



Belief Function Definition for Ensemble Methods - Application to Pedestrian Detection in Dense Crowds

Jennifer Vandoni, Sylvie Le Hégarat-Mascle, Emanuel Aldea

► To cite this version:

Jennifer Vandoni, Sylvie Le Hégarat-Mascle, Emanuel Aldea. Belief Function Definition for Ensemble Methods - Application to Pedestrian Detection in Dense Crowds. 21st International Conference on Information Fusion (FUSION), Jul 2018, Cambridge, United Kingdom. 10.23919/icip.2018.8455313 . hal-01790288

HAL Id: hal-01790288

<https://hal.science/hal-01790288>

Submitted on 11 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Belief Function Definition for Ensemble Methods - Application to Pedestrian Detection in Dense Crowds

Jennifer Vandoni, Sylvie Le Hégarat-Masclé, Emanuel Aldea
SATIE Laboratory, CNRS UMR 8029
Paris-Sud University, Paris-Saclay University, Orsay 91405, France
Email: {jennifer.vandoni,sylvie.le-hegarat,emanuel.aldea}@u-psud.fr

Abstract—Large scale social events are characterized by very high densities (at least locally) and an increased risk of congestions and fatal accidents. Our work focuses on the specific problem of pedestrian detection in high-density crowd images, denoted by strong homogeneity and clutter. We propose and compare different evidential fusion algorithms which are able to exploit multiple detectors based on different gradient, texture and orientation descriptors. The evidential framework allows us to model spatial imprecision arising from each of the detectors, both in the calibration and in the spatial domains. Moreover, we propose a *bba* allocation that takes into account both types of imprecision. Results on difficult high-density crowd images acquired at Makkah during the Muslim pilgrimage show that the proposed combined fusion algorithm leads to better results than taking into account only individual sources of imprecision.

I. INTRODUCTION

For video surveillance, the automatic detection of pedestrians is a fundamental task which is directly related to applications such as tracking or action recognition. Recently, the accurate detection of pedestrians in high-density scenes gained traction due to the increased frequency of large scale social events, and due to the safety risks linked to them. Although a significant effort has been devoted in the last decade to pedestrian detection, the advances proposed in the literature are not always applicable to crowd detections for multiple reasons such as the absence of background, the heavy occlusion of body parts, the high visual homogeneity and the small size of the targets. It becomes therefore essential to rely on multiple independent visual detectors which are able to provide different interpretations of the input data. A global decision must then be made by taking into account the partial verdicts provided by the individual detectors. Classifier fusion is a well known problem in artificial intelligence, and for the specific task of pedestrian detection a number of solutions have been proposed over the years. Again, fusion strategies exhibiting a good performance for standard pedestrian detection may behave differently in a crowded context where individual detectors operate poorly.

In this study, we propose to use the belief function framework [1] to perform fusion between different SVM-based pedestrian detectors. The evidential framework is indeed able to naturally take into account the concept of imprecision, that in our case can arise in two different and complementary ways: firstly, in the derivation of posterior probability values from SVM decision scores, and later, from the spatial layout of the

detections in the output image space. We aim therefore at comparing two fusion solutions based on these different definitions of imprecision, and to propose a valid *bba* allocation that can take into account both of them. The experiments show that the proposed combined fusion algorithm outperforms the approaches considering only individual sources of imprecision.

II. RELATED WORK

For the detection task in high-density crowds, it is not immediately clear which detectors are the most adapted or discriminative, and which fusion strategy is the most effective.

A. Detectors

In the difficult context of high-density crowds, simple solutions relying on appearance cues, such as local color histogram which may be associated to skin, hair or clothes are not well suited. These approaches are indeed limited by multiple factors: the object resolution needs to be relatively high, the color spaces are not discriminative enough, and lastly many surveillance cameras provide gray level data. Likewise, common face detectors such as Viola-Jones [2] are unsuited, since pedestrian faces are not detailed enough.

Among some descriptors which are widely used in pedestrian detection, those which are the best suited for detection in high-density crowds have been recently highlighted in [3].

Related to the image gradient, the Histogram of Oriented Gradients (HOG) descriptor [4] is very popular and has exhibited in various contexts excellent performances. The contour related to the specific shape of the head and shoulders is indeed highly discriminative, but it may fade away due to clutter. For this reason, it is important to consider as well some descriptors aimed at other characteristics than shape.

Traditionally employed in texture classification, the Local Binary Pattern (LBP) operator [5] has been successfully used in pedestrian detection due to its reasonable robustness to occlusion provided by its local sampling strategy. Some alternative solutions are the covariance matrix based descriptors [6], but at the expense of low compactness and higher computational cost. Also related to texture representation, Gabor filter banks have been used for head detection [7] to encode the local frequencies and orientations.

Popular particularly in the field of stereo matching, the DAISY [8] descriptor has been successfully employed for the first time for head detection in difficult crowd scenes in [3].

Its Gaussian smoothing, along with the sampling overlap, naturally enforces spatial consistency.

B. Fusion

In order to perform the fusion of detectors based on different features, there exist in the literature various approaches depending on the considered problem. Let us mention the more adapted to our task, namely pedestrian detection.

In order to avoid fusing different detectors altogether, a straightforward solution is to simply concatenate all the available representations before classification. The fundamental limitations of this solution are the suboptimal projection of all the features into a common discriminative space that may be not well adapted, and the risk of overfitting due to the length of the final descriptor with respect to the number of available training samples.

Another popular simple strategy is to use a cascade of detectors, where the first detector has a low computational cost and a low false negative rate, while the subsequent detection is costlier and more accurate [6]. Alternatively, a hybrid detector may be built by switching entirely among more available detectors based on the current estimated difficulty of the scene [9]. In the context of online multi-target tracking, [10] proposes to exploit a batch of classifiers and obtain strong and weak detections on the basis of a confidence score. Although these strategies are beneficial for real-time performance, they are not well suited for difficult settings such as high-density crowds, as the individual classifiers exhibit rather modest performance, and the final decision does not necessarily exploit all the available information.

In order to benefit simultaneously from all the available features, AdaBoost remains very popular for application with real-time constraints and for architectures with limited power. However, nonlinear classification exploits better the available data [11], [12], with a drawback of higher computational cost.

Alternatively, multiple kernel learning (MKL) is a well established methodology which aims to combine different kernels relying on different data representations as a linear combination, by casting this information fusion task as a convex optimization problem [13]. The problem scales very well with the number of individual classifiers, but the main limitation of MKL is the difficulty to interpret the final decision and to take into account the imprecision coming from the different sources.

Another established framework able to benefit from the information provided by multiple features is the decision tree analysis. Recent work highlighted that intrinsic uncertainty related to learning as well as uncertainty due to imprecise data may be jointly managed inside the decision tree by defining entropy intervals from evidential likelihood [14].

Fusion of the classifiers may be performed also by recovering their output in a probabilistic form, and then applying some combination rules which are more or less ad-hoc, such as a product, or a pseudo-likelihood as the averaged sum of the individual likelihoods over the detectors [15] in order to cope better with individual missed detections. In the context

of SVM binary classification, posterior probabilities from confidence scores are usually obtained with a calibration process based on logistic regression [16].

One limitation of probabilistic calibration methods however is that the uncertainty due to the number of samples in the calibration set is not taken into account. For this reason in [17] the authors propose a calibration process based on the evidence theory, a popular framework for reasoning under uncertainty, that can represent more accurately the imprecision of the calibration procedure especially in presence of few data. This evidential calibration has been already proved to be effective in many fields, from multi-class classification with hyperspectral data [18] to pedestrian detection [19]. In this latter work however, the considered scenes are not very dense and the pedestrians are entirely visible, so that the detection is done at object-level and the fusion is performed among the bounding boxes provided by each detector. The authors of [20] extended the method to handle sources providing pixel-level information, for the specific application of face blurring, but also in this case the considered data have high resolution with respect to crowded scenes. Thus, the integration into the global system of reliable detectors such as the ones based on skin hue becomes possible and lessens the necessity of a complex fusion algorithm.

In the context of high-density crowd pedestrian detection, in [3] we propose a robust fusion strategy also based on the belief functions framework, that is able to take into account the spatial imprecision of each different classifier, and we show the improved performance of this method with respect to MKL and the straightforward product of probabilities. Spatial imprecision is introduced following the work of [21], that allows to define a *valid bba* using morphological operators, exploiting the duality property between erosion and dilation (respectively opening and closing) and *Bel* and *Pl* values for the final *bba*.

This work aims at comparing this latter fusion strategy with a similar approach applied in the score space instead (i.e. where SVM scores are projected with respect to their label and their distance to the hyperplane boundary in the feature space), at the moment of the derivation of the initial mass functions, where imprecision can arise from possible errors in the parameters estimation during the calibration step. Moreover, the proposed *bba* allocation procedure combines the two methods, aiming to be more robust to possible mislocation of the calibration functions from which the mapping from SVM scores to probability values is made, while at the same time taking into account the information coming from neighbors pixels in the image space.

III. OUR METHOD

A. Belief functions defined from mathematical morphology

To handle both uncertainty and imprecision, belief functions are defined on a larger hypothesis set than in the case of the probabilistic framework. Specifically, if Θ denotes the discernment frame, i.e. the set of mutually exclusive hypotheses, belief functions are defined on the set of the subsets of Θ ,

noted 2^Θ in reference to its number of elements: $2^{|\Theta|}$ where $|\Theta|$ is the cardinality of Θ .

In our case, denoting by H and $\bar{H}$ the two singleton hypotheses, *head* and *not_head*, $\Theta = \{H, \bar{H}\}$, and the set of hypotheses is:

$$2^\Theta = \{\emptyset, H, \bar{H}, \{H, \bar{H}\}\}. \quad (1)$$

Classically, the *mass* function noted m is the *basic belief assignment* (bba) that satisfies $\forall A \in 2^\Theta, m(A) \in [0, 1]$, $\sum_{A \in 2^\Theta} m(A) = 1$. The hypotheses for which the mass function is non null are called *focal elements*. Then, other belief functions are used either for decision, namely the *plausibility* and the *credibility* functions noted Pl and Bel respectively, or for some computations. In this particular setting, in which we have only two singleton hypotheses and $m(\emptyset) = 0$, they are defined as:

$$Bel(A) = m(A), \forall A \in \{H, \bar{H}\}, \quad (2)$$

$$Pl(A) = m(A) + m(\Theta), \forall A \in \{H, \bar{H}\}. \quad (3)$$

It is important to notice that Pl and Bel functions may also be interpreted as upper and lower probabilities [1] and they check the duality property: $\forall A \in 2^\Theta, Pl(A) = 1 - Bel(\bar{A})$ (where $\bar{A}$ denotes the complement of A with respect to Θ).

As pointed out in [21], (fuzzy) erosion and dilation (respectively opening and closing) are also dual with respect to complementation, and they can be interpreted as belief and plausibility functions. The imprecision is modeled through a structuring element and, using a *discounting* operator, the mass on Θ can be directly reinforced according to the discounting factor.

Let us define an initial bba m_0 derived from the output of a classifier. Then, the following property holds:

$$Pl(A) = 1 - Bel(\bar{A}) \leftrightarrow \delta_v(m_0(A)) = 1 - \mathcal{E}_v(m_0(\bar{A})), \quad (4)$$

$\forall A \in 2^\Theta$, where δ_v and $\mathcal{E}_v$ are the dilation and erosion operators respectively, with structuring element v .

B. BBA allocation based on calibrated scores

Given training samples $x_j \in \mathbb{R}^n$, $j = 1, \dots, l$, labeled by $y_j \in \{+1, -1\}$, a binary Support Vector Machine (SVM) computes a decision function $f(x)$ such that $sign(f(x))$ is used to predict the label of unseen test samples. In order to obtain class probability $P(y = 1|x)$, a well established method proposed by Platt [16] approximates the posterior probability by learning a logistic sigmoid function

$$P(y = 1|x) \approx Sigm_{\lambda_0, \lambda_1}(f) = \frac{1}{1 + e^{\lambda_0 f + \lambda_1}}. \quad (5)$$

The optimal parameter configuration $(\lambda_0^*, \lambda_1^*)$ is then determined by solving a regularized maximum likelihood problem, with respect to a calibration set independent from the training data.

However, in difficult settings such as our application, a robust estimation of the sigmoid parameters is almost impossible to achieve, and few changes in the calibration set (cardinality or in the samples within it) can cause the sigmoid to appear very different. In presence of a steep transition between the two classes particularly, even a slight shift of the sigmoid may induce different decisions for quite numerous samples, especially in presence of strong overlap between the two classes.

Now, belief functions are a natural framework to take into account the imprecision inherent to the sigmoid learning process. Instead of deriving a simple probabilistic value through logistic regression, we aim at associating a bba to each test sample directly from its score and from the estimated sigmoid (from calibration process). Then we propose to apply mathematical morphology-based discounting to derive the imprecision value, like in [21]. This allows us to define two new sigmoid functions corresponding to lower and upper bounds of probability with respect to the learned sigmoid, that can be interpreted as Bel and Pl functions.

For each different test sample, given its score s_i , namely its distance to the hyperplane boundary defined by classifier i , with $i = 1 \dots N$, we define an associated bayesian bba $m_0^{s,i}$ (i.e. bba having only singleton focal elements), from posterior probability given by calibration step:

$$\begin{aligned} m_0^{s,i}(H) &= Sigm_{\lambda_0^*, \lambda_1^*}(s_i), \\ m_0^{s,i}(\bar{H}) &= 1 - Sigm_{\lambda_0^*, \lambda_1^*}(s_i), \\ m_0^{s,i}(\Theta) &= 0, \\ m_0^{s,i}(\emptyset) &= 0. \end{aligned} \quad (6)$$

This initial bayesian bba is only able to model the uncertainty about the class the sample belongs to, so that the imprecision modeling is introduced as follows. Let us consider a flat structuring element of width w . Then, erosion and dilation (respectively opening and closing) morphological operators are applied to the $m_0^{s,i}$ value of the considered focal element in order to get Pl and Bel values for the target bba:

$$\begin{aligned} m_1^{s,i}(A) &= \mathcal{E}_w(m_0^{s,i}(A)), \forall A \in \{H, \bar{H}\}, \\ m_1^{s,i}(\Theta) &= 1 - m_1^{s,i}(H) - m_1^{s,i}(\bar{H}), \\ m_1^{s,i}(\emptyset) &= 0, \end{aligned} \quad (7)$$

where $\mathcal{E}_w$ is the erosion operator with a structuring element w , applied in the score space (i.e. the 2D space where the sigmoid function is learned, where the scores are projected with respect to their label and their distance to the hyperplane boundary in the feature space). Due to the fact that we consider a flat structuring element, and to the intrinsic monotonic increasing behavior of the sigmoid function,

$$Bel^{s,i}(H) = Sigm_{\lambda_0^*, \lambda_1^*}(s_i - \frac{w}{2}), \quad (8)$$

$$Pl^{s,i}(H) = Sigm_{\lambda_0^*, \lambda_1^*}(s_i + \frac{w}{2}). \quad (9)$$

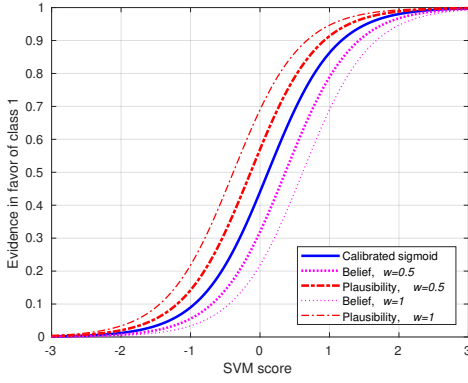


Fig. 1. Example of a sigmoid function obtained with calibration, and derived Belief and Plausibility bounds at different structuring element w sizes.

Figure 1 shows an example of a sigmoid function learned on the calibration set, as well as the two derived sigmoid functions (for two different given structuring elements w), that represent *Bel* and *Pl* functions and provide the interval of imprecision.

The interval between *Bel* and *Pl* functions embeds thus the amount of imprecision in the calibration step we have to cope with. It takes low values for points far from the hyperplane boundary for which the decision is already pretty sure, whereas on the contrary it takes high values in the area near to the hyperplane boundary, where even a little difference in the parameters of the sigmoid can change the decision.

This approach is similar to the calibration process proposed by [17], in the sense that both works aim at defining bbas from SVM scores, but the difference resides in the modeling of what we consider as *imprecision*. Indeed, in [17] the imprecision comes from a low number of samples per score, whereas in our case we want to model the fact that the calibration function may be not perfectly fitted due to the difficulty in the definition of a robust calibration set. Besides, the calibration in [17] is not directly applicable to our problem for two main reasons. Firstly, it is useless (i.e. the Θ mass is almost null) when the number of samples per score is high, and we have indeed a consistent number of calibration samples at our disposal. Secondly, in difficult settings such as our application, where it is hard for SVM to find a very large margin between the two classes, there can be a consistent overlap between samples with different labels for the same score. However, since the number of samples per score would be high, we would paradoxically not assign a high value of imprecision to them. With our proposed method instead, we are able to allocate higher values of imprecision to the samples having their correspondent score within the SVM margin, in the overlapping area.

C. BBA definition based on pixel neighborhood information

The previous modeled imprecision was derived from information obtained during the calibration step (that determines the sigmoid function). However, it is quite usual, in case of an image data, to derive information also from the neighbors of the considered pixel. Therefore, as our application is a

classification problem based on images, we aim at taking into account also the imprecision detected by a spatial analysis. In the context of high-density crowd pedestrian detection, strong occlusions make the head of each pedestrian barely visible. Besides, due to the specific geometry of the recordings, each head corresponds to few pixels. The most effective head detectors are based on features computed in sub-windows around the pixel of interest, which further increases the spatial imprecision of the detection.

For this reason, following the preliminary work we introduced in [3], we use bbas to model the spatial imprecision due to the close resolutions of object (head) and descriptor respectively. Applied in the spatial domain, the work of [21] allows indeed to associate to every pixel p of the image a bba $m_2^{p,i}$, for each classifier i .

Also in this case initial bbas $m_0^{p,i}$ for each detector i are bayesian, defined from posterior probabilities after the calibration step, so that initial bbas have only two focal elements, and $m_0^{p,i} = m_0^{s,i}$, with s being the score associated to pixel p given by descriptor i .

Now, to introduce spatial imprecision, we derive bbas performing erosion and dilation (respectively opening and closing), but this time in the image domain, that is to say taking into account neighbor pixels based on the assumption that they are likely to belong to the same class. The bba allocation in this case would be:

$$\begin{aligned} m_2^{p,i}(A) &= \gamma_a \left(m_0^{p,i}(A) \right), \forall A \in \{H, \bar{H}\}, \\ m_2^{p,i}(\Theta) &= 1 - m_2^{p,i}(H) - m_2^{p,i}(\bar{H}), \\ m_2^{p,i}(\emptyset) &= 0, \end{aligned} \quad (10)$$

where γ_a is the opening operator of parameter a . As in [3], a spatial Gaussian structuring element fitted in a window of radius a is considered, to better take into account the spatial consistency.

This latter bba allocation allowing to derive $m_2^{p,i}$ is completely independent from the first bba allocation $m_1^{s,i}$, although they both take into account possible sources of spatial imprecision, in the location of the estimated sigmoid function or in the image domain.

At this point, a straightforward solution could be to simply combine all the obtained sources $m_1^{s,i}$ and $m_2^{p,i}$ for each test sample of the image. A drawback of this approach is that even if the total ignorance is globally reduced thanks to the combination, strong conflict may arise due to a lack in modeling a part of the imprecision for each bba (either the calibration one or the spatial one) and due to the important number of bbas involved in the final combination.

For this reason we propose a third bba definition, that is naturally able to take into account both types of imprecision in bba allocation. Firstly, we define bbas following $m_1^{s,i}$, in order to take into account the imprecision due to possible errors in the calibration. Then, we increase the mass on Θ discounting previous bbas by performing opening in the image space, decreasing the mass on the singleton hypotheses and

increasing the mass on Θ accordingly. The difference here is that the initial bba is not bayesian, but we start already from less committed bbas that can have a mass also on the compound hypothesis Θ :

$$\begin{aligned} \forall A \in \{H, \bar{H}\}, \\ m_3^{s,p,i}(A) &= \gamma_a \left(m_1^{s,i}(A) \right), \\ &= \gamma_a \left(\mathcal{E}_w \left(m_0^{s,i}(A) \right) \right), \\ m_3^{s,p,i}(\Theta) &= m_1^{s,i}(\Theta) + \left(1 - m_1^{s,i}(\Theta) + \right. \\ &\quad \left. - m_3^{s,p,i}(H) - m_3^{s,p,i}(\bar{H}) \right), \\ &= 1 - m_3^{s,p,i}(H) - m_3^{s,p,i}(\bar{H}), \\ m_3^{s,p,i}(\emptyset) &= 0, \end{aligned} \quad (11)$$

where $\mathcal{E}_w$ is the erosion operator with a structuring element w applied in the score space, while γ_a is the opening operator of parameter a applied in the image space. These two morphological operations are clearly not commutative, and we find it more natural to firstly consider the imprecision due to the calibration step and later consider the imprecision in the spatial context. Finally, we note that $m_1^{s,i}$ and $m_2^{p,i}$ allocations are two specific cases of the proposed $m_3^{s,p,i}$ bba definition, namely in presence of null-size structuring elements a or w respectively.

D. BBAs combination

Considering the N descriptors in every pixel p , N bbas are defined that represent the soft output of each of the N binary classifiers. According to bba associated to descriptor i , the uncertainty of a head presence in a pixel p ranges between $Bel^{s,p,i}(H) = m_3^{s,p,i}(H)$ and $Pl^{s,p,i}(H) = m_3^{s,p,i}(H) + m_3^{s,p,i}(\Theta)$, so that $m_3^{s,p,i}(\Theta)$ represents the imprecision on the uncertainty value provided by i^{th} descriptor at pixel location p . In the proposed model, the uncertainty comes from the binary classifier score whereas the imprecision comes both from the initial score calibration and from spatial heterogeneity of uncertainty values within the considered structuring element.

Lastly, the combination between bbas can be performed. As the descriptors are considered *cognitively* independent, the orthogonal sum or its unnormalized version, the conjunctive combination rule [22], are well-suited for this task. For two sources m^1 and m^2 , the conjunctive combination rule is defines as

$$\forall A \in 2^\Theta, m^{1 \odot 2}(A) = \sum_{\substack{(B,C) \in 2^\Theta \times 2^\Theta, \\ B \cap C = A}} m^1(B) m^2(C). \quad (12)$$

In our case where $|\Theta| = 2$, and considering $m_3^{s,p,i}$ bbas

allocation, the analytical result may be easily derived:

$$\begin{aligned} \forall A \in \{H, \bar{H}\}, \\ m^{\odot 1^N}(A) &= \sum_{\substack{(B_1, \dots, B_N) \in \{A, \Theta\}^N, \\ \exists i \in [1, N] \text{ s.t. } B_i = A}} \prod_{i=1}^N m_3^{s,p,i}(B_i), \\ m^{\odot 1^N}(\Theta) &= \prod_{i=1}^N m_3^{s,p,i}(\Theta), \\ m^{\odot 1^N}(\emptyset) &= 1 - m^{\odot 1^N}(H) - m^{\odot 1^N}(\bar{H}) + \\ &\quad - m^{\odot 1^N}(\Theta). \end{aligned} \quad (13)$$

Finally, in every pixel, the decision is taken from $m^{\odot 1^N}$. Several rules have been proposed in the literature. Most popular ones only consider singleton hypotheses (in order to avoid ambiguous decision) and are based on functions that have a probabilistic interpretation: maximum of plausibility, credibility, or pignistic probability [22].

Pignistic probability in particular can be used to give a probabilistic interpretation to the bbas. Since in our setting $|\Theta| = 2$, $\forall A \in \Theta$

$$BetP(A) = \frac{1}{1 - m^{\odot 1^N}(\emptyset)} \cdot \left(m^{\odot 1^N}(A) + \frac{m^{\odot 1^N}(\Theta)}{2} \right). \quad (14)$$

This allows us to assign a probabilistic interpretation to the resulting bba associated to each pixel p , so that at every pixel its $BetP(H)$ value will be differently normalized on the basis of its conflict value, represented by the mass on the empty set. Then, from the $BetP(H)$ map we obtain, we will derive statistics for the quantitative evaluation of the detection results.

IV. EXPERIMENTS

For the experiments we rely on SVM, using the descriptors (with adapted kernels) pointed out by [3], already proven to be effective in the hard setting of high-density crowds, namely HOG [4], LBP [5], DAISY [8] and Gabor filter banks [7]. We tested our proposed fusion method on difficult high-density crowd images acquired at Makkah during Hajj [3], [23], [24].

Figure 2 shows the sigmoid functions obtained with the calibration step, for each descriptor considered, and the Bel and Pl functions that define the interval obtained by erosion using a structuring element of size $w = 0.5$, with the first proposed bba allocation $m_1^{s,i}$ (in our case, $i = 1 \dots 4$). Two important considerations can be made. Firstly, it becomes clear that with a larger structuring element we introduce more imprecision, allowing for less committed bbas. On the contrary, a smaller structuring element introduces less imprecision. Secondly, considering the same size of the structuring element w , the interval between Bel and Pl functions, namely the imprecision we introduced, is bigger in presence of the steeper sigmoid functions like HOG and Daisy ones, for which even a small shift in the location would possibly cause the final decision to be different for the same score, resulting in less committed bbas associated to each possible score. On

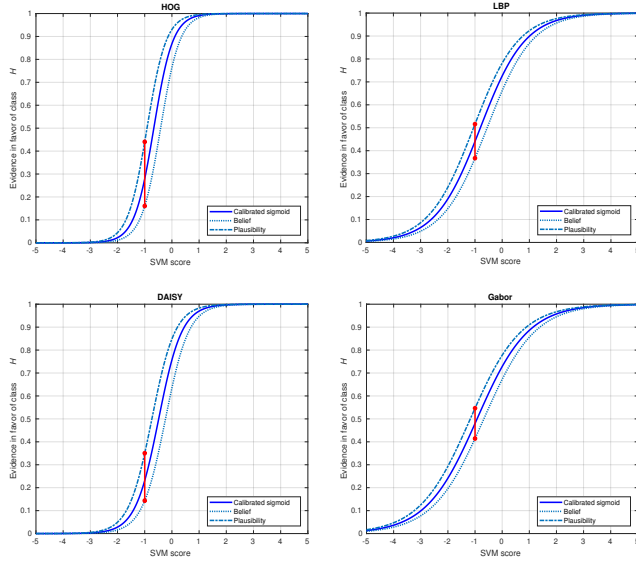


Fig. 2. Sigmoid functions obtained with the calibration step and associated *Bel* and *Pl* sigmoids, with $w = 0.5$ size of structuring element. In red: example of the imprecision interval at $\text{SVM score} = -1$.

the contrary, scores from LBP and Gabor sigmoid functions have higher absolute values and less overlap, meaning that the decision about them is pretty stable, resulting in a smaller imprecision interval and thus in more committed bbas.

Considering the effective choice of w value, we noticed that increasing the size of the structuring element w , the precision in the fusion results consistently increases, stressing the importance of the introduction of imprecision during calibration. However, increasing too much the size of the structuring element is detrimental for the detections, because they become too much uncertain and thus recall tends to decrease. For the experiments, we set $w = 2$, a good compromise that allows us to reach high precision without harming the overall recall.

Figure 3 shows an example patch from an image of the dataset with its relative ground-truth, where heads are highlighted in green. Then, for each different descriptor, $m_3^{s,p,i}$ bbas allocation is shown through the mass on the head hypothesis and total ignorance arising after discounting both in the calibration and image space. The last row presents the results after the conjunctive combination rule. We notice that each source has a specific behavior, which underlines their complementarity. HOG and Gabor provide more localized detections, visible in the mass on H hypothesis, but more noise is present. On the contrary, LBP and DAISY provide larger and rougher results. Each descriptor then has higher values of ignorance in the pixels corresponding to the border of the heads, since their neighborhood is not homogeneous in the image space, and their correspondent score is probably in the area of uncertainty during calibration. With the conjunctive combination rule however, we are able to consistently reduce the total ignorance as shown in Figure 3(l), and the shape of the head detections becomes very clear, as depicted in Figure 3(k).

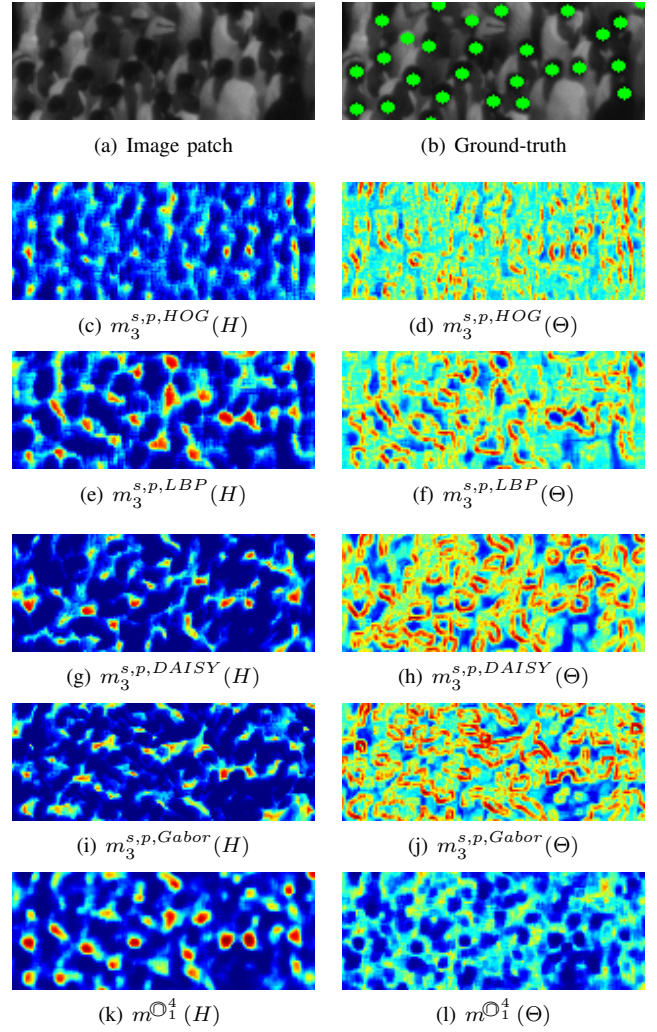
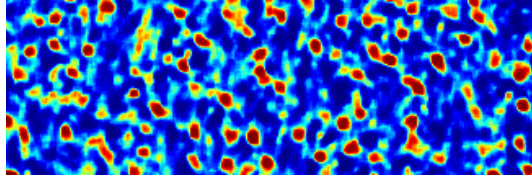
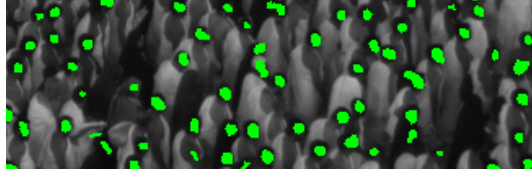


Fig. 3. Sigmoid functions obtained with the calibration step and associated *Bel* and *Pl* sigmoids, at different structuring element w size. In red: example of the imprecision interval for $w = 0.5$ at $\text{SVM score} = -1$.

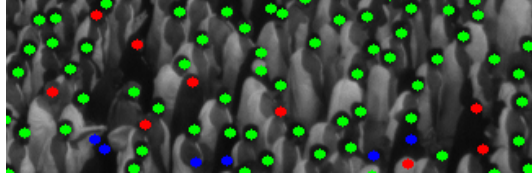
Figure 4 shows an example of the classification result on the basis of $\text{BetP}(H)$ value at every pixel obtained with the conjunctive combination rule after $m_3^{s,p,i}$ allocations, both in terms of colormap and detections at a reasonable threshold $th = 0.8$. This particular threshold choice has been made in order to be able to recover the most confident detections while at the same time keeping them localized at the center of the head. Nevertheless, the learning process works at pixel level, with a balanced training set, while in the testing image the number of pixels corresponding to the two classes is unbalanced, hence the need for a quite high threshold. Results after non maxima suppression (NMS) are then presented in Figure 4(c), setting the radius of a head $r = 3$, with $2r + 1$ minimum distance between two maxima (head centers) in order to avoid overlapping detections, highlighting in green True Positives (TPs), in red False Negatives (FNs) and in blue False Positives (FPs). Most of the heads are correctly detected even in this condition of extreme density, while the number of false detections is kept low. False negative heads can be



(a) $BetP(H)$ map after $m_3^{s,p,i}$ allocations.

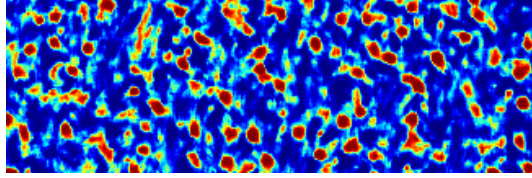


(b) Detection at $th = 0.8$.

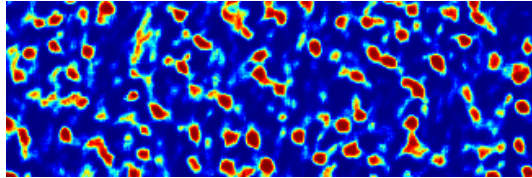


(c) Non-maxima suppression.

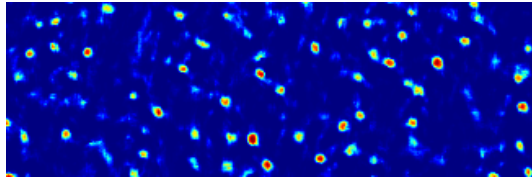
Fig. 4. Fusion results after $m_3^{s,p,i}$ allocations, in terms of (a) $BetP(H)$ colormap, (b) detections at a given threshold and (c) non-maxima suppression.



(a) $BetP(H)$ map after $m_1^{s,i}$ allocations.



(b) $BetP(H)$ map after $m_2^{p,i}$ allocations.



(c) Product of probabilities map.

Fig. 5. Comparison between fusion results after $m_1^{s,i}$ and $m_2^{p,i}$ allocations in terms of $BetP(H)$ colormap, and simple product of probabilities.

explained by the presence of dark heads or low contrast at the border.

Figure 5(a) and 5(b) provide a visual comparison of the fusion results obtained taking into account spatial imprecision

in the calibration and in the image domain *separately*, namely using $m_1^{s,i}$ and $m_2^{p,i}$ allocations respectively. The detections obtained with the two approaches are similar in their locations, but are a bit larger taking into account imprecision during the calibration, while they are spatially more consistent considering imprecision in the spatial domain. Problematic areas are in both cases mostly at the boundary of the detections, that corresponds to pixels having their related score at lower distance from the hyperplane in the first case, and to pixels on which neighborhood disagrees the most in the second case. Considering the proposed $m_3^{s,p,i}$ bba definition, whose result has been previously shown in Figure 4, we are able to take the best out of the two approaches, obtaining larger while at the same time spatially homogeneous detections.

Figure 5(c) provides a visual comparison with a straightforward fusion solution which simply performs the product of the probabilities given by each independent detector at every pixel, without considering imprecision from calibration nor neighborhood information. There are just few heads for which the detection is pretty sure, and the size of the detections is always underestimated, since to have a confident detection all the sources must agree. Instead, taking into account possible sources of imprecision as proposed, we obtain more committed and smoother detections, that can be more useful as starting point for later stages such as tracking applications.

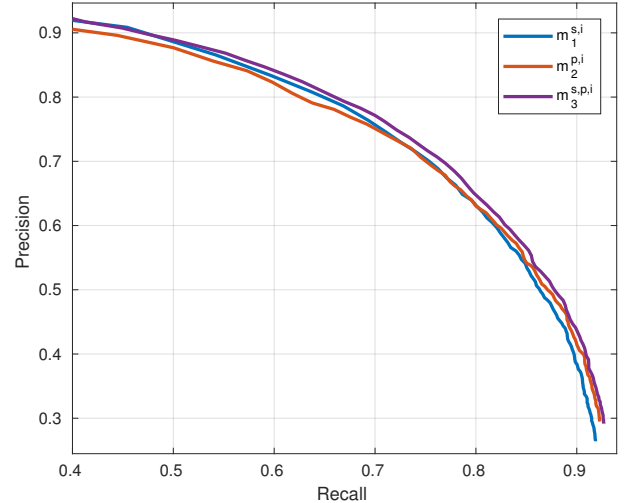


Fig. 6. Fusion results in terms of PR-curves, after conjunctive combination rules with the three investigated bba allocations.

A quantitative study using the three different proposed approaches separately is given by Figure 6, where PR-curves are derived after non-maxima suppression. A flat structuring element of size $w = 2$ is used for $m_1^{s,i}$, while a spatial Gaussian structuring element of size $a = 2$ is employed for $m_2^{p,i}$, as in [3]. Even if the three curves are similar, as the results are obtained after the non-maxima suppression operation that flatten the already highlighted visible differences between the methods, the plot stresses the complementarity between the first two approaches, that combined together through $m_3^{s,p,i}$

allocation give the best result. Indeed, we are able to tackle the problem of sparse false positives due to unreliability of the descriptors while at the same time increasing the number and homogeneity of the detections.

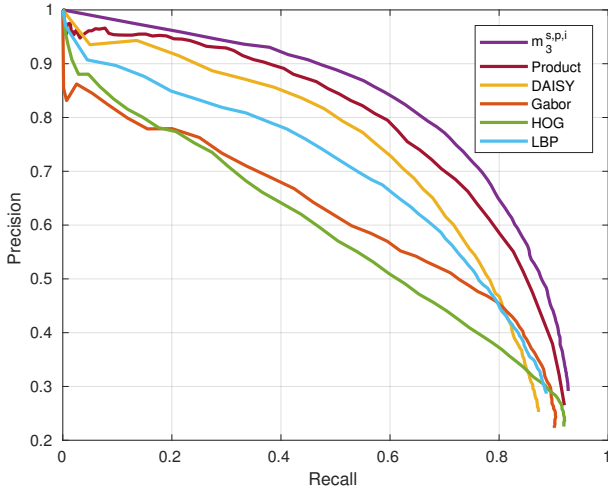


Fig. 7. PR-curves of the comparison of the proposed fusion with $m_3^{s,p,i}$ bba allocation, product of probabilities and the original four detectors.

Figure 7 shows a comparison of the proposed fusion approach with respect to the simple product of probabilities and the original four sources, i.e. detectors. The results obtained with the proposed method provides overall better values both for precision and recall, highlighting once again the importance of considering imprecision both in the calibration and in the image space.

V. CONCLUSION

In this paper we proposed a belief function definition to perform fusion exploiting an ensemble of SVM-based pedestrian detectors relying on different gradient, texture and orientation descriptors. Two possible sources of imprecision are taken into account, namely in the interpretation of SVM decision scores, for which instead of assigning a posterior probability we derive a bba that accounts for the possible mislocation of the estimated sigmoid curve, and later in the image space, where imprecision arises due to the spatial scale of the considered descriptors and may be detected by the heterogeneity of neighboring pixels. The experiments show that the proposed combined fusion algorithm performs better than by taking into account individual sources of imprecision, and by considering a simple fusion solution based on the product of probability values. Finally, since our solution is based on the evidential framework that naturally provides clues about the amount of ignorance as well as conflict between sources, we can use those indicators in the future to refine the selection of the training samples for the individual detectors.

ACKNOWLEDGMENT

This work was partly funded by ANR grant ANR-15-CE39-0005 and by QNRF grant NPRP-09-768-1-114.

REFERENCES

- [1] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton university press Princeton, 1976, vol. 1.
- [2] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *CVPR*, 2001, pp. 511–518.
- [3] J. Vandoni, E. Aldea, and S. Le Hégarat-Masclé, “An evidential framework for pedestrian detection in high-density crowds,” in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.
- [4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*. IEEE, 2005, pp. 886–893.
- [5] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [6] R. Hu, R. Wang, S. Shan, and X. Chen, “Robust head-shoulder detection using a two-stage cascade framework,” in *ICPR*, 2014, pp. 2796–2801.
- [7] M. Li, S. Bao, W. Dong, Y. Wang, and Z. Su, “Head-shoulder based gender recognition,” in *ICIP*, 2013, pp. 2753–2756.
- [8] E. Tola, V. Lepetit, and P. Fua, “Daisy: An efficient dense descriptor applied to wide-baseline stereo,” *IEEE TPAMI*, vol. 32, no. 5, pp. 815–830, 2010.
- [9] A. Bera, N. Galoppo, D. Sharlet, A. Lake, and D. Manocha, “Adapt: real-time adaptive pedestrian tracking for crowded scenes,” in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1801–1808.
- [10] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, “Online multi-target tracking with strong and weak detections,” in *European Conference on Computer Vision*. Springer, 2016, pp. 84–99.
- [11] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detection: Survey and experiments,” *IEEE trans. on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.
- [12] R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Ten Years of Pedestrian Detection, What Have We Learned?” in *ECCV Workshops*, 2014, pp. 613–627.
- [13] M. Gönen and E. Alpaydm, “Multiple kernel learning algorithms,” *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2211–2268, 2011.
- [14] L. Ma, S. Destercke, and Y. Wang, “Online active learning of decision trees with evidential data,” *Pattern Recognition*, vol. 52, pp. 33–45, 2016.
- [15] V. Eiselein, D. Arp, M. Pätzold, and T. Sikora, “Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors,” in *AVSS*, 2012, pp. 325–330.
- [16] J. Platt *et al.*, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
- [17] P. Xu, F. Davoine, H. Zha, and T. Denoeux, “Evidential calibration of binary svm classifiers,” *International Journal of Approximate Reasoning*, vol. 72, pp. 55–70, 2016.
- [18] M. Lachaize, S. Le Hégarat-Masclé, E. Aldea, A. Maitrot, and R. Reynaud, “Evidential multi-class classification from binary classifiers: application to waste sorting quality control from hyperspectral data,” in *Thirteenth International Conference on Quality Control by Artificial Vision 2017*, vol. 10338. International Society for Optics and Photonics, 2017, p. 103380V.
- [19] P. Xu, F. Davoine, and T. Denoeux, “Evidential combination of pedestrian detectors,” in *British Machine Vision Conference*, 2014, pp. 1–14.
- [20] P. Minary, F. Pichon, D. Mercier, E. Lefevre, and B. Droit, “Face pixel detection using evidential calibration and fusion,” *International Journal of Approximate Reasoning*, vol. 91, pp. 202–215, 2017.
- [21] I. Bloch, “Defining belief functions using mathematical morphology—application to image fusion under imprecision,” *Int. journal of approximate reasoning*, vol. 48, no. 2, pp. 437–465, 2008.
- [22] P. Smets and R. Kennes, “The transferable belief model,” *Artificial intelligence*, vol. 66, no. 2, pp. 191–234, 1994.
- [23] J. Vandoni, E. Aldea, and S. Le Hégarat-Masclé, “Active learning for high-density crowd count regression,” in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.
- [24] E. Aldea, D. Marastoni, and K. H. Kiyani, “Spatio-temporal consistency for head detection in high-density scenes,” in *Asian Conference on Computer Vision*. Springer, 2014, pp. 665–679.