Cleaning Up Multiple Detections Caused by Sliding Window Based Object Detectors

Arne Ehlers, Björn Scheuermann, Florian Baumann, and Bodo Rosenhahn

Institut für Informationsverarbeitung (TNT)
Leibniz Universität Hannover, Germany
lastname@tnt.uni-hannover.de

Abstract. Object detection is an important and challenging task in computer vision. In cascaded detectors, a scanned image is passed through a cascade in which all stage detectors have to classify a found object positively. Common detection algorithms use a sliding window approach, resulting in multiple detections of an object. Thus, the merging of multiple detections is a crucial step in post-processing which has a high impact on the final detection performance. First, this paper proposes a novel method for merging multiple detections that exploits intra-cascade confidences using Dempster's Theory of Evidence. The evidence theory allows hereby to model confidence and uncertainty information to compute the overall confidence measure for a detection. Second, this confidence measure is applied to improve the accuracy of the determined object position. The proposed method is evaluated on public object detection benchmarks and is shown to improve the detection performance.

1 Introduction

Object detection is a widely used application in computer vision and has been intensively studied. Most detectors used in computer vision have been trained by a machine learning algorithm. Especially the cascaded object detector proposed by Viola & Jones [1] which employs the AdaBoost [2] machine learning algorithm is very successful. Object detectors are commonly applied by a sliding window which scans the scene image on shifted positions and varied scales. This frequently results in multiple detections of an object at slightly shifted and scaled positions. In a post-processing step, these multiple detections have to be combined to determine the final object position and scale. Often only little effort is spend on detection merging and simple methods are applied. Although this subtask has a strong impact on the overall accuracy of the detection framework and the results achieved in benchmarks. E.g., Viola & Jones in [1] merge all overlapping detection windows to one detection. But this approach easily leads to worse results in case of increasing numbers of detections, in particular if detections on large scales are involved. Everingham et al. [3] thus reported in the PASCAL VOC Challenge that the measured average precision steeply dropped for all participating methods when they tightened the tolerances for correct detections on the "car" class.

In this work, a novel method for merging multiple detections is proposed. Dempster's Theory of Evidence is applied to combine confidence values similar to Real AdaBoost [4] and uncertainty information that is available in a cascaded detector. In this

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 456–463, 2013.

way intra-cascade information is exploited in an improved merging of multiple detections during post-processing. Huang et al.[5] introduced a nested classifier to inherit classification confidences in detection cascades. But their approach is confined to the classification step and requires a retraining. This paper proposes a novel confidence measure which is in addition applied to refine the position and scale of merged detections. It is shown that the proposed confidence gives an appropriate measure to distinguish the reliability of detections. As a post-processing step, the proposed method is easily applicable in other object detection frameworks without the need of retraining the object classifiers. Hence, other object detection frameworks could benefit from the proposed detection merging.

2 Merging Multiple Detections Based on Dempster's Theory

In this Section, the proposed strategies on merging detections are described in detail. The required methods of machine learning, object detection and evidence theory are briefly discussed in advance.

2.1 Cascaded Classifier

The object detection framework used in this work utilizes a cascaded classifier as introduced by Viola & Jones [1] and illustrated in Figure 1. Each stage of this cascaded classifier consists of a strong classifier that is created using the AdaBoost machine learning algorithm [2]. Hence in a cascade of S stages, S strong classifier have to decide positively for a scanned sub-window x to be classified as an object. Any of these candidate sub-windows is then further processed in the post-processing step in which the merging of multiple detections is done.

Each strong classifier $H_s(x) = \sum_{t=1}^{T_s} \alpha_{s,t} h_{s,t}(x), s \in 1 \dots S$ is composed of an ensemble of T_s weak classifiers $h_{s,t}$ which have been selected in the training phase of the AdaBoost algorithm. Each weak classifier returns 0 or 1 in case of a negative or positive classification, respectively. These ensembles decide in a weighted majority vote in which each weak classifier $h_{s,t}$ supports its decision by an assigned weight $\alpha_{s,t}$ that represents the classification error of that weak classifier in training. Thus, the maximum positive classification of a strong classifier is given by $H_{s,max} = \sum_{t=1}^{T_s} \alpha_{s,t}$ and the decision threshold of AdaBoost is the weighted majority $\tau_s = \frac{1}{2} \sum_{t=1}^{T_s} \alpha_{s,t}$.

AdaBoost's decision threshold aims at a low error rate on the training set without differentiating between positive and negative training examples. But due to the rejection opportunity of each cascade stage, a very high true positive rate is primarily desired. Hence according to [1], a subsequently adjusted threshold τ_s is used to maintain a very high true positive rate accepting an also high false positive rate.

2.2 Dempster-Shafer Theory of Evidence

In this section Dempster's theory of evidence is briefly described. It is utilized in the proposed method to model intra-cascade decision confidences and uncertainties. The Dempster-Shafer theory of evidence was introduced in 1968 by A. P. Dempster [6] and

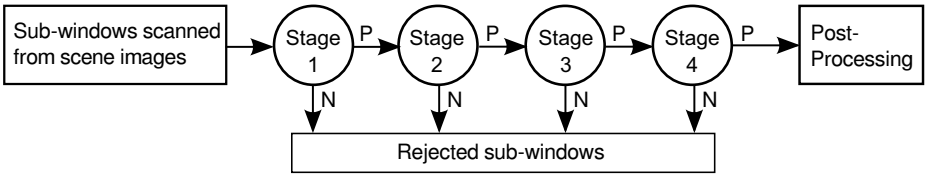


Fig. 1. Detection cascade: Evaluated sub-window have to be positively classified (P) and passed by all cascade stages to be considered as a found object. Each cascade stage can reject a sub-window if it is negatively classified (N) and thus prevents its processing by the following stages.

later in 1976 expanded by G. Shafer [7]. Evidence theory can be interpreted as a generalization of Bayesian theory that directly allows the representation of uncertainty and inaccuracy information. The key element of the evidence theory is the definition of a mass function on a hypotheses set Ω. Let a hypotheses set be denoted by Ω and composed of n single mutually exclusive subsets Ω_i written as $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_n\}$. For each element A of the power set $\wp(\Omega)$ a mass function $m(A)$ is defined that expresses the proportion of all evidence assigned to this hypothesis. Hence, the mass function m represents a degree of confidence and is defined as $m : \wp(\Omega) \rightarrow [0, 1]$. Furthermore, the following conditions have to be fulfilled by the mass function:

$$(i) \quad m(\emptyset) = 0 \quad (ii) \quad \sum_{A_n \subseteq \Omega} m(A_n) = 1\,. \tag{1}$$

Mass functions in evidence theory describe the totality of belief as opposed to Bayesian probability functions. This belief can be associated with single and composed sets of hypotheses allowing for a higher level of abstraction. The so-called additivity rule $p(A) + p(\overline{A}) = 1$ is in contrast to Bayesian theory not generally valid in Dempster-Shafer evidence theory. This means that if $m(A) < 1$, the remaining evidence $1-m(A)$ does not necessarily claim its negation $\overline{A}$.

Dempster's Rule of Combination. In order to combine information from different stages of the detection cascade, *Dempster's rule of combination* is applied. Dempster's rule combines two mass functions that are defined within the same frame of discernment but belong to independent bodies of evidence. Let m_1 and m_2 be two mass functions associated to such independent bodies of evidence. Then Dempster's rule defines the new body of evidence by the mass function

$$m(A) = m_1(A) \otimes m_2(A) = \frac{\sum\limits_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum\limits_{B \cap C = \emptyset} m_1(B) m_2(C)}\,. \tag{2}$$

The denominator in Equation (2) works as a normalization factor that ignores the conflicting evidence. Hence, Dempster's rule of combination focuses on the measure of agreement between two bodies of evidence. Dempster's rule is associative and thus can be used to iteratively combine evidences obtained from arbitrary number of classifiers.

2.3 Joint Confidence Based on Dempster-Shafer

In the proposed application of joining intra-cascade confidences, the frame of discernment is defined as $\Omega = \{TP, FP\}$ containing the set of hypotheses supporting a true positive (TP) and a false positive (FP) decision, respectively. The uncertainty of each cascade stage s is modeled by $m_s(\Omega)$ with respect to its size:

$$m_s(\Omega) = 1 - \frac{T_s}{\sum_{s=1}^{S} T_s} \tag{3}$$

This leads to a higher belief into stages that consist of larger number of weak classifiers.

The mass functions, expressing the proportion of evidence of a stage s, for true positive or false positive decisions are defined by:

$$m_s(TP) = \frac{H_s(x) - \tau_s}{H_{s,max} - \tau_s}(1 - m_s(\Omega)), \tag{4}$$

$$m_s(FP) = \Big(1 - \frac{H_s(x) - \tau_s}{H_{s,max} - \tau_s}\Big)(1 - m_s(\Omega)) \tag{5}$$

This results in higher stage confidence when the difference between the response of the strong classifier and the decision threshold grows. Using Dempster's rule of combination the stage confidences for a detection D_i are joined by

$$m_{D_i}(TP) = m_1(TP) \otimes m_2(TP) \otimes \cdots \otimes m_S(TP) \tag{6}$$

to gain an overall detection confidence.

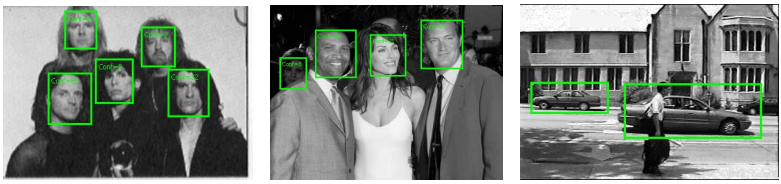
2.4 Confidence-Based Detection Merging

Merging of multiple detection commonly takes place in the post-processing step of an object detection framework. The position and scale information of the candidate subwindows has to be processed to determine the true object location.

In this work, the candidate detections are first clustered using the Meanshift algorithm [8,9] as the number of true objects and thus desired clusters is unknown in advance. The i-th candidate detection is hereby defined as a four-dimensional vector $D_i = (x_i, y_i, \gamma_i, \delta_i)^\top$ which represents the combined position $(x_i, y_i)^\top$ and scale $(\gamma_i, \delta_i)^\top$ in x and y-dimension. The set of n candidate detections is partitioned by the Meanshift algorithm in four-dimensional space into $k \leq n$ sets $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ of clusters. The merged detections are then set as the cluster centers of the k clusters in $\mathbf{C}$ and a simple confidence of the k-th cluster is given by its cluster size $|C_k|$.

To improve the performance of the object detector, this paper proposes two enhancements to the detection merging. First, the detection confidences given by Equation (6) are exploited to define the Dempster-Shafer based confidence of the k-th cluster as $\Gamma_k = \sum_{D_i \in C_k} m_{D_i}(TP)$. Second, these confidences of detections associated to one cluster are utilized to refine the position and scale of the cluster center. In this way the Dempster-Shafer refined position/scale of the k-th cluster is defined by:

$$D'_k = \frac{1}{\Gamma_k} \sum_{D_i \in C_k} D_i m_{D_i}(TP) \tag{7}$$



(a) (b) (c)

Fig. 2. Example images showing detections of our method on the three evaluated data sets: MIT+CMU [11], FDDB [10] and UIUC lateral car database [12]

3 Experimental Results

In this section, cascaded classifiers are applied by a sliding window to data sets for face and lateral car detection. The acquired multiple detections are post-processed using different merging strategies and results are presented for the Face Detection Data Set and Benchmark (FDDB) [10], the MIT+CMU frontal face database [11] and the UIUC lateral car database [12]. Figure 2 exemplary shows detections found by our method in the evaluated data sets.

3.1 Face Detection

For the detection of faces, a classifier is trained on the ”MPLap GENKI-4K” database from the Machine Perception Laboratory in California [13] that consists of 4000 faces under different facial expressions. The obtained strong cascaded classifier consists of 10 stages and 593 weak classifiers in total.

Experiments Incorporating Confidence. The first experiments are conducted using the Face Detection Data Set and Benchmark [10] that contains 5171 faces in 2845 images. This data set also provides an evaluation tool for a consistent comparison of the performance of competing methods. Evaluations generated by this tool for different face detectors are available on the project web page[1]. The evaluation procedure requires multiple detections to be priorly merged to single detections that have an assigned confidence value. In descending order, each unique confidence value is then selected as a threshold and the true positive rate and total false positives are calculated considering all merged detections that have a greater confidence. In this way, a ROC curve is constructed that presents the detection performance.

The inspection of the detection confidence enables the separate evaluation of two contributions in the proposed approach: The confidence computation based on Dempster-Shafer theory of evidence and the position and scale refinement using these confidences.

Figure 3 presents the detection results for different strategies on merging multiple detections. The performance of the Viola & Jones detector in OpenCV, supplied by the FDDB project page, is presented as a baseline result. But the primary topic of this work

[1] `http://vis-www.cs.umass.edu/fddb/results.html`

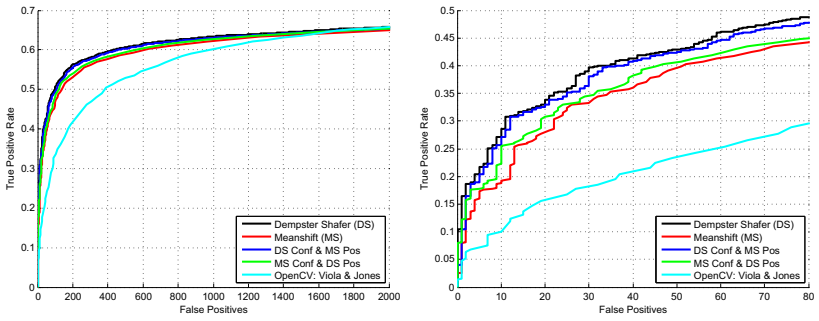
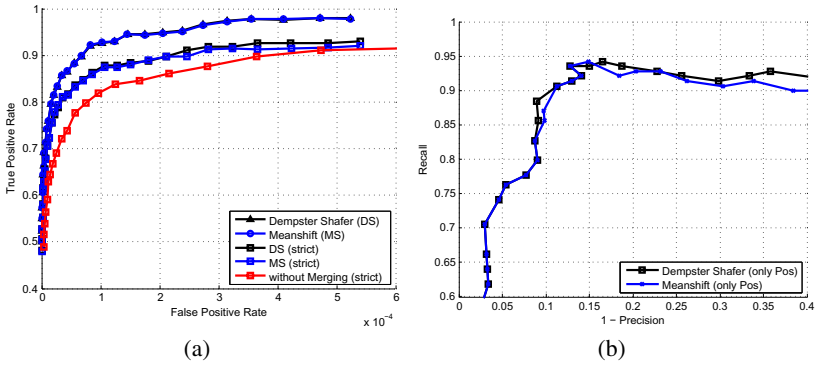


Fig. 3. ROC curve presenting the detection performance on FDDB [10] for different approaches on merging multiple detections. Confidence calculation and position/scale refinement based on Dempster-Shafer (DS) is compared to Meanshift-based confidence and position/scale (MS) and mixed approaches using Dempster-Shafer only for confidence and position/scale, respectively. The performance of the Viola & Jones implementation in OpenCV is presented as a baseline result. The shown range is (a) up to saturation and (b) a detailed view.

is the impact of the pre-processing step of multiple detection merging and not the comparison to different object detection methods. The proposed method (DS) is compared to an approach that only exploits the preceding Meanshift clustering (MS). For this, the number of detections forming each cluster is utilized as the confidence value. In addition, the results of two mixed approaches are presented that use Dempster-Shafer only for confidence calculation and position/scale refinement, respectively. The detailed view in Figure 3(b) demonstrates that, although the same detector is used, the performance can be significantly improved by about 5% in terms of true positive rate. It can be also observed from the blue curve in Figure 3(b) that the proposed confidence computation causes the biggest part of the improvement. This demonstrates that the Dempster-Shafer confidence gives an appropriate measure to distinguish the reliability of detections. The position/scale refinement slightly improves the detection performance, indicating that the trained classifier is not detecting symmetrically around the true object location. The proposed refinement can rectify that bias presenting improved results in the green curve of Figure 3(b).

Experiments on Position/Scale Refinement. Additional experiments are performed on the MIT+CMU frontal face database [11] which consists of 130 grayscale images containing 511 faces. The image database is partially noisy and blurred and contains several difficult samples like comics, line drawings and a binary raster image and thus is, despite its age, still challenging. This test set gives ground truth information on the position and scale of the faces but no evaluation tool is provided. Hence, the evaluation against ground truth is done by a built-in function of the detection framework that governs the ROC curve by a threshold multiplier in the detection process instead of exploiting confidence values.



(a) (b)

Fig. 4. (a) ROC curve presenting the detection performance on the MIT+CMU frontal face database [11]. The effect of the position/scale refinement using Dempster-Shafer is compared to Meanshift clustering in the case of loosened and stricter ground truth tolerances. Additionally results when omitting multiple detection merging are presented. (b) ROC curve presenting the detection performance on the UIUC lateral car database [12]. The effect of the additional position/scale refinement using Dempster-Shafer is compared to merging multiple detections by Meanshift clustering.

For this reason, Figure 4(a) shows only the impact of the position/scale refinement. In addition, results for completely omitting the post-processing are presented as the built-in evaluation does not require the merging of multiple detections. The general benefit of the post-processing can be observed from the improved results compared to the approach without merging multiple detections. During the merging process detection outliers are suppressed that are outside the ground truth tolerances. The detector performance only slightly benefits from the position/scale refinement. This is partly a consequence of the properties of the MIT+CMU frontal face database that contains many very small faces but provides no subpixel accuracy in the ground truth data. As the accuracy of the detections position and scale has no influence on the ROC curve as long as they are inside the tolerances, additional results for stricter tolerances are presented by the curves labeled as strict. These curves reveal a slight improvement due to the proposed position/scale refinement even on this unfavourable test set.

3.2 Lateral Car Detection

To evaluate an additional object class, experiments are conducted on the UIUC lateral car database [12]. This database provides a training set containing 1050 grayscale images (550 cars and 500 non-car images). In addition, images for single and multi-scale tests are contained as well as an evaluation tool for the calculation of precision and recall. Figure 4(b) compares the detection results achieved when merging multiple detections by Meanshift clustering and the proposed position/scale refinement using Dempster-Shafer confidences. The evaluation tool does not consider detection confidences but requires multiple detections to be merged to a single detection in advance. Hence, a concentration on only the impact of the position/scale refinement is

predetermined. In this experiment, that utilizes a different object class, an improvement of the detection performance can be observed due to the position/scale refinement. This indicates that the car classifier as well does not detect symmetrically around the true object location but introduces a bias that can be rectified by the proposed method.

4 Conclusion

This paper presents a novel method for merging multiple detections which exploits classification information available in cascaded detectors. Two enhancements are proposed. First, Dempster-Shafer theory of evidence is applied to model a confidence measure which incorporates intra-cascade decision confidences and uncertainties. Second, a method is presented to refine the position and scale of merged detections based on these confidence measures. These methods can be easily integrated in existing detection frameworks to improve performance without retraining of typical cascaded detectors. Results are presented for a recent benchmark on unconstrained face detection (FDDB), the MIT+CMU face and the UIUC car database. The refinement of position and scale solely results in a slight improvement in detection performance. In addition, the proposed confidence measure shows an improvement of 5% in true positive rate for applications that consider detection confidences. This demonstrates that Dempster-Shafer theory of evidence is a powerful technique to model and exploit intra-cascade confidences.

References

1. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
2. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the Thirteenth International Conference, pp. 148–156 (1996)
3. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2) (2010)
4. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning 37(3), 297–336 (1999)
5. Huang, C., Al, H., Wu, B., Lao, S.: Boosting nested cascade detector for multi-view face detection. In: Pattern Recognition, ICPR 2004 (2004)
6. Dempster, A.P.: A generalization of bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological) 30(2), 205–247 (1968)
7. Shafer, G.: A mathematical theory of evidence, vol. 1. Princeton University Press, Princeton (1976)
8. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. 17(8), 790–799 (1995)
9. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24(5), 603–619 (2002)
10. Jain, V., Learned-Miller, E.: Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst (2010)
11. Sung, K.K., Poggio, T., Rowley, H.A., Baluja, S., Kanade, T.: MIT+CMU frontal face dataset a, b and c. MIT+CMU (1998)
12. Agarwal, S., Awan, A., Roth, D.: UIUC image database for car detection (2002)
13. TheMPLab GENKI Database, u.S., `http://mplab.ucsd.edu`