Inter-stage Feature Propagation in Cascade Building with AdaBoost

Jan Šochman Jiří Matas
Centre for Machine Perception, Dept. of Cybernetics, Faculty of Elec. Eng.
Czech Technical University in Prague, Karlovo nám. 13, 121 35 Prague, Czech Rep.
{sochmj1,matas}@cmp.felk.cvut.cz

Abstract

A modification of the cascaded detector with the AdaBoost trained stage classifiers is proposed and brought to bear on the face detection problem. The cascaded detector is a sequential classifier with the ability of early rejection of easy samples. Each decision in the sequence is made by a separately trained classifier, a stage classifier. In proposed modification the features from one stage of training are propagated to the next stage classifier. The proposed intra-stage feature propagation is shown to be greedily optimal, does not increase computational complexity of the stage classifier and leads to shorter stage classifiers and accordingly to faster detectors.

A cascaded face detector is built with the intra-stage feature propagation and is compared with the Viola and Jones approach. The same detection and false positive rates are achieved with a detector that is 25 % faster and consists of only two thirds of the weak classifiers needed for a cascade trained by the Viola and Jones approach. The latter property facilitates hardware implementation, the former opens scope for the increase in the search space, e.g. the range of scales at which faces are sought.

1. Introduction

The AdaBoost algorithm [1] has become a very popular pattern recognition technique. In computer vision, Viola and Jones proposed a method combining advantages of the AdaBoost algorithm and the cascaded decision making to build a face detector [5].

In the cascaded decision making the cascaded classifier is composed of several stage classifiers. Evaluation of the cascaded classifier is sequential. When the current stage classifier rejects a hypothesis, the decision process is terminated and the object is marked as a non-face. Otherwise, the next stage classifier is run. An object is declared a face if it is accepted by all stage classifiers in the cascade. In the cascaded classifier training a new training set is constructed after each stage classifier training. The examples already rejected by the cascaded classifier are removed from the training set and new still non-decided examples are added.

The advantage of the cascaded classifier is twofold. First, unlike a complex monolithic detector, the cascaded detector can be evaluated in real-time. This is a consequence of the highly desirable property that only a few relatively simple classifiers of the first stages are evaluated on average and the complete complex cascaded classifier is evaluated only on rare difficult examples. Second, during training, a much larger training sets can be explored. Face detection is an example of a class of problems where one of the classes is extremely diverse (here the non-face class) and can not be represented by a small number of training examples. In this setup, the cascade training works as an efficient non-face class pruning method and a relevant example selector.

In this paper, a novel interpretation of the cascade training is proposed. In the Viola and Jones approach, each stage of the cascade classifier is trained from scratch. Each stage classifier training is treated as an independent learning problem. The partially trained cascade classifier under construction is used to generate new training data for next stage training and the effort already put in finding the classification boundary is abandoned. Unlike in the Viola and Jones approach, in the proposed approach, a new stage training is seen as search for a more precise approximation of the decision boundary. This novel view can be naturally integrated into the AdaBoost learning of each cascade stage by an inter-stage feature propagation.

The face detection problem is used to experimentally verify a hypothesis that inter-stage feature propagation leads to shorter stage classifiers and consequently to faster face detector.

The main contribution of the paper is a proposal of a novel view of the cascade building process of Viola and Jones. An inter-stage feature propagation is propounded for the cascade building with the AdaBoost learning algorithm. A hypothesis of shortening the stage classifiers is experimentally verified on the face detection problem. The stage classifiers shortening results in a faster face detector.

In the next section, the method of inter-stage feature

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

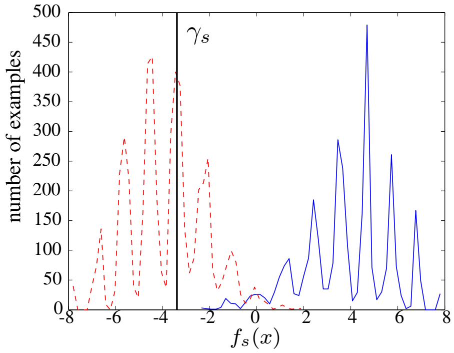

Figure 1. Histogram of values of $f_s(x)$ in a model stage s (solid: faces, dashed: non-faces). A threshold γ_s determines the false positive and false negative rate of the stage (see equation (1)).

propagation in the cascade building is explained in more detail. Section 3 describes experiments on the face detection problem and results are given in Section 4. The paper is concluded in Section 5.

2. Inter-stage feature propagation

The AdaBoost decision rule used in the Viola and Jones cascade building algorithm is of the form

$$\begin{aligned} H_s(x) &= sign(f_s(x) - \gamma_s) && (1) \\ f_s(x) &= \sum_{t=1}^{T} \alpha_t h_t(x) && (2) \end{aligned}$$

where s is the stage number, T is the length of the classifier, h_t is combined weak classifier and α_t its coefficient. The γ_s parameter is a threshold adjusted to reach the specified detection rate and false positive rate. $H_s(x) = -1$ implies rejection of the sample x and $H_s(x) = 1$ its acceptance.

In the Viola and Jones approach, after a stage classifier is trained, a new training set is generated and the AdaBoost algorithm is run on this *independent* problem to generate a classifier. AdaBoost generates a classifier as if no previous training has been done.

The parameter γ_s in each stage is set so that the majority of the face examples is kept and a substantial part of the non-face ones is rejected. Nevertheless, the non-decided part of the non-face examples is kept for training of the next stage. The current stage classifier can still be very good on these non-face examples. However the bias towards correct face example classification leaves this information unused.

Figure 1 displays a typical situation during the cascade building, at stage s. Face and non-face examples are already sufficiently separated, so γ_s can be found such that only small fraction of the face examples is misclassified and, concurrently, large part of the non-face examples is correctly classified. The non-face examples with $f_s(x) < \gamma_s$ already classified as non-face are replaced by newly sampled examples (such that $f_s(x) > \gamma$) in the next stage training. It can be seen that the stage s classifier is still reasonably good classifier on the new training set, only a new threshold have to be found. The previous stage classifier can be therefore further used as a very good starting point for the next stage training.

The AdaBoost algorithm runs in cycles. In each cycle, a new weak classifier with the smallest weighted error on the training set is added to the sum in equation (2). The proposed algorithm uses the features propagated from the previous stage to construct an additional weak classifier. Since the previous-stage classifier is fully represented by the f_s function, this function is used with different threshold to originate the weak classifier. The previous-stage weak classifier is inserted into the classifier at the beginning of the AdaBoost learning with coefficient α_0 found by the AdaBoost algorithm. The AdaBoost decision rule then becomes

$$\begin{aligned} H_s(x) &= sign(f_s(x) - \gamma_s) && (3) \\ f_s(x) &= \alpha_0 H'_{s-1}(x) + \sum_{t=1}^{T} \alpha_t h_t(x) && (4) \end{aligned}$$

where

$$\alpha_0 = \frac{1}{2} ln(\frac{1-\epsilon_0}{\epsilon_0})$$

and H'_{s-1} is the previous-stage weak classifier with a threshold changed to τ_s

$$H'_{s-1}(x) = sign(f_{s-1}(x) - \tau_s).$$

The weighted error ϵ_0 is computed as for an ordinary weak classifier as

$$\epsilon_0 = \sum_{i=1}^{m} D_0(i)[sign(f_{s-1}(x_i) - \tau_s) \neq y_i]$$

where $D_0(i)$ is a weight of the i^{th} example x_i with label y_i and m is the size of the current training set. The threshold τ_s is set to minimise ϵ_0.

Decision to insert the previous-stage weak classifier to the stage classifier as first has several reasons. First, since the weights of the training examples are initialised to the uniform distribution ($D_0(i) = 1/m$), they do not influence the performance of this weak classifier. Therefore, the weak classifier is very strong compared to the other simple weak classifiers (it is already boosted). Second, it also helps to focus AdaBoost learning to parts of the problem not learned by the previous stages.

An important property of such weak classifier is its zero evaluation cost. In both training and detection phase, a stage has to be evaluated to find out whether the next stage should be trained (in training) or used (in detection) on a given example. On the example x, $H_s(x)$ has to be evaluated and ergo the value of $f_s(x)$ is known. Comparing this value with a different threshold τ_s is very cheap operation compared to the evaluation of any other weak classifier and can be regarded as a zero-cost.

3. Experiments

The performance of the cascade building with inter-stage feature propagation and the cascade building of Viola and Jones was compared on the face detection problem. The training dataset and training process are discussed next. The performance evaluation concentrates on the speed and complexity of the learned cascaded classifiers.

Training data. The data for training were collected from various sources. Face images are taken from the MPEG7 face dataset [2]. The dataset contains face images of variable quality, different facial expressions and taken under wide range of lightning conditions, with uniform or complex background. The pose of the heads is generally frontal with slight rotation in all directions. Eyes and the nose tip are aligned in all images. The dataset contains 3176 images, one image was removed due to severe distortion.

Pose variability was added synthetically to the data. The images were randomly rotated by up to 5°, shifted up to one pixel and the bounding box was scaled by a factor of 1 ± 0.05. Two datasets, training and validation, of the same size as the original dataset were created by the perturbations.

Non-face images were collected from the web. Images of diverse scenes were included. The dataset contains images of animals, plants, countryside, man-made objects, etc.. More than 3000 images were collected and random sub-windows used as non-face examples.

Training process. During the training process, the training and validation dataset are updated for each stage (cf. [5], Table 2). The non-face part of the training and validation datasets consist of 5000 randomly selected regions from the non-face images. Only regions that were not rejected by previous stages of the cascade are included. The face set remains almost the same over the whole training. The faces rejected by some of the stage classifiers are removed, but the cascade is build to ensure that these false rejects are just a small fraction of the face data.

The process is driven by the stage false positive, detection and final false positive rates. In the reported experiments, the values were set to 0.4 stage false positive rate, 0.999 detection rate and 0.0001 the final false positive rate. The final false positive rate (a product of the stage false positive rates) was reached in the stage ten in both algorithms.

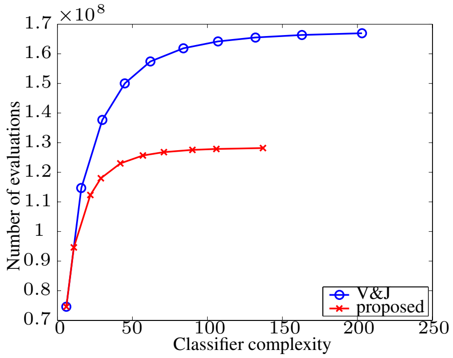

Figure 2. Selectivity comparison. Horizontal axis: the complexity of the cascaded classifier expressed by the number of weak classifiers used. Vertical axis: number of weak classifier evaluations on the MIT+CMU dataset.

4. Results

The classifiers were tested on the MIT+CMU dataset [3]. This dataset has been widely used for comparison of face detectors [3, 4, 5]. The main objective of the experiments is to demonstrate the detection speedup in comparison with the classical Viola and Jones approach, rather than improvement of the detection rate per se. This means that we did not try to find e.g. the optimal sets of weak classifiers since this is not important for a fair comparison of the methods.

The results for the cascades trained by the Viola and Jones approach and by the proposed inter-stage feature propagation approach are summarized in Table 1. For each number of stages in the cascade, the following quantities are recorded (left to right in Table 1): the number of weak classifier forming a stage in the cascade (the previous-stage weak classifier is not included, since its evaluation costs nothing), the total number of evaluations of each stage, and the false negative and false positive rates on the MIT+CMU dataset.

It can be observed that inter-stage feature propagation leads to the shorter stage classifiers. The saving is 25–50 %. Only the first stage remains the same, since there are no features to propagate. Obvious improvement is to shorten this stage to a minimal length and use its features in the next stage building. Similar approach was taken by Viola and Jones, but without the inter-stage feature propagation. Another observation can be made about the false negative and the false positive rates. The false negative rate is almost the same for both methods. However, the false positive rate is slightly better when the inter-stage feature propagation is used. This is consequence of integration of the very strong previous-stage weak classifier which uses the classification

Number of stages	Stage classif. length		Number of evaluations		False negatives		False positives	
	V&J	ISFP	V&J	ISFP	V&J	ISFP	V&J	ISFP
1	6	6	12431151	12431151	0	0	3930473	3930473
2	10	5	4009205	4009205	0	0	1598933	1571172
3	14	11	1643072	1609004	0	0	795262	780704
4	15	7	823246	806185	2	1	415751	368852
5	17	13	435483	385159	4	5	189902	168917
6	22	15	201982	179603	11	9	92226	73603
7	23	14	100887	80642	17	14	46499	34325
8	25	19	52867	39324	26	25	22966	16148
9	31	16	27504	19671	35	39	11262	7653
10	40	31	14818	10179	53	63	5070	1686

Table 1. Comparison of the cascade building with inter-stage feature propagation (ISFP) and the original Viola and Jones algorithm (V&J) performance on MIT+CMU dataset.

boundary found by the previous stages.

To compare the speed of the cascades trained with the inter-stage feature propagation and by the Viola and Jones approach, the number of weak classifiers evaluated on MIT+CMU dataset was measured. All regions have to be evaluated by the first stage classifier. The number of evaluations is consequently a product of the number of regions and the length of the first stage classifier. The same holds for the second (and higher) stage classifier, but only regions not rejected by the first (previous) stage(s) are evaluated. Summing the numbers evaluations of the first and the second stage gives the number of evaluations of the two-stage cascade classifier. The result for all lengths of the cascade and for both algorithms is depicted in Figure 2.

Figure 2 demonstrates two important phenomena. First, the complexity of the cascade classifiers with the same number of stages is about 30 % smaller when the inter-stage feature propagation is used. Second, the number of evaluations needed in original cascade building is higher by 25 % than in the inter-stage feature propagation approach.

5. Conclusions

A modification of the cascade building for the AdaBoost algorithm was proposed and compared with the Viola and Jones algorithm. The proposed algorithm is based on a novel view of the cascade building process which is seen as an algorithm for gradually finding more precise decision boundary in each stage. The inter-stage feature propagation was proposed to better focus the learning on the examples on which a decision has not been reached by the previous stage classifier. As was shown, the inter-stage feature propagation leads to shorter stage classifiers with the same false positive and false negative rate without increase in their computational complexity. The cascade trained by the proposed method was about 30 % shorter and 25 % faster compared to the cascade trained by the Viola and Jones algorithm.

Since the proposed algorithm gives almost the same results as the Viola and Jones algorithm and the resultant classifier is faster, it could be used instead of the original one without any disadvantages. The reduction of the number of weak classifiers can be important in areas where the weak classifiers are expensive to compute or to implement, e.g. on smart cards or other special purpose hardware.

Acknowledgments

The authors were supported by The Czech Science Foundation under project GACR 102/02/1539 and by The Grant Agency of the Czech Technical University under project CTU 0307313 and by The STINT under project Dur IG2003-2 062.

References

[1] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *ECCLT*, pages 23–37, 1995.

[2] ISO/IEC JTC 1/SC 29/WG 11 Moving Picture Experts Group.

[3] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *PAMI*, 20(1):23–38, January 1998.

[4] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, 2000.

[5] P. Viola and M. Jones. Robust real-time object detection. In *SCTV*, Vancouver, Canada, 2001.