Person Head Detection in Multiple Scales Using Deep Convolutional Neural Networks

Muhammad Saqib*, Sultan Daud Khan†, Nabin Sharma* and Michael Blumenstein*
*Centre for Artificial Intelligence, School of Software, FEIT, University of Technology Sydney, Australia
† University of Hail, Saudi Arabia
*muhammad.saqib@student.uts.edu.au, †su.khan@uoh.edu.sa, *{nabin.sharma, michael.blumenstein}@uts.edu.au

***Abstract*—Person detection is an important problem in computer vision with many real-world applications. The detection of a person is still a challenging task due to variations in pose, occlusions and lighting conditions. The purpose of this study is to detect human heads in natural scenes acquired from a publicly available dataset of Hollywood movies. In this work, we have used state-of-the-art object detectors based on deep convolutional neural networks. These object detectors include region-based convolutional neural networks using region proposals for detections. Also, object detectors that detect objects in the single-shot by looking at the image only once for detections. We have used transfer learning for fine-tuning the network already trained on a massive amount of data. During the fine-tuning process, the models having high mean Average Precision (mAP) are used for evaluation of the test dataset. Experimental results show that Faster R-CNN [18] and SSD MultiBox [13] with VGG16 [21] perform better than YOLO [17] and also demonstrate significant improvements against several baseline approaches.**

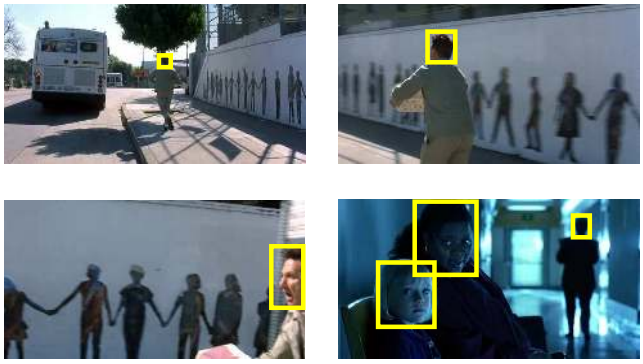

Fig. 1: Sample images from HollywoodHeads(HH) dataset overlaid with ground-truth annotations

I. Introduction

Object detection is a major part of many practical applications of computer vision such as face detection, pedestrian detection, vehicle detection, and video surveillance. In classification, a classifier is trained to categorize and label the content of the image globally. While in detection, detector not only categorizes and labels the content but also localize the bounding box of the object. The object detector faces similar challenges to that of classification. These challenges include viewpoint variation, scale changes, illumination changes, occlusion, background clutter, and deformation. Also, a good object detector is required to find precise bounding boxes for the detected objects. Person detection is a popular research topic due to its applications in safety and security. The technology has reached its maturity level in face detection and recognition task. Person detection still poses many challenges due to the articulated nature of the human body. However, in this paper, the focus is on head detection in the images taken from video. The choice of head detection for detecting the number of people is because it is visible most of the time even when other body parts are fully occluded. The feature extracted for head detection are not discriminative enough to be used alone for person detection. Previously, contextual information in conjunction with the features extracted for the head is used for detection [24]. The contextual information provides additional cues and useful information for detection and recognition. The use of contextual information is in contrast to the traditional way of extracting features for an object inside the bounding box.

Traditionally, machine learning approaches applied to computer vision tasks used hand-crafted features. The features are less robust and discriminative as compared to the learned features. Recently, learned feature using deep learning techniques had been successfully applied for major computer vision tasks such as segmentation [14], classification [10], detection and recognition [18]. The deep learning techniques have been winning the classification and detection competitions such as ImageNet [4] for so many consecutive years. The structure of typical deep Convolutional Neural Network (CNN) consists of several layers followed by classification layer. However, object detector based on deep CNN differs in such a way that it not only classify but also localize the bounding box containing the object. Each layer in the deep CNN extract features at different abstraction level. The lower layers extract features representing more fine details in the image like edges and texture. As we go across the neural network towards the final layers, more complex and abstract features are extracted. By truncating the last layer, most of the previous layers can represent input images in abstract space of network parameters or weights. Deep CNNs are data-hungry architectures, to work properly for particular task requires a massive amount of data. So in most of the cases, more real data is useful if available. However, data augmentation is used to fulfill the requirement for deep CNN. If the last classification layer is removed,

the models obtained can be used as pre-trained models for other tasks. These models can be generalized to many other similar tasks. The pre-trained weights/parameters can be used for fine-tuning the network for a new task. Thus the resultant network converges faster and performs well instead of training from scratch. Most of the pre-trained models are available and trained on a hugely popular dataset such as ImageNet [4]. In this study, we have extensively carried out experimentation with state-of-the-art object detectors based on deep learning to detect head in the videos as shown in the Fig 1.

The remainder of the paper is organized as follows. In section II, we discuss the current state-of-the-art object detectors. The section III describes the proposed approach which discusses the use of the pre-trained networks for fine-tuning the networks for person head detection task. The section IV describes the experimental results and discussion followed by conclusion in section V.

II. Literature Review

The efficient object detector needs to detect an object of different scale and aspect ratio anywhere in the image. Traditionally, the concept of the image pyramid is used to represent the image at a different scale and then sliding window approach is applied to search object of various scale anywhere in the image. The features are extracted from sliding window position on the image and passed on to the classifier for detection. Much of the previous research extract hand-crafted feature such as LBP [1], HOG [3], and SIFT [15] for describing the objects. Due to sliding window approach, classifier detects multiple bounding boxes of an object. Therefore, a non-maximum suppression is applied to remove redundant and overlapping bounding boxes [3]. However, all of these approaches are computationally expensive, and furthermore, the training process is complex, not end-to-end and often done in stages. The complex training process makes these approaches not suitable for real-time applications.

The current resurgence of neural networks especially CNN has shown remarkable results in image classification challenges surpassing human-level performance on certain tasks [20]. The success of the CNN can be attributed to the availability of huge amount of data as well as the processing power in the form of GPUs to process the data. The features extracted from CNN are powerful, robust and expressive as compared to its traditional counterpart like DPM [7], HOG [3], SIFT [15]. A series of convolution layers followed by non-linear activation units constitutes the convolutional neural network. In CNN, regions in one layer called receptive field are connected to another layer. A series of filters are applied to each layer to derive the output. The weights or parameters of the filter are learned during the training process. In contrast to the feedforward neural network in which every layer is fully connected to other layers. The features extracted from CNN architecture are locally invariant and compositional, in which higher more complex features like objects and shapes are constructed from low-level features such as edges. Most of the popular architectures are based on CNN e.g. AlexNet [12], ZF [25], GoogLeNet [22], VGG16 [21] and ResNet [10].

In classification, sliding windows applied on image require much computation for scanning each possible location in the image. To avoid scanning all possible location, Selective Search (SS) [23] computes region proposals of different sizes and start merging from pixel level into objects. The merging of pixels is based on similarity in low-level features such as texture and color. The SS is a class agnostic method which detects multiple foreground objects without knowing the class of the object. The computation of SS is fast and act as a pre-processing step for the more highly specialized classifier to find out the exact class within the bounding box. The initial implementation of the selective search was not optimized for GPU based implementation and thus being a major bottleneck in the object detection framework. Edgeboxes [26] relatively takes less time to compute region proposal as compared to Selective Search. A specialized classifier called Region-based Convolutional Neural Network (R-CNN) [9] is used to classify an object into their specific classes. The R-CNN expects fixed size images, i.e., $227X227$ as R-CNN uses Alex Net [12] as its backbone CNN architecture. Therefore, the region proposals ($\sim$2k) are warped before passing it on to the R-CNN which uses SVM to classify the object and linear regression to find out more exact bounding box boundaries for the object. Despite the fact that R-CNN works very well, however, the computation is very slow. For each region proposal, the image has to be passed on through the network. Moreover, the region proposals are computed in an offline manner. Thus end-to-end training of R-CNN is very complex. In the next iteration of the R-CNN called Fast R-CNN [8], there is a change in network architecture; region proposals are extracted from convolutional feature map using Region of Interest Pooling (RoI) rather than directly from images. The fact that most of the region proposals are overlapped. Thus convolutional feature map is computed once and shared across all computation of the region proposals. This change in architecture drastically reduces the overall time for detection. Furthermore, the Fast R-CNN architecture unified the network for extracting the region proposals followed by its classification and regression to its tighter boundaries. The SVM classifier is replaced by softmax layer in Fast R-CNN [8]. In Fast R-CNN [8], Spatial Pyramid Pooling (SPP) make sure Fully Connected layer (FC) get fixed-size feature vector. The major shortcoming of Fast R-CNN is the time taken for region proposals at the test time. Also, the training pipeline of the architecture is complex. In yet another iteration of Faster R-CNN [18], a fully convolutional neural network is used as a Region Proposal Network (RPN). The cost of computation is very low compared to SS for region proposals. The use of RPN as region proposal made Faster R-CNN a streamlined end-to-end trainable network for object detection.

All the previously discussed methods consider detection as a classification problem. The detection is carried out in stages. The first stage computed proposals and classified in the second stage into object categories. However, there are

some methods, e.g., YOLO [17], SSD MultiBox [13], which consider detection as a regression problem and glance at an image once for detections. Therefore called single shot detectors. The YOLO [17] divides the image into a grid of $13x13$ resolution. Each cell in the grid predicts the bounding box of the object along with its confidence score. The bounding boxes with low confidence score are removed by setting a threshold. The YOLO [17] performance is fast but not as accurate as its other counterparts. The reason behind its low accuracy is that it only predict one type of class in a grid. The Single Shot Detector MultiBox [13] compute a convolutional feature map by looking only once at the image and predicts bounding boxes at multiple of these feature maps for the objects of various scales.

Previous studies used different models to capture various aspects of head detection in contextual reasoning [24]. The Global model predicts the scale and coarsely localize the object using full image, while Local CNN captures the object appearance. Moreover, the Pairwise model captures the relationship between the objects. All the above models are jointly optimized for contextual head detection. However, such models do not provide a real-time unified architecture for head detection. Similarly, the performance of methods used for face detection [16] on head detection task is very low. The choice of particular object detection depends upon the trade-off between speed and accuracy. The single shot detectors are fast and suitable for the embedded real-time applications. However, region proposals technique such as Faster R-CNN [18] is more accurate and intended for applications where accuracy is more important than speed. Furthermore, the size of an object is another consideration for the selection of detector to be used for detection. Both types of detector perform poorly on the very small scale objects.

III. Proposed Methodology

Transfer learning is an approach widely used in deep learning applications in which pre-trained CNN is used as a feature extractor for a given dataset. The features are typically extracted from the last layer in the form of activations and cached to the disk. Then a standard machine learning classifier such as SVM is used for training and testing on the features. There is another type of transfer learning, more powerful than just feature extraction approach is called fine-tuning. In fine-tuning, the last fully-connected layers are replaced with the new set of fully-connected layers suitable for a given task. Typically the added fully-connected start learning from the previous convolutional layers while keeping the convolutional layers frozen so that their weights cannot be modified. However, in this paper, all the layers are included in the fine-tuning process. The process involves the state-of-the-art neural network architecture already trained on a large collection of images such as ImageNet [12] and PascalVOC [6]. These network architectures include VGG16 [21], ZF [25], VGG_CNN_1024 [2] which contains rich and discriminative filters. A very small learning rate is used during retraining of the architecture making sure that already learned *CONV* filters do not deviate dramatically. For Faster R-CNN [18] and SSD MultiBox [13] detector, experiments have been carried out in Caffe [11] framework. The first approach proposes regions in the bounding boxes and then applying the high-quality classifier to classify into object class categories. These approaches are called Region-based convolutional Neural network. In Faster R-CNN, the convolutional layers are shared by both the detection network and by the small Region Proposal Network (RPN) as shown in Fig. 2. The RPN is a small network computes region proposal from the last shared convolutional layers. The RPN consist of few CNN layer which does not add to the overall computation. The RPN can be trained in end-to-end fashion for object detection proposals. The concept of anchor boxes is introduced to cater for the object of different scale and aspect ratio. In anchor boxes, a single image and single filter size are used whereas for each location of feature map several anchor boxes of different scale and aspect ratio are used. In RPN, the sliding window on the last shared convolutional feature map encodes the output into $256-d$ feature vector and $512-d$ feature vector in the case of ZF [25] and VGG16 [21] respectively followed by non-linear activation unit of ReLu [18]. The extracted features are fed into regression for finding bounding box coordinates and classification layer for classifying the bounding box region into one of the class.

The Single Shot detector detects the object by a single pass of the image through deep Convolutional Neural Network without explicitly proposing the more probable regions. SSD MultiBox [13] detector is one of the techniques which uses the same powerful CNN network architecture VGG16 [21] generally used for classification. However, the last classification layers are truncated, and some auxiliary layers are added for object detection as shown in Fig. 3. The network architecture without additional layers is called base network. For all the experiments, VGG16 [21] has been used as a base network. The additional layers are organized in such a way that the CNN feature map size decrease down in size until the extraction of final feature vector. These layers produce a fixed-size collection of bounding boxes by applying a convolutional filter on these feature maps. As a result, there are some redundant boxes which are suppressed by Non-Maxima Suppression (NMS) for boxes having Intersection over Union (IoU) less than 0.7 with the ground-truth bounding boxes. The usage of all feature map of additional layers serves an opportunity to predict the detections at various scales.

The usage of several feature maps is in contrast to another state-of-the-art object detector, YOLO [17], where single feature map is used as shown in Fig. 4. The feature map is divided into a grid of cells. The set of default bounding boxes are associated with each feature map cell position. The position of default box is fixed relative to the cell. These default boxes act like anchor boxes in Faster R-CNN [18] to capture the object of different aspect ratio. At each cell location, position offset is predicted along with the class score of an object present in the box. The offset indicates the coordinates how much far away is the predicted bounding box from one of the default

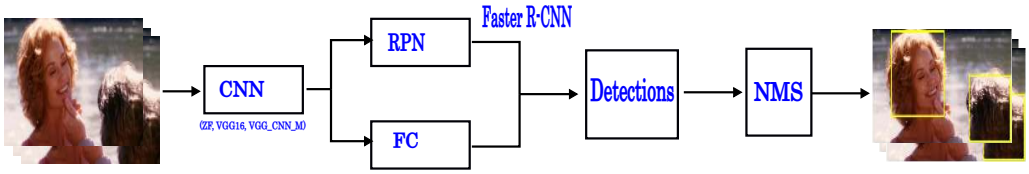

Fig. 2: Faster R-CNN detection framework

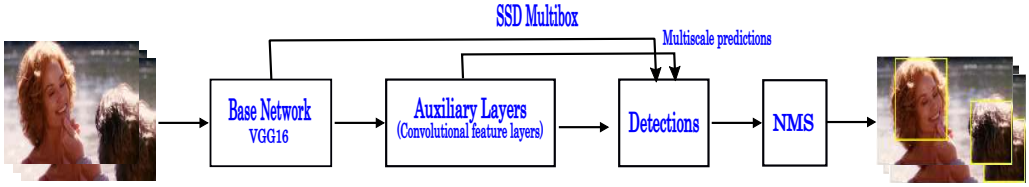

Fig. 3: Single Shot Detector MultiBox

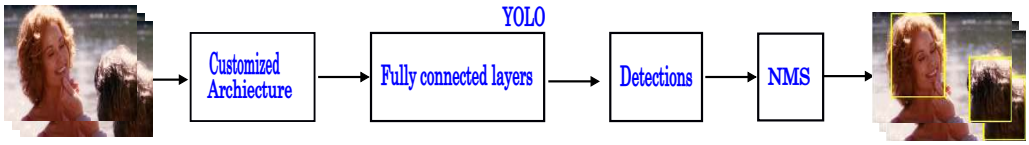

Fig. 4: YOLO framework

bounding boxes.

IV. Experimental Results and Discussion

We trained our models with Nvidia Quadro $P6000$ GPU with the common implementation settings of weight decay, momentum, learning rate and batch size of 0.0005, 0.9, 0.0001 and 64 respectively for all the detection frameworks.

A. Datasets

We have carried out several experiments to evaluate the performance of various detectors on publicly available datasets. In this section, we discuss the two datasets that have been used for experimentation namely HollywoodHeads (HH)) [24] and Casablanca datasets [19].

1) HollywoodHeads (HH) [24]: The HH dataset contains $369,846$ human heads annotated in $224,740$ frames taken from 21 Hollywood movies [24]. In the annotations, keyframes are annotated with the bounding box for the head. The remaining frames are linearly interpolated to head position and verified manually. However, some of the interpolated frames are not correctly annotated. Upon visualization, some of the XML annotation files corresponding to the frames are found to be empty. Therefore, the empty frames were not included in the fine-tuning process. The dataset is divided into the splits of training, test and validation sets of 216719, 1302, and 6719 frames. The training, test and validation sets are taken from 15, 3, and 3 movies respectively.

2) Casablanca [19]: The Casablanca dataset is taken from the movie Casablanca . The dataset contains 1466 frames in which head is annotated with bounding boxes [19]. However, the annotations only consider the frontal face part as a human head. The annotations were corrected for scale and aspect ratio like HH. This dataset is mainly used for evaluating the models fined-tuned on HH dataset.

B. Experimental Results

In Faster R-CNN we have followed the same implementation setting as that of the original paper [18]. However, few changes were made while fine-tuning the network from ImageNet trained models. The images are rescaled to 600 pixels on the shorter-side of dimension. The anchor boxes of 3 different scale and aspect ratio are used to capture various scales in the process of region proposals. During training, the cross-boundary anchors are causing the network to converge therefore ignored, while in testing phase the cross-boundary anchors are clipped to the image boundary. A Non-Maxima Suppression (NMS) approach is applied to reduce overlapping proposals based on their IoU intersection with ground-truth bounding boxes. To evaluate the detection performance, we have adapted the average precision (AP) a standard metric calculated from the area under the Precision-Recall (PR) curve [5]. The mean Average Precision (mAP) is the mean over classes for the set of detections. In true positive detections, there is more than 70% overlap calculated as Intersection over Union (IoU) between the detected class and their corresponding ground-truth bounding box as shown in Eq. (1) [5]. The remaining detections are considered as false negative. Similarly, the detections having no overlap with any corresponding ground-truth box are false negative or background.

$$IoU = area\left(B_{pred} \cap B_{gt}\right) / area\left(B_{pred} \cup B_{gt}\right) \quad (1)$$

Where B_{pred} and B_{gt} denotes predicted bounding box and ground truth bounding box respectively. In the first set of experiments using Faster R-CNN [18], we have used three architectures ZF [25], VGG16 [21], and VGG_CNN_1024 [2]. In the process of fine-tuning, the networks were trained for $100k$ iterations. The snapshot of fine-tuned models is saved at an interval of $10k$ iteration. The performance analysis of each of these saved models was tested on test dataset as shown in the Fig. 5, 6, and 7 for Faster R-CNN [18], SSD MultiBox [13], and YOLO [17] respectively. Faster R-CNN [18] with VGG16 [21] and VGG_CNN_1024 [2] converge at $70k$ iteration while ZF [25] converge at $100k$ as reported in Table I. These models having high mAP's are considered best models later on used for evaluation on Casablanca dataset. Furthermore, in experiments with Single Shot Detectors, YOLO [17] has been fine-tuned for HH dataset according to the original data splits [24]. The performance of YOLO is worst amongst all the detectors with mAP of 0.63 which is lower by 13.34% than the current baseline of mAP 0.72 as reported in the Table I. The SSD MultiBox [13] has also been used for experiments with VGG16 [21] as its base network architecture. The performance of SSD MultiBox is comparable with the performance of Faster R-CNN [18] as reported in Table I. SSD MultiBox with $300X300$ input achieves 0.788 mAP outperforming the current best approach by 8.4% as reported in Table III. Similarly, Faster R-CNN achieves 0.791 mAP on HH dataset outperforming the baseline approach by 8.8% as shown in Table III.The existing baseline approaches include DPM Face [16] and deep learning based approach such as R-CNN [9] with a different combination of local, global, and pairwise models [24].The best-fine-tuned models for HH dataset are used to evaluate on Casablanca dataset without fine-tuning on Casablanca. The results are reported in Table II shows the superior performance and generalization of Faster R-CNN for the unknown dataset. In Fig. 8 some detection samples are taken from a test set of HH and Casablanca datasets for each of the network architectures used in the experiments.

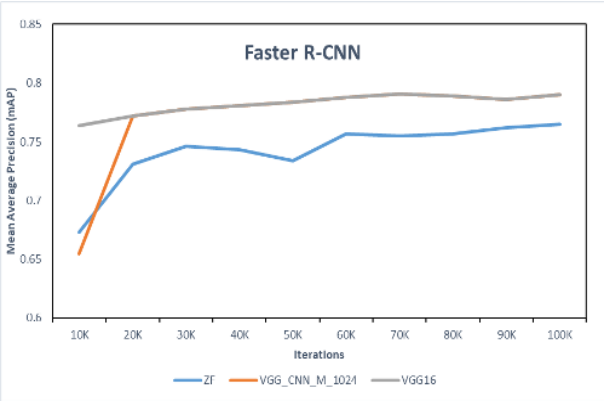

Fig. 5: Performance of network architectures at every $10k$ iteration

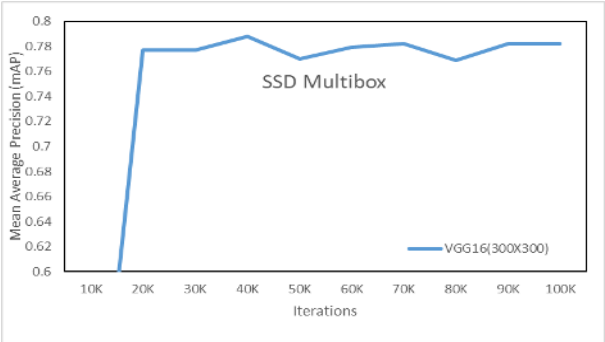

Fig. 6: Performance of network architecture at every $10k$ iteration

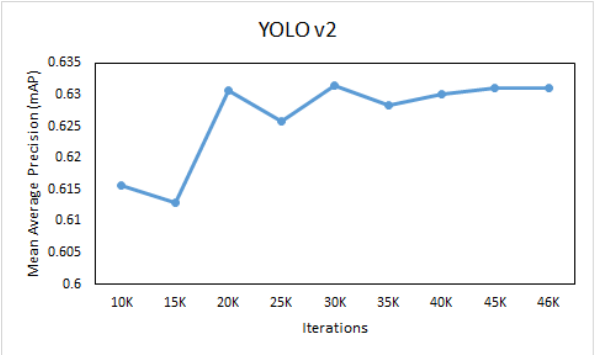

Fig. 7: Performance of network architecture at every $10k$ iteration

TABLE I: Testing error on HollywoodHeads Dataset

Detection frameworks	CNN architectures	Iteration	mAP
Faster R-CNN [18]	ZF [25]	$100k$	0.765
	VGG16 [21]	$70k$	0.791
	VGG_CNN_M_1024 [2]	$70k$	0.791
YOLO v2 [17]	13-layered architecture	$46k$	0.631
SSD MultiBox [13]	VGG16 [21]	$40k$	0.788

TABLE II: Selected best models fine-tuned on Hollywood-Heads(HH) dataset performance on Casablanca dataset

Detection frameworks	CNN architectures	mAP
Faster R-CNN [18]	ZF [25]	0.486
	VGG16 [21]	0.556
	VGG_CNN_M_1024 [2]	0.513
YOLO v2 [17]	13-layered architecture	0.518
SSD MultiBox [13]	VGG16 [21]	0.52

V. Conclusion and Future Scope

In this work, state-of-the-art object detectors have been used for people detection. The detectors were trained and evaluated on head detection task. It is demonstrated through

TABLE III: Comparative analysis with other approaches

Detection frameworks	CNN architectures	mAP
DPM Face [16]		0.374
R-CNN [9]		0.671
Local model [24]		0.718
Local+Global+Pairwise model [24]		0.727
Faster R-CNN [18]	ZF [25]	0.765
	VGG16 [21]	0.791
	VGG_CNN_M_1024 [2]	0.791
YOLO v2 [17]	13-layered architecture	0.631
SSD MultiBox [13]	VGG16 [21]	0.788

experiments that Faster R-CNN [18] with VGG16 [21] works very well for head detection as compared to the single shot detectors. However, all of the object detectors perform poorly for the small-scale and high-dense head detections. Moreover, it is observed that region-based CNN are more suitable for situations where high accuracy is required. While single shot detectors are fast although less accurate, thus more suitable for real-time and mobile applications. This paper investigated the state-of-the-art object detectors thoroughly for head detection task. However, head detection is a difficult task and lacks discriminative features. Therefore there is still room for performance improvement. The future work involves detections combined with the tracker for enhanced performance.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006. 2

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 3, 5, 6

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1, 2

[5] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 4

[6] M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, et al. The pascal visual object classes challenge 2007 (voc2007) results. 2007. 3

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[8] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 2

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2, 5, 6

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 2

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 3

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2, 3

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 1, 3, 5, 6, 7

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[15] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 2

[16] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014. 3, 5, 6

[17] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 1, 3, 5, 6, 7

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 3, 4, 5, 6, 7

[19] X. Ren. Finding people in archive films through tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 4, 7

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 3, 5, 6

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[23] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[24] T.-H. Vu, A. Osokin, and I. Laptev. Context-aware cnns for person head detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2893–2901, 2015. 1, 3, 4, 5, 6, 7

[25] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2, 3, 5, 6

[26] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014. 2

Fig. 8: Qualitative Results: The first column shows the sample frames from test sequences overlaid with the ground-truth annotations. The second, third and fourth column shows the detection results using models fine-tuned on SSD MultiBox [13], YOLO [17] and Faster R-CNN [18] respectively. The rows from $1-4$ and $5-7$ show the results from HollywoodHeads [24] and Casablanca [19] respectively.