Pedestrian Detection via Mixture of CNN Experts and thresholded Aggregated Channel Features

Ankit Verma, Ramya Hebbalaguppe, Lovekesh Vig, Swagat Kumar, and Ehtesham Hassan
TCS Innovation Labs,
New Delhi

Abstract

In this paper, we propose a two stage pedestrian detector. The first stage involves a cascade of Aggregated Channel Features (ACF) to extract potential pedestrian windows from an image. We further introduce a thresholding technique on the ACF confidence scores that seggregates candidate windows lying at the extremes of the ACF score distribution. The windows with ACF scores in between the upper and lower bounds are passed on to a Mixture of Expert (MoE) CNNs for more refined classification in the second stage. Results show that the designed detector yields better than state-of-the-art performance on the INRIA benchmark dataset[5] and yields a miss rate of 10.35% *at FPPI=*10^{-1}.

1. Introduction

Pedestrian detection is an important problem in computer vision which finds application in several areas such as monitoring, surveillance, smart vehicles, healthcare and assistive robotics. The problem becomes challenging owing to several factors such as variation in appearance, pose and size of pedestrians, illumination, backgound and degree of occlusion. Autonomous vehicles are crucial for improving the quality of life for disabled. These require reliable pedestrian detection to ensure safety during autonomous navigation.Autonomous medical robots are currently providing healthcare assistance and require pedestrian detection for effective navigation in hospitals [1, 25].

Recent years have seen significant progress in this field with the appearance of several state-of-the-art pedestrian detectors [12] [16] [10]. Most of these methods use some hand-crafted features such as Haar [23], HOG [5] along with a cascade of (boosted) classifiers to detect pedestrians. The detection rate was further enhanced by using deformation part based model (DPM) [15] where a human is represented as a collection of parts in a deformable configuration.

Recently, deep learning techniques has been shown to provide remarkable results in large scale object recognition tasks [18] [17]. This has encouraged researchers to apply deep learning techniques to the pedestrian detection problem as well. For instance, Sermanet *et al.* [21] proposed a two layer convolutional model where the layers were pretrained by convolutional sparse coding. Chen *et al.* [4] used a pre-trained Deep Convolutional Neural Networks (DCNN) to learn features from windows obtained from an ACF detector [6]. These features are then applied to a SVM classifier to detect pedestrians. Zeng *et al.* [28] propose a deep model to automatically learn scene-specific features so that a generic detector can work satisfactorily over different datasets. Similarly, Ouyang *et al.* [20] propose a deep learning technique that jointly learns the four key components in pedestrian detection simultaneously: feature extraction, deformation handling, occlusion handling and classifiers. Recently, Cai *et al.* [3] have proposed a cascade design procedure where an optimization algorithm is used to select complex features which are progressively pushed to later stage of cascades. This allows a single detector to work with a wide variety of features having different complexities.

Our work is primarily motivated by Chen's approach [4] where a DCNN is used to extract features from the proposal candidate windows obtained from an ACF detector. These features are then fed to a SVM classifier to confirm the presence of pedestrians in these windows. Essentially, it uses three stages for pedestrian detection, namely, an ACF detector, a CNN feature extractor and a SVM classifier. We propose to simplify this architecture by using only two stages comprising of an ACF detector and a CNN Mixture of expert (MoE) module which itself is used as a classifier thereby obviating the need for a separate classifier module. This Mixture of Experts trained on ACF detected windows is used for classifying the difficult candidates accurately for which the ACF confidence scores lie within a statistically identified band. This band represents those cases where the correct decisions could not be made by merely using thresholds on the ACF confidence scores. We use a feed-forward neural network to provide the adaptive combination of experts for classification similar to conventional adaptive boosting frameworks. The difference lies in the fact that the input itself decides the weightage for a given CNN expert in the ensemble. The resulting detector is shown to provide lowest miss-rate at FPPI=10^{-1} on the INRIA dataset as compared to the existing methods.

The main contributions made in this paper are as follows. (1) A mixture of experts (MoE) is used as an adaptive combination of CNN classifiers for improving the performance of standard ACF detectors. Here, CNN is used as classifier unlike [4] where it is used as feature extractor. Use of MoE in the context of CNN-based pedestrian detectors is novel. (2) We present novel use of ACF confidence scores for identifying the difficult samples which are then used for training the MoE CNNs. The upper and lower bounds on the confidence scores computed for these difficult cases are shown to augment the performance of ACF detectors.

The remainder of the paper is organized as follows. An overview of related work is provided in Section 2. The details of the proposed method is provided in Section 3. This is followed by Section 5 which provides details of various experiments conducted. Section 6 outlines the results. Finally, the conclusion and discussion on future work is provided in Section 7.

2. Related Work

Pedestrian detection is a well defined problem with established benchmarks and evaluation metrics. It has attracted a considerable amount of attention in the recent years which has resulted in development some of the best detectors in literature. Viola and Jones [23] were the first to demonstrate that a cascade of weak classifiers could be used for complex object detection tasks. They used AdaBoost algorithm to select critical features from a larger set to build efficient classifiers. On the other hand Dalal and Triggs [5] used complex features such as HoG with a simple classifier such as SVM to achieve detection. Either of these two frameworks were extensively used along with a number of features to detect pedestrians - Haar [24], LBP [26], ICF [9] and recently, Aggregated Channel Features (ACF) [6]. Felzenszwalb *et al.* proposed the deformable part model (DPM) [15] which is a considered to be a breakthrough in pedestrian detection. Readers are referred to survey papers [12] [16] [10] to get an overview on the current state-of-the-art in pedestrian detection.

The success of deep learning techniques in object detection [18] has prompted researchers to apply it to other problems including pedestrian detection [20] [28] [4] [3]. Among these models, Convolutional Neural Networks (CNN) have been found to provide state-of-the-art performance in pedestrian detections. Convolutional Neural Nets combine features at each layer using a non-linear transformation and generate hierarchies of features. Sermanet et al [21] obtained state of the art results using a convolution network whose filters were pre-trained using convolutional sparse coding, with feedforward connections across multiple layers. Chen *et al.* [4] used an Aggregated Channel Feature (ACF) Detector [6] whose candidate windows are fed into rich and deep Convolutional neural network pretrained on the ImageNet dataset. The resulting features were then classified using an SVM. Zeng et al [28] propose a deep joint network for learning scene specific features using a novel objective function which allows them to learn both unsupervised and discriminative feature representations. They address the problem of transferring detector performance from one dataset to another. In order to improve the detection performance further, Ouyang and Wang [20] propose a unified deep model for jointly learning feature extraction, a part deformation model, an occlusion model and classification. Recently, Cai *et al.* [3] proposed a cascade design procedure where an optimization module is used to select features based on complexity which are pushed to later stages of the cascade. Their focus was to develop an optimal cascade learning for deep networks.

Mixture of experts (MoE) [19] is an ensemble of classifiers which can be utilized for improving the performance of pedestrian detection [13]. Readers can refer to [27] for a survey in this area. In this paper, we use a MoE CNNs to augment the detection capabilities of a standard ACF detector unlike the above approaches where CNNs have been primarily used as feature extractors and classifiers. The proposed method is explained next in this paper.

3. Pedestrian Detection Framework

Our pedestrian detection utilises a combination of an ACF detector and a Mixture of Expert (MoE) CNNs. Fig 1 depicts our MoE combining five CNNs with varying architectures. The CNNs are trained using the augmented train set as explained in the Section 4. The trained network hyperparameters are determined via error minimization on a validation set. Subsequently, a four stage ACF detector is trained on our train set. The lower (l_t) and upper (u_t) ACF score cut-off thresholds are determined via five fold cross validation. At each fold of the cross validation, candidate windows are extracted via ACF, and different thresholds are imposed on the ACF scores for the candidate windows. Windows with ACF scores higher than the current upper threshold are classified as pedestrian and windows with ACF scores lower than the current lower threshold are classified as non-pedestrian. The windows with intermediate ACF score are classified based on the MoE train as described above. The thresholds that yield the best miss rate at 0.1 FPPI are selected for that fold. The mean thresholds over all the folds are selected as the final threshold values l_t and u_t to be applied for testing.

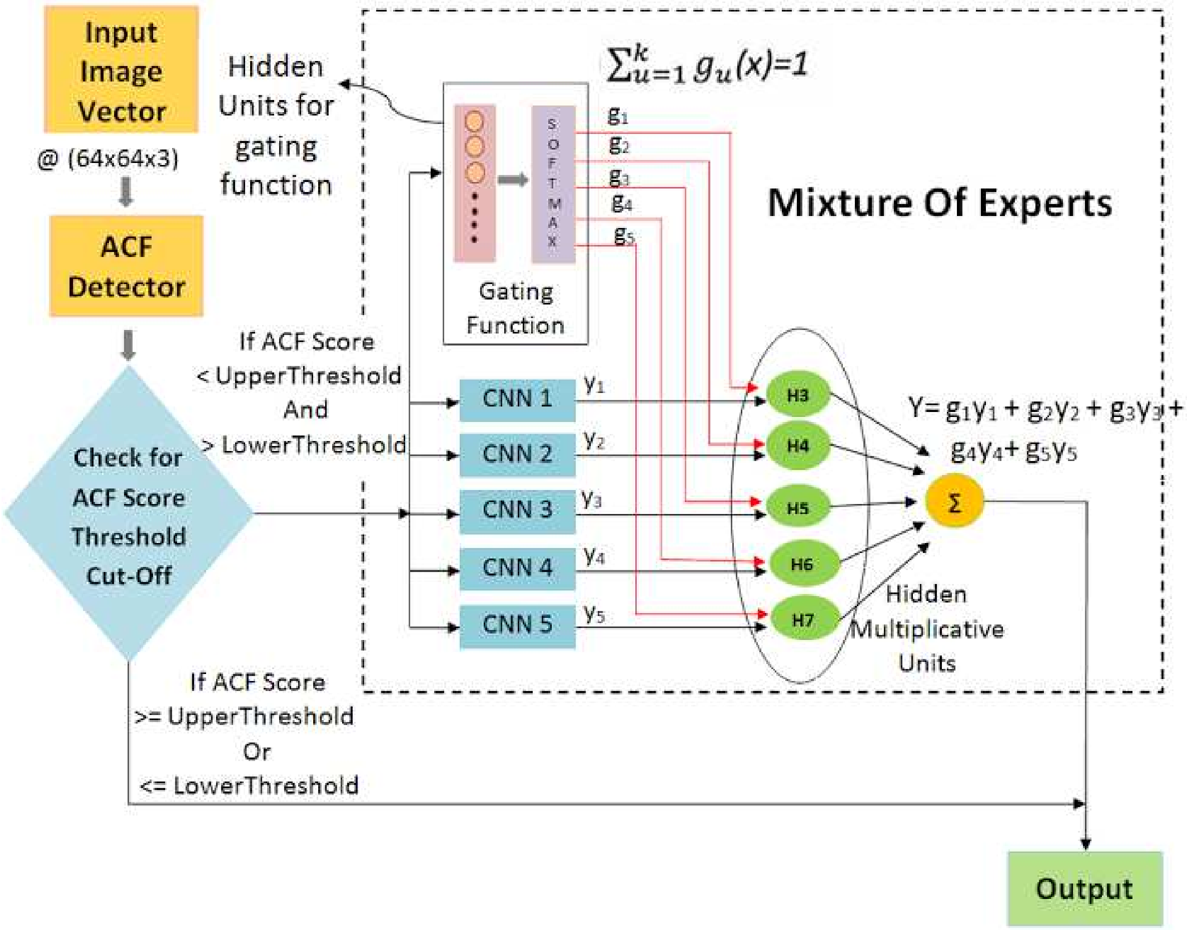

Input Image Vector
@ (64x64x3)
ACF Detector
Hidden Units for gating function
$\sum_{u=1}^{k} g_u(x)=1$
S O F T M A X
Gating Function
g1
g2
g3
g4
g5
Mixture Of Experts
Check for ACF Score Threshold Cut-Off
If ACF Score < UpperThreshold And > LowerThreshold
CNN 1
CNN 2
CNN 3
CNN 4
CNN 5
y1
y2
y3
y4
y5
H3
H4
H5
H6
H7
Y= g1y1 + g2y2 + g3y3 + g4y4+ g5y5
Σ
Hidden Multiplicative Units
If ACF Score >= UpperThreshold Or <= LowerThreshold
Output

Figure 1. The proposed approach consists of a combination of ACF and Mixture of Expert(MoE) CNNs. The output of ACF are the candidate windows and their respective confidence scores. The ACF scores are used to segment candidate windows into pedestrian, non-pedestrian and uncertain cases. The windows with confidence scores below the lower threshold are treated as non-pedestrians and the windows with confidence scores above upper threshold are classified as pedestrians. The ACF candidate windows that are within the threshold limits are passed onto the MoE for fine classification. The MoE learns to weigh the different CNNs dynamically based on the candidate windows.

3.1. Detector

We have trained the ACF detector with our train set images for localizing candidate windows. The three channel features considered are normalized gradient magnitude, HOG (Histogram of Oriented Gradients) and LUV color channels. These channels are generated from the input image and summed up in a 5×4 pixel grid. Next, a bootstrapping iteration is performed to construct a cascade of classifiers. We have used a cascade of 4 stages that combines $32, 128, 512$ and 2048 classifiers at each stage.

3.2. CNN Architecture

Figure 2 depicts a sample CNN architecture. The input images are fed to alternating convolutional and pooling layers as in a standard CNN. The outputs of the final pooling layers are fed into local layers. Local layers are similar to the convolutional layers except that there is no weight sharing at local layers. The activation function used at convolutional and local layers is ReLu (i.e. Rectified Linear Units). The CNN takes a 64×64 color image $(3$ channels$)$ as input and produces two output values, these two values are the probabilities of the input image belonging to the pedestrian or non-pedestrian class. Mini-batch gradient descent was used to train the CNNs using the cuda-convnet library[18] on an NVIDIA Quadra 4000 GPU processor. Table 1 lists the parameters of the CNN architectures used as inputs for the MoE model.

The gating function in the MoE was trained via a single hidden layer with sigmoidal activation function, which was fed into a softmax layer to generate a probability distribution over the different classifiers. The MoE network

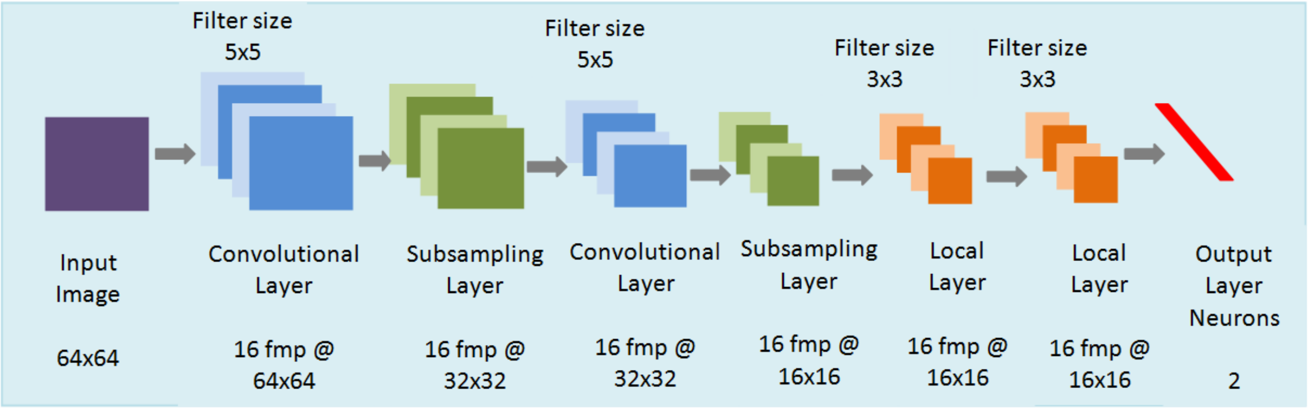


Figure 2. Illustration of CNN Architecture used: The windows with intermediate ACF scores (between the upper and lower bounds) are passed to multiple CNNs (having different architectures) for classification. The different CNNs yield different output probabilities for the candidate pedestrian windows. These output probabilities are merged through a gating function Mixture of Experts (MoE) to get the final output.

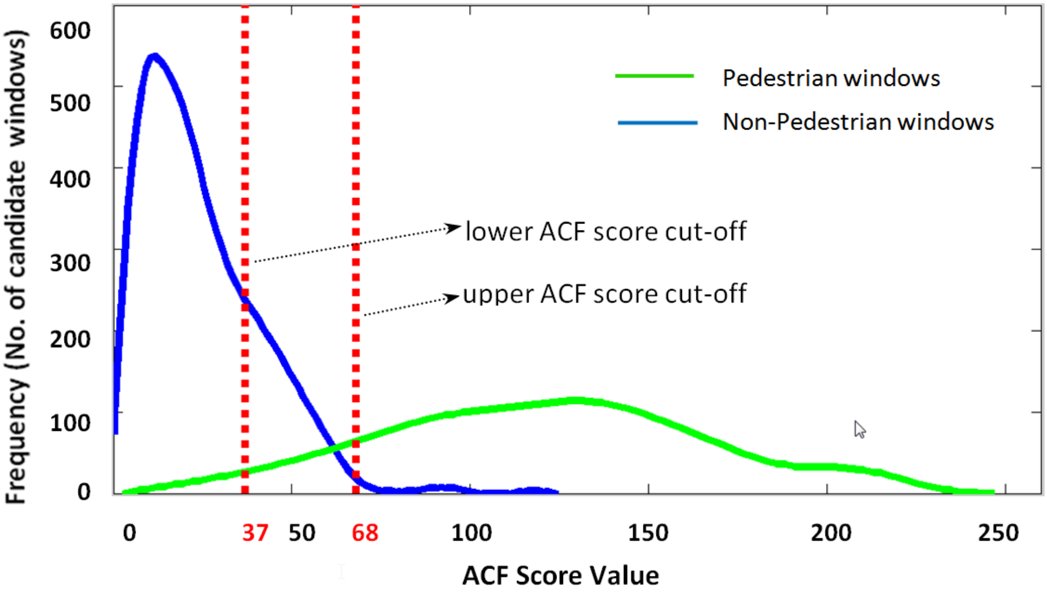


Figure 3. ACF confidence score histogram: Aggregated Channel Features (ACF) are extracted from the input images. The output of ACF are the candidate windows and their respective confidence values. The ACF score is used to decide on probability of candidate windows being pedestrian. The windows with confidence score below the lower threshold (37) are non-pedestrian windows and the windows with confidence score above upper threshold (68) are pedestrian windows. The windows in the band between upper and lower bounds are probabilistic and they need to be passed to the CNN network. Grid search to decide the best lower and upper ACF score cutoff. This greatly reduces the additional overhead on the CNN. The upper and lower thresholds are determined by varying thresholds from minimum to maximum value and determining when a best missed rate at 0.1 FPPI is acheived.

was trained over the same training and validation sets used to train the CNNs. The weights for the gating function of the MoE are adaptive to the input data i.e. the weights dynamically change depending on the input image, thereby allowing the MoE to adapt to changes in data distributions.

3.3. Combining ACF and MoE

The windows with intermediate ACF scores (between u_t and l_t) are passed to the MoE for classification. The different CNNs yield different output probabilities for the candidate pedestrian windows. These output probabilities are merged through a Mixture of Experts (MoE) to get the final

Table 1. Architectures for 5 layer CNN's

CNN	# Conv. layers	# Pooling layers	# Feature maps	Filter size (Conv.)
1	1	1	[64]	[3×3]
2	1	1	[16]	[11×11]
3	1	1	[16]	[15×15]
4	2	2	[16 16]	[5×5 5×5]
5	2	2	[32 32]	[3×3 3×3]

Table 2. Parameters for Mixture of Experts (MoE) training

Sl. No.	Parameter	Numeric Value
1	No. of CNN's used	5
2	No. of hidden units	51
3	Momentum	0.02
4	Weight Decay	0.01
5	Learning rate	0.02
6	Learning decay	1.0

score. Table 2 provides the paremeters of the MoE architecture that could be used to reproduce the pedestrian detection results on INRIA dataset.

4. Dataset

We utilize the INRIA pedestrian dataset in our experiments. The dataset is very popular in the pedestrian detection community, both for training detectors and reporting results. It offers high quality annotations of pedestrians in diverse settings and poses. The details of the INRIA training, test images and annotations are provided in Table 3

Table 3. INRIA dataset

	Training	Test
Positive images	614	288
Annotated windows	1237	589
Negative images	1218	453

The INRIA pedestrian dataset has only train and test sets. To optimally train our detector and to prevent overfitting, a validation set was constructed by randomly selecting 25% of the training set data. The remaining 75% of the training data is used for training the ACF detector and the CNN's.

The train set is prepared to train the ACF detector and CNN. To train ACF detector, the annotated pedestrian windows (provided with the INRIA train set) were extracted from the positive images of train set. These images were used as positive samples and the ACF windows generated from the negative images were used as negative samples.

To train the CNNs, the same data used for training the ACF detector is used. Additionally, we have augmented the negative set with images which were obtained as a result of running the trained ACF detector on the negative images. The negative set is further augmented with false positive windows extracted after running ACF on the positive images of INRIA train set. Finally we have also augmented the train set with the horizontal flipped images of both positive and negative samples. Flipping and augmentation makes the training more robust and less sensitive to noise. Finally, the prepared training set contains 1750 positives and 4902 negatives. The test set remains unchanged from the original INRIA test dataset, i.e., it comprised of 288 positive images and 453 negative images.

5. Experiments

The ACF detector generates many overlapping candidate pedestrian windows. Non maximal supression in the context of object detection is used to transform a smooth response map that triggers many imprecise window hypotheses into a single candidate window. The necessity stems from the inability of detection algorithms to localize the concept of interest resulting in groups of multiple detections near the real location.

5.1. Non-Maximum Suppression (NMS)

The ACF detector generates many overlapping candidate pedestrian windows. Non-Maximal supression is used to filter out these overlapping bounding boxes. The object detection community typically uses the PASCAL overlap criteria to determine overlap between two bounding boxes as per equation 1.

$$OverlapArea = \frac{(IntersectionArea)}{(UnionArea)} \tag{1}$$

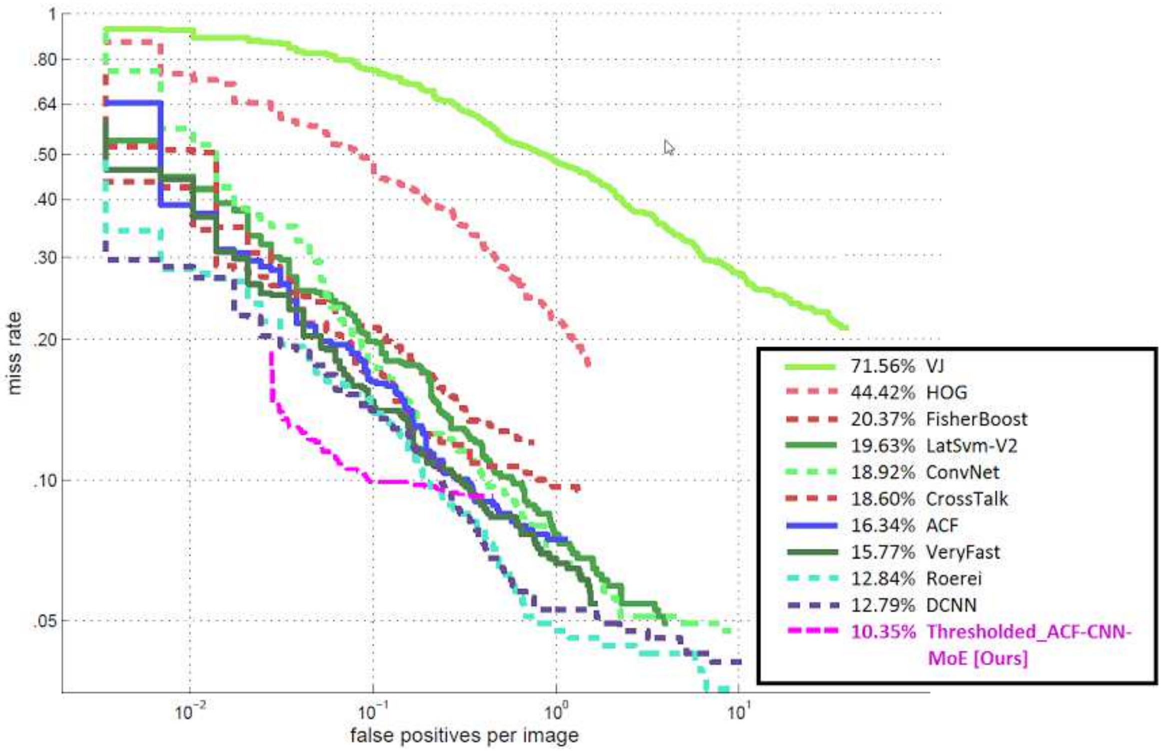


Figure 4. Comparison of miss rate of various detectors with our method (Thresholded ACF-CNN-MoE). We obtain the state of the art miss rate of 10.35% at 0.1 FPPI.

If two bounding boxes overlap by more than 60% then the bounding box with highest ACF score is selected.

We have used the PASCAL overlap criteria to determine overlap between two bounding boxes. To suppress the resulting false positives that arise due to partially occluded pedestrians, we have augmented our training set for the CNNs with these type of false positives (discussed in Section 4), so that our CNN is better trained to filter them out.

5.2. Determining ACF threshold via Cross Validation

After applying ACF detection with NMS on validation images we obtained candidate pedestrian windows. The overlap of these candidate windows with ground truth windows are used to validate the performance of the ACF detector. Again the PASCAL overlap criteria is used to determine overlap. The ground truth window is declared to be detected if the overlap area of the candidate window and the ground truth window exceeds 50%. In our case the ACF after thresholding is able to detect 92.8% ground truth pedestrian windows, in addition to false positives. We pass the candidate windows to the trained MoE to obtain probabilities that these candidate windows belong to the pedestrian class.

At this stage, we have the ACF scores and CNN classification probabilities of all candidate windows generated by the ACF detector on the validation set. We utilize grid search to obtain the optimal thresholds on the ACF score which minimize miss rate on the valdation set at 0.1 FPPI. The candidate windows having ACF scores greater or equal than the upper threshold are classified as pedestrian and the candidate windows having ACF score lesser or equal than the lower threshold are classified as non-pedestrian. For the windows having ACF scores between the optimal thresholds, the MoE output probabilities are used to compute the miss rate at 0.1 FPPI.

5.3. Final Testing

The purpose of dividing the original given train set into – train and validation sets was to determine the optimal ACF score thresholds. Now for final testing we use the complete INRIA train set for training the ACF detector and the MoE. We train a 4 stage ACF detector on the complete INRIA train set. We also train the CNNs and the MoE with same original INRIA train set images augmented as described in Section 4.

After training the ACF detector and CNN, we give the entire test set to the ACF detector which extracts candidate pedestrian windows. To filter out the duplicate overlapping candidate windows we have applied NMS by utilizing the

PASCAL overlap criteria to determine overlap between two bounding boxes as discussed above for validation set evaluation. After applying ACF with NMS on the test set images we obtained the candidate pedestrian windows. The overlap with ground truth windows are used to check the performance of ACF. Again the PASCAL overlap criteria is used to find overlap area. In our case ACF is able to provide a cover rate of 93.21% on the test set. Now we apply the lower and upper ACF score cutoffs l_t and u_t as described in Section 5.2. Figure 3 shows the application of l_t and u_t ACF score cutoff thresholds on the test set. The windows with intermediate scores were assigned probability scores by the MoE.

The experiments were run on Intel Xeon CPU ES-2650 running at 2GHz, 40 GB RAM, and GPU card NVIDIA GF100GL Quadro 4000 on an Ubuntu 14.04 64 bit system. For implementation of the ACF detector we used Piotr Dollar's Toolbox[1]. For implementation of the MoE we used the cuda-convnet library [18]. The average time for processing an image was found to be 0.6 secs. which is comparable to state of the art. The current code is in MATLAB and is yet to be optimised to work on a real-time hardware.

6. Results

Table 4. Miss rate for various detectors on INRIA dataset

Sl. No.	Technique used	Missed Rate
1.	Voila -Jones [23]	71.56%
2.	HOG [5]	44.42%
3.	Fisher Boost [22]	20.37%
4.	LatSVM-v2[14]	19.63%
5.	ConvNet [18]	18.92%
6.	Cross Talk[7]	18.60%
7.	ACF [6]	16.34%
8.	Very Fast [8]	15.77%
9.	Roerei [2]	12.84%
10.	ACF-DCNN-SVM [4]	19.79%
11.	**Thresholded ACF-CNN-MoE (Ours)**	**10.35%**

The plot in Figure 4 compares our method, named Thresholded-ACF-CNN-MoE, against the other leading methods in pedestrian detection on the INRIA dataset. Table 4 denotes the performance of various detectors on the INRIA dataset at 0.1 FPPI. After applying lower and upper ACF score thresholds on the candidate windows, and further classification by the MoE, we obtained the state of the art miss rate of 10.35% at 0.1 FPPI. Figure 5 shows examples of both the pedestrian and non-pedestrian detections with the addition of MoE CNN architecture with thresholded ACF detector. These cases would fail when ACF detector was used stand-alone.

7. Conclusion and Future Work

We conclude by summarising our key contributions:

1. We present a technique involving the combination of an aggregated channel feature based detector and a mixture of CNN experts for pedestrian detection and demonstrate its efficacy on the INRIA dataset. The proposed detector acheives state of the art performance in terms of miss rate.

2. We simplify the architecture proposed by Chen et al [4] by using only two stages comprising of an ACF detector and a CNN Mixture of expert (MoE) module which itself is used as a classifier thereby obviating the need for a separate SVM classifier module listed in their work.

3. We present the novel use of ACF confidence scores for identifying the difficult samples which are then used for training the MoE CNNs. The threshold scores are determined by using grid search on the bimodal ACF score histogram.

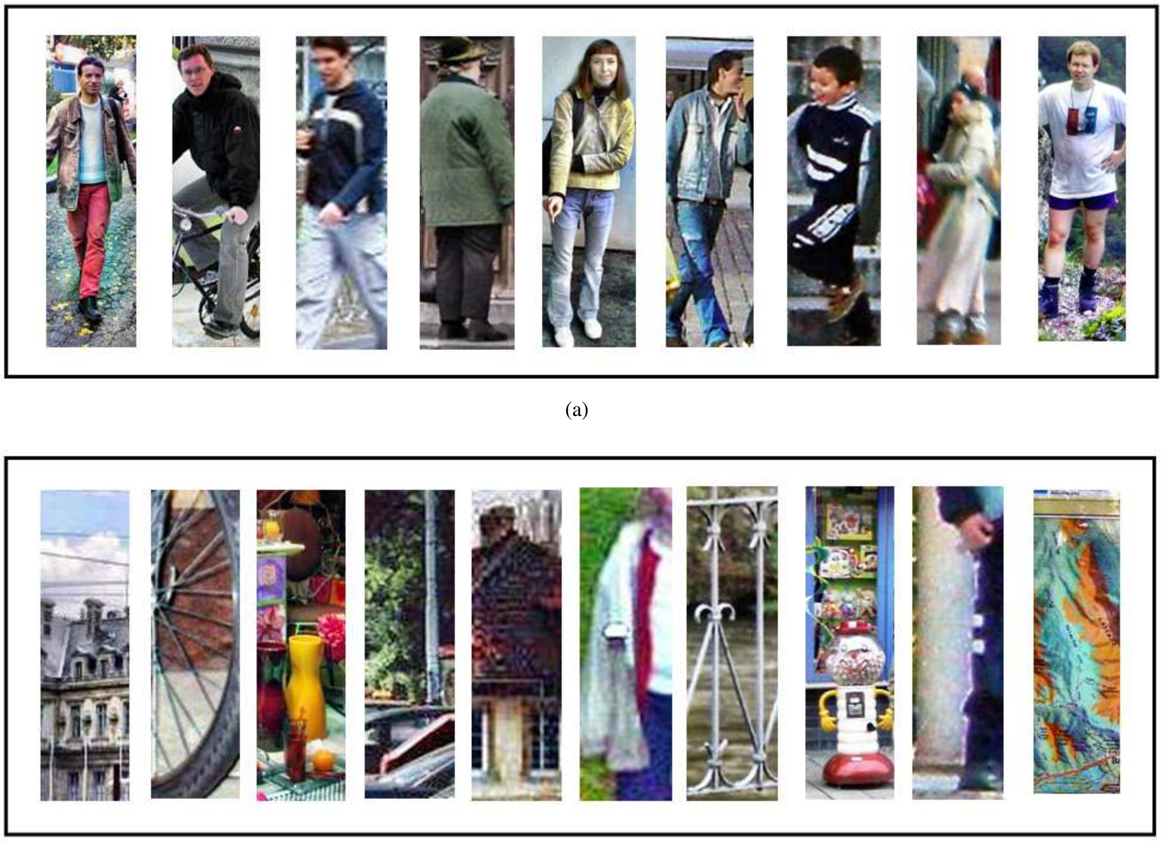
The research directions are in the broad areas of changing MoE CNN architecture for transfer learning, exploring combinations of feature detectors that yeild better cover rate and miss-rate. In the future, we intend to follow the research directions listed below.

1. The MoE is robust to changes in data distributions and thus may be extremely useful when transferring classifiers from one dataset to the another. We would like to test the proposed method across all pedestrian benchmark datasets such as ETH, CALTECH, and KITTI etc.

2. We plan to fuse the detectors that could potentially provide complementary descriptions and better detection rates in combination with the ACF detector. We also intend to explore feature detectors and object proposal techniques such as the Edgebox [11] could yield better true positive rates.

3. It may be possible to feed the ACF as a competing classifier in MoE instead of using it only to generate windows. We will experiment with such architectures.

[1] https://github.com/pdollar/toolbox



(a)

(b)

Figure 5. Example of correctly classified pedestrians and non-pedestrian windows are shown in subfigures (a) and (b) respectively. These windows would otherwise be incorrectly classified by ACF detector alone without MoE CNN architecture.

References

[1] G. A. Bekey. On autonomous robots. *The Knowledge Engineering Review*, 13(02):143–146, 1998. 1

[2] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3666–3673. IEEE, 2013. 7

[3] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. *arXiv preprint arXiv:1507.05348*, 2015. 1, 2

[4] X. Chen, P. Wei, W. Ke, Q. Ye, and J. Jiao. Pedestrian detection with deep convolutional neural network. In *Computer Vision-ACCV 2014 Workshops*, pages 354–365. Springer, 2014. 1, 2, 7

[5] 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. 1, 2, 7

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1532–1545, 2014. 1, 2, 7

[7] P. Dollár, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *Computer Vision–ECCV 2012*, pages 645–659. Springer, 2012. 7

[8] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, volume 2, page 7. Citeseer, 2010. 7

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, volume 2, page 5, 2009. 2

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):743–761, 2012. 1, 2

[11] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1841–1848. IEEE, 2013. 7

[12] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2179–2195, 2009. 1, 2

[13] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *Image Processing, IEEE Transactions on*, 20(10):2967–2979, 2011. 2

[14] 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. 7

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 1, 2

[16] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1239–1258, 2010. 1, 2

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. 1

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1, 2, 3, 7

[19] S. J. Nowlan and G. E. Hinton. Evaluation of adaptive mixtures of competing experts. In *NIPS*, volume 3, pages 774–780, 1990. 2

[20] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2056–2063. IEEE, 2013. 1, 2

[21] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multistage feature learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3626–3633. IEEE, 2013. 1, 2

[22] C. Shen, P. Wang, S. Paisitkriangkrai, and A. van den Hengel. Training effective node classifiers for cascade classification. *International journal of computer vision*, 103(3):326–347, 2013. 7

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001. 1, 2, 7

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

[25] K. Wada, T. Shibata, T. Saito, and K. Tanie. Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proceedings of the IEEE*, 92(11):1780–1788, 2004. 1

[26] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009. 2

[27] S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(8):1177–1193, 2012. 2

[28] X. Zeng, W. Ouyang, M. Wang, and X. Wang. Deep learning of scene-specific classifier for pedestrian detection. In *Computer Vision–ECCV 2014*, pages 472–487. Springer, 2014. 1, 2