution model was alternatively trained on small faces of the Wider Face set then larger faces from the AFLW, with weight parameter only being update on the corresponding convolutional layer for that resolution. Each model was trained for 20 epochs using a learning rate parameter of 0.001 for the initial 10 epochs and 0.0001 for the final 10.

Figures 3,4 and 5 show the results of our experiments on each of the image test sets. Our results highlight the importance of how training data can highly influence the model accuracy. While our multi-resolution method does not outright perform the others on the AFLW and AFW test sets, it is significantly better by 14% in terms of average precision. The slight decrease in performance on larger face datasets is likely due to the reduction of filters that are employed within the network to detect larger faces at the final convolutional layer. Note that when using only large face datasets like the AFLW they are very poor when generalizing to face detection tasks which require detection of smaller faces as shown by the 9% average precision on the Wider Face set. There are still challenges to overcome these primarily are heavily occluded faces and extremely small faces which are the primary reason for the poor performance on the Wider Face set as it is extremely challenging (see image shown in figure 6). The network is also computationally efficient with an average time to process an image 0.065 seconds.

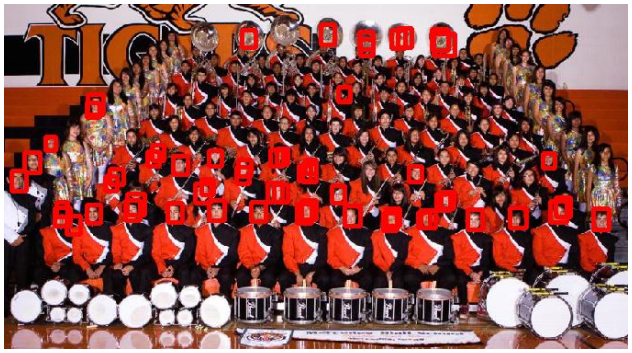

Figure 6: Example of poor detection when presented with extremely small faces. The first rows are detected but those faces further back are missed.

CONCLUSION

In this paper we present a multi-resolution deep neural network which has potential for the use within smart cities for tasks such as pedestrian counting or the face detection component of a facial biometrics service. We show that our method generalizes better across the three test sets then when using applying a general faster R-CNN network with the same training data. This highlights the importance that training set data has on the performance of face detection architectures.

While our work is currently specifically focused on the task of face detection, the general principle of object detection networks can be applied other areas of interest within smart cities such as vehicle and bicycle monitoring and counting. The prospect of super-resolution which can upscale tiny faces within images using generative adversarial networks (GAN) is an area which could be applied to further enhance the network specifically in the challenging area of super small faces as discussed with the paper.

REFERENCES

[1] I. Celino and S. Kotoulas, "Smart Cities," *Internet Comput. IEEE*, vol. 17, no. 6, pp. 8–11, 2013.

[2] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban Computing: Concepts, Methodologies, and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, 2014.

[3] J. Rose, "Biometrics as a Service: The next giant leap?," *Biometric Technol. Today*, vol. 2016, no. 3, pp. 7–9, 2016.

[4] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2009.

[6] D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," pp. 2879–2886, Jun. 2012.

[7] 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," Mar. 2016.

[8] P. Hu and D. Ramanan, "Finding Tiny Faces," Dec. 2016.

[9] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2144–2151.

[10] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016–Decem, pp. 5525–5533.

[11] "People Counting System Market by Technology (IR Beam, Video Based, Thermal Imaging), Application (Retail, Transportation, Banking & Finance, Hospitality, Sports & Entertainment, Government), and Geography - Global Forecast to 2022," *Markets and Markets*, 2017. [Online]. Available: https://www.researchandmarkets.com/research/rmzj2n/people_counting. [Accessed: 27-Feb-2018].

[12] V. Talreja, T. Ferrett, M. C. Valenti, and A. Ross, "Biometrics-as-a-Service: A Framework to Promote Innovative Biometric Recognition in the Cloud," Oct. 2017.

[13] 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, 2013.

[14] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, 2012.

[15] C. L. Zitnick and P. Doll, "Edge Boxes:Locating Object Proposals from Edges," *Eur. Conf. Comput. Vis.*, pp. 1–15, 2014.

[16] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov,

"Scalable Object Detection Using Deep Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2155–2162.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jun. 2015.

[18] M. Jones and P. Viola, "Fast Multi-view Face Detection," *Mitsubishi Electr. Res. Lab TR2000396*, no. July, 2003.

[19] P. Jonathon Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, 2000.

[20] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *Int. J. Comput. Vis.*, vol. 74, no. 2, pp. 167–181, 2007.

[21] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 11–18–Dece, no. 3, pp. 3676–3684, 2016.

[22] S. S. Farfade, M. Saberian, and L. J. Li, "Multi-view Face Detection Using Deep Convolutional Neural Networks," *Int. Conf. Multimed. Retr. 2015*, p. 19, 2015.

[23] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, 2002.

[24] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," *Microsoft Res.*, no. June, p. 17, 2010.

[25] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Underst.*, vol. 138, pp. 1–24, 2015.

[26] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8694 LNCS, no. PART 6, pp. 94–108.

[27] G. Gkioxari, R. Girshick, and J. Malik, "Contextual Action Recognition with R*CNN," May 2015.

[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," Mar. 2016.