Improving Performance and Applying Cascades Visual Classification

Master Thesis

Michal Kiwkowitz

Contents

ABSTRACT

Computer-based object classification has improved consistently over the last decade, the performance of current computational schemes is still significantly lower than hu classification performance. In addition, the errors made by current classifiers are c unreasonable by human standards. Since improvements to performance bec increasingly difficult to achieve when absolute performance is already high, it is curre unclear which future directions will be useful for reaching truly human level performa

In this work I examined two general directions for future improvements in cur classification scheme. One general direction assumes that current methods extract or part of the available information for classification from the training set, and attemp identify the main possible sources of additional information. I examined a numbe plausible sources for possible improvements. Some of the methods under study, bu concluded that they are unlikely to be sufficient by themselves to reach human-l performance.

The second general direction assumes that a major limitation comes from the size o training set: training sets used in practice may be inherently insufficient to capture al necessary variations needed to learn a truly high-performance classifier. Our experim suggested that classification performance indeed increase monotonically and wit clear saturation as the training set increases in size. We also found that the set of feat used for classification needs to be increased to capture the additional information in increased training set.

These results have two implications to the classification scheme. First, they raise advantage of constructing classifiers in an on-line manner, in which the classifie continuously improved by new examples. Second, they raise the problem of increa computational load as the number of features used for classification increases.

To deal with these problems we developed and tested two classification schemes. Both multi-stage, or 'cascade' methods in which all images are analyzed in the first stage, depending on the results, only some of them are analyzed further by subsequent sta The first of these methods was a so-called 3-way cascade, which is an extensio previously used cascade methods. In this method, a first-level classifier is first applie the new image to be classified. If the response is high or low enough a decision is m otherwise, a second-level classifier, which uses more features, is applied to the imag allow a more confident response. The second multi-stage scheme we developed is the called configuration cascade. In this approach, the decision to process an imag additional stages is based on the particular feature configuration discovered in the im Using this scheme, erroneous configurations of the first-level classifier are processed second-level classifier which uses more features and achieves better performance. showed that the number of possible erroneous configurations is bounded in practice, that the error on the erroneous configurations is reduced after additional processing. scheme leads to a continuous improvement in performance, with only a small increas computational cost.

1 Introduction

The field of visual recognition has been making continues progress in the are classification and recognition. However, up to date, no classifier for a real complex c has achieved classification without errors or fully human-level performance. In this v we examine the sources of these errors, and possible methods to improve perform without a large increase in computational cost.

We start by investigating the type of errors made, using as an example the fragment b approach [2], to classify faces and non faces, and evaluate in Section 2 the classif performance and the errors it makes. We then survey existing classifiers and advantages and drawbacks, in Section 3 . In Section 4, we examine different ideas improving the performance of the fragment based classifier. We continue by identif the training set size as an inherent boundary to the performance, and the implicatior this observation on classifier's computational cost in Section 4.2 . We suggest in Secti two methods of cascades, which can improve performance with relatively low additi computational cost. The first, a three-way cascade described in Section 5.1 , w performs further processing on images near the margin, and the second, a configura cascade described in Section 5.2 which uses the configuration of the detected featur determine whether an image requires additional processing in the cascade.

In the next section we examine more closely the performance of current classifiers and goals and motivations for the current work.

2 Evaluating Classification

The goals of our work are to answer the following questions:

- Can the performance of current state-of-the-art visual classifiers be consid satisfactory?
- Are we learning all we can learn from the training data?
- What are possible ways of improving performance further?

In this section we survey methods of evaluating classification performance. We consider the performance for practical applications and examine the amount of e

made by machine classifiers compared with human performance. The errors ca generally divided into two types: false negative errors, images that have been reje falsely, and false positive errors, images in which the class has been falsely detected. begin with an overview of the methods used to evaluate a classifier.

2.1 Evaluation Methods

Judging the performance of a classifier is not a trivial question. Many authors plot compare the Receiver Operating Characteristic (ROC) curve (seen in Figure 1) of diff classifiers, and although error rates may be small and satisfactory, the error insta themselves are quite alarming, Figure 2 and Figure 3 show some examples. This g enables understanding of the relationship between the different errors made by classifier. A threshold on the response of the classifier determines a desired hit rate respectively a false positives rate that depends on the classifier's ROC curve. In an i world we would be able to set the threshold so that our classifier will get 100% hits 0% false positives, a perfect classifier with no errors. In reality, there is no pe threshold, and the possibility is to select some acceptable mixture of false positives false negatives.

The ROC, although very useful, is not a simple parameter to compare. A measure th easier to compare is the Equal Error Rate (EER) of a classifier. This is the location o ROC curve where the percent of false positives equals the percent of false negatives. perfect classifier the value of the EER will be zero and therefore there will be a thres that allows the classifier to make a perfect decision every time. A similar measure is Minimum Error which provides the minimum total error (false positives and negatives). The EER and the Minimum Error are usually found at close locations o ROC curve. Another way to appraise the ROC is to measure the percent of the area b the ROC curve; the perfect ROC will have a value of one.

For most real world classification problems, there are no perfect classifiers; n classifiers improve the classification by improving the ROC curve at the top left corne many cases the changes are very subtle and the comparison of EER, area or even different ROC curve misses a lot of information about the classifier. Given two classi with the same ROC or very close ROC how can they be compared? Consider a vendor

is purchasing a classifier that uses images to alert him when a police car passes by property. He is offered two classifiers with very good ROC curves but imperfect c Shouldn't he consider the type of false positives made by these as an additional metho rating their performance? Wouldn't a classifier that detects a cab as a police car e once in a while be better than one that detects a cat or just the wind in the trees? Simil when a child calls a cow a horse, his parents will happily just let him know that l wrong, but shouldn't they pay a little more attention if their child is using the word when he encounters shoes or when staring at a clear wall? The error instances themse are important and there is much to learn from them. We next examine the EER anc type of errors made by current classifiers.

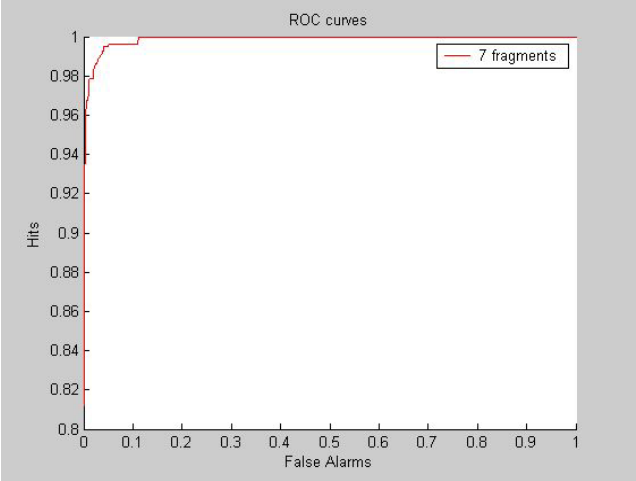

Figure 1: Classifier Performance. The Reciever Operating Characteristic curve of a classifier using seven fragments as described in [3]. Although the ROC is good, imperfect.

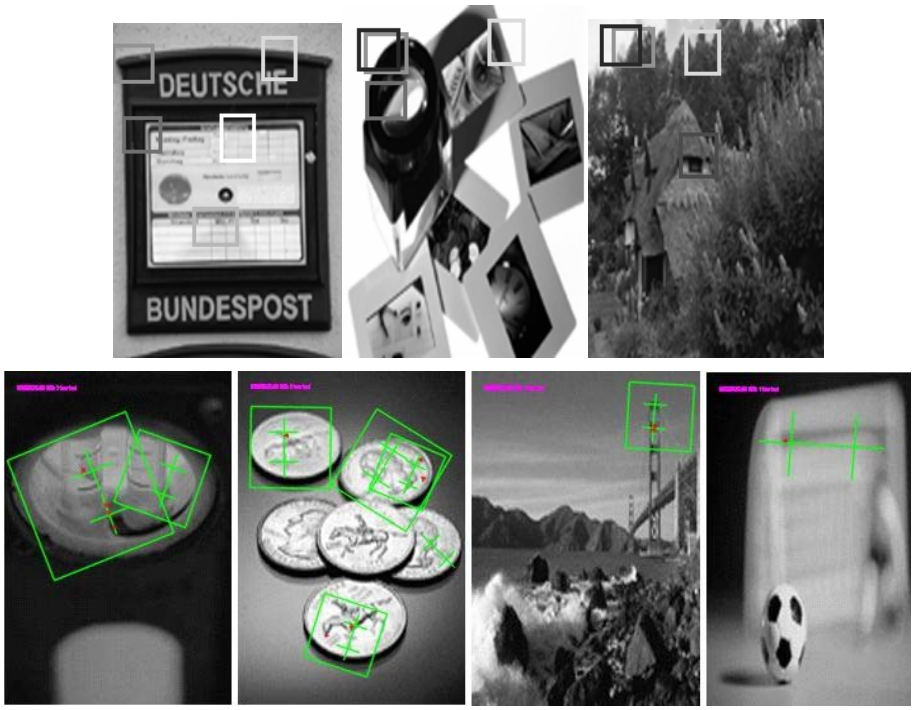

Figure 2: Faces Found by Face Classifiers. Images detected as faces by two diffe face classifiers with high performance. On the top row are results from a fragı classifier as in [3], the fragments detected are marked. On the bottom row are false f detected by the Betaface classifier [8]. This classifier detects the eyes in an image marks them in green.

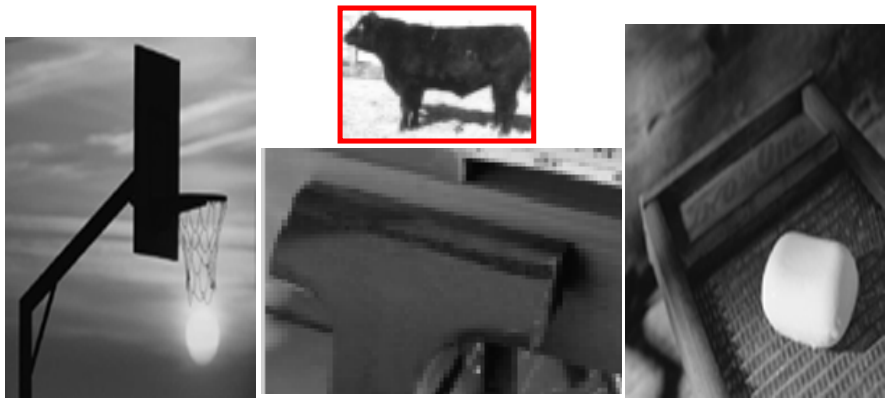

Figure 3: Cows Found using SIFT. These images recived very high scores and v classified as cows by a classifier that uses SIFT features [7], we show a sample cow ir red rectangle.

Figure 4: **Classifier Performance on Specific Images**. On the top row are some misse
the classifier, in the center are some false alarms or false positives. The bottom row sh
some images that we believe may serve as more reasonable false alarms.

2.2 Practical Application of Classification

Although the equal error rate in some of the best classifiers can reach 2-6 %, this is stil
quite high. Consider searching a natural scene and trying to determine if there is a hors
the image. Suppose that the image is 1000 pixels high by 1000 pixels wide, and we exp
the horse object to be found in a window of around 50 by 50 pixels. Searching the ima
serially with local windows with a 50% overlap, we will need to check 40x40 location
which is 1600 sub windows. With an error rate of 2%, this means that 32 sub windows
will be falsely detected as horses, which is not very practical. In the next section we
contrast human performance with the performance of machine classifiers.

2.3 Machine Classification vs. Human Classification

If humans had the false alarm rate mentioned in the previous section, they would horses everywhere, or in other words they would be prone to mistakes and vi hallucinations. Compared with State-of-the-art classifier's performance, humans perf better: for most databases or classification tasks there is no question whether mac performance is better or poorer than that of humans, humans outperform cu classifiers. When comparing machine performance to human performance, the tas classification is harder on machines, given that the in-class variability is greater. Mac software is closer to human performance in somewhat simpler tasks like recognition, are controlled for rotation and lighting. Not many real comparisons have been n between human and machine categorization, perhaps because it is common knowl that humans are better. A comparison study by Adler and Schuckers (2006) comp automatic face recognition technologies available in 1999, 2001, 2003, 2005 and 2 with human performance and showed that as automatic face detection improves, average human performance is no longer better than the machine performance. Th impressive, yet we need to consider that some human performance is nevertheless b than the best automatic face recognition, and the task at hand is answering 'sam 'different' to pair of images from the NIST Mugshots Identification Database, where match is between images of the same person at different age points which may cause l facial differences, see Figure 5 and [1] for more details. The field of face recognition classification is one of the better studied classes, on other classes (planes, motorb dogs, etc.) performance is usually poorer. As mentioned in the previous section, hur do not make as many false positive errors as machine classifiers; this fact is importan understanding the human visual system and for improving machine classification. In next section we take a closer view at some classification error dividing them into negative and false positive errors.

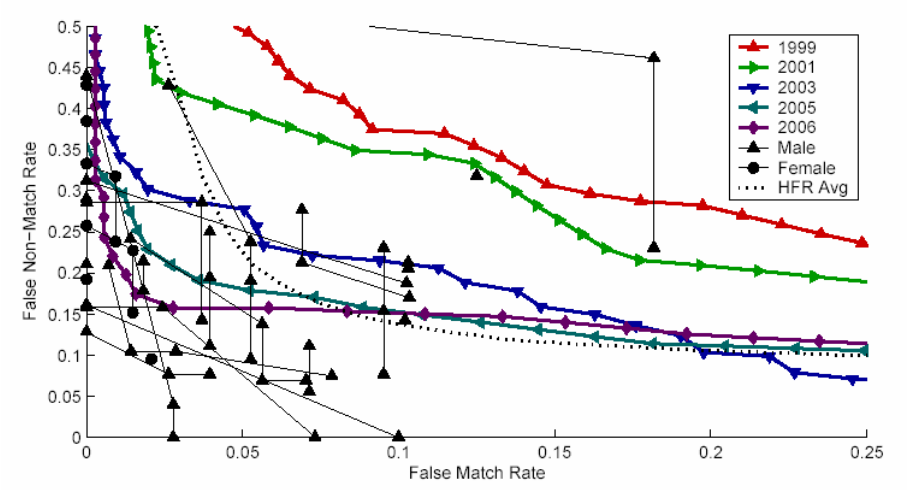

Figure 5: Comparison of face recognition software and human performance from The black circles and triangles indicate the human female and male perform respectively, and because they could respond on a scale of 1-5 conveying their confid in the match, an curve was drawn for each human by adjusting the threshold on confidence, the continues curves show the results for the highest performing soft available to the authors of [1] in the specified years, and the dotted line is an averag the human performance, that is highly affected by the outliers. Most human perform remained at the bottom left corner with the lowest error rates, and since where the hu curves do not completely outperform an automatic curve, the authors counted it a intermidiate performance (33.3% comparing to the 2006 condition), they got a good s for the 2006 comparison. This may be challanged if the statistics where different.

2.4 False Negative Errors

When we examine the false negatives (misses) that current face classifiers make we often understand them (see Figure 4): many times misses occur in challenging face im with long hair, high foreheads, glasses, or in images that contain non frontal faces, st shadows or noise. This seems intuitive and is consistent with human performance: inf are known to sometimes burst in tears when seeing a familiar kin that has grown a bea put on new glasses. We can assume that large changes in a face that were not seen be may not be encoded as part of the stored features and may cause unfamiliarity. T errors are usually understandable; furthermore, algorithms have been suggested to re the rate of false negatives by adding more training samples, selecting more feature boosting the classifier [13]. The false positive errors, discussed in the next section more interesting.

2.5 False Positive Errors

False positives errors are those images that the classifier falsely detects as class but are not. They can be divided into two main types of errors, which I call multi-class e and single-class errors. As multi-class errors we consider those errors that are cause confusion between similar classes. For example, when trying to classify horses, we recognize some cows as horses. Since both classes are four legged animals, we con these errors reasonable; extending the classifier into a multi class classifier to learn similar class, or adding the similar class to the training samples labeled as non-class w resolve this problem. Similar errors are also made by human infants before they lear name of the new class.

The single-class errors are those images that do not resemble the class and are neverth classified as the class; see Figure 4 for some examples. Although the classifier shown a fairly good ROC curve, it makes unreasonable mistakes that a two-year old would make, failing to see for instance that a register is not a face. This problem of unreason false alarm (compared with human judgment) is not easily avoided, it persists in n different types of classifiers and when using different kind of features. Using a N Bayes classifier that extracts SIFT features as in [7] the problem persists, examples their likelihood scores can be seen in Figure 6. In Figure 2 we show another algori Betaface [8] that detects faces by finding the eye region, and fails as well when prov with queries of some non-class images that are hard for the fragments classifier.

The single-class false positives are interesting since it is clear they are unreasonable.

We continue to discuss some popular classification methods, their relative advantages shortcomings in section 3 . In Section 4 we use our observations to suggest and com possible general methods for potential improvements in performance of the classifiers.

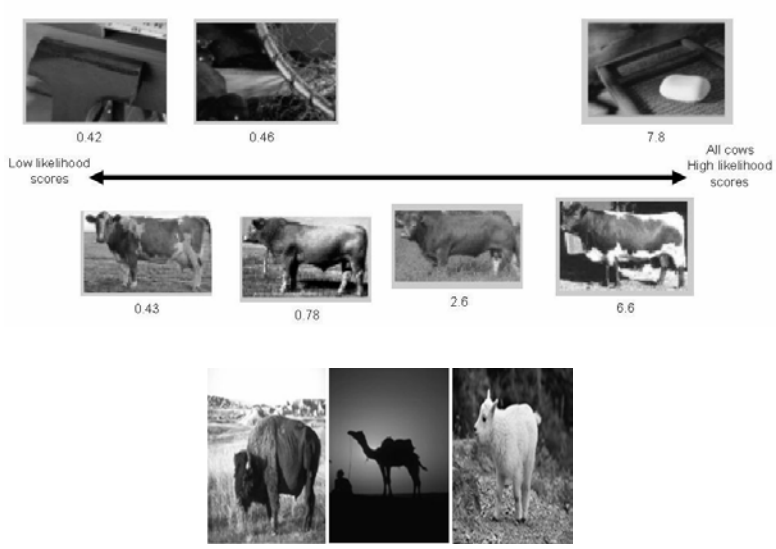

Figure 6: Cow Database Likelihood. On the top some examples of the likelihood sc produced by a cow classifier based on SIFT features [7]. The false positives are a irrelevant to the class of cows. On the bottom some examples of images that we cons more relevant as false positives**.**

3 Existing Classification Methods

In this section I survey the main relevant classification methods, the features they use, how they combine features to reach a decision. The goal is to identify possible reason failure and possible methods for improvement.

Classification is the process of determining a label Y_i for each sample X_i for s unknown function $F\ X_i \quad Y_i$