A Modular Approach to Training Cascades of Boosted Ensembles

Teo Susnjak, Andre L. Barczak, and Ken A. Hawick

Institute of Information and Mathematical Sciences,
Massey University, Albany, New Zealand
teo.susnjak.1@uni.massey.ac.nz
http://iims.massey.ac.nz

Abstract. Building on the ideas of Viola-Jones [1] we present a framework for training cascades of boosted ensembles (CoBE) which introduces further modularity and tractability to the training process. It addresses the challenges faced by CoBE frameworks such as protracted runtimes, slow layer convergences and classifier optimization. The framework possesses the ability to bootstrap positive samples and may in turn be extended into the domain of incremental learning. This paper aims to address our framework's susceptibility to overfitting with possible solutions. Experiments are conducted on face detectors using the bootstrapping of large positive datasets and their accuracy, with respect to overfitting, is examined.

Keywords: cascades of boosted ensembles, AdaBoost, classification, classifier training, face detection.

1 Introduction

Face detection has received much attention in recent years in the field of computer vision. Though a number of notable face detectors with accurate and fast execution runtimes in controlled environments have been developed, the problem of developing robust face detectors that operate in variable environments is still an open problem.

The most successful methods so far have been extensions of the seminal work by Viola-Jones [1], which combined AdaBoost as the learning algorithm together with Haar-like features that can be computed rapidly through integral images. The key feature of this detector was the decomposition of a monolithic ensemble of boosted weak classifiers into cascades.

Despite the successes achieved using cascades of boosted ensembles in both accuracy and real-time performance, one of the greatest obstacles to their wider proliferation when deployed in face detection or similarly computationally intensive domains, lies in their protracted training runtimes [2]. Though massive feature spaces are an obvious contributing factor, particularly as dataset sizes increase [3], other factors are slow training convergences [1] and limited classifier optimization capabilities [4]. Additionally, the lack of positive sample boostrapping capabilities of CoBEs has meant that all positive samples needed to be

E.R. Hancock et al. (Eds.): SSPR & SPR 2010, LNCS 6218, pp. 640–649, 2010.

learned simultaneously, thus prohibiting the usage of massive positive datasets. Lastly, the limited abilities of the CoBE frameworks to learn incrementally also leads to significant total training runtime overheads in instances where it is not feasible, requiring the re-training of entire classifiers each time new and *relevant* datasets become available.

[3] minimize the problem of massive feature spaces by applying statistical methods and assumptions to it regarding its distribution and achieve a dramatic reduction in the amount of time required to train each weak classifier while [5] employs feature filtering. [6] attempted to accelerate the cascade layer convergence speed by strengthening the discriminatory ability of the feature types. Alternatively, [7] and others have modified the AdaBoost learning algorithm to produce variants with same intentions, however none have significantly contributed to a training runtime reduction in respect to faster layer convergences. Automating the optimization of cascade parameters remains an unsolved problem though [4] provided significant contributions.

Only recently has research [8,9] surfaced with methods to enable positive sample bootstrapping. While, [10] introduces on-line incremental learning using AdaBoost implemented using neural networks rather than CoBEs.

The PSL (*Parallel Strong* classifier within the same *Layer*) training framework introduced by Barczak et al [11] originally sought to address the convergence bottleneck during the training of cascade layers. However, the modularity of the approach also simplified cascade optimization. Moreover, it provided the basis for addressing the issue of bootstrapping positive samples, seen in initial experiments on the Bootstrapped Dual-Cascaded framework (BDC) [12], as well as for further extensions that enable incremental learning.

The shortcomings of the PSL-based frameworks, have been an elevation in false detection rates due to a tendency to overfit. This characteristic has been more evident in rare-event domains like face detection where exceptionally low false positive rates are needed in order to produce practical detectors.

The purpose of this paper is to explore the causes of overfitting in PSL-based frameworks and to present modifications to them which preserve their ability to rapidly train real-time execution-capable classifiers. In order to provide a thorough analysis of the overfitting issue, this paper will make use of the face detection classifiers from [12], which were created using the positive sample bootstrapping method (BDC) and will compare them with classifiers trained using the modified BDC framework designed to address the overfitting.

The structure for this paper is as follows: Section 2 sets forth the fundamental ideas of modularizing CoBE training using the PSL-based method. Section 3 discusses extensions to PSL which led to the development of the BDC framework that enabled positive sample bootstrapping. The same section explores the framework's ability to implement incremental learning. Following sections present the analysis of the occurrence of overfitting in these frameworks and propose a solution to it. Subsequent sections explain the implementation of the experiments followed by their analysis and a conclusion.

2 PSL Training Framework

The architecture of the PSL framework can be seen in Figure 1b and is contrasted with the standard cascading approach of Viola-Jones in Figure 1a. The PSL architecture extends the standard cascading structure by introducing an additional nested cascade within each layer of a strong classifier, thus creating a quasi two dimensional cascade structure. While the Viola-Jones approach executes an independent round of AdaBoost training for each layer, the PSL framework executes multiple independent rounds of AdaBoost within each layer and in the process constructs a complementing cascade with an alternate goal. We refer to each layer of an internal cascade as an *intra-layer stage*.

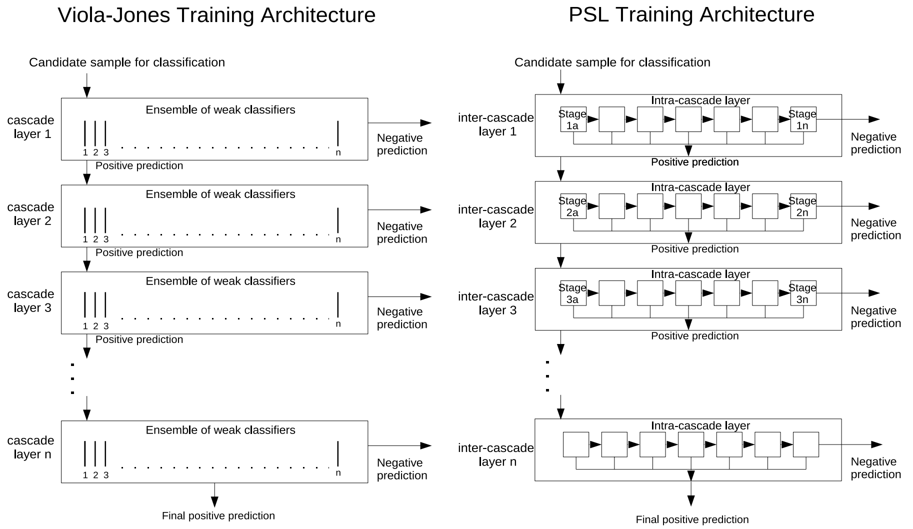

Fig. 1. a) The standard cascade structure of Viola-Jones. b) the PSL structure [11].

Whereas the cascading of the Viola-Jones method focuses on rejecting negative training samples, the intra-layer cascading of the PSL framework focuses on correctly predicting positive samples. Thus, the underlying principle found in the Viola-Jones method with respect to its approach to handling more difficult negative samples with each succeeding layer, is replicated to the positive samples in the internal stage-to-stage propagation. The propagation of the positive training samples of the PSL framework is seen in Figure 2a. As the intra-layer cascade of stages is constructed, correctly predicted positive samples are removed from succeeding stages while the misclassified positives are retained until all the positive samples have been correctly predicted. By removing correctly predicted positives, faster layer convergences are realized, while 100% hit rates are attained without artificial threshold adjustments, thus ultimately resulting in accelerated overall training runtimes.

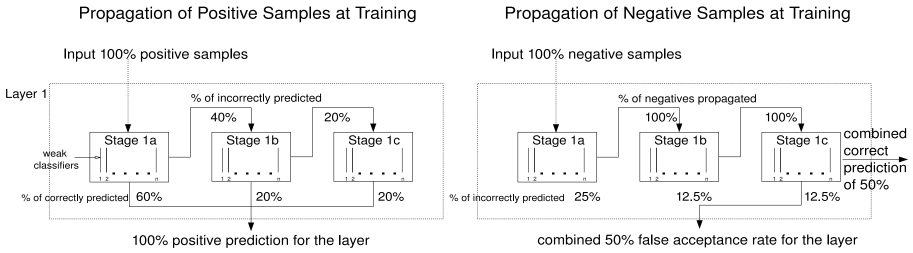
Propagation of Positive Samples at Training
Input 100% positive samples
Layer 1
% of incorrectly predicted
40%
20%
Stage 1a
Stage 1b
Stage 1c
weak classifiers
1 2 n
% of correctly predicted
60%
20%
20%
100% positive prediction for the layer
Propagation of Negative Samples at Training
Input 100% negative samples
% of negatives propagated
100%
100%
Stage 1a
Stage 1b
Stage 1c
combined correct prediction of 50%
% of incorrectly predicted
25%
12.5%
12.5%
combined 50% false acceptance rate for the layer

Fig. 2. a) The propagation of positive training samples within the cascade of PSL stages inside a layer. 1 b) The usage of negative training samples within the cascade of PSL stages inside a layer.

During training, all negative training samples propagate to each stage irrespective of how successfully previous stages have learned to predict them as seen in Figure 2b. Each stage is assigned a target to learn to reject 50% of the negative samples and to achieve a 100% hit rate. However, a key constraint in the form of a maximum number of weak classifiers is added to each stage which not only accelerates layer convergences but also simplifies classifier optimization through the variation in size of this constraint at different layers.

At detection time, the classification process also becomes modularized and more efficient. A candidate sample is predicted as a negative by a layer only if *all* nested stages within it classify it as a negative. A sample is predicted as a positive once *any* nested stage predicts it as a positive thereby not requiring the computation of the remaining internal stages.

3 Positive Sample Bootstrapping

The BDC framework builds upon the concepts of the PSL structure and extends it in order to implement a positive sample bootstrap capability. Unlike the positive sample bootstrapping approaches of [8,9], the BDC training framework utilizes the modularity offered by the PSL's nested cascade-of-stages to achieve further malleability. Through a strategy of divide and conquer, massive positive datasets can be employed while only a fraction of its samples undergo training at each stage.

The whole negative dataset and a subset of the entire positive dataset constitute the training sets used for each stage of a BDC nested cascade. The positive sample subset which the learning algorithm *sees* and trains on explicitly we call the *base set*. The entire positive dataset from which new positive samples are bootstrapped is referred to as the *reserve set*.

The procedure for intra-layer cascade training can be seen in Algorithm 1. The training of an intra-layer cascade initiates with randomly selecting a comparably small subset of positive samples from the reserve set in order to construct the base set. The base set is then trained against the negative dataset to produce

```
Given:
$C_n = n_{th}$ inter-layer layer sub-classifier
$S_i = i_{th}$ intra-layer stage sub-classifier
$PB_i$ = positive base set used on $S_i$
$PR$ = positive reserve set
$f_{min}$ = minimum false acceptance rate
$d_{min}$ = minimum required hit rate set at 100%
$WK_{max}$ = max number of weak classifiers
1 randomly select positive samples from $PR$ to create $PB_i$
2 train $C_nS_i$ against $PB_i$ until $f_{min}$ and $d_{min}$ or $WK_{max}$
3 validate $PR$ using $C_nS_i$ and remove from it correctly classified samples
4 if all samples in $PR$ have been correctly predicted then start a new layer
  $C_{n+1}$ otherwise start new stage $S_{i+1}$ repeat step 1
```

Algorithm 1. BDC bootstrapping method for each cascade layer

individual stages. Each stage of this nested cascade is trained with a target hit rate of 100% and a high rejection rate. As in PSL, the size of each nested stage is restricted by the maximum number of weak classifiers that can comprise it. Once this maximum number has been reached, the training for that stage ceases and a new intra-layer stage begins. The positive bootstrapping procedure is then initiated. The positive samples in the reserve set are validated against the resulting stage classifier and all correctly predicted samples are removed from training subsequent nested stages. The remaining positive samples are randomly selected to comprise the new *base set* for the next intra-layer stage together with all the incorrectly predicted positive samples from the previous stage's base set.

3.1 Incremental Learning with PSL

The modular nature of the PSL framework, combined with the ideas from BDC leads to the possibility of implementing effective incremental learning in a novel approach. Incremental learning can be achieved in this scenario by constructing additional intra-layer stages trained on new positive samples which are incorrectly predicted at each layer. The new stages can then either be appended to the existing cascade-of-stages or a strategy can be devised to replace less accurate existing stages with new ones. The incremental training would be initiated on batches of incorrectly predicted positive samples once they reach the minimum required number for each *base set.* The composition of the negative set is less trivial and has to consist of similar patterns which previous stages in a layer have learned to predict otherwise false detection rates for a layer would increase. It is proposed that the negative set comprises of those images which have up until that cascade layer been misclassified and that a substantially larger negative dataset be used for incremental learning than that of the initial off-line training phase.

3.2 PSL Framework and Overfitting

Experiments in [12] have demonstrated the capability of the BDC framework to potentially train classifiers on massive positive datasets with relatively small increases in training runtimes whilst maintaining 100% layer hit rates on the training data. The face detectors trained in those experiments showed that the training runtimes using the BDC bootstrapping method result in a fraction of the runtimes required by standard training structures without bootstrapping. However, the framework also exhibited a susceptibility to elevate false acceptance rates which makes it less suitable for rare-event operating domains like face detection.

Further analysis of the classifiers obtained in [12] has identified varying degrees of overfitting occurring in final intra-layer stages. The nature of the BDC training approach delays training most difficult positive samples until the last stages. These stages often tend to be trained on positive datasets that comprise of a small number of samples which mostly contain highly unrepresentative patterns in respect to the overall positive dataset. Figure 3 shows examples of images trained by first intra-layer stages and contrasted with those learned in latter stages. The figures point to large concentrations of positive images in final stages which exhibit extensive variations in illumination and also occlusions of vital facial features.

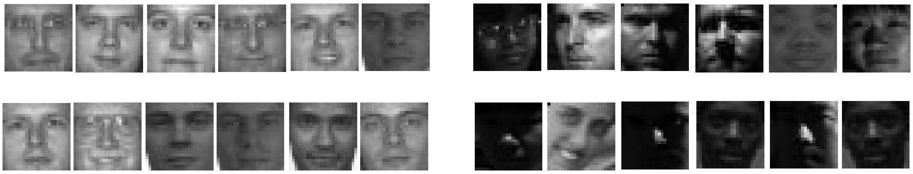

Fig. 3. Examples of positive images learned at different points within the cascade-of-stages on a 15000 sample BDC classifier. Cluster a) stage 1 layer 30 b) last stage layer 30 c) stage 1 layer 42 d) last stage layer 42.

The accuracy of training that takes place in the trailing stages is further compromised by high weights assigned to the positive samples initially before each round of boosting. This occurs since an even 50%-50% distribution of total weights is shared between the positive and negative training sets irrespective of their sizes. Consequently, as the number of stages in each layer grows, fewer positive samples remain. This leads to a proportional increase in their weights, while at the same time, their patterns also become less representative of the whole dataset.

In order to demonstrate the effects of the final stages on classifiers' accuracy, we compared the generalization patterns of classifiers, with and without their last stages, using receiver operating curves (ROC) in Figures 4a-c. The data shows improved generalization of truncated classifiers particularly for segments of the ROC graphs which portray the lower end of false acceptance rates.

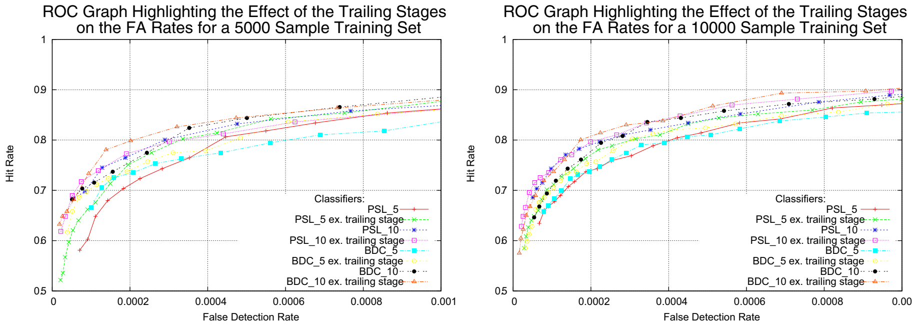

Fig. 4. ROC graphs displaying the generalization patterns of the BDC classifiers with their final intra-layer stages excluded

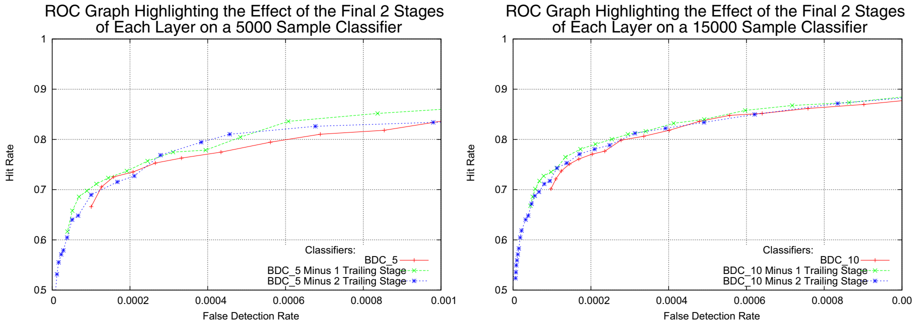

Fig. 5. ROC graphs displaying the generalization patterns of a 5000 and 15000 BDC classifiers with their final two intra-layer stages excluded

The ROC graphs in Figures 5a-b go a step further and demonstrate the effects of excluding the last two stages of each cascade. In both instances, an improvement in the generalization of the truncated classifiers is observed indicating that a degree of overfitting is occurring.

It can be concluded that the effectiveness of each cascade layer is only as strong as the accuracy of its weakest stage. Overall, the generalization ability of a BDC classifier can be summarized as being only as strong as the combined accuracy of all its weakest stages from each layer.

3.3 Anti-overfitting Modifications

Our proposed solution to the problem of overfitting found in the underlying foundation of the BDC structure, is based on incorporating additional positive samples into the datasets of the trailing stages of each layer. With this strategy, our intent is to offset the overfitting brought on by a high concentration of less representative positives. We propose augmenting the trailing stages of each layer with positive samples which have already been correctly predicted by previous stages. By including these samples into the dataset a degree of protection against overfitting is expected to be achieved and thus the likelihood of producing more generalizable intra-layer stages.

The inclusion of redundant positive samples is also expected to have negative effects. The learning process will become more complicated since the convergence speed of layer targets towards required 100% hit rates will decrease and a degree of weak classifier redundancy is likely to be introduced. In order to assist rapid layer convergences, greater weights are initially assigned to *relevant* positive samples at the beginning of each boosting round. Additionally, to counterbalance the generation of an exceeding number of intra-layer stages, the requirement to maintain fixed stage sizes is removed. Instead, the maximum number of weak classifiers is increased as the number of the misclassified positive samples, in respect to the size of the base set, decreases. Since generating a greater number of weak classifiers on a small base set can itself result in overfitting, we also increased the base sets from 500 in prior experiments [12] to 2000 positive samples.

4 Method

The experiments consisted of training face detection classifiers using the modified BDC structure and comparing it to the classifiers trained by the original *naive* BDC structure in [12]. The datasets used on all training were identical as were the parameters. The total of 15000 facial images were collected from various publicly available datasets; FERET, Yale *Face Database B* [13] and the face database from the Vision Group of Essex University. Three main groups of classifiers were trained which were divided into 5000, 10000 and 15000 sample datasets. For each dataset, a classifier was trained with a flexible stage size of 10 weak classifiers. All classifiers were trained with a base set size of 2000 positives against 2000 negatives extracted from a total of 2500 images which generated millions of negative sub-windows. An additional set of classifiers using the naive BDC were trained on base sets of 2000 positive samples in order to isolate the proposed increase in base sets as the determining factor in addressing overfitting. Finally, classifiers were trained to attain a 0% training error in a maximum of 100 layers, using no more than 50 stages per layer.

Testing was performed on the CMU MIT image dataset containing 130 images which contain 506 positive face images.

5 Results

The BDC classifiers with overfitting adjustments generated training runtimes that were 15%-20% longer than those of the naive BDC, however they were still significantly lower than those of the PSL and Viola-Jones. Classifiers trained on naive BDC with base sets of 2000 produced shortest training runtimes, thus highlighting a modest additional cost involved in our approach.

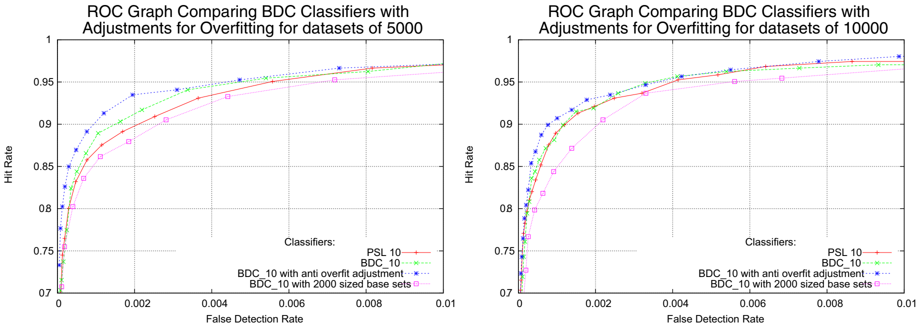

Fig. 6. ROC graphs displaying the generalization patterns of the modified BDC structure

Additionally, the size of the modified BDC classifiers increased over the naive implementation by 15%-20% extra weak classifiers which is likely to incur a larger detection runtime cost too. Both structures generated similar numbers of stages per layer, which ranged from three in earlier layers, through to six as training became more difficult.

Figures 6a-b show the generalization patterns of the classifiers. In both figures, it is evident that the modified BDC classifiers have achieved a superior generalization over all other classifiers on the CMU MIT test dataset. It is worth noting that the weakest accuracy was exhibited by the naive BDC classifiers trained on base sets of 2000 positive samples. This eliminates the possibility of attributing improvements in accuracy of the modified BDC to solely its increase in base set sizes, but instead demonstrates that the solution to overfitting was the result of a combined new strategy.

6 Conclusion

In this paper we demonstrated how classifier training using CoBEs can be modularized using the PSL framework, thereby addressing issues of slow convergence rates and protracted training runtimes, while eliminating many of the post-training classifier optimization overheads. The framework's ability to implement positive sample bootstrapping on large datasets was put forward and its potential to enable incremental learning was also introduced. A thorough analysis of the framework's susceptibility to overfit data was presented, to which an effective solution was proposed.

Future research will focus on extending PSL to enable incremental learning.

References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001, Kauai, HI, December 2001, vol. I, pp. 511–518. IEEE, Los Alamitos (2001)

2. Brubaker, S.C., Mullin, M.D., Rehg, J.M.: Towards optimal training of cascaded detectors. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 325–337. Springer, Heidelberg (2006)
3. Pham, M.T., Cham, T.J.: Fast training and selection of haar features using statistics in boosting-based face detection. In: IEEE 11th International Conference on Computer Vision, 2007, October 2007, pp. 1–7 (2007)
4. Brubaker, S.C., Wu, J., Sun, J., Mullin, M.D., Rehg, J.M.: On the design of cascades of boosted ensembles for face detection. Int. J. Comput. Vision 77(1-3), 65–86 (2008)
5. Wu, J., Rehg, J.M., Mullin, M.D.: Learning a rare event detection cascade by direct feature selection. In: NIPS Advances in Neural Information Processing Systems 2003, Vancouver, Canada (2003)
6. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: ICIP 2002, Rochester, NY, September 2002, vol. I, pp. 900–903 (2002)
7. Viola, P.A., Jones, M.J.: Fast and robust classification using asymmetric adaboost and a detector cascade. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) NIPS, pp. 1311–1318 (2001)
8. Xiao, R., Zhu, H., Sun, H., Tang, X.: Dynamic cascades for face detection. In: IEEE 11th International Conference on Computer Vision 2007, October 2007, pp. 1–8 (2007)
9. Yan, S., Shan, S., Chen, X., Gao, W., Chen, J.: Matrix-Structural Learning (MSL) of cascaded classifier from enormous training set (2007)
10. Huang, C., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Incremental learning of boosted face detector. In: ICCV, pp. 1–8. IEEE, Los Alamitos (2007)
11. Barczak, A.L.C., Johnson, M.J., Messom, C.H.: Empirical evaluation of a new structure for adaboost. In: SAC 2008, pp. 1764–1765. ACM, New York (2008)
12. Susnjak, T., Barczak, A.L.C., Hawick, K.A.: A novel bootstrapping method for positive datasets in cascades of boosted ensembles. Research Letters in the Information and Mathematical Sciences Vol. 14, pp.17-24, Institute of Information and Mathematical Sciences, Massey University Albany (2010) ISSN 1175-2777
13. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intelligence 23(6), 643–660 (2001)