Evidential combination of pedestrian detectors

Philippe Xu[1]
https://www.hds.utc.fr/~xuphilip
Franck Davoine[12]
franck.davoine@gmail.com
Thierry Denœux[1]
https://www.hds.utc.fr/~tdenoeux

[1] UMR CNRS 7253, Heudiasyc, Université de Technologie de Compiègne, France

[2] CNRS, LIAMA, Beijing, P. R. China

Abstract

The importance of pedestrian detection in many applications has led to the development of many algorithms. In this paper, we address the problem of combining the outputs of several detectors. A pre-trained pedestrian detector is seen as a black box returning a set of bounding boxes with associated scores. A calibration step is first conducted to transform those scores into a probability measure. The bounding boxes are then grouped into clusters and their scores are combined. Different combination strategies using the theory of belief functions are proposed and compared to probabilistic ones. A combination rule based on triangular norms is used to deal with dependencies among detectors. More than 30 state-of-the-art detectors were combined and tested on the Caltech Pedestrian Detection Benchmark. The best combination strategy outperforms the currently best performing detector by 9% in terms of log-average miss rate.

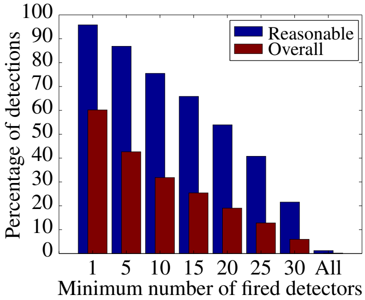
1 Introduction

Object detection is one of the most important and challenging tasks in computer vision. More and more sophisticated and efficient algorithms are proposed every year. Several benchmarks have appeared in the past decade. The PASCAL VOC Challenge [17] is certainly the most popular one. In the last VOC 2012 edition [18], the organizers built a *super-classifier* over the seven methods that were submitted to the classification challenge. The scores returned by the classifiers were concatenated into a single vector and a linear SVM was trained with it. An increase of more than 10% in terms of average precision was reported for certain object classes such as “bottle” or “pottedplant”. Performance losses were observed for five classes out of twenty, but they remained relatively limited. One main drawback of such an approach is the difficulty to include new methods as a whole new classifier needs to be trained every time.

In this paper, we use this idea for pedestrian detection, which is the most studied case. There exist many pedestrian datasets; INRIA [7], ETH [16], TUD-Brussels [44] and Caltech Pedestrian Detection Benchmark [15] are among the most popular ones. The last one is the largest. More than 30 state-of-the-art detectors were tested on it and their outputs are publicly

#	Algorithm	Features	Classifier	Training
1	'VJ' [40]	Haar	AdaBoost	INRIA
2	'HOG' [7]	HOG	linear SVM	INRIA
3	'HikSvm' [25]	HOG	HIK SVM	INRIA
4	'LatSvm-V1' [19]	HOG	latent SVM	PASCAL
5	'LatSvm-V2' [19]	HOG	latent SVM	INRIA
6	'MultiResC' [32]	HOG	latent SVM	Caltech
7	'MultiResC+2Ped' [29, 32]	HOG	latent SVM	Caltech
8	'MT-DPM' [illegible]	HOG	latent SVM	Caltech
9	'MT-DPM+Context' [illegible]	HOG	latent SVM	Caltech
10	'PoseInv' [22]	HOG	AdaBoost	INRIA
11	'MLS' [27]	HOG	AdaBoost	INRIA
12	'DBN-Isol' [28]	HOG	DeepNet	INRIA
13	'DBN-Mut' [30]	HOG	DeepNet	INRIA/Caltech
14	'HOG-LBP' [42]	HOG+LBP	linear SVM	INRIA
15	'MOCO' [illegible]	HOG+LBP	latent SVM	Caltech
16	'pAUCBoost' [31]	HOG+COV	pAUCBoost	INRIA
17	'FtrMine' [11]	channels	AdaBoost	INRIA
18	'ChnFtrs' [12]	channels	AdaBoost	INRIA
19	'FPDW' [13]	channels	AdaBoost	INRIA
20	'CrossTalk' [14]	channels	AdaBoost	INRIA
21	'Roerei' [illegible]	channels	AdaBoost	INRIA
22	'ACF' [illegible]	channels	AdaBoost	INRIA
23	'ACF-Caltech' [illegible]	channels	AdaBoost	Caltech
24	'ACF+SDt' [33]	channels	AdaBoost	Caltech
25	'MultiFtr' [43]	multiple	AdaBoost	INRIA
26	'MultiFtr+CSS' [41]	multiple	linear SVM	TUD-Motion
27	'MultiFtr+Motion' [41]	multiple	linear SVM	TUD-Motion
28	'MF+Motion+2Ped' [29, 41]	multiple	linear SVM	TUD-Motion
29	'FeatSynth' [2]	multiple	linear SVM	INRIA
30	'AFS' [23]	multiple	linear SVM	INRIA
31	'AFS+Geo' [23]	multiple	linear SVM	INRIA
32	'Pls' [37]	multiple	PLS+QDA	INRIA
33	'Shapelet' [36]	gradients	AdaBoost	INRIA
34	'ConvNet' [38]	pixels	DeepNet	INRIA

(a)



(b)

Figure 1: (a) List of algorithms evaluated on the Caltech Pedestrian Benchmark. (b) Percentage of detected pedestrians by at least $k \in \{1, 5, \ldots, 34\}$ detectors at 1 FPPI. The detections were done on the Caltech-Test dataset with the "Reasonable" and "Overall" scenarios.

available. Moreover, the high diversity of the evaluated methods makes their combination an ever more interesting issue. Figure 1 (a) lists the detectors evaluated on the Caltech dataset.

Diversity, and thus potential complementarity of the detectors exist because of mainly three reasons. The first one is related to the features used to represent pedestrians. Haar-like features [40], shapelets [36], shape context [26] and histogram of oriented gradient (HOG) [7] features are commonly used. The last one is the most popular and almost all detectors use it in some forms. Wojek and Schiele [43] concatenated all the previously mentioned features and trained a new model outperforming all individual ones. Other features such as local binary pattern (LBP) [42] or motion features [41] were also considered in addition to HOG. However, even though the HOG feature is used in those methods, it is not guaranteed that a pedestrian detected by the original 'HOG' detector [7] would still be detected by the other methods. Nevertheless, the use of multiple types of features as in [12, 13, 33] or features learned in very large spaces [2, 11] have led to significant improvements.

The second source of diversity comes from the classifier. Linear SVM and AdaBoost are often considered. The use of latent variables in SVM has been popularized by Felzenszwalb *et al.* [19] for part-based approaches. Non-linear SVM [25], Partial Least Squares analysis [37] or boosting optimizing directly the area under the ROC curve [31] were also used. More recently, deep learning was also considered [28, 30, 38]. Finally, the choice of the training data, if not the same for all detectors, is an additional source of diversity.

Different forms of detectors combination can be found in the literature. The use of multiple sensors in robotics has often led to the combination of several detectors. The easiest way is to use a first weak detector to gather a set of regions of interest, which are then more deeply analyzed by a more efficient one. The 'FeatSynth' [2] algorithm actually only processes the detections returned by 'FtrMine' [11]. Some works make use of other object detectors such as cars [45] or 2-pedestrians detectors [29]. Recently, Denoeux *et al.* [10] applied an optimal object association algorithm to combine the outputs of two object detectors in polynomial time. However, the optimal association problem with more than two detectors is NP-hard.

To figure out the potential gain from combining multiple detectors, we show in Fig. 1 (b) some detection statistics for the Caltech dataset. We can see that, at one False Positive Per Image (FPPI), more than 95% of the pedestrians in the "Reasonable" scenario were detected by at least one detector. The "Reasonable" scenario corresponds to pedestrians over 50 pixels tall and with an occlusion rate lower than 35%. As a comparison, the currently best performing algorithm ('ACF+SDt' [33]) has a recall rate of about 80% at 1 FPPI. Similarly, in the "Overall" scenario where all the pedestrians were considered, about 60% of the pedestrians were detected by at least one detector. The 'MT-DPM+Context' [45] algorithm, which outperforms 'ACF+SDt' in this scenario, hardly reached a 40% recall rate. The potential gain of combining in a proper way all those detectors is thus fairly significant.

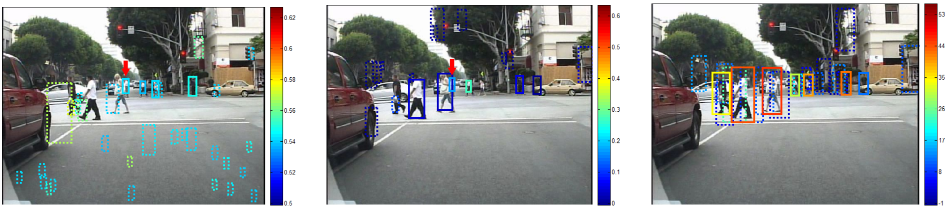
In this paper, we propose a combination framework that models the outputs of the detectors with the theory of belief functions and combines them with a pre-defined rule. A pedestrian detector is seen as a black box that only outputs a set of bounding boxes (BB) with associated scores. In Sec. 2 we propose a new way to associate the BBs returned by multiple detectors; we then show how the scores are calibrated and combined from a probabilistic point of view. In Sec. 3 we present different combination strategies using belief functions. Finally, in Sec. 4 we compare the different combination methods using the Caltech dataset.

2 Combination of pedestrian detectors

The outputs of most pedestrian detectors are given as bounding boxes. To each of them is associated a score representing the confidence of the detector. The range of those scores depends on the features and the classifier used for detection. Figure 2 shows some detection results from three algorithms, applied to one particular image frame. The 'VJ' algorithm gives pretty poor results with a very high false detection rate. Even worse, the BBs with the highest scores are actually false positives. The 'HOG' algorithm gives relatively good results with few false positives. It can be noticed that the two detected pedestrians in the foreground have very low scores. The 'ACF+SDt' algorithm is the one with the highest recall rate. Even though it returns more false positives than 'HOG', most of the true positives have a higher score than the false negatives. It is, however, interesting to notice that on the particular image shown in Figure 2, the only pedestrian missed by 'ACF+SDt' was actually detected by both the 'VJ' and 'HOG' algorithms. The aims of combining the algorithms are thus to obtain a higher overall recall rate and to increase the confidence in the pedestrians that are detected by multiple detectors. For this purpose, two main issues have to be solved. First of all, it is necessary to appropriately associate the BBs returned by the detectors. The scores from the different algorithms have then to be made comparable and combined.

2.1 Clustering of bounding boxes

In a sliding windows approach, a single pedestrian is often detected at several nearby positions and scales. A non-maximal suppression (NMS) step is often needed in order to select only one BB per pedestrian. In our context, the same issue occurs but instead of having multiple detections from a single detector they are returned by several ones. As reported



(a) 'VJ' [40] (b) 'HOG' [7] (c) 'ACF+SDt' [33]

Figure 2: Pedestrian detection results from three algorithms. The colour of the bounding boxes represents the score. Solid boxes are true positives and doted boxes are false positives. The red arrow points to a pedestrian detected by both 'VJ' and 'HOG' but not 'ACF+SDt'.

by Dollár *et al.* [15], there exist two dominant NMS approaches: mean shift mode estimation [7] and pairwise maximum suppression [19]. For the former it is necessary to define a covariance matrix representing the uncertainty in position and size of the BBs. This can be difficult considering the high variety of detectors. Felzenszwalb *et al.* [19] proposed a simpler way by suppressing the least confident of any pair of BBs that overlap sufficiently. Given two bounding boxes BB_i and BB_j, their area of overlap is defined as

$$a_{\text{union}} = \frac{\text{area}(BB_i \cap BB_j)}{\text{area}(BB_i \cup BB_j)}. \tag{1}$$

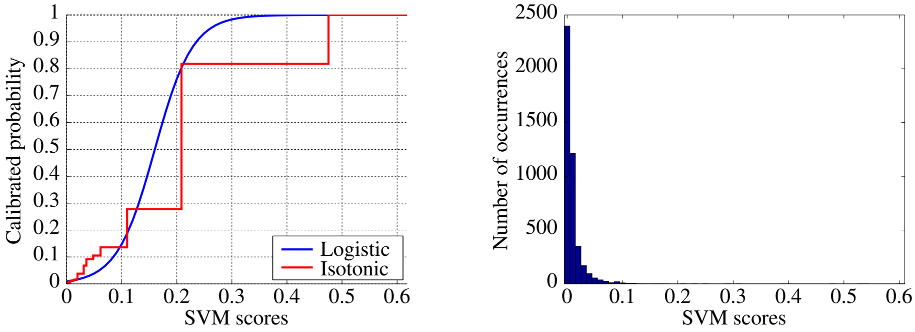
Dollár *et al.* (see addendum to [12]) proposed to replace the above definition with

$$a_{\min} = \frac{\text{area}(BB_i \cap BB_j)}{\min(\text{area}(BB_i), \text{area}(BB_j))}. \tag{2}$$

By using a_{union} or $a_{\min}$ as a distance measure between BBs, a simple hierarchical clustering can be used to group them until the overlap exceeds a certain threshold. The distance between two clusters is defined as the maximum distance between every pairs of BBs. This guarantees that within a cluster the overlapping area between two BBs is always sufficient. Dollár *et al.* [12] showed that proceeding greedily leads to the best results. They processed the detections in decreasing order of scores; when two BBs are associated, the one with the lowest score would no longer be used for further associations. In our clustering formulation, this later point is equivalent to defining the distance between two clusters as the distance between their respective highest-scored BBs. One issue with this approach is that the scores from the different detectors have to be comparable.

2.2 Score calibration

As stated earlier, the scores from the detectors can be of very different natures and a probability measure is often used as a common representation. The transformation from a classifier score into a probability is referred to as calibration. Given a BB with a score $s \in \mathbb{R}$ and an unknown label $y \in \{0,1\}$, the calibration aims at finding a function $f : \mathbb{R} \to [0,1]$ so that $f(s)$ is an estimator of $P(y=1|s)$. Several calibration methods can be found in the literature [3]. As the scores are supposed to represent the confidence of the classifier, the calibration function f is assumed to be non-decreasing. Logistic regression [34] and isotonic regression [46] are two popular calibration methods that use this assumption. Given a set of n detected BBs



(a) (b)

Figure 3: (a) Logistic and isotonic calibration of the scores from the 'HOG' pedestrian detector. (b) Histogram of the scores.

with scores $s_1, \ldots, s_n \in \mathbb{R}$ and known labels $y_1, \ldots, y_n \in \{0, 1\}$, Platt [34] proposed to use logistic regression and to fit a sigmoid function $g : \mathbb{R} \to [0, 1]$ defined as

$$g(s) \quad = \quad \frac{1}{1+\exp(A+Bs)}, \qquad \forall s \in \mathbb{R}, \tag{3}$$

where the parameters $A, B \in \mathbb{R}$ are determined by minimizing the negative log-likelihood function on the training data

$$\min_{A,B\in\mathbb{R}} -\sum_{k=1}^{n} y_k \log\left(g\left(s_k\right)\right) + \left(1-y_k\right)\log\left(1-g\left(s_k\right)\right). \tag{4}$$

Zadrozny and Elkan [46] proposed a non-parametric calibration method by fitting a stepwise-constant non-decreasing, *i.e.* isotonic, function $h : \mathbb{R} \to [0, 1]$ that directly minimizes the mean-squared error

$$MSE(h) = \frac{1}{n}\sum_{k=1}^{n}[h(s_k) - y_k]^2. \tag{5}$$

This function can be computed efficiently using the pair-adjacent violators algorithm [1]. Figure 3 (a) shows two calibrations of SVM scores computed from the 'HOG' algorithm.

2.3 Probabilistic combination of bounding boxes

One particularity of object detection is the relatively high false positive rate. For example with the 'HOG' algorithm, more than 99% of the detections have a score less than 0.1, as illustrated on Figure 3 (b). Less than 0.1% of these detections are true positives. As a result, most detections have an associated probability lower than 0.1. From a Bayesian perspective, multiple sources of information returning low probabilities would actually lead to an even lower one. Let $s_{(1)}, \ldots, s_{(k)}$ be the scores returned by k detectors for a cluster of BBs of label y. By using Bayes' rule and assuming conditional independence, the following equation holds:

$$P\left(y=1|s_{(1)},\ldots,s_{(k)}\right) = \frac{P(y=1)}{P\left(s_{(1)},\ldots,s_{(k)}\right)}\prod_{i=1}^{k}P\left(s_{(i)}|y=1\right) \tag{6}$$

$$= \frac{P\left(s_{(1)}\right)\cdots P\left(s_{(k)}\right)}{P\left(s_{(1)},\ldots,s_{(k)}\right)P(y=1)^{k-1}}\prod_{i=1}^{k}P\left(y=1|s_{(i)}\right) \tag{7}$$

$$\propto P(y=1)^{1-k}\prod_{i=1}^{k}P(y=1|s_{(i)}). \tag{8}$$

By using different approximations on this product rule, Kittler *et al.* [22] derived several classical combination rules such as the minimum, average, or maximum rules. Those rules have the following relations:

$$\prod_{i=1}^{k}P\left(y=1|s_{(i)}\right) \leq \min_{i=1,\ldots,k} P\left(y=1|s_{(i)}\right) \leq \frac{1}{k}\sum_{i=1}^{k}P\left(y=1|s_{(i)}\right) \leq \max_{i=1,\ldots,k} P\left(y=1|s_{(i)}\right). \tag{9}$$

Kittler *et al.* [22] reported the superiority of the average combiner. A popular variant is the weighted average combiner. Bella *et al.* [4] showed that using a weighting in addition to calibration often leads to better results. Another classical combination strategy is the voting rule and its weighted variant.

3 Theory of belief functions

The theory of belief functions, also known as Dempster-Shafer theory [39], is a generalization of probability theory. It is commonly used as an alternative to probability theory as it is especially well adapted for information fusion [21].

3.1 Information representation

Let $\Omega = \{\omega_1,\ldots,\omega_N\}$ be a finite set of classes and 2^Ω its powerset. A *mass function* is defined as a function $m : 2^\Omega \mapsto [0,1]$ verifying

$$\sum_{A\subseteq\Omega} m(A) = 1, \qquad m(\emptyset) = 0. \tag{10}$$

Given a sample of class $\omega \in \Omega$ and a subset $A \subseteq \Omega$, the quantity $m(A)$ represents the amount of belief strictly supporting the hypothesis $\omega \in A$. In the case of pedestrian detection where $\Omega = \{0,1\}$, when a detector returns a BB, it actually only supports the hypothesis that a pedestrian is present. The output of a calibration function f for a detection score s is thus interpreted as the following mass function:

$$m(\{1\}) = f(s), \qquad m(\{0,1\}) = 1 - f(s). \tag{11}$$

Such a mass function is said to be *simple* and will be noted as $\{1\}^{1-f(s)}$. More generally, given $A \subseteq \Omega$ and $x \in [0,1]$, A^x refers to the mass function $m(A) = 1-x$, $m(\Omega) = x$. The reliability of a source of information can be encoded through a discounting factor $\delta \in [0,1]$, which leads to a discounted mass function ${}^\delta m$ defined as

$${}^\delta m(A) = \begin{cases} (1-\delta)m(A), & \text{if } A \subsetneq \Omega, \\ (1-\delta)m(\Omega)+\delta, & \text{otherwise.} \end{cases} \tag{12}$$

This factor can be seen as a weight in a combination. The mass function is kept unchanged if $\delta = 1$ and becomes the total ignorance, *i.e.* $m(\Omega) = 1$, when $\delta = 0$.

3.2 Evidential combination rules

There exists several ways to combine two mass functions m_1 and m_2. The most basic way is Dempster's rule of combination [39] which defines a mass function $m_1 \oplus m_2$ as follows:

$$(m_1 \oplus m_2)(\emptyset) = 0, \qquad (m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C} m_1(B) m_2(C), \ \forall A \subseteq \Omega, \ A \neq \emptyset, \quad (13)$$

with

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \quad (14)$$

For two simple mass functions $\{1\}^{\alpha_1}$ and $\{1\}^{\alpha_2}$, it leads to

$$\{1\}^{\alpha_1} \oplus \{1\}^{\alpha_2} \quad = \quad \{1\}^{\alpha_1 \alpha_2}. \quad (15)$$

This combination rule is based on the assumption that the mass functions to combine are generated by independent sources of information. In particular, Dempster's rule is not idempotent: a mass function combined with itself leads, in general, to a different mass function. In our context, some detectors are clearly not independent, for example, the 'FPDW' [13] algorithm is just an accelerated variant of 'ChnFtrs' [12]. To deal with such issues, Denœux [8, 9] proposed the cautious rule, which in the case of simple mass functions is defined as

$$\{1\}^{\alpha_1} \owedge \{1\}^{\alpha_2} \quad = \quad \{1\}^{\alpha_1 \wedge \alpha_2}, \quad (16)$$

where $\wedge$ denotes the minimum operator. In practice, the cautious rule simply keeps the most confident mass function. It thus leads to the same results as a NMS procedure.

Quost *et al.* [35] proposed to optimize an operator that generalizes both Dempster's rule and the cautious rule. They used the Frank's family of t-norms, which is defined as

$$\alpha_1 \top_p \alpha_2 \quad = \quad \begin{cases} \alpha_1 \wedge \alpha_2 & \text{if } p = 0, \\ \alpha_1 \alpha_2 & \text{if } p = 1, \\ \log_p \left(1 + \frac{(p^{\alpha_1} - 1)(p^{\alpha_2} - 1)}{p - 1} \right) & \text{otherwise.} \end{cases} \quad (17)$$

For any $p \in [0, 1]$, $\alpha_1 \top_p \alpha_2$ returns a value between $\alpha_1 \alpha_2$ and $\alpha_1 \wedge \alpha_2$. Using this family of t-norms, they finally define the following combination rule:

$$\{1\}^{\alpha_1} \circledT_p \{1\}^{\alpha_2} \quad = \quad \{1\}^{\alpha_1 \top_p \alpha_2}. \quad (18)$$

To choose the value of the parameter p, Quost *et al.* [35] proposed to group the detectors into clusters and optimize the value of p for each cluster independently.

3.3 Clustering detectors

Detectors that return similar mass functions are likely to be using similar information. Their combination should thus be handled more cautiously. To define a measure between classifiers, a distance between mass functions has to be defined first. A survey of such distances can be found in [20]. A commonly used distance measure between two mass functions m_1 and m_2 is defined as

$$d(m_1, m_2) = \sqrt{\frac{1}{2} \sum_{A,B \subseteq \Omega \setminus \{\emptyset\}} \frac{|A \cap B|}{|A \cup B|} (m_1(A) - m_2(A))(m_1(B) - m_2(B))} \tag{19}$$

For two simple mass functions, we get

$$0 \leq d\left(\{1\}^{\alpha_1}, \{1\}^{\alpha_2}\right) = \frac{|\alpha_1 - \alpha_2|}{\sqrt{2}} \leq \frac{1}{\sqrt{2}}. \tag{20}$$

The average distance for all detections is then used as a distance between the detectors $\mathcal{C}_{(k)}$ and $\mathcal{C}_{(\ell)}$:

$$\mathcal{D}(\mathcal{C}_{(k)}, \mathcal{C}_{(\ell)}) = \frac{1}{n} \sum_{i=1}^{n} d\left(m_{(k),i}, m_{(\ell),i}\right), \tag{21}$$

where $m_{(k),i}$ and $m_{(\ell),i}$ refers to the mass functions associated to the i-th BB cluster provided by $\mathcal{C}_{(k)}$ and $\mathcal{C}_{(\ell)}$, respectively. The above definition actually assumes that, for every BB returned by $\mathcal{C}_{(k)}$ there is an associated one returned by $\mathcal{C}_{(\ell)}$. It is actually not the case. When one of the detector does not provide any BB, the distance is set to

$$\frac{1}{\sqrt{2}} \leq d\left(\{1\}^{\alpha_1}, \emptyset\right) = d\left(\{1\}^{\alpha_1}, \{0\}^{0}\right) = \sqrt{\frac{1 + (1 - \alpha_1)^2}{2}} \leq 1. \tag{22}$$

Using this pairwise distance, the detectors can be grouped through hierarchical clustering.

4 Experimental results

We conducted our experiments on the Caltech Pedestrian Detection Benchmark. The dataset consists in six training sets (set00-set05) that have been used to train detectors (see Figure 1 (a)), and five testing sets (set06-set10). For our experiments we kept one of the testing sets (set06) as a validation set for calibration and the remaining four sets were used for testing. A five fold cross-validation step was also conducted on the validation set to tune the different parameters of the combination system. As a performance measure, we used the *log-average miss rate* as proposed in [15].

Figure 4 (a) shows the influence of the threshold used for BBs association. The results were much better when doing the association greedily after score calibration. The best performance was obtained using a_{union} with a threshold of 0.45, although using a_{min} with a threshold of 0.8 gave very close results. We then compared the probabilistic combination rules (Sec. 2.3) with the evidential ones (Sec. 3.2). Figure 4 (b) shows the results obtained from a logistic calibration on the "Reasonable" case scenario. For the weighted version of the combinations, the average precision estimated on the validation set was used as weight. In the evidential method, this weight was used as a discounting factor δ (Eq. 12). We can see that the product and minimum rules performed very poorly. The average rule performed better than the majority vote. The cautious rule, which is equivalent to the maximum rule, performed better than all the other probabilistic rules but worse than Dempster's rule and the t-norm based rule. Using an additional weight led to better results for all combination

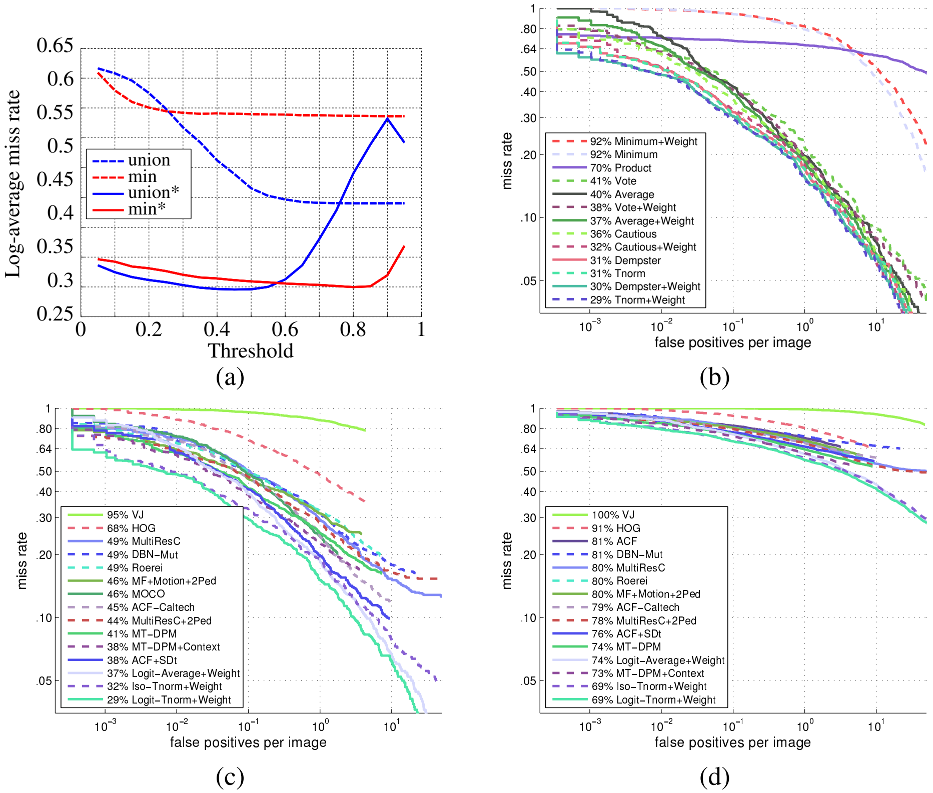


Figure 4: (a) Log-average miss rate for different values of the overlapping threshold. The methods marked with a star correspond to greedy box association after a logistic calibration. (b) Results of different combination strategies using a logistic regression calibration on the "Reasonable" scenario. (c) Results on the "Reasonable" scenario. (d) Results on the "Overall" scenario.

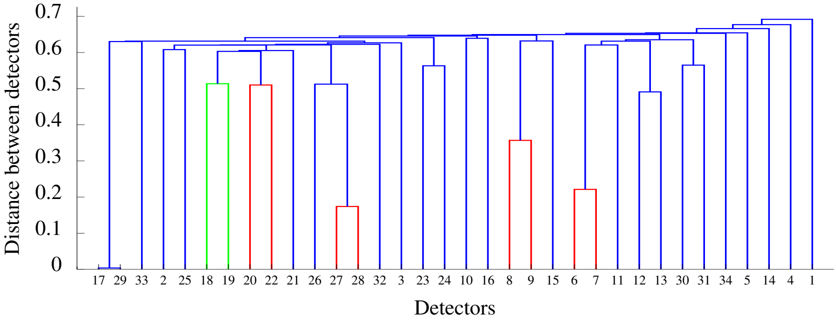


Figure 5: Detectors hierarchical clustering. The colors show how the detectors are to be combined. The blues branches correspond to Dempster's rule, the red ones to the cautious rule and the green ones to a t-norm rule with parameter different from 0 and 1.

methods except the minimum combination rule. Similar conclusions were reached by using an isotonic calibration.

For the t-norm based rule, the detectors were combined following the hierarchical clustering shown on Figure 5. For each pairwise combination the parameter of the t-norm was computed from the validation set. For most pairs of clusters, the best results were obtained using Dempster's rule ($p = 1$). The detectors #7 and #26 are, respectively, the combination of #6 and #27 with a 2-pedestrian detector while #9 uses a car detector with #8. For those three pairs, the cautious rule ($p = 0$) was optimal. The only case where the t-norm parameter was different from 0 and 1 was the combination between 'ChnFtrs' and 'FPDW'. The relatively high diversity of the evaluated detectors explains the limited gain from the t-norm rule compared to Dempster's rule.

Figure 4 (c-d) compares the 12 best detectors, including 'VJ' and 'HOG', to the logistic and isotonic weighted t-norm and the logistic weighted average. In the "Reasonable" scenario, the logistic weighted t-norm led to an improvement of 9% in terms of log-average miss rate and 6% for the isotonic one. The weighted average only led to 1% improvement. In the "Overall" scenario, the logistic and isotonic t-norm have very similar results with a performance improvement of 4% while the weighted average performed worse than the 'MT-DPM+Context' alone. Results for the other scenarios, details of the isotonic combinations, detection examples and source code are supplied as supplementary materials on the authors' website[1].

5 Conclusions and perspectives

In this paper, we proposed and evaluated an evidential framework for combining pedestrian detectors, noticing it could also be applied directly to detect other classes of objects. The use of belief functions and evidential combination rules yielded much better results than classical probabilistic approaches. One novelty of our approach relies on the use of an optimized t-norm rule, which can take into account the dependencies between detectors. This property can become critical if many new detectors are to be added. As optimizing pairwise rules may provide only sub-optimal results, a global optimization will be investigated in future work. An important advantage of the proposed approach is that it allows us to easily include a new detector regardless of the features, training data and classifier it uses. Moreover, this modularity allows new detectors to rely on existing state-of-the-art ones. Therefore one may focus future research on the development of detectors specially designed to detect *hard* examples without risking an overall recall loss.

Acknowledgments

This research was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). It was supported by the ANR-NSFC Sino-French PRETIV project (Reference ANR-11-IS03-0001).

[1] https://www.hds.utc.fr/~xuphilip/dokuwiki/en/data

References

[1] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 5:641–647, 1955.

[2] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg. Part-based feature synthesis for human detection. In *Proceedings of the European Conference on Compute Vision*, pages 4.5–4.11, Crete, Greece, 2010.

[3] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. Calibration of machine learning models. In Emilio Soria Olivas, José David Martín Guerrero, Marcelino Martinez-Sober, Jose Rafael Magdalena-Benedito, and Antonio José Serrano López, editors, *Handbook of Research on Machine Learning Applications and Trends*, pages 128–146. IGI Global, 2009.

[4] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. On the effect of calibration in classifier combination. *Applied Intelligence*, 38(4):566–585, 2013.

[5] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3666–3673, Portland, USA, 2013.

[6] G. Chen, Y. Ding, J. Xiao, and T. Han. Detection evolution with multi-order contextual co-occurrence. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1805, Portland, USA, 2013.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, San Diego, USA, 2005.

[8] T. Denœux. The cautious rule of combination for belief functions and some extensions. In *Proceedings of the 9th International Conference on Information Fusion*, pages 1–8, Florence, Italy, 2006.

[9] T. Denoeux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

[10] T. Denoeux, N. El Zoghby, V. Cherfaoui, and A. Jouglet. Optimal object association in the Dempster-Shafer framework. *IEEE Transactions on Cybernetics*, 2014. (accepted for publication), doi:10.1109/TCYB.2014.2309632.

[11] P. Dollár, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[12] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *Proceedings of the British Machine Vision Conference*, pages 91.1–91.11, London, England, 2009.

[13] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *Proceedings of the British Machine Vision Conference*, pages 68.1–68.11, Aberystwyth, Wales, 2010.

[14] P. Dollár, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *Proceedings of the European Conference on Compute Vision*, pages 645–659, Florence, Italy, 2012.

[15] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[16] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, USA, 2008.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[19] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[20] A. Jousselme and P. Maupin. Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, 53(2):118–145, 2012.

[21] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14:28–44, 2013.

[22] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[23] D. Levi, S. Silberstein, and A. Bar-Hillel. Fast multiple-part based object detection using kd-ferns. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–954, Portland, USA, 2013.

[24] Z. Lin and L. Davis. A pose-invariant descriptor for human detection and segmentation. In *Proceedings of the European Conference on Compute Vision*, pages 423–436, Marseille, France, 2008.

[25] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, USA, 2008.

[26] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837, 2005.

[27] W. Nam, B. Han, and J. H. Han. Improving object localization using macrofeature layout selection. In *ICCV Workshop on Visual Surveillance*, Barcelona, Spain, 2011.

[28] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3258–3265, Providence, USA, 2012.

[29] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3198–3205, Portland, USA, 2013.

[30] W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship with a deep model in pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3222–3229, Portland, USA, 2013.

[31] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Efficient pedestrian detection by directly optimize the partial area under the roc curve. In *Proceedings of the IEEE International Conference on Compute Vision*, pages 1057–1064, Sydney, Australia, 2013.

[32] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *Proceedings of the European Conference on Compute Vision*, pages 241–254, Crete, Greece, 2010.

[33] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár. Exploring weak stabilization for motion feature extraction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2882–2889, Portland, USA, 2013.

[34] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, 1999.

[35] B. Quost, M.-H. Masson, and T. Denœux. Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. *International Journal of Approximate Reasoning*, 52(3):353–374, 2011.

[36] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[37] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *Proceedings of the IEEE International Conference on Compute Vision*, pages 24–31, Kyoto, Japan, 2009.

[38] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, Portland, USA, 2013.

[39] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[40] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2012.

[41] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13–18, San Francisco, USA, 2010.

[42] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *Proceedings of the IEEE International Conference on Compute Vision*, pages 32–39, 2009.

[43] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. In *DAGM Symposium Pattern Recognition*, pages 82–91, Munich, Germany, 2008.

[44] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 794–801, Miami, USA, 2009.

[45] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li. Robust multi-resolution pedestrian detection in traffic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3033–3040, Portland, USA, 2013.

[46] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, New York, USA, 2002. ACM.