It takes two to tango: Cascading off-the-shelf face detectors

Siqi Yang , Arnold Wiliem and Brian C. Lovell
The University of Queensland
siqi.yang@uq.net.au, a.wiliem@uq.edu.au, lovell@itee.uq.edu.au

Abstract

Recent face detection methods have achieved high detection rates in unconstrained environments. However, as they still generate excessive false positives, any method for reducing false positives is highly desirable. This work aims to massively reduce false positives of existing face detection methods whilst maintaining the true detection rate. In addition, the proposed method also aims to sidestep the detector retraining task which generally requires enormous effort. To this end, we propose a two-stage framework which cascades two off-the-shelf face detectors. Not all face detectors can be cascaded and achieve good performance. Thus, we study three properties that allow us to determine the best pair of detectors. These three properties are: (1) correlation of true positives; (2) diversity of false positives and (3) detector runtime. Experimental results on recent large benchmark datasets such as FDDB and WIDER FACE support our findings that the false positives of a face detector could be potentially reduced by 90% whilst still maintaining high true positive detection rate. In addition, with a slight decrease in true positives, we found a pair of face detector that achieves significantly lower false positives, while being five times faster than the current state-of-the-art detector.

1. Introduction

Face detection has been studied for decades in the computer vision domain, and it is one of the fundamental problems for many facial analysis tasks.

The goal of face detection is to obtain a high detection rate with extremely low false positives. There are two concurrent challenges for developing this: 1) how to detect more faces in the unconstrained environment to improve the accuracy; 2) how to reduce false positives to achieve efficiency. Recent works [15, 18, 30, 4] primarily aim at developing face detectors in the unconstrained environment which is closely aligned with the real world applications, such as video surveillance. In this scenario, faces may have large variations in shape and appearance, *e.g.,* large head pose and expression variations, illumination variations, low resolution, out-of-focus blur, motion blur, and occlusions.

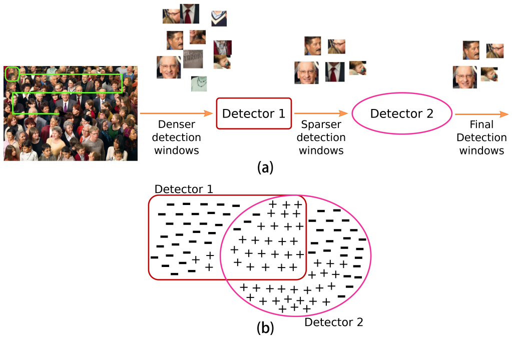


Figure 1: (a) We propose a two-stage cascade framework which cascades a face detector with a second face detector as a post-processing classifier to remove the false positives. Compared with the dense patches produced by the sliding window approach or the region proposal network, it is still very efficient for the second detector to process these sparse patches. (b) We propose three cascade properties that the second detector must have 1) a high correlation of true positives ('+'); 2) a high diversity of false positives ('-') with the first detector and 3) less runtime than the first detector to achieve an overall fast speed.

Although these state-of-the-art methods achieve high detection rates, they also generate many false positives. As face detection is the first step in many facial analysis tasks, the unexpected false positives will affect the accuracy and speed of the subsequent tasks and the overall system. In a surveillance video, the majority of video frames are occupied by the background (*i.e.,* non-faces) with much fewer faces appearing. The imbalanced distribution of non-faces and faces increases the probability of generating false positives for a face detector. Therefore, a face detector with low false positive rate is highly desirable.

Indeed, the trade-off between speed and accuracy always exists in developing a face detector. For a detector, more true faces can be detected by accepting more false positives. The challenge now is how to reduce the false positives whilst still maintaining the high true positive rates of the state-of-the-art detectors. As all existing state-of-the-art face detectors generate false positives, any method aimed at reducing false positives has significant potential to improve their performance.

With the seminal work of Viola and Jones [24], a cascading approach has shown its advantages in achieving both accuracy and efficiency by quickly rejecting a large number of easy negatives in the earlier stages and training stronger classifiers for the hard negatives in the later stages. More recently, the idea of cascade has also been applied to the Convolutional Neural Network (CNN) based methods [13, 30] to save on the high computational expense.

Although a vast number of easy negatives are discarded in the earlier stages, there still remains a large number of false positives after the final stage. This is because of the huge imbalance in the distribution of faces and non-faces which makes classifiers not sufficiently discriminative to distinguish these hard negatives. To address this, *bootstrapping* or *hard negative mining* is frequently used when training face detectors. In a nutshell, this method first trains the detector with an initial training set. Once trained, the method iteratively adds any false positive detected by the current detector model into the training set and then retrains the model.

Most of the state-of-the-art face detectors [15, 18, 30, 4] utilise this bootstrapping scheme, however, the methods still generate excessive false positives. Due to the features, classifiers and training samples, every face detector has its own theoretical limits. In other words, the classifiers in the final cascade stage are simply not powerful enough to distinguish between faces and non-faces. Recent work [29] calls this problem the Hard Face/non-Face problem.

To address the problem, one needs to use completely different architectures and features due to the discrimination limits of each detector. This motivates the use of ensemble/fusion methods combining different methods based on orthogonal features. However, rather than training several models in parallel, several works [29, 2, 12, 23] propose cascading a *post-processing* classifier using quite different features and classifiers.

Furthermore, the effort to train a new face detection model is enormous, *e.g.,* large training data and some face detectors do not provide open source training codes. In this work, we propose a method that addresses the problem of reducing false positives of existing face detectors without spending the tremendous effort of retraining them or developing an entirely new detector. One possible way is to cascade two pre-trained/off-the-shelf face detectors. Here, we consider the face detectors as black boxes characterised by their face/non-face performance.

In this work, we propose to cascade two pre-trained face detectors, where the first detector can be considered as a region-proposal detector and the second one as a post-processing classifier. In the two-stage framework, the second detector is expected to have the ability to detect and pass through all the faces output from the first detector, while being able to remove the false positives at the same time. For this reason, it is crucial to determine the set of properties that allow us to optimise which two detectors can be cascaded and in which order they should be cascaded. Inspired by the fusion approaches in pedestrian and object detection [25, 6, 10], we study the properties of existing face detectors.

Some fusion approaches in the domain of pedestrian detection and object detection [25, 6, 10] propose to exploit the complementary information from multiple existing detectors by combining the results from these detectors. However, different from these fusion approaches which are primarily aimed at increasing true positives by using the complementary information, we argue that it is still possible to develop a cascade method to reduce the false positive using the complementary information. Inspired by this, we propose to study the cascade properties by analysing the *correlation* and *diversity* in the true positives and false positives respectively as well as the runtime.

We then validate our findings by cascading various recent state-of-the-art face detector methods and evaluate the efficacy of the proposed properties in selecting pairs of face detectors. These validations are performed using the FDDB [5] and WIDER FACE [28] datasets.

Contributions - We list our contributions as follows: 1) To reduce the false positives of the existing face detectors, we propose a two-stage cascade framework that cascades two pre-trained detectors (refer to Fig. 1). 2) We propose three essential properties that guide us in determining the efficacy of the cascaded detector. These properties are based on the correlation and diversity of both true and false positives from the two face detectors as well as the runtime. 3) With the proposed cascade properties, we study twelve pairs of detectors. The experimental results show our proposed framework is able to remove a large number of false positives with an insignificant loss of true positive rate. 4) We found a pair of face detectors that achieves significantly lower false positive rate with competitive detection rate, which is five times faster than the current state-of-the-art detector described in [4].

2. Related Work

Face detection methods can be roughly grouped into three families: 1) boosting based methods, 2) Deformable Parts-based Models (DPM) methods and 3) deep learning based methods. Viola and Jones (VJ) [24] are the first to propose Haar-like features and use the AdaBoost learning algorithm to train weak classifiers. They proposed to cascade the face/non-face classifiers which discard the easy negatives quickly whilst spending more computation on more face-like samples. Due to the high efficiency, the VJ made face detection ubiquitous for many real-time applications. However, it has been shown that in unconstrained scenarios, the VJ detector is not effective in detecting faces with large head pose variations and occlusions [7]. Similar to VJ's framework, Mathias *et al.* [18] introduced an integral channel features detector with boosting, called HeadHunter. HeadHunter is essentially a multi-scale detector model based on 22 rigid templates. For each scale, there are 11 templates: 5 templates for frontal faces and 6 for rotated faces. In an entirely different approach, Liao *et al.* [15] developed an unconstrained face detector by proposing a novel feature, called Normalised Pixel Difference (NPD). A deep quadratic tree is proposed to learn and combine the features and a single soft-cascade boosting classifier is further applied to learn the trees, without resorting to pose-specific cascade structures or pose labelling. In addition to the VJ's framework, there are several face detection methods [32, 18] based on Deformable Part Models (DPM) to model potential deformations between facial parts.

Recently, deep learning methods have shown exceptional performance in object detection [22, 9], so they have been extended to face detection [30, 4, 1, 20, 21, 14, 13, 27, 19]. Zhang et al. [30] proposed a deep multi-task framework, called Multi-task CNN (MTCNN), in a three-stage cascaded structure. In each stage, face classification, facial landmarks localisation, and bounding box regression are trained jointly. The state-of-the-art face detection method proposed by Hu *et al.* [4] is able to find tiny faces. Separate detectors are trained for different resolutions in a multi-task fashion and therefore it is referred to as 'hybrid-resolution' (HR). This research argues that the context information is crucial for detecting small faces and therefore, they associate the receptive fields over the features extracted from different layers of the network.

Even though the current face detectors can achieve very high recall, they also generate false positives. Yang *et al.* [29] introduced the Hard Face/Non-Face (HFnF) problem that embodies the challenge of reducing the false positives generated by the existing face detectors. Solving this problem is critical as the solution could have a significant impact on all the existing face detectors. Previous works [12, 2] showed that in conjunction with facial landmarks based features, an SVM classifier [3] can be used as a post-processing classifier. As discussed in [29], although the above methods demonstrate their effectiveness in reducing false positives, they are shown to be insufficient due to their high dependence on face alignment accuracy. Besides using the cues of facial landmarks, Li *et al.* [11] construct a contour-based classifier to reduce the false positives after the VJ detector. However, the contour features will not perform well when faces have large head pose variations or occlusions; thus, cannot be used to detect faces in the unconstrained environment.

Similarly, several recent works [31, 17] aim at achieving both high speed and accuracy by utilising the pre-trained detectors. Zhou *et al.* propose to train an Adaptive Feeding (AF) classifier to determine a given image is easy or hard by a linear SVM. An "easy" image is then fed into a fast but less accurate detector, whereas a "hard" image is by an accurate but slow detector. Different from the ensemble methods, both their work and ours do not run the two detectors in parallel, which saves enormous computations and time. However, their AF classifier is like a "switch" which decides one out of the two detectors to process the image, while our work is a cascade of two detectors. Moreover, their "easy" and "hard" labels do not explicitly explore the correlation and diversity of true and false positives between different detectors. Perhaps the most relevant work to ours is proposed by Marčetić *et al.* [17], which cascades two detectors: NPD [15] and DPM [32]. In fact, their method is a special case of our proposed framework. Both theirs and ours use a two-stage cascade model to reduce false positives. Unlike their method which only shows the efficacy of cascading NPD and DPM, in our work, we show that there are more effective pairs of cascaded face detectors. More importantly, we propose the cascade properties that can determine the pair of cascaded detectors. These properties allow us to sidestep the expensive detector retraining step.

3. Proposed approach

We first discuss our two-stage framework and then we describe the properties used to find the most effective pairs of face detectors to cascade.

3.1. Two-stage cascade framework

The two-stage framework is shown in Fig. 1. Two pre-trained face detectors are represented by rectangular and elliptical shapes, respectively. Let $\mathrm{h}_1 : \mathcal{R}^i_1 \mapsto \mathcal{R}^o_1$ and $\mathrm{h}_2 : \mathcal{R}^i_2 \mapsto \mathcal{R}^o_2$ be the first and second face detectors, respectively. $\mathcal{R}^i$ and $\mathcal{R}^o$ are the set of input and output image regions. In the absence of ambiguity, to simplify the notation, we will drop the subscript for $\mathcal{R}^i$ and $\mathcal{R}^o$. Hence, the two-stage cascade detector can be denoted as

$$\mathrm{f}(\mathcal{R}^i) = \mathrm{h}_2(\mathrm{h}_1(\mathcal{R}^i)) \; . \tag{1}$$

Remarks. The set of input image regions, $\mathcal{R}^i$ is first generated by using a sliding window approach [24, 15, 18, 13, 30] or the region proposal network [1, 4]. Some face detectors will also consider multiple resolutions of image regions. Generally, the sliding window approach or region proposal network will generate hundreds of proposals ($|\mathcal{R}^i_1| \approx 500$) while only output a few of them, *e.g.,* $|\mathcal{R}^o_1| \approx 5$. Since it is assumed that the number of regions containing faces will be much smaller than the number of input regions, $|\mathcal{R}^i| >> |\mathcal{R}^o|$, a face detector can be considered as a set reduction function that reduces $\mathcal{R}^i$ into $\mathcal{R}^o$. From this relationship we have the following proposition.

Proposition 3.1 *The cardinality of the set of input regions of the second detector is always far smaller than the cardinality of the input regions of the first detector,* $|\mathcal{R}^i_2| << |\mathcal{R}^i_1|$.

Proof. To prove this, note that $\mathcal{R}^i_2 = \mathcal{R}^o_1$. As we know that

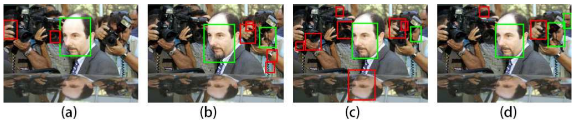


Figure 2: Detections from four different face detectors on the FDDB dataset [5]: (a) NPD [15], (b) HeadHunter [18], (c) MTCNN [30] and (d) HR [4]. Green: true positives, red: false positives.

$|\mathcal{R}_1^i| >> |\mathcal{R}_1^o|$, therefore $|\mathcal{R}_2^i| << |\mathcal{R}_1^i|$ must be true.

Remarks. The computational complexity of fusion-based detectors (placed in parallel) increase linearly according to the number of detectors and the overall running time is constrained by the slowest detector. Unlike fusion-based detectors, cascading detectors (*i.e.,* placing them in series) will not significantly increase the overall running time and computations due to a much smaller number of regions for the second detector to process. If we use a slower but more complex face detector as the second detector, it is potential to achieve a faster overall speed than the slowest detector. It is also noteworthy to mention that the cascading of two face detectors will not create a face detector that has better detection rate than the weaker of the two detectors. This is due to Proposition 3.1 which essentially limits the ability of the second detector to detect more faces. In other words, the second detector will not be able to detect faces not detected by the first detector. However, when carefully selected, the cascaded face detector could outperform the first detector with respect to the low false positive rate without adding much computational time due to the efficiency introduced by the cascade structure.

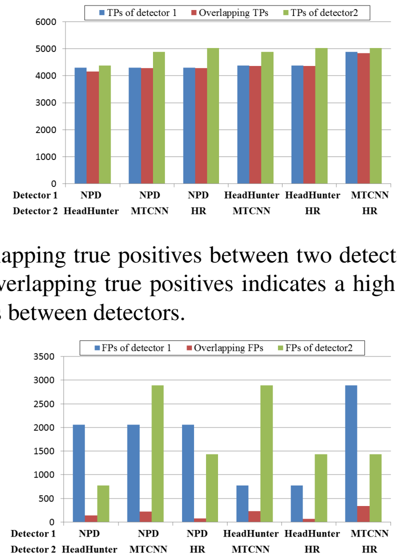
3.2. Diversity and correlation metrics

Given the same image, different face detectors will produce different detection results, as shown in Fig. 2. This is caused by the various training samples, features and classifiers used by the detectors, as shown in Table 1. Even though every face detector has its own theoretical limits, the cascade framework is able to utilise the different features and classifiers to improve the discrimination.

In the two-stage cascade framework, two questions naturally arise: 1) which pair of detectors should be cascaded; 2) which order should they be cascaded. We need a set of properties that will guide us to address these questions.

The proposed properties are derived from the distribution of the true and false positives between the detectors. More specifically, we define a detected window $bb = [x, y, w, h]$ as a true positive when its intersection-over-union ratio with the ground truth window is greater than 0.5 and otherwise, as false positive.

Given two pre-trained face detectors, named as detector 1 and detector 2, we collect the detection windows $\mathcal{R}_1^o = \{bb_1^m\}_{m=1}^{N_1}$ and $\mathcal{R}_2^o = \{bb_2^n\}_{n=1}^{N_2}$, where N_1 and N_2 are the numbers of detections. The set $\mathcal{R}_1^o$ comprises the set of true positives $\mathcal{T}_1$ and false positives $\mathcal{F}_1$. Thus, $\mathcal{R}_1^o = \mathcal{T}_1 \cup \mathcal{F}_1$ and $\mathcal{R}_2^o = \mathcal{T}_2 \cup \mathcal{F}_2$.



(a) The overlapping true positives between two detectors. The large number of overlapping true positives indicates a high correlation of true positives between detectors.

(b) The overlapping false positives between two detectors. The small number of false positives indicates a high diversity of false positives between detectors.

Figure 3: The distribution of overlapping detections between some face detectors on the FDDB [5]. We can see that only a small number of false positives are detected by both detectors, whereas a majority of true positives overlap.

To measure the overlap ratio α between two bounding boxes bb_1^m and bb_2^n, we adopt the commonly used intersection-over-union (IoU): $\alpha = \frac{area(bb_1^m \cap bb_2^n)}{area(bb_1^m \cup bb_2^n)}$.When the overlap ratio α of two bounding boxes is larger than 0.3, we consider them as an overlapping pair. Then we collect the overlapping true positives $\mathcal{T}_o$ and overlapping false positives $\mathcal{F}_o$ respectively:

$$\mathcal{T}_o = \mathcal{T}_1 \cap \mathcal{T}_2 \, , \qquad \mathcal{F}_o = \mathcal{F}_1 \cap \mathcal{F}_2 \, . \tag{2}$$

Fig. 3 shows the distribution of the overlapping true and false positives of the four face detectors on the FDDB dataset [5]. To quantify the distribution, we define the *correlation* and the *diversity* to measure the overlapping and non-overlapping detections between two detectors. Since the ability of different detectors to detect faces varies, the correlation and diversity towards different detectors need to be considered individually. In this work, we denote the correlation of detector 2 to detector 1 as $c_{2 \rightarrow 1}$, which is the ratio of the number of overlapping detections to the total number of detections from the detector 1. The diversity $d_{2 \rightarrow 1}$ is defined as the ratio of the number of non-overlapping detections to the total number of detections.

As the detections consist of true positives and false positives, we argue that it is necessary to formulate the correlation and diversity with true and false positives separately:

$$c^T_{2\to1} = \frac{|\mathcal{T}_o|}{|\mathcal{T}_1|}\,, \qquad c^F_{2\to1} = \frac{|\mathcal{F}_o|}{|\mathcal{F}_1|}\,,$$
$$d^T_{2\to1} = 1 - \frac{|\mathcal{T}_o|}{|\mathcal{T}_1|}\,, \qquad d^F_{2\to1} = 1 - \frac{|\mathcal{F}_o|}{|\mathcal{F}_1|}\,, \qquad (3)$$

where $c^T_{2\to1}$ is the ratio of the number of overlapping true positives to the number of detections of detector 1, and $d^T_{2\to1}$ is the ratio of the number of non-overlapping true positives to the number of detections of detector 1. $c^F_{2\to1}$ and $d^F_{2\to1}$ are used to measure the ratio of overlapping or non-overlapping false positives in a similar way.

In our proposed two-stage cascade framework, only the detections agreed by both detectors can pass all the stages. Therefore, we assume that the final detection result is the intersection set of detector 1 and detector 2 as stated in Eq. 2. The high diversity of false positives shows that we can utilise the conflicting decision on the false positives from the two detectors to let the different detectors compensate for their own mistakes.

3.3. Cascade Properties

For face detection, both accuracy and efficiency are the most critical concerns. To choose the best pairs of two detectors to cascade and determine the cascade order, we propose three cascade properties: 1) correlation of true positives; 2) diversity of false positives; and 3) detector runtime.

On one hand, in order to improve the accuracy, the detector in the late stage is expected to have the ability to remove the false positives as well as maintaining the detection rate unchanged at the same time. The detection rate is often referred to recall. To this end, there are two properties are required to hold when we would like to cascade two detectors. The recall, R_c, and precision, P_c, of the cascade detector are used to evaluate the accuracy with respect to true positives and false positives respectively. Let us introduce the following proposition.

Proposition 3.2 *In the cascade framework, when the two properties are maximised: 1) $c^T_{2\to1} \approx 1$; 2) $d^F_{2\to1} \approx 1$, the performance of the cascaded detector will always be better than the performance of the first detector. That is, $P_c > P_1$, and $R_c \approx R_1$.*

Proof. To prove this, let us denote the ground truth as $\mathcal{T}_*$, then $R_c = (|\mathcal{T}_c|/|\mathcal{T}_*|)$. As we know from the Eq. 3 that $|\mathcal{T}_c| = |\mathcal{T}_o| = |\mathcal{T}_1| \times c^T_{2\to1}$, then the recall becomes $R_c = (|\mathcal{T}_1| \times c^T_{2\to1})/|\mathcal{T}_*|$. Therefore, when the $c^T_{2\to1} \approx 1$, the following relationship $R_c \approx R_1$ must be true. Similary, since $|\mathcal{F}_c| = |\mathcal{F}_o| = |\mathcal{F}_1| \cdot (1 - d^F_{2\to1}))$, when $d^F_{2\to1} \approx 1$, $|\mathcal{F}_c| >> |\mathcal{F}_1$. Hence, $P_c = |\mathcal{T}_c|/(|\mathcal{T}_c|+|\mathcal{F}_1|\cdot(1-d^F_{2\to1})) > P_1$.

Remarks. In our proposed two-stage cascade framework, if the second face detector has a high correlation of true positives to the first detector, $c^T_{2\to1}$, the recall, R_c, will tend to be maintained; otherwise the recall will drop significantly. Meanwhile, the high diversity of false positives $d^F_{2\to1}$ achieves the goal of reducing false positives, which results in an increase of precision, P_c. In this way, cascading the second detector can reduce a large number of false positives while maintaining the true positives detected by the first detector. As a result, the precision of the first detector, P_1, will be increased to P_c with the recall well maintained, $R_c \approx R_1$. In other words, by visualising it with the Receiver Operating Characteristic (ROC) curve, we can see from Fig. 4 that at a specific number of false positives, the true positive rate can be increased significantly.

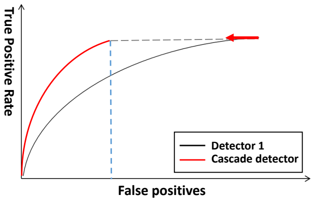


Figure 4: By cascading a second detector, a large number of false positives can be removed while the recall is well maintained. As a result, at a low number of false positives, the true positive rate can be increased significantly.

On the other hand, to achieve the high efficiency, the running time is the third property to be considered. As discussed in Proposition 3.1, vastly fewer image regions are fed to the second stage, and therefore the expected computational load of the detector in the second stage is much smaller than the first stage. In terms of the two-stage cascade framework, to achieve overall fastest speed, we propose to use the faster face detector in the first stage and the slower detector in the second stage.

4. Experiments

4.1. Datasets

The experiments employed two datasets:

FDDB dataset [5]. The dataset includes images of faces with a wide range of difficulties such as occlusions, difficult poses, low resolution and out-of-focus faces. The images are collected from the Yahoo! news website. It contains 2,845 images with a total of 5,171 faces labelled.

WIDER FACE dataset [28]. The dataset is currently the largest face detection dataset, which contains 32,203 images and 393,703 annotated faces based on 61 events from the Internet. The dataset contains faces with various appearance, poses, and scales. It divides the test protocols into three levels of difficulties: 'Easy', 'Medium' and 'Hard'.

4.2. Implementation Details

Note that our proposed method does not need retraining. However, we still need to select the best pair of detectors to cascade. In this work, we explore the cascade framework with four face detectors in Table 1: NPD [15],

Table 1: Comparisons of face detectors with regard to the features, classifiers and training sets.

Method	Features	Backbone Machinery	Training Set
NPD [15]	Normalised pixel difference	AdaBoost	AFLW [8]
HeadHunter [18]	Integral channel features	AdaBoost	AFLW [8]+Pascal Faces [26]
MTCNN [30]	Deep CNN features	DeepNet	CelebA [16]+WIDER FACE [28]
HR [4]	Deep CNN features	DeepNet	ImageNet pre-training+WIDER FACE [28]

HeadHunter [18], MTCNN [30], HR [4]. The first two detectors are tree-based detectors whilst the last two detectors are based on deep networks. Twelve possible pairs of detectors can be constructed from these four detectors. We do not evaluate the cascade framework with the VJ [24] as its true positive rate is half of the other detectors and its correlation with the other detectors will be very low. We use the FDDB dataset [5] as a validation set to select which two detectors can be cascaded and their order according to the cascade properties in Section 3.3: *i.e.,* high correlation of true positives, high diversity of false positives and runtime. Once the best pairs of detectors are determined, we test them on the WIDER FACE dataset's validation set [28] as its test set does not provide ground truth information.

Following the FDDB dataset [5], we compute the Intersection-over-Union (IoU) as the evaluation metric. When the IoU is larger than 0.5, the detection is considered as true positive; otherwise, false positive. For the evaluation on FDDB, we plot the Receiver Operating Characteristic (ROC) curves. For the WIDER FACE dataset [28], we follow their evaluation metric and plot the Precision and Recall (PR) curves.

4.3. Evaluation on FDDB dataset

Before evaluating the proposed cascade framework, we calculate the correlation and diversity of true positives and false positives between detectors to decide the pairs to cascade and their orders. Runtime analysis is then conducted.

4.3.1 The correlation and diversity

From Fig. 3, we can see that only a small number of false positives are detected by both detectors, whereas a majority of true positives overlap as expected. We calculate the correlation and diversity metrics defined in Section 3.2 on these detector pairs. The results on the FDDB [5] dataset are shown in Table 2 and 3. In Table 2, a high correlation of true positives c^T corresponds to the large number of overlapping true positives. It is not surprising to see that most detectors overlap on the true positives as they are designed to detect true faces. On the contrary, there is a high diversity of false positives d^F, which is caused by the various training samples, features and classifiers used by the detectors (see Table 1).

Table 2: The correlation of true positives $c^T_{2\to1}$.

	Detector 2			
Detector 1	NPD [15]	HeadHunter [18]	MTCNN [30]	HR [4]
NPD [15]	1	0.9683	0.9970	0.9967
HeadHunter [18]	0.9487	1	0.9959	0.9961
MTCNN [30]	0.8755	0.8926	1	0.9900
HR [4]	0.8523	0.8694	0.9640	1

Table 3: The diversity of false positives $d^F_{2\to1}$.

	Detector 2			
Detector 1	NPD [15]	HeadHunter [18]	MTCNN [30]	HR [4]
NPD [15]	0	0.9339	0.8916	0.9645
HeadHunter [18]	0.8236	0	0.7030	0.9170
MTCNN [30]	0.9228	0.9207	0	0.8826
HR [4]	0.9491	0.9554	0.7636	0

It is noteworthy to mention that the order of detectors is important. Table 2 and Table 3 show that the correlation metric $c^T_{2\to1}$ may decrease when the order is reversed. This is due to the inability of the second detector to detect the true positives detected by the first detector. According to the cascade properties in Section 4.2, we select 6 out of 12 possible pairs of cascade detectors: NPD-HeadHunter, NPD-MTCNN, NPD-HR, HeadHunter-MTCNN, HeadHunter-HR and MTCNN-HR.

4.3.2 Evaluating the two-stage cascade framework

Fig. 5 shows the discrete ROC curves of our proposed 12 different pairs of cascade detectors as mentioned in Section 4.2 and the individual face detectors: NPD [15], HeadHunter [18], MTCNN [30] and HR [4]. For practical purposes, a good face detector is considered to have not only a high true positive rate but also a low false positive rate. Therefore, we plot the True Positive Rate (TPR) at the same number of false positives in the legend of Fig. 5 and Table 4. As the FDDB dataset contains 2,845 images, the specific number of false positives is selected as 284, which corresponds to 1 false positive per image (*i.e.,* a False Positives Per Image (FPPI) of 0.1).

As shown in Fig. 5, compared with the performance of the individual NPD detector [15], the TPR (FPPI=0.1) of NPD can be significantly increased from 80% to 81%, 84% or 84% by cascading HeadHunter, MTCNN or HR respectively. The reasons are two-fold.

First, due to the high diversity of the false positives between the two detectors $d^F_{2\to1}$, the false positives of NPD can be reduced from $2,058$ to 200 by cascading HeadHunter, MTCNN or HR (*i.e.,* yielding a 10 times reduction of false positives!).

Second, the high correlation of the true positives $c^T_{2\to1}$ ensures the overall true positive rate is well preserved. Both NPD-MTCNN and NPD-HR have higher TPR (FPPI=0.1) than NPD-HeadHunter. It is because the $c^T_{2\to1}$ of NPD-MTCNN (0.997) and NPD-HR (0.997) are higher than that of NPD-HeadHunter (0.96).

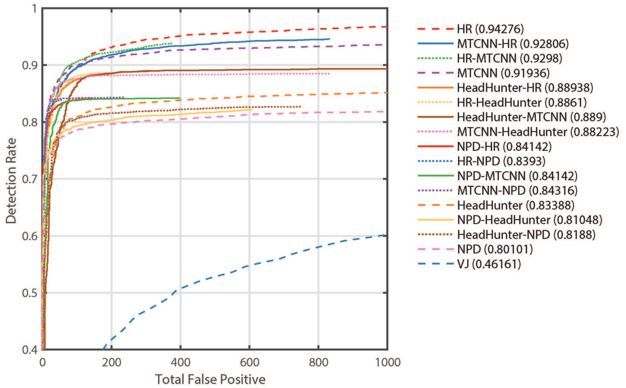


Figure 5: The comparisons between our proposed two-stage cascade detectors and individual face detectors on the FDDB dataset [5].

With the large reduction of false positives as well as the overall TPR maintained, the ROC curve is shifted to the left, which results in an increase of TPR at the low number of false positives when compared with the individual detector.

In addition, when comparing NPD-MTCNN with NPD-HR, NPD-HR can reduce more false positives than NPD-MTCNN with the same overall TPR. This owes to the higher diversity of false positives: $d^F_{HR \to NPD}$ (0.965) which is higher than $d^F_{MTCNN \to NPD}$ (0.8916). It is noteworthy to mention that NPD-HR can even have a slightly higher true positive rate than NPD itself. It is because the HR needs to utilise the context information to classify faces/non-faces so that we expand the detections before forwarding to the second stage. As such, the false positives with localisation error can be corrected by the bounding box regression scheme of HR. Similarly, HeadHunter-MTCNN and HeadHunter-HR can improve the HeadHunter with a higher detection rate at the low number of false positives because of the high correlation of true positives (0.996) and high diversity of false positives.

For the cascade detectors using all deep learning detectors, we still increase the performance. For instance, the MTCNN TPR at FPPI = 0.1 increases from 92% to 93% when it is cascaded with HR. These results suggest that when the proposed cascade properties are satisfied, the first detector TPR can potentially be improved by cascading it with a stronger second detector.

As discussed, the six above-mentioned pairs of cascade detectors are chosen and ordered according to the proposed cascade properties. Beside these pairs, we evaluate another six pairs of which the order is reversed. Fig. 5 shows that the performance (with regard to TPR) of a reversed pair is on par with the corresponding original pairs. For example, both HR-MTCNN and MTCNN-HR have the same TPR at FPPI=0.1 (93%). These pairs are on par with the current state-of-the-art detector, HR [4] (94%).

Table 4: The runtime of detectors evaluated in this work on the FDDB dataset.

Method	CPU time (SPF*)			TPR (FPPI#=0.1)
	1st stage	2nd stage	total time	
VJ [24]	0.271	-	0.271	0.462
NPD [15]	0.678	-	0.678	0.801
NPD-HeadHunter	0.678	988	988.678	0.810
NPD-MTCNN	0.678	0.073	0.751	**0.841**
NPD-HR	0.678	2.678	3.356	**0.841**
HeadHunter [18]	1961	-	1961	0.834
HeadHunter-NPD	1961	0.404	1961.404	0.819
HeadHunter-MTCNN	1961	0.116	1961.116	**0.889**
HeadHunter-HR	1961	3.648	1964.648	**0.889**
MTCNN [30]	0.355	-	0.355	0.919
MTCNN-NPD	0.355	0.220	0.575	0.843
MTCNN-HeadHunter	0.355	456	456.355	0.882
MTCNN-HR	0.355	3.496	3.851	**0.930**
HR [4]	17.687	-	17.687	**0.943**
HR-NPD	17.687	0.170	17.857	0.839
HR-HeadHunter	17.687	794	811.687	0.886
HR-MTCNN	17.687	0.076	17.763	0.930

* SPF–Seconds Per Frame # FPPI–False Positives Per Image

Table 5: Comparison of our proposed framework and the state-of-the-art face detector.

Method	CPU time (SPF*)	TPR (FPPI#=0.1)
HR [4]	17.687	**0.943**
MTCNN-HR (ours)	**3.851**	0.930

* SPF–Seconds Per Frame # FPPI–False Positives Per Image

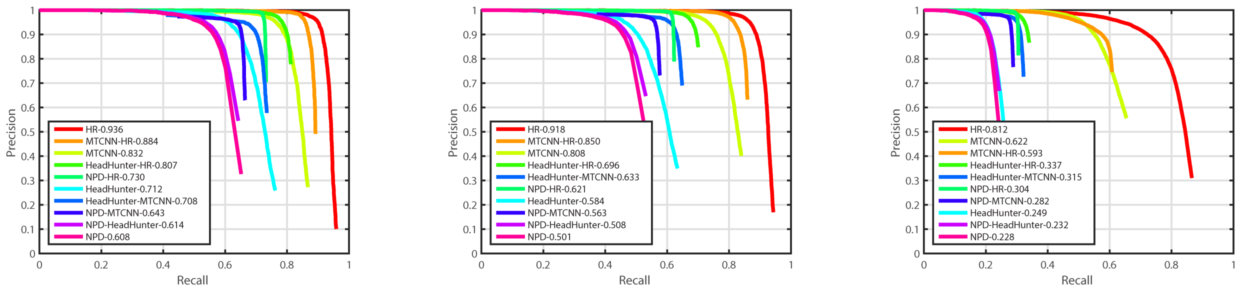
Since HR is the best detector, it is not possible to improve its accuracy. Nevertheless, it is still possible to significantly decrease its running time. This can be observed from the MTCNN-HR which runs five times faster than HR (in Table 5). We will discuss this in details in the next subsection.

4.3.3 Runtime Analysis

In face detection, both detection accuracy and running time are critical factors. Therefore, we evaluate the runtime in a video surveillance scenario, where all images are resized to 640 × 480 VGA images. To make a fair comparison, all detectors are tested on a E5-1620@3.5 GHz CPU with only a single thread. In this work, the GPU time is not evaluated as the detector, NPD [15], is not implemented with GPU. The minimum face sizes of all the detectors are set to 20×20 pixels. The CPU time in Table 4 is the average time per image on the FDDB dataset [5].

We first benchmark the runtime of each individual face detectors in the first column of Table 4. It is worth noting that the MTCNN [30], a CNN based face detector, runs faster than NPD, on CPU. Unlike the reported time of the NPD in [15], which is 30 ms per image with a single thread of CPU, it is 67.81 ms in our experiments. The following experimental settings may lead to the different speed of the NPD: 1) our minimum face size (20 × 20) is smaller than theirs (80 × 80) and the test images are different; 2) we use the unoptimized MATLAB code.

Compared with running individually, when a face detector is cascaded as the second detector, the runtime of this second detector is much smaller. It can be seen from the sec-



(a) Val:easy.

(b) Val:medium.

(c) Val:hard

Figure 6: Comparisons on the WIDER FACE validation set [28]. The precision and recall curves of different subsets: easy, medium and hard.

ond and third columns of Table 4 that the runtime of HR [4] is 17.6 s, whereas when HR [4] is performed as a second detector in MTCNN-HR, the runtime is only 3.5 s, which is five times faster. This is because only a small amount of candidate windows are fed into the second detector. This means less data to process for the second detector compared to running it as an individual detector.

With a small price of slight runtime increase, the performance of current face detectors can be significantly improved by cascading with a second detector, as shown in the last column of Table 4.

According to the proposed properties that the TPR and FPR are not affected significantly no matter which order is chosen for cascading a pair of detectors, this leaves the decision to the runtime of the resulting cascade. From our experiments, we found that using the faster detector as the first detector will not significantly increase the overall runtime.

In Table 5, we compare the best pair of detectors, MTCNN-HR, with the state-of-the-art face detector, HR [4], with regard to runtime and TPR. It is noteworthy that the MTCNN-HR achieves five times less runtime than the HR while maintaining a competitive accuracy, with a TPR of 93%. The results demonstrate that our proposed two-stage cascade framework can not only improve the accuracy of current face detector by removing false positives but also achieve high computational efficiency.

4.4. Evaluation on WIDER FACE dataset

In the WIDER FACE dataset [28] evaluation, we only test the six pairs of cascade face detectors satisfying the cascade properties. Fig. 6 reports the performance of the proposed cascade detectors and the individual detectors. The curve labels in the legend are sorted according to the average precision (AP).

It can be seen from the Fig. 6 that the proposed two-stage cascade detectors can have larger AP than the individual detectors in these three different subsets. This demonstrates that our proposed two-stage cascade detectors successfully reduce a large number of false positives while maintaining the true positive rates. In the comparison of each two-stage cascade detector and the individual detectors, the performance is consistent with the FDDB dataset [5] which indicates that the correlation and diversity of the true and false positives between the face detectors still exist on the WIDER FACE dataset. This suggests that it is possible to optimise the pair of face detectors using a dataset.

5. Conclusions

The central goal of this work was to improve the existing face detectors' performance by reducing their false positives whilst maintaining high true positive rate. To this end, a two-stage cascade framework, cascading two pretrained face detectors, was proposed. The cascade framework showed its efficiency and effectiveness as fewer detections are passed onto the second detector and there is no significant increase in the overall runtime. In this two-stage framework, the cascade properties were studied by exploring the correlation and diversity between the face detectors. We further showed that to improve a face detector, the second detector must have a high correlation of true positives and a high diversity of false positives with respect to the first face detector. Our experiments showed that our proposed cascade framework improves existing face detectors significantly by removing a large number of false positives with minor loss of true positives. The improvement is shown as an increasing detection rate at low numbers of false positives. In addition, we showed that the diversity and correlation metrics are consistent between datasets. This suggests, it is possible to find the best pair of detectors using a pilot dataset and apply it this to another dataset. In this way, we can avoid retraining the detectors. In this work, we successfully found a pair of face detector that achieves significantly lower false positives with competitive detection rates, and five times greater speed than the current state-of-the-art detector described in [4].

Acknowledgements This work has been partly funded by Sullivan Nicolaides Pathology, Australia and the Australian Research Council (ARC) Linkage Projects Grant LP160101797. Arnold Wiliem is funded by the Advance Queensland Early-Career Research Fellowship. We thank Dr Shaokang Chen for the useful discussions.

References

[1] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *ECCV*, 2016. 3

[2] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*, 2014. 2, 3

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008. 3

[4] P. Hu and D. Ramanan. Finding tiny faces. In *CVPR*, 2017. 1, 2, 3, 4, 6, 7, 8

[5] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010. 2, 4, 5, 6, 7, 8

[6] S. Karaoglu, Y. Liu, and T. Gevers. Detect2rank: Combining object detectors using learning to rank. *IEEE Transactions on Image Processing*, 25(1):233–248, 2016. 2

[7] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015. 2

[8] 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 *ICCV Workshops*, 2011. 6

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[10] H. Lee, H. Kwon, R. M. Robinson, W. D. Nothwang, and A. M. Marathe. Dynamic belief fusion for object detection. In *WACV*, 2016. 2

[11] H. Li and L. Chen. Removal of false positive in object detection with contour-based classifiers. In *ICIP*, 2010. 3

[12] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *ICCV*, 2013. 2, 3

[13] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, 2015. 2, 3

[14] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *ECCV*, 2016. 3

[15] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *PAMI*, 38(2):211–223, 2016. 1, 2, 3, 4, 5, 6, 7

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6

[17] D. Marčetić, T. Hrkać, and S. Ribarić. Two-stage cascade model for unconstrained face detection. In *Sensing, Processing and Learning for Intelligent Machines (SPLINE), First International Workshop on*. IEEE, 2016. 3

[18] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. 1, 2, 3, 4, 6, 7

[19] H. Qin, J. Yan, X. Li, and X. Hu. Joint training of cascaded cnn for face detection. In *CVPR*, 2016. 3

[20] 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. *PAMI*, 2016. 3

[21] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. *arXiv preprint arXiv:1611.00851*, 2016. 3

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

[23] M. Tapaswi, C. Ç. Çörez, M. Bäuml, H. K. Ekenel, and R. Stiefelhagen. Cleaning up after a face tracker: False positive removal. In *ICIP*, 2014. 2

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 2, 3, 6, 7

[25] P. Xu, F. Davoine, and T. Denœux. Evidential combination of pedestrian detectors. In *BMVC*, 2014. 2

[26] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014. 6

[27] S. Yang, P. Luo, C.-C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, 2015. 3

[28] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *CVPR*, 2015. 2, 5, 6, 8

[29] S. Yang, A. Wiliem, and B. C. Lovell. To face or not to face: Towards reducing false positive of face detection. In *Image and Vision Computing New Zealand (IVCNZ), International Conference on*. IEEE, 2016. 2, 3

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1, 2, 3, 4, 6, 7

[31] H.-Y. Zhou, B.-B. Gao, and J. Wu. Adaptive feeding: Achieving fast and accurate detections by adaptively combining object detectors. In *ICCV*, 2017. 3

[32] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 3