Enhanced Face Detection: An Adaptive Cascade-Mixture Approach for Large-Scale Detection

Michael J. Taylor
taylorm@cs.man.ac.uk
Tim Morris
morris@cs.man.ac.uk

School of Computer Science
University of Manchester
Manchester, UK

Abstract

Programmatically detecting faces within images is a problem that has a multitude of solutions. Typically, it is achieved through the detection of facial features, whereby a certain number of features being found within close proximity of one another would indicate the presence of a face. However, if the features of a face are largely indiscernible, how can we still confidently classify it as a face? This is an issue we attempt to solve with our new approach, which mixes large numbers of feature detections, and filters them according to size, grouping, and colour. In this paper, we present the concept behind each component of our system, demonstrate its superiority to existing work over a dataset we have constructed, draw conclusions from our findings, and discuss a number of improvements we intend to make.

1 Introduction

As the power and ubiquity of computers, be they desktop machines or smaller electronic devices, rapidly increases, so too does the potential for them to be used for beneficial purposes. Sensing technologies have certainly been subject to great advancements over recent years, strengthening the efficacy and accessibility of human-computer interaction (HCI) applications. A fundamental and extremely prominent aspect of this interaction is face detection. The detection of faces is the gateway to a multitude of other face-based processes, including recognition, modelling, authentication, tracking, expression analysis, and numerous others.

Given an arbitrary image, it is the objective of a face detection algorithm to ascertain whether or not faces are present, and, if so, to return information pertaining to the apparent size and location of any existent faces. This is not a trivial problem to overcome for computers, because of the extreme variabilities involved. Even two separate instances of the same person's face can be subject to changes in scale, lighting conditions, occlusion, location, orientation, and expression, among other factors.

There exist many different approaches to solving the issue [1]. In 2002, Yang et al. [2] devised four categories of detection techniques: knowledge-based methods, which use pre-defined rules to classify faces based upon human knowledge; feature-invariant methods, which aim to find certain facial features that are robust to variations; template-matching

methods, which compare images to pre-stored templates of known faces in order to determine their presence; and appearance-based methods, which learn the characteristics of faces from representative training image sets.

No matter the type of process used, the developmental priority for any face detection system will be on confidently detecting faces whilst avoiding false positive results. The development of our system is no different, although the specific types of images we intend to detect faces within can be considered more complex than the arbitrary images most techniques are tested using, meaning that our classification process to separate faces from false positives must also be more complex (in terms of the number and types of criteria utilised).

It is our intention to develop a system that is capable of "large-scale" face detection, by which we mean face detection in images that exhibit a large range of face scales, such as those taken in lecture theatres. We consider such inputs to be complex because we can reasonably expect them to contain not only large numbers of faces, but also for many of those faces to be "weak". This weakness can relate to a number of factors, including small size, a lack of focus (which will often go hand in hand with the former), unideal pose, or partial occlusion.

Throughout our research, we have yet to come across an adequate solution to the problem. This is somewhat understandable, as the vast majority of detection systems would rightly only classify image regions as faces when they exhibit compelling face-likeliness according to certain criteria. Positively classifying regions that do not necessarily do so would most certainly lead to increases in false positive rates, which would be damaging to overall efficacies. This does not preclude, however, the notion that those unconvincing regions could become so provided additional criteria. Given the nature of the inputs we are focusing on, it is our desire to find methods to overcome the outlined obstacles to weak-face detection, in order to achieve confident, accurate classifications.

A sample of the type of images we will primarily be building our solution for can be seen in Fig. 1, which is also the image we will be using as input to demonstrate how our process works throughout this paper.

Figure 1: A typical image of a lecture theatre, with many visible faces at various scales.

The aim of the PhD that this system is a part of is to develop an adequate solution to the problem we have identified, and to implement it as both a mobile application for a blind lecturer and as a tool for centrally monitoring lecture theatre occupancies. We place great value in the adaptability of methods, so any processes we develop will have been built with a wide range of potential use-cases in mind. This is particularly true of the adaptive skin colour modelling and segmentation approach that we have previously published [3], which we will be adopting for this face detection system.

In this paper, we will take a brief look at some of the more prominent face detection technologies (Section 2), explain how each component of our system works in the context of solving our outlined problem (Section 3), present results that quantify the extent to which our solution can outperform certain existing techniques (Section 4), draw a number of conclusions from our findings that pertain to the effectiveness of the system we have developed (Section 5), and discuss a number of issues with our solution in its current state that we intend to improve upon (Section 6).

2 Previous work

Of the types of detectors identified by Yang et al. [2], it is the appearance-based methods that have risen to great prominence over recent years, as significant advancements to their efficiency have made them suitable for real-world applications. Much of the recent work in the field owes greatly to the foundations laid down by Viola and Jones in 2001 [4], as they introduced three extremely important innovations to world of Haar-like feature-based object detection. The first of these was the integral image, which is a system for rapidly calculating the values of image sub-regions, affording significant reductions to computational costs. The second was the AdaBoost (Adaptive Boosting) algorithm, a process by which a very strong classifier can be constructed through the combination of numerous weak classifiers. Finally, the attentional cascade classifier structure ensures that clearly negative image sub-regions are swiftly discarded by the given detection process, further improving efficiency. Viola and Jones went on to successfully apply their methodologies directly to the problem of real-time face detection [5].

Since then, a multitude of researchers have attempted to improve upon the work. Lienhart and Maydt [6], for instance, extended the Haar-like feature set of the original cascades by introducing 45˚-rotated rectangular features. Similarly, Mita et al. [7] proposed the introduction of "joint" Haar-like features, which would be used to specify the co-occurrence of multiple individual features, resulting in a stronger classification system. Fröba and Ernst [8] did not believe that the original features were robust enough to handle extreme lighting conditions, so new, illumination-invariant features were derived using a modified census transform in conjunction with a boosting algorithm. Away from using Haar-like features entirely, Ojala et al. [9] proposed using the features of the local binary patterns (LBP) to detect faces, which have also been successfully applied to the problem of face recognition.

As well as direct improvements to the original work, many face detection systems have been developed that combine the Viola-Jones detector [5] with alternative modalities to reinforce results. For example, Burgin at al. [10] combine face detection results with context and depth information in order to improve human-computer interaction in robots. Nanni et al.[11] propose the use of a pre-trained colour classifier to filter face candidates, as well as the use of a depth map to investigate the unevenness of those regions, creating a process that takes multiple face properties into account, improving overall detection accuracies. Shieh and Hsieh [12] use the IR information provided by a Kinect sensor in order to refine detection results through structured light analysis. Alternative hardware is also employed by Ruiz-Sarmiento et al. [13], as they initially detect face candidates by using the Viola-Jones detector on intensity data provided by a time-of-flight (ToF) camera, then filter them according to three-dimensional information, such as region flatness and size-to-distance ratio. Seguí et al. [14] found that integrating content-based face detection information with contextual body part information could yield significant improvements to accuracy, provided the availability of such information.

3 Methodology

3.1 Overview

A typical feature-based face detector will classify faces by simply applying a group size threshold to clusters of features detections. However, in instances where facial features may be largely indiscernible, we must look for alternative methods of classification. Given our intention to build a solution for large-scale detection problems, wherein poorly expressed features are commonplace, developing such a method becomes a necessity.

The approach we have developed is based around the concept of mixing a multitude of individual feature detections for a given image, and intelligently filtering them to leave an accurate set of faces. Our system is comprised of a multi-stage process, wherein each stage is designed to filter out detections that are unlikely to be faces according to certain criteria, all the while adding confidence to those detections that remain. The pipeline of our new system can be seen in Fig. 2.

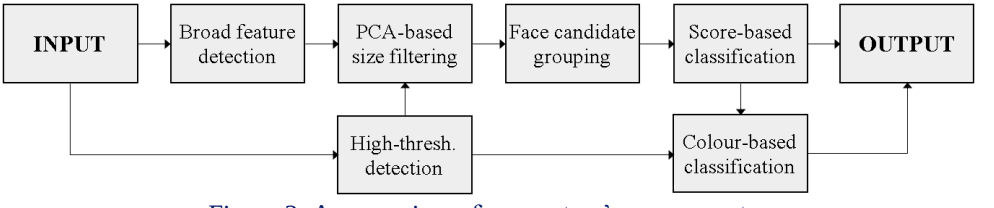


Figure 2: An overview of our system's components.

3.2 Broad feature detection

Our process begins with the aforementioned collation of feature detections, which involves mixing the results of multiple Viola-Jones cascades [5], run with no classification threshold (preventing detection elimination). The OpenCV implementation of their system comes packaged with four Haar-like feature cascades trained for frontal face detection, as well as one trained to detect faces from a profile point-of-view. Additionally, the LBP features relating to frontal faces (as used by Ojala et al. [9]) are also available. The features of a single cascade would usually yield acceptable detection rates for simpler tasks, but where our problem is concerned, it is paramount that as much information as possible is made available about potential faces, so we utilise all six of these cascades.

The detections themselves are simply image regions that exhibit a feature that would also be exhibited by a human face under normal circumstances. Usually, if enough of these features are detected within the same region of an image, a decision is made that a face is present there. Conversely, unless an actual face region has a prerequisite number of feature detections associated with it, it will not be classified as such. This is the main issue we wish to address. Fig. 3 illustrates how we retain and mix every single feature detection for further processing.

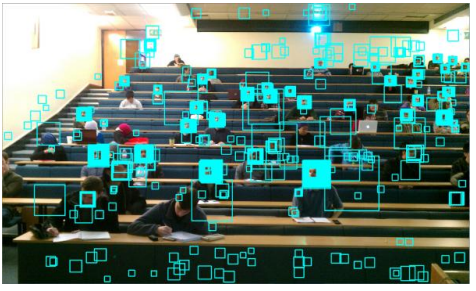
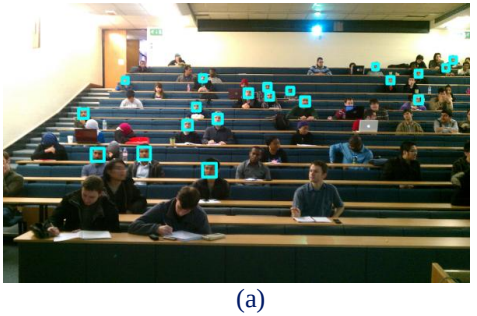


Figure 3: A typical result of broad feature detection, with thousands of potential face candidates found.

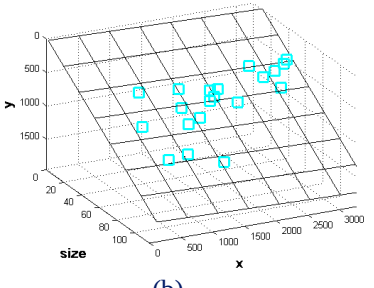
3.3 PCA-based size filtering

The first method by which we filter out detections is according to their size. This cannot be as simple as declaring any detection "larger than x" or any detection "smaller than y" to definitely not be a face, however, as this would not take into account the distribution of potential faces within the given image, or the circumstances under which the image has been captured. On the other hand, it is plain to see in Fig. 3 that there are large discrepancies in the sizes of detections within close proximity in numerous instances, and, assuming a reasonably ordinary distribution of people, not all of them can be faces.

Getting a sense of this distribution is key to constructing a size filter for a given image, and it is towards this end that principal component analysis (PCA) is used. If we were to build a small set of high-confidence faces (as in Fig. 4(a), achieved by applying a large group size threshold to detection clusters), then PCA would allow us to establish an accurate relationship between image coordinates and expected face sizes for the given image. The high-confidence aspect of this process is extremely important, as the results of using PCA are highly susceptible to false positive data. For our purposes, "establishing a relationship" essentially means building a two-dimensional plane in our three-dimensional (x,y,size) space, the result of which can be seen in Fig. 4(b).



(a)



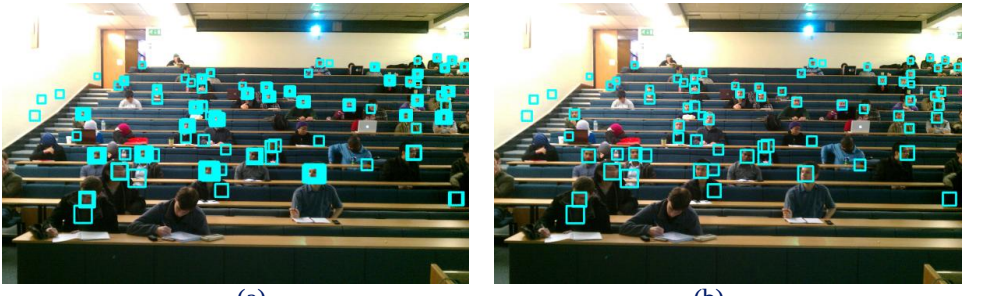
(b)

Figure 4: (a) The result of high-threshold face detection, and (b) the result of using PCA to build a model relating image coordinates to face sizes.

Interestingly, using this process, we have derived a three-dimensional partial representation of the environment being depicted by the given two-dimensional image. With such a model, we can simply deem any detections that do not fall within error bounds of their expected size (given their coordinates) to not be representative of actual faces, and discard them. The result of size filtering can be seen in Fig. 5(a).

3.4 Face candidate grouping

Typically, a size-filtered set of detections will still be very large, primarily because most of the faces present within a given image will have multiple detections associated with them. It is therefore pertinent to group these detections, and form a set of actual face candidates. This is a relatively simple process, whereby detections are consolidated based upon their locations and sizes meeting certain similarity criteria. The result of grouping detections can be seen in Fig. 5(b).



(a) (b)

Figure 5: (a) Our feature detections are consolidated into (b) face candidates.

3.5 Score-based classification

An incredibly useful by-product of detection grouping is that we can assign a "score" to every face candidate. The score of a candidate is simply the number of individual detections that were consolidated to form it. This metric gives an insight into how face-like a candidate is based solely upon its features. Although a tenet of our approach is to not rely on detection thresholding alone in order to find faces, it would be negligent to discard the information entirely. Therefore, we classify any candidates with a score greater than a very high threshold as faces, as we believe this constitutes a stronger classification than our subsequent processing stages can offer.

Through experimentation, we have found that attempting to establish a globally applicable threshold yields rather inconsistent results. This is a result of variations in image quality, as, for instance, small degrees of blur will reduce the overall number of feature detections, and, consequently, the scores of otherwise-clear faces.

Therefore, we calculate thresholds on an image-by-image basis by using the given maximum candidate score as a measure of image quality. The score threshold for classifying candidates as faces is then simply a fraction of that score. For example, if the maximum score for a given set of candidates is 300, which would indicate an image of good quality, then we would classify any candidate with a score over 30 as a face. For an image of lower quality, a threshold that high would be inappropriate, as almost-certain faces could ultimately be misclassified. The result of applying score-based classification can be seen in Fig. 6(a).

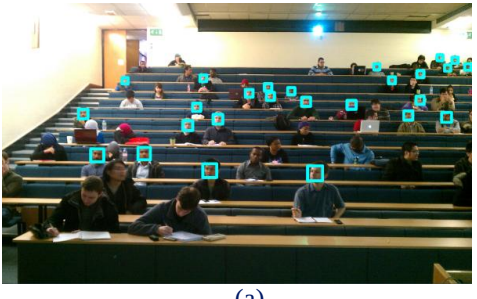
3.6 Colour-based classification

Further processing is required to determine the nature of candidates with a score below the given face threshold, since it is entirely plausible for many of them to still be faces, only without particularly strongly expressed features. Since feature-based techniques have been

used to identify such candidates, but have proven insufficient for classifying them, we must make use of alternative criteria. We have chosen to classify them based upon colour, as it is a property we can relatively simply investigate for any given candidate. We believe that we can classify candidates that exhibit some face-like features and are composed of skin-like colours as faces with a high degree of confidence.

To this end, we adopt our own adaptive skin segmentation system [3]. It uses information acquired by a feature-based face detector to build a unimodal Gaussian model of skin colour for the given image in the normalised rg space. Our previous work has shown that the system can comfortably outperform a wide range of other techniques, and given that the requirement of having high-confidence face data available has already been met by the PCA-based construction of the size filter, it is ideally suited to meet our needs now, without the incurrence of any avoidable computational overheads.

With a colour model built, we examine each candidate individually. Rather than take the entirety of candidate regions into account, we examine circular sub-regions around their centre, with the expectation that if a candidate does indeed represent a face, then any apparent skin is likely to be found within a certain radius of the centre. This factor, combined with the colour modelling approach's tendency towards precision over sensitivity, means that we use a relatively low threshold on the fraction of skin pixels within a given sub-region required for a candidate to be classified as a face of 0.3. The result of this classification process can be seen in Fig. 6(b).



(a)

(b)

Figure 6: Faces found through (a) score-thresholding and (b) colour-based classification.

The combination of the two sets of faces we have produced, through score-based and colour-based classification respectively, constitutes the final result of our face detection process.

4 Results

In order to demonstrate the effectiveness of our system at this time, we will evaluate it against the six individual detection cascades upon which it is built. This is the simplest way for us to prove the value of our approach and quantify the extent to which it outperforms that highly regarded work. We will present the results having run the detectors with their default parameters.

For this analysis, we will be using our own dataset, which is comprised of 18 separate images, exhibiting a total of 803 faces. These images all pertain to the problem we have outlined, and have been captured in different lecture theatre environments. We consider this test data to be extremely rigorous, as many of the faces we include as "true positive"

are very much out of focus (primarily due to their distance from the capturing device, and, in a number of cases, due to momentary movement), or are being severely occluded. We felt it pertinent to include these faces, however, to see just how effectively our system could detect faces under non-ideal circumstances, and to discover in which cases we could succeed where the existing approaches have failed. Our findings can be seen in Table 1.

	CASCADE						OUR SYSTEM
	Default	Alt	Alt2	Alt_Tree	Profile	LBP	
Precision	75.17%	95.03%	91.85%	98.41%	94.32%	86.32%	**95.87%**
Detection rate	55.42%	52.43%	53.30%	38.48%	41.34%	47.95%	**69.36%**
Accuracy	46.84%	51.03%	50.89%	38.24%	40.34%	44.56%	**67.35%**

Table 1: The results achieved over our 18-image dataset.

5 Conclusions

As our results have shown, the system we have developed has yielded results that are far greater than those produced by the work we have based it on. The detection rate of our approach is particularly encouraging, as it is almost 14% better than the nearest competitor for what is a supremely challenging dataset. All the while, we have managed to maintain an extremely high precision of almost 96%, resulting in an overall accuracy that is more than 16% greater than the most effective of the individual cascades.

We regard the superior detection rate of our system as a result of two key factors. Firstly, our approach's broad detection of features gives us the largest possible pool of potential faces to work with. Secondly, we treat any visual feature that could potentially suggest the presence of a face as a candidate until we can say otherwise with some confidence. Some detections that ultimately do not pertain to actual faces can potentially make it to the point of colour classification without being discarded, which gives actual, but weakly expressed, faces maximum opportunity to be classified as such.

We also attribute our low susceptibility to false positive results to the adaptability of our system. Using high-confidence face data from within images to build the size filters and colour models means that we can establish highly specific conditions for candidates to be identified as faces.

With regards to the problem we have outlined, we believe where the alternative detectors we have experimented with have failed lies in their restricted sets of expected facial features, and in their rudimentary method of translating feature detections into actual faces.

Comparative assessment against a wider range of alternative techniques is difficult, largely because of a general lack of standardisation in the evaluation of face detection approaches. However, given the absence of identified solutions to our specific problem, we believe the individual cascades are sufficient for giving us a sense of what existing work in the field is capable of, especially given how successfully they have previously been applied to more common issues. Although, even if there was greater conformity in evaluation, it is unlikely that popular datasets would be exploring the same detection capabilities as the one we have put together does.

It is important to note that we have also tested our system on a wide range of more typical images, but our findings have been less significant than those we have reported

here. Although we have found our approach to still be very effective for simpler tasks, its benefits are far less profound, purely because the alternative detectors themselves perform more effectively.

6 Further work

Despite achieving such encouraging results, there remain several aspects of our system that we intend to improve upon, a couple of which we have alluded to already.

First of all, modelling face distributions according to linear, two-dimensional planes can, in theory, lead to misrepresentation, as they simply may not suit the environment in which an image has been captured. Empirically, however, we have found that face distributions do conform to the models we build in the vast majority of cases, even where simpler, arbitrary images are concerned. Nonetheless, we will attempt to find a more flexible method of distribution modelling to ensure that every function we build is suitable for the set of faces it pertains to.

Furthermore, in its current state, our system automatically discards any candidate with a score of just 1 after grouping. This is regrettable, as we have found on numerous occasions that an actual face can have only a single initial detection associated with it, and yet still a strong enough colour representation to be positively classified. However, the number of false positives we eliminate through discarding single-detection candidates is far greater than the number of faces we typically lose (which we find is especially true where environments with a lot of skin-coloured surfaces are concerned), meaning the decision is beneficial for overall detection accuracy. This is not ideal, however, and we will investigate alternative methods for differentiating such candidates.

The aspect of our approach that we believe is most in need of attention is its computational efficiency. The system currently takes about six times longer to process a given image than any of the individual cascade detectors, which is to be expected when we consider that we are searching for six times as many features. This would be of little consequence when solving any offline problem that prioritises detection accuracy over any other performance factor, but we feel that there is plenty of room for improvement. Streamlining the process and achieving results at much closer to "real-time" would be vital to maximising the potential applicability of our system.

One method by which we could go a long way towards improving efficiency would be to consolidate the features of the individual cascades into a single set. Undoubtedly, there is significant overlap in the features searched for by our system now, which results purely in the duplication of detections and unnecessary time expenditure. In creating a single, broad feature set, we could eliminate these issues entirely, without any loss of potentially useful information, and we would also be making it possible to introduce additional features for detection. This would be greatly beneficial, as we have found during our experimentation that a much greater number of false negatives are a result of no features pertaining to a face being found whatsoever than them being later classified as non-faces.

References

[1] Zhang, C. and Zhang, Z., "A Survey of Recent Advances in Face Detection," Tech. Rep. MSR-TR-2010-66 (2010).

[2] Yang, M.-H., Kriegman, D. J. and Ahuja, N., "Detecting faces in images: A survey," IEEE Trans. Pattern Analysis and Machine Intelligence 24(1), 34-58 (2002).

[3] Taylor, M. J. and Morris, T., "Adaptive skin segmentation via feature-based face detection," Proc. SPIE Photonics Europe 2014 – Real-time Image and Video Processing (9139) 27 (2014).

[4] Viola, P. and Jones, M., J., "Rapid object detection using a boosted cascade of simple features," Proc. Computer Vision and Pattern Recognition 2001, 511-518 (2001).

[5] Viola, P. and Jones, M. J., "Robust Real-Time Face Detection," Int. J. Computer Vision 57(2), 137-154 (2004).

[6] Lienhart, R. and Maydt, J., "An extended set of Haar-like features for rapid object detection," Proc. 2002 Int. Conf. Image Processing, 900-903 (2002).

[7] Mita, T., Kaneko, T. and Hori, O., "Joint Haar-like features for face detection," Proc. 2005 Int. Conf. Computer Vision 2, 1619-1626 (2005).

[8] Fröba, B. And Ernst, A., "Face detection with the modified census transform," 6th Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, 91-96 (2004).

[9] Ojala, T., Pietikäinen, M. and Mäenpää, T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Analysis and Machine Intelligence 24(7), 971-987 (2002).

[10] Burgin, W., Pantofaru, C. and Smart, W. D., "Using depth information to improve face detection," Proc. 6th ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 119-120 (2011).

[11] Nanni, L., Lumini, A., Dominio, F. and Zanuttigh, P., "Effective and precise face detection based on color and depth data," J. Applied Computing and Informatics 10(1-2), 1-13 (2014).

[12] Shieh, M. Y. and Hsieh, T. M., "Fast facial detection by depth map analysis," Mathematical Problems in Engineering 2013, ID: 694321 (2013).

[13] Ruiz-Sarmiento, J. R., Galindo, C. and Gonzalez, J., "Improving Human Face Detection through TOF Cameras for Ambient Intelligence Applications," J. Ambient Intelligence – Software and Applications: Advances in Intelligent and Soft Computing 92, 125-132 (2011).

[14] Seguí, S., Drozdzal, M., Radeva, P. and Vitrià, J., "An Integrated Approach to Contextual Face Detection," Proc. 2nd Int. Conf. Pattern Recognition Applications and Methods, 90-97 (2012).