Face pixel detection using evidential calibration and fusion

Pauline Minary, Frédéric Pichon, David Mercier, Eric Lefevre, Benjamin Droit

Face pixel detection using evidential calibration and fusion

Pauline Minary[1,2], Frédéric Pichon[1], David Mercier[1], Eric Lefevre[1], and Benjamin Droit[2]

[1]Univ. Artois, EA 3926,
Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
Béthune, F-62400, France.
{frederic.pichon,david.mercier,eric.lefevre}@univ-artois.fr
[2]SNCF Réseau,
Département des Télécommunications,
La Plaine Saint Denis, France.
{pauline.minary,benjamin.droit}@reseau.sncf.fr

Abstract. Due to legal reasons, faces on a given image may have to be blurred. This may be achieved by combining several information sources, which may provide information at different levels of granularity; for instance face detectors return bounding boxes corresponding to assumed positions of faces, whereas skin detectors may return pixel level information. A general, well-founded and efficient approach to combining box-based information sources was recently proposed in the context of pedestrian detection. This approach relies on evidence theory to calibrate and combine sources. In this paper, we apply this approach to combine face (rather than pedestrian) detectors, in order to obtain a state-of-the-art face blurring system based on multiple detectors. Then, we propose another approach to tackle the blurring problem, which consists essentially in applying at the pixel-level the central idea – combining evidentially calibrated information sources – of the preceding box-based approach. This shift of focus induces several conceptual advantages. In addition, the proposed approach shows better performances on a classical face dataset, as well as on a more challenging one.

Keywords: Belief functions, Information fusion, Evidential calibration, Face blurring.

1 Introduction

Due to legal reasons, faces on a given image may have to be blurred. Yet, it can rapidly become a tedious task if it is done manually, especially if there is a large amount of images to process. A solution may consist in using a face detection system, which aims to automatically find the positions of the faces in a given image.

Since the early 2000s, there has been significant research on face detection and many algorithms have been proposed, in particular based on machine learning

techniques, such as the well-known Viola and Jones approach [34] or the neural network-based approach proposed by Rowley *et al.* [27]. Recently, more elaborate algorithms based on deep convolutional neural networks [10, 37, 40] made a major breakthrough in the field. Yet, another path of research consists in merging information given by multiple sources, whether situated at the pixel level or directly on the faces [1, 11, 23, 32]. Indeed, since sources, such as face detectors, generally provide complementary information, using several of them is a means to improve overall performance.

There are many different ways to perform the fusion of some given information. Among them, in the context of pedestrian detection, Xu *et al.* [35] recently proposed a well-founded and general approach. In this approach, for a given image each used detector provides a set of bounding boxes corresponding to the assumed positions of the pedestrians, as well as a confidence score for each of these boxes. The main idea is then to use a step called score calibration [26], in order to be able to combine these calibrated scores afterwards, and to obtain better detection performance. Of particular interest is that the calibration and combination steps of this approach rely on a framework for reasoning under uncertainty called evidence theory [28, 29]. This theory is a generalization of probability theory, which enables to account for uncertainties due to randomness and incompleteness. As a matter of fact, Xu *et al.* [36] subsequently proposed more elaborate calibration procedures than those used in [35], which exploit more fully the expressive power of this theory and as a result, model more precisely the uncertainties inherent to the calibration process. Hence, by replacing the calibration procedure in [35] by one of the most efficient ones studied in [36], and by applying to faces the general detection approach introduced in [35], one obtains what may be considered presently as a state-of-the-art face detection system based on multiple detectors. Nonetheless, despite its appeals, we note that such a system suffers from two main limitations inherited from Xu *et al.*'s approach [35]. First, it is designed to handle only detectors providing bounding boxes, *i.e.*, it can not integrate directly sources providing information at the pixel level. Second, this approach relies on a parameter (so-called overlap threshold) necessary in the handling of boxes.

Using a face detection system is a natural means to solve the face blurring problem. However, we may remark that this problem is not exactly equivalent to face detection: face blurring amounts merely to deciding whether a given pixel belongs to a face, whereas face detection amounts to determining whether a given set of pixels corresponds to the same face. This remark opens the path for a different approach to reasoning about blurring, which may then be situated at the pixel-level. Within this scope, we propose in this paper a face blurring system, which consists essentially in applying at the pixel-level the central idea and contributions of Xu *et al.* [35, 36], *i.e.*, combining evidentially calibrated information sources. As it will be seen, this pixel-level perspective presents several conceptual advantages over operating at the box-level. In particular, sources providing pixel-level information can be directly integrated and the parameter necessary in the handling of boxes can be avoided. Nonetheless, let us note that while our

approach presents some interests over box-based methods for the problem of face blurring, these latter methods provide more information (specifically, they isolate faces) and are thus relevant for other problems, such as face recognition.

This paper is organized as follows. Section 2 recalls necessary background on evidence theory as well as on calibration. Section 3 exposes what may be considered as a state-of-the-art face detection system based on multiple detectors, that is, a system performing face detection using Xu *et al.*'s evidential box-based detection approach [35], improved using evidential calibration [36]. In Section 4, our proposed pixel-based face blurring system is detailed and its fundamental differences with respect to blurring using Xu *et al.*'s box-based approach are discussed. The performances of the box-based and pixel-based approaches, given the same input information, are then compared in Section 5 on two datasets (one from the literature and one composed of railway platforms images coming from the French railway company SNCF). The ability of the proposed approach to integrate directly pixel-level information is illustrated in Section 6 on these same two datasets. Finally, conclusions and perspectives are given in Section 7.

2 Background

In this section, necessary concepts of evidence theory, such as combination and decision-making schemes, are recalled. Classical calibration methods based on probability theory are then described, followed by their extensions to the evidential framework.

2.1 Evidence theory

The theory of evidence is a framework for reasoning under uncertainty. Let Ω be a finite set called the frame of discernment, which contains all the possible answers to a given question of interest Q. In this theory, uncertainty with respect to the answer to Q is represented using a *Mass Function* (MF) defined as a mapping $m^\Omega : 2^\Omega \to [0, 1]$ that satisfies $m^\Omega(\emptyset) = 0$ and

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{1}$$

The quantity $m^\Omega(A)$ corresponds to the share of belief that supports the claim that the answer is contained in $A \subseteq \Omega$ and nothing more specific. A mass function can be equivalently represented by the *plausibility function*, defined by

$$Pl^\Omega(A) = \sum_{B \cap A \neq \emptyset} m^\Omega(B), \quad \forall A \subseteq \Omega. \tag{2}$$

It represents the amount of evidence which does not contradict the hypothesis $\omega \in A$. The plausibility function restricted to singletons is called the *contour function*, denoted pl^Ω and defined by

$$pl^\Omega(\omega) = Pl^\Omega(\{\omega\}), \quad \forall \omega \in \Omega. \tag{3}$$

Given two independent MFs m_1^Ω and m_2^Ω about the answer to Q, it is possible to combine them using *Dempster's rule of combination*. The result of this combination is a MF $m_{1\oplus 2}^\Omega$ defined by

$$m_{1\oplus 2}^\Omega(\emptyset) = 0, \tag{4}$$

$$m_{1\oplus 2}^\Omega(A) = \frac{1}{1-\kappa} \sum_{B\cap C=A} m_1^\Omega(B) m_2^\Omega(C), \tag{5}$$

where

$$\kappa = \sum_{B\cap C=\emptyset} m_1^\Omega(B) m_2^\Omega(C), \tag{6}$$

represents the *degree of conflict* between m_1^Ω and m_2^Ω. If $\kappa = 1$, there is a total conflict between the two pieces of evidence and they cannot be combined.

Different decision strategies exist to make a decision about the true answer to Q, given a MF m^Ω on this answer [6]. In particular, the answer having the smallest so-called *upper expected cost* may be selected, where the upper expected cost $R^*(\omega)$ of some answer $\omega \in \Omega$ is defined as

$$R^*(\omega) = \sum_{A\subseteq\Omega} m^\Omega(A) \max_{\omega'\in A} c(\omega, \omega'), \tag{7}$$

with $c(\omega, \omega')$ is the cost of deciding ω when the true answer is ω'.

2.2 Probabilistic calibration of sources

Consider an object whose true label Y is such that $Y \in \mathbb{Y} = \{0, 1\}$. Furthermore, suppose that after observing this object, a source returns a piece of information of the form $X \in \mathbb{X}$ for some domain $\mathbb{X}$. To learn how to interpret what this piece of information tells us about Y, a step called calibration may be used, which consists in estimating the probability distribution $p^{\mathbb{Y}}(\cdot|X)$. This step relies on a training set $\mathcal{L}$, which contains n other objects for which the variable Y is known, and for which we observed what the source returned on $\mathbb{X}$, *i.e.*, $\mathcal{L} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ where $X_i \in \mathbb{X}$ represents the information given by the source for the i^{th} object whose true label is $Y_i \in \mathbb{Y}$. Domain $\mathbb{X}$ can be either discrete or continuous, thus different calibration procedures are necessary depending on the output type of the source. For instance, a face detector may return a score associated to a box, which is a continuous piece of information, while a skin detector may return a binary decision for a pixel, which is a discrete piece of information.

Probabilistic calibration of discrete information Let $\mathbb{X} = \{1, 2, ..., M\}$. The probability that the label $Y = 1$ given information $X = j \in \mathbb{X}$ may be estimated by

$$P^{\mathbb{Y}}(1|j) = \frac{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j, Y_i = 1\}|}{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j\}|}. \tag{8}$$

Probabilistic calibration of continuous information Let $\mathbb{X} = \mathbb{R}$. The probability that the label $Y = 1$ given information $X = S \in \mathbb{R}$, *i.e.*, $P^{\mathbb{Y}}(1|S)$, may be estimated using three main approaches defined to calibrate sources returning real-valued confidence scores, which will be the case in this paper. These approaches are binning [38], isotonic regression [39] and logistic regression [26].

The binning approach consists in dividing the score spaces into different bins, for example $]-3;-2],]-2;-1]$, etc. For each bin j, the number k_j of pairs $(X_i, Y_i) \in \mathcal{L}$ such that $Y_i = 1$ and X_i in bin j, and the number n_j of pairs (X_i, Y_i) such that X_i in bin j, can be obtained. Then, for a score $X = S$ such that S belongs to bin j, we have

$$P^{\mathbb{Y}}(1|S) = \frac{k_j}{n_j}. \tag{9}$$

Yet, the accuracy of binning highly depends on the number and size of the bins.

Isotonic regression can be seen as a sort of binning, where the size and the boundaries of the bins are dynamically created. It relies on the pool adjacent violators algorithm [2], which consists in fitting a non-decreasing function to the training data by minimizing the mean-squared error [39].

Logistic regression is a more elaborate and accurate method based on a fitting of a sigmoid function h defined by

$$P^{\mathbb{Y}}(1|S) \approx h_S(\theta) = \frac{1}{1 + e^{(\theta_0 + \theta_1 S)}}, \tag{10}$$

where the parameter $\theta = (\theta_0, \theta_1) \in \Theta = \mathbb{R}^2$ is chosen as the one maximizing the following likelihood function:

$$L(\theta) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i}, \tag{11}$$

with

$$p_i = \frac{1}{1 + e^{(\theta_0 + \theta_1 X_i)}}. \tag{12}$$

Note that some score values may be less present than others in the training set, thus some estimated probabilities may be less accurate than others. To address this issue, Xu *et al.* proposed to refine the above three calibrations using the theory of evidence [36], in order to manage these inaccuracies. The following section deals with the evidential versions of calibration procedures.

2.3 Evidential calibration of sources

As for the probabilistic analysis, evidential calibration procedures can be defined differently depending on the type of outputs returned by the considered source. Any of these evidential calibration procedures yields a MF $m^{\mathbb{Y}}(\cdot|X)$ (rather than a probability distribution) accounting explicitly for uncertainties in the calibration process.

Evidential calibration of discrete information Evidence theory provides different models to extend a probabilistic approach into an evidential one, as detailed in [36]. Hence, in the discrete case where the received information from the source is such that $X = j$, there are different ways to obtain the MF $m^{\mathbb{Y}}(\cdot|j)$. In particular, one may use the model of Dempster [5], which leads to the following MF:

$$\begin{aligned} m^{\mathbb{Y}}(\{0\}|j) &= \frac{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j, Y_i = 0\}|}{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j\}| + 1}, \\ m^{\mathbb{Y}}(\{1\}|j) &= \frac{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j, Y_i = 1\}|}{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j\}| + 1}, \end{aligned} \tag{13}$$

and

$$m^{\mathbb{Y}}(\{0,1\}|j) = \frac{1}{|\{(X_i, Y_i) \in \mathcal{L} | X_i = j\}| + 1}.$$

Hereafter, we will refer to this type of calibration as evidential Dempster calibration.

Evidential calibration of continuous information Xu *et al.* [36] proposed several evidential extensions of probabilistic calibration methods of scores. This paper focuses on the extension of the logistic regression based on the so-called likelihood model [7, 18, 17], as Xu *et al.* showed that this is the one presenting overall the best performances of all methods [36].

In this extension of the logistic regression, one first represents knowledge about parameter $\tau = h_s(\theta) \in T = [0, 1]$ after observing $X = S$, in the form of a consonant belief function $Bel^T(.|S)$ with contour function $pl^T(.|S)$ defined by

$$pl^T(\tau|S) = \sup_{\theta_1 \in \mathbb{R}} pl^{\Theta}(\ln(\tau^{-1} - 1) - \theta_1 S, \theta_1), \forall \tau \in (0, 1), \tag{14}$$

with pl^{Θ} the function defined on Θ by

$$pl^{\Theta}(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \qquad \forall \theta \in \Theta, \tag{15}$$

with L the likelihood function given in Eq. (11), and $\hat{\theta} = (\hat{\theta_0}, \hat{\theta_1})$ the maximum likelihood estimate of θ. Then, viewing the label of an object after observing its score S as the realisation of a random variable Y with a Bernoulli distribution $\mathcal{B}(\tau)$, one uses the solution proposed by Kanjanatarakul *et al.* [18, 17] to make statements about Y. In a nutshell, it consists in using the fact that $Bel^T(.|S)$ is equivalent to a random set [24], and in using the sampling model of Dempster [5] to deduce a belief function on $\mathbb{Y}$. As shown by Xu *et al.* [36], this belief function obtained given information S has associated mass function $m^{\mathbb{Y}}(.|S)$ defined by

$$\begin{aligned} m^{\mathbb{Y}}(\{0\}|S) &= 1 - \hat{\tau} - \int_{\hat{\tau}}^{1} pl^T(u|S)du, \\ m^{\mathbb{Y}}(\{1\}|S) &= \hat{\tau} - \int_{0}^{\hat{\tau}} pl^T(u|S)du, \end{aligned} \tag{16}$$

and

$$m^{\mathbb{Y}}(\{0,1\}|S) = \int_0^1 pl^T(u|S)du,$$

where $\hat{\tau}$ maximizes the contour function pl^T.

3 An evidential box-based face detection approach

Face blurring may be achieved using simply the boxes returned by a face detection system. In this section, we present such a system, which may be considered as a state-of-the-art system with respect to face detection based on multiple detectors returning box information. In a nutshell, this system is merely Xu *et al.* [35] evidential box-based detection approach, whose calibration step has been replaced by the evidential likelihood-based logistic regression calibration procedure proposed in [36] and recalled in the previous section. This section first provides an overview of this approach and then details some of its steps.

3.1 Overview of the approach

Let us consider a given image and assume that J face detectors are run on this image. Formally, each detector D_j, $j = 1, ..., J$, provides N_j pairs $(B_{i,j}, S_{i,j})$, where $B_{i,j}$ denotes the i^{th} box, $i = 1, ..., N_j$, returned by the j^{th} detector and $S_{i,j}$ is the confidence score associated to this box.

Through a calibration procedure using a training set that will be described in Section 3.2, score $S_{i,j}$ is transformed into a MF $m^{\mathcal{B}_{i,j}}$ defined over the frame $\mathcal{B}_{i,j} = \{0, 1\}$, where 1 (resp. 0) means that there is a face (resp. no face) in box $B_{i,j}$.

Then, using a clustering procedure detailed in Section 3.3, all the boxes $B_{i,j}$ returned by the J detectors for the considered image, are grouped into K clusters C_k, $k = 1, ..., K$, each of these clusters being represented by a single box B_k.

In addition, for each box $B_{i,j} \in C_k$, its associated MF $m^{\mathcal{B}_{i,j}}$ is assumed to represent a piece of evidence regarding the presence of a face in B_k, that is, $m^{\mathcal{B}_{i,j}}$ is converted into a MF $m_{i,j}^{\mathcal{B}_k}$ on $\mathcal{B}_k = \{0, 1\}$ defined by $m_{i,j}^{\mathcal{B}_k}(A) = m^{\mathcal{B}_{i,j}}(A)$, for all $A \subseteq \{0, 1\}$. These pieces of evidence are then combined using Dempster's rule:

$$m^{\mathcal{B}_k} = \bigoplus_{i,j} m_{i,j}^{\mathcal{B}_k}. \tag{17}$$

The combination results in a MF $m^{\mathcal{B}_k}$ representing the overall system uncertainty with respect to the presence of a face in B_k. We note that the use of Dempster's rule is appropriate when the sources may be considered to be independent and reliable. More complex combination schemes are also considered in [35]. However, only Dempster's rule, which presents good performance in [35], is considered here.

The three main steps of the approach, namely calibration, clustering and fusion, are illustrated in Figure 1. For the sake of simplicity only two detectors,

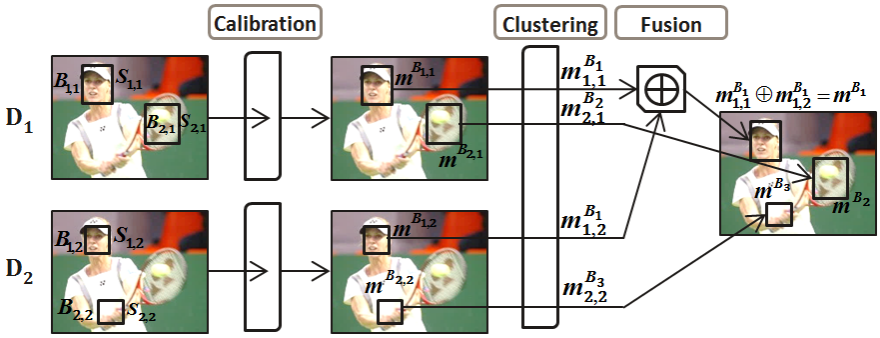


Fig. 1: Illustration of the box-based approach

each returning two boxes, are considered in this example. $B_{i,j}$ corresponds to the i^{th} box, $i = 1, 2$, returned by the j^{th} detector $j = 1, 2$, and which has $S_{i,j}$ as associated score. In this scenario, the boxes $B_{1,1}$ and $B_{1,2}$ are grouped into the same cluster C_1, represented by the box B_1. Their associated scores, transformed into mass functions, are combined and result in the final mass function $m^{\mathcal{B}_1}_{1,1} \oplus m^{\mathcal{B}_1}_{1,2}$, which is denoted by $m^{\mathcal{B}_1}$. The other boxes $B_{2,1}$ and $B_{2,2}$ form their own clusters, respectively represented by B_2 and B_3. Finally, for each resulting box with its associated MF, a decision has to be made whether the box has to be blurred or not; it may be done using the decision strategy given in Section 2.1 and in particular using Eq. (7) for some cost function c.

3.2 Box-based score calibration for a detector

In order to transform the score $S_{i,j}$ associated to a box $B_{i,j}$ into a MF $m^{\mathcal{B}_{i,j}}$, detector D_j needs to be calibrated. In particular, the evidential likelihood-based logistic regression calibration procedure recalled in Section 2.3 may be used instead of the cruder procedures used in [35]. This procedure requires a training set, which we denote by $\mathcal{L}_{cal,j}$. We detail below how $\mathcal{L}_{cal,j}$ is built.

Assume that L images are available. Besides, the positions of the faces really present in each of these images are known in the form of bounding boxes. Formally, this means that for a given image ℓ, a set of M^ℓ boxes $G^\ell_r, r = 1, ..., M^\ell$, is available, with G^ℓ_r the r^{th} bounding (ground truth) box on image ℓ.

Furthermore, detector D_j to be calibrated is run on each of these images, yielding N^ℓ_j pairs $(B^\ell_{t,j}, S^\ell_{t,j})$ for each image ℓ, where $B^\ell_{t,j}$ denotes the t^{th} box, $t = 1, ..., N^\ell_j$, returned on image ℓ by detector D_j and $S^\ell_{t,j}$ is the confidence score associated to this box.

From these data, training set $\mathcal{L}_{cal,j}$ is defined as the set of pairs $(S^\ell_{t,j}, YB^\ell_{t,j})$, $\ell = 1, ..., L$, and $t = 1, ..., N^\ell_j$, with $YB^\ell_{t,j} \in \{0, 1\}$ the label obtained by evaluat-

ing whether box $B^{\ell}_{t,j}$ "matches" some face in image ℓ, *i.e.*,

$$YB^{\ell}_{t,j} = \begin{cases} 1 \text{ if } \exists G^{\ell}_r, r = 1, ..., M^{\ell}, \text{ such that } ov(G^{\ell}_r, B^{\ell}_{t,j}) \geq \lambda, \\ 0 \text{ otherwise,} \end{cases}$$

where λ is some threshold in $(0, 1)$ and $ov(G^{\ell}_r, B^{\ell}_{t,j})$ is a measure of the overlap between boxes G^{ℓ}_r and $B^{\ell}_{t,j}$. It is defined by [9]

$$ov(B_1, B_2) = \frac{area(B_1 \cap B_2)}{area(B_1 \cup B_2)}, \tag{18}$$

for any two boxes B_1 and B_2. Informally, $\mathcal{L}_{cal,j}$ stores the scores associated to all the boxes returned by detector D_j on images where the positions of faces are known, and records for each score whether its associated box is a true or false positive. It is then clear that the MF $m^{\mathcal{B}_{i,j}}$ associated to a new score $S_{i,j}$ and obtained from calibration relying on $\mathcal{L}_{cal,j}$, represents uncertainty toward box $B_{i,j}$ containing a face.

3.3 Clustering of boxes

As several detectors are used, some boxes may be located in the same area of an image, which means that different boxes assume that there is a face in this particular area. The step of clustering allows one to group those boxes and to retain only one per cluster. A greedy approach is used in [35], based on the work of Dollar *et al.* [8]: the procedure starts by selecting the box $B_{i,j}$ with the highest mass of belief on the face hypothesis, *i.e.,* the box $B_{i,j}$ such that $m^{\mathcal{B}_{i,j}}(\{1\}) > m^{\mathcal{B}_{u,v}}(\{1\}), \forall (u, v) \neq (i, j)$, and this box is considered as the representative of the first cluster. Then, each box $B_{u,v}$, $\forall (u, v) \neq (i, j)$, such that the overlap $ov(B_{i,j}, B_{u,v})$ is above the threshold λ, is grouped into the same cluster as $B_{i,j}$, and is then no longer considered for further associations. Among the remaining boxes, the box $B_{i,j}$ with the highest $m^{\mathcal{B}_{i,j}}(\{1\})$ is selected as representative of the next cluster, and the procedure is repeated until all the boxes are clustered.

4 Proposed evidential pixel-based approach

The approach exposed in the previous section is general and well-founded. It is designed for detectors returning boxes, but it does not allow to directly integrate pixel-based information. Besides, as explained in the introduction, for the purpose of blurring it seems interesting to work at the pixel level rather than box level. Thus, the idea of the approach proposed in this section is to use elements from the previous system, in particular the evidential calibration and fusion, and to apply them at the pixel level. This section first exposes an overview of the proposed approach. Then, in order to be able to compare subsequently the proposed pixel-based approach to the previous system, we detail how the same input information as in the previous section, *i.e.*, boxes and scores returned by detectors, can be used within our pixel-based approach. Finally, fundamental differences between the two approaches are discussed.

4.1 Overview of the approach

To each pixel $p_{x,y}$ in an image, we associate a frame of discernment $\mathcal{P}_{x,y} = \{0, 1\}$, where x and y are the coordinates of the pixel in the image and 1 (resp. 0) means that there is a face (resp. no face) in pixel $p_{x,y}$. For the pixel $p_{x,y}$, N mass functions are obtained on $\mathcal{P}_{x,y}$ from N independent sources. They are then combined using Dempster's rule of combination, resulting in the MF denoted $m^{\mathcal{P}_{x,y}}$, *i.e*,

$$m^{\mathcal{P}_{x,y}} = \bigoplus_{k=1}^{N} m_k^{\mathcal{P}_{x,y}}, \tag{19}$$

with $m_k^{\mathcal{P}_{x,y}}$ the MF representing the uncertainty with respect to the presence of a face in the pixel $p_{x,y}$ for the k^{th} source. Each MF $m_k^{\mathcal{P}_{x,y}}, k = 1, \ldots, N$, is obtained using the calibration method corresponding to the type of the outputs of the k^{th} source. Specifically, if the source gives a score information, the MF is obtained through the evidential likelihood-based logistic regression calibration, using a training set $\mathcal{L}$ composed of pairs (X_i, Y_i), with X_i the score associated to the i^{th} object which is now a pixel, and Y_i its true label. Otherwise, if the source gives discrete information, the evidential Dempster calibration is used, with a training set $\mathcal{L} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ where X_i is the discrete information provided by the source for the i^{th} pixel.

4.2 Face detections as inputs to our approach

Consider strictly the same input information as in Section 3, that is J detectors each returning a set of bounding boxes with associated scores corresponding to the assumed positions of the faces. This section exposes how our approach can be applied in that case.

For a given pixel in an image and a given detector, two exclusive situations occur: either the pixel $p_{x,y}$ is contained by one of the box $B_{i,j}$ returned by the detector, or it is not. If it is contained by a box $B_{i,j}$, the score $S_{i,j}$ of the box is associated ("transferred") to the pixel. If the pixel does not belong to any box, no score is associated to it. As a consequence, the considered pixel either has an associated score, or it does not. These two situations are now detailed.

In the first case, when a score is available for the considered pixel, it is transformed into a MF using the evidential logistic regression and a training set, that we denote $\mathcal{L}_{calP,j}$. Let us describe this set $\mathcal{L}_{calP,j}$ underlying the transformation using calibration of a score $S_{i,j}$ associated to a pixel $p_{x,y}$ by a detector D_j, into a MF $m_{i,j}^{\mathcal{P}_{x,y}}$. For a given image ℓ, each pair $(B_{t,j}^{\ell}, S_{t,j}^{\ell})$ introduced in Section 3.2 yields, *via* "transfer", $\left|B_{t,j}^{\ell}\right|$ pairs $(p_{d,t,j}^{\ell}, S_{t,j}^{\ell})$, with $d = 1, \ldots, \left|B_{t,j}^{\ell}\right|$, and $\left|B_{t,j}^{\ell}\right|$ the number of pixels in box $B_{t,j}^{\ell}$, and where $p_{d,t,j}^{\ell}$ denotes the pixel in d^{th} position in box $B_{t,j}^{\ell}$. From these data, we define $\mathcal{L}_{calP,j}$ as the set of pairs $(S_{t,j}^{\ell}, YP_{d,t,j}^{\ell})$, with $\ell = 1, \ldots, L$, $t = 1, \ldots, N_j^{\ell}$, and $d = 1, \ldots, \left|B_{t,j}^{\ell}\right|$, with $YP_{d,t,j}^{\ell} \in \{0, 1\}$ the label simply obtained by checking whether pixel $p_{d,t,j}^{\ell}$ belongs to some ground

truth box G_r^ℓ in the image ℓ, *i.e*,

$$YP_{d,t,j}^{\ell} = \begin{cases} 1 \text{ if } \exists\ G_r^\ell,\ r = 1, \ldots, M^\ell, \text{ such that } p_{d,t,j}^{\ell} \in G_r^\ell, \\ 0 \text{ otherwise.} \end{cases} \tag{20}$$

$\mathcal{L}_{calP,j}$ may pose a complexity issue as $|\mathcal{L}_{calP,j}| = \sum_{\ell=1}^{L} \sum_{t=1}^{N_j^\ell} \left|B_{t,j}^{\ell}\right|$. To avoid this, one may use a smaller set $\mathcal{L}'_{calP,j} \subset \mathcal{L}_{calP,j}$, which represents roughly the same information as $\mathcal{L}_{calP,j}$ and built as follows: for each triple (ℓ, t, j), only 10 pairs among the pairs $(S_{t,j}^{\ell}, YP_{d,t,j}^{\ell})$, $d = 1, \ldots, \left|B_{t,j}^{\ell}\right|$, are selected such that the ratio

$$\frac{\left|\{YP_{d,t,j}^{\ell} | d = 1, ..., \left|B_{t,j}^{\ell}\right|, YP_{d,t,j}^{\ell} = 1\}\right|}{\left|\{YP_{d,t,j}^{\ell} | d = 1, ..., \left|B_{t,j}^{\ell}\right|, YP_{d,t,j}^{\ell} = 0\}\right|} \tag{21}$$

is preserved. $\mathcal{L}'_{calP,j}$ has then a size of $\left|\mathcal{L}'_{calP,j}\right| = 10 \sum_{\ell=1}^{L} N_j^\ell$.

Let us now consider the second situation, where a pixel is not contained by any of the boxes and thus does not have an associated score. Since it should be taken into account that detectors do not present the exact same performances (in particular, some may have many more pixels not in boxes than others), it seems interesting to calibrate this kind of outputs from detectors. We propose to do so by viewing this information of score absence as a discrete information, and thus by applying the evidential Dempster calibration. Specifically, the training set, denoted $\mathcal{L}_{*,j}$, necessary for this calibration is obtained using L images on which the detector D_j is applied. The number n_j of pixels of these images, which are not contained by any of the boxes returned by the detector D_j, can be obtained. For the i^{th} of these n_j pixels, the absence of score can be encoded by the discrete information $X_i = 1$. As the ground truth of these L images is known, its associated true label Y_i is available. It is then possible to obtain a MF, denoted $m_{*,j}^{\mathcal{P}_{x,y}}$ and calculated using $\mathcal{L}_{*,j}$ and Eq. (13) with $j = 1$, representing the uncertainty with respect to the presence of a face on pixel $p_{x,y}$, when this pixel is not included in a box of detector D_j.

4.3 Comparison of both approaches

The proposed pixel-based approach presents several advantages over the one of Section 3. First, as can be seen in Section 4.2, the construction of the training set for calibration in case of pixels avoids the use of the parameter λ, whose value needs to be fixed either *a priori* (but then it is arguably arbitrary) or empirically.

Furthermore, our approach avoids the use of the clustering step, which also involves the parameter λ and that may behave non optimally in a multi-object situation, especially when they are close to each other, which may be the case with faces in a crowd.

In addition, it allows us to have an arguably more consistent modelling of box absence than the box-based method. Indeed, in this latter method, for a given

area in an image, there are two different modellings of box absence depending on the situation: either none of the detectors has provided a box, in which case the area is considered as non face, which amounts to considering that the detectors know that there is no face; or only a subset of the detectors has provided a box, in which case the other detectors are ignored, which is equivalent (under Dempster's rule) to considering that these detectors know nothing. By contrast, in the proposed method, the use of calibration enables us to take into account in a consistent manner the information of score absence into the fusion process, as when a detector D_j does not return a box for a given pixel $p_{x,y}$, its associated MF $m_{*,j}^{\mathcal{P}_{x,y}}$ is considered regardless of the outputs of the other detectors for this pixel. Thus, all detectors are involved in each fusion. Figure 2 illustrates this point, highlighting the differences with the previous approach. For the sake of

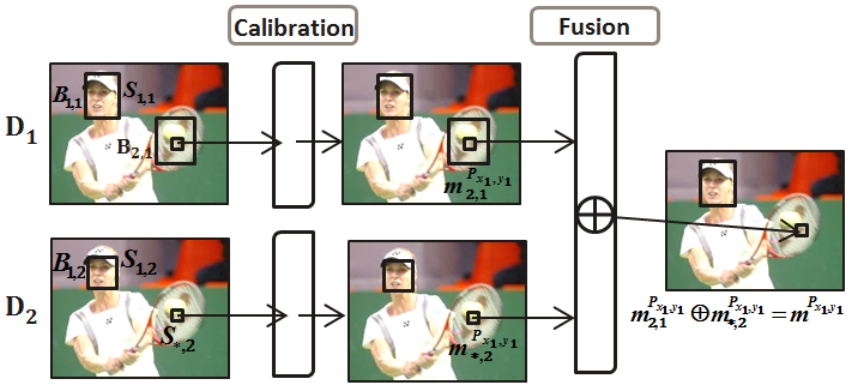


Fig. 2: Illustration of the pixel-based approach

simplicity only one pixel, at the position (x_1, y_1), is considered here. Pixel p_{x_1,y_1} is contained by the box $B_{2,1}$, with $S_{2,1}$ as associated score, so the corresponding mass function is obtained trough the evidential logistic regression. However, there is no box containing p_{x_1,y_1} for the second detector, and thus it does not have an associated score. Yet, the opinion of the second detector is still taken into account *via* the MF $m_{*,2}^{\mathcal{P}_{x_1,y_1}}$ defined in Section 4.2.

As explained before, one of the disadvantages of the box-based approach is that the integration of a pixel-based information is not straightforward. In the proposed system however, a source of information which gives pixel-based information can be integrated into the fusion process as easily as a box-based information. It will be illustrated with an experiment in Section 6.

Finally, we note that locating the approach at the pixel level brings potentially a complexity issue. This will be discussed in the next section.

5 Experimental comparison of both solutions

In this section, the results of the proposed approach are presented and compared to those of the box-based method, when all available inputs are box-based information. The experiment is performed on a literature dataset as well as on another dataset, composed of images coming from cameras filming railway platforms. The experiment is first described, then the results are discussed.

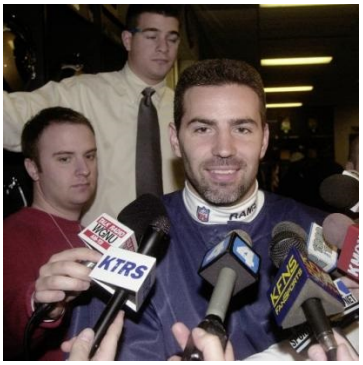
5.1 Description of the experiment

We selected four face detectors based on machine learning techniques for which an open source implementation was available. The first detector is the one proposed by Viola and Jones [34], which is based on a classification algorithm called Adaboost and that uses Haar feature extraction. The second detector is a variant of the previous one: the same classification algorithm is used but with Local Binary Patterns (LBP) feature extraction [13]. The third detector relies on Support Vector Machine (SVM) and uses Histogram of Oriented Gradients (HOG) features [4, 25]. It was provided by the DLIB library [21]. Following the current popularity of deep learning techniques, the fourth selected face detector is the deep neural network[1] proposed in [16], which is based on a compact design of a convolutional neural network and a cascade approach.

We used a literature dataset called Face Detection Data Set and Benchmark (FDDB) [14]. It contains the annotations (ground truth) for 5171 faces in a set of 2845 images. We trained both Adaboost-based detector with the same 2000 images of this dataset; the third and fourth detector were already trained. 200 others images were used for the calibration of the four detectors. The performances of the box-based and pixel-based approaches were then evaluated over the remaining 645 images.

Although the FDDB dataset presents various situations, we note that on the whole the images are generally of good quality and the faces of reasonable size. Thus, we also considered a more challenging dataset, composed of low-quality images and with situations where faces are more difficult to detect. These images, which we refer to as SNCF images, are extracted from video footage provided by video-protection cameras filming some railway platforms. We created a dataset of 600 images, containing multiple different conditions such as indoor and outdoor environment, different light settings and low image quality. The true positions of the 1089 faces on these images were manually annotated. Figure 3 shows an example of images extracted from the two datasets. As there were not enough face examples in the SNCF dataset to train detectors, we used the two Adaboost-based detectors trained with the 2000 images of the FDDB dataset and the other two already trained detectors. Nonetheless, we calibrated these detectors using 100 annotated SNCF images. Performance tests were then conducted over the remaining 500 images.

[1] Available at https://github.com/Bkmz21/CompactCNNCascade



(a) FDDB image

(b) SNCF image

Fig. 3: Example of FDDB image (3a) and SNCF image (3b).

The box-based approach returns MFs associated to boxes while our approach gives an MF for each pixel. Whatever the approach, to decide if a given pixel or a given box has to be blurred or not, we use the decision procedure relying on upper expected costs recalled in Section 2.1; in a binary case, they are simply defined by

$$R^*(\{0\}) = m^{\Omega}(\{1\})c(0,1) + m^{\Omega}(\{0,1\})c(0,1), \tag{22}$$

$$R^*(\{1\}) = m^{\Omega}(\{0\})c(1,0) + m^{\Omega}(\{0,1\})c(1,0), \tag{23}$$

by considering that the cost is equal to zero when the answer is correct ($c(0,0) = c(1,1) = 0$). As our purpose is to minimize the number of non-blurred faces, it is worse to consider a face as non-face than the opposite. In other words, decisions were made with costs such that $c(1,0) <= c(0,1)$. More specifically, we fixed $c(1,0) = 1$ and gradually increased $c(0,1)$ starting from $c(0,1) = 1$, to obtain different performance points. To quantify performances, we used the recall rate (proportion of pixels correctly blurred among the pixels to be blurred) and the precision rate (proportion of pixels correctly blurred among blurred pixels).

5.2 Results

Figure 4 compare the results of the four selected detectors taken alone to that of our approach relying on a combination of their outputs, on the FDDB dataset. As it can be seen, the fusion of the four detectors outputs considerably increased the performances, as for example a precision of 80% gives a recall of around 52% for the Haar/Adaboost detector instead of 77% for the combination result. Let us note that the performances of the deep neural network face detector are only represented by a point because all the scores returned by this detector were similar, thus all the boxes have the same associated MF and increasing the cost c(0,1) (the cost of deciding not to blur a pixel while it has to be) does not gradually increase the number of blurred pixels.

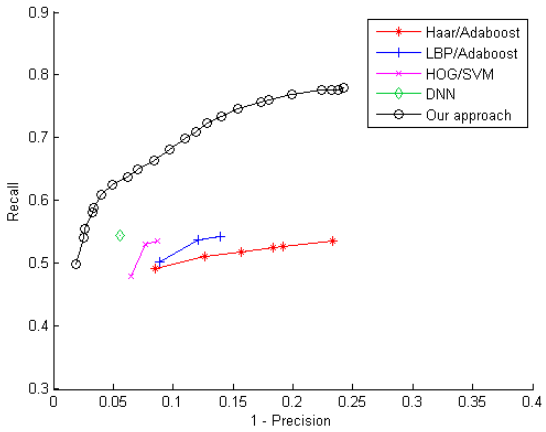


Fig. 4: Pixel-based approach vs detectors on FDDB.

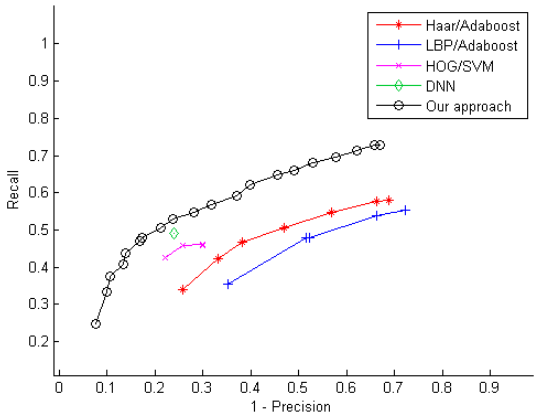


Fig. 5: Pixel-based approach vs detectors on SNCF dataset.

Figure 5 shows the result for the same experiment but this time on the SNCF dataset. The conclusion is the same as the proposed approach has better performances than the detectors taken alone. Let us remark that their performances could be improved by training them with face and non-face images closer to those encountered in the SNCF dataset.

Comparison on the FDDB dataset between the box-based approach used with different values of the overlap threshold λ and our approach is shown in Figure 6. As it can be noticed, for a same precision rate, the recall of our approach is always the highest. Figure 7 shows the results of this comparison on the SNCF dataset; the conclusions are the same.

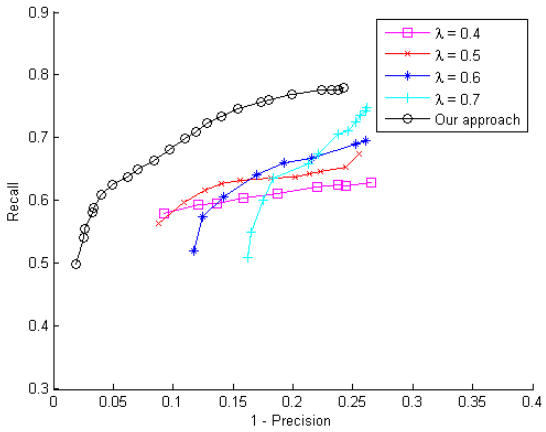


Fig. 6: Pixel-based approach vs box-based approach on FDDB.

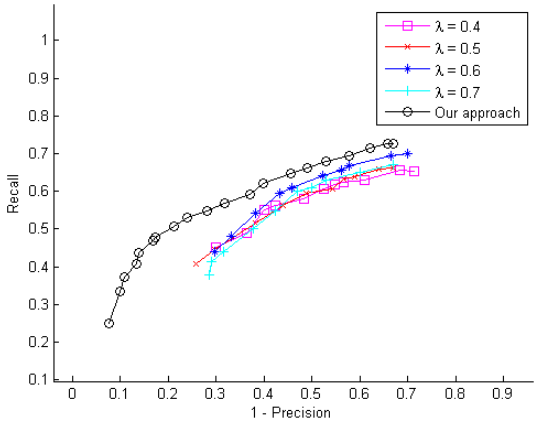


Fig. 7: Pixel-based approach vs box-based approach on SNCF dataset.

5.3 Discussion

Reasoning at the pixel level rather than with boxes as in the box-based approach may involve a complexity issue. Indeed, as the fusion is performed on every pixel instead of on sets of boxes, the proposed approach has *a priori* a higher complexity. For the pixel approach and for a given image, the number of operations is equal to $J \times a$, where J the number of fusion operations (which is equal to the number of used detectors) and a the number of pixels in the image. By contrast, in the box-based approach, the complexity is $O(b^2)$, with b the total number of boxes returned by J detectors. Indeed, at worst the clustering procedure is $O(b^2)$ [8] and this is the most costly step. Thus, at first glance, it seems that

the complexity is much higher for the proposed approach as a is generally significantly higher than b^2. However, any two pixels $p_{x,y}$ and $p_{x',y'}$ that do not belong to any box of D_j have associated MFs with the same definitions, *i.e.*, we have $m_{*,j}^{\mathcal{P}_{x,y}}(A) = m_{*,j}^{\mathcal{P}_{x',y'}}(A)$, for all $A \subseteq \{0,1\}$. Thus, pixels that do not belong to any of the returned boxes by the detectors have the same resulting MF. This latter case happens often in practice, hence this allows us to have a common processing. For instance, in a set of 200 images of FDDB, with the four face detectors considered in our experiment, it corresponds on average at around 80% of the pixels of the image. In terms of time processing, an image takes on average around 120 milliseconds to process (including the time of detection of the four detectors) for the box-based approach and 150 milliseconds for the proposed system; we consider that it is a reasonable difference.

This section showed that given the same information, *i.e.*, detectors returning boxes, the proposed approach gives better results than the box-based approach. Our approach is a little more time-consuming but the difference is reasonable. The following section illustrates another advantage of our approach, which is its ability to integrate directly sources providing pixel-based information.

6 Using pixel-based information

Color information can be useful for the face blurring problem as the color of the faces, the skin tone, is very distinct from others colors. It is thus an interesting information that can be used to detect skin, and thus faces, in complex scene images. It is actually a widely studied subject [3, 15, 33]. We used in this paper the same detector as in [30], with the same parameters, and which gives a classification of pixels as skin or non skin.

In order to combine this color information with the others detectors, a mass function has to be associated to each pixel of the image. As the used skin detector returns a binary decision, either skin or not skin, it returns a discrete information. Thus, this information can be calibrated using the evidential Dempster calibration. When a pixel $p_{x,y}$ is classified as skin by the skin detector, it is possible to obtain a MF representing the uncertainty with respect to the presence of a face on pixel $p_{x,y}$. The necessary training set is obtained using L images on which the skin detector is applied; the process is the same as in Section 4.2. The numbers n of pixels which have been classified as skin can be obtained, and for the i^{th} of these n pixels the classification of this pixel as skin can be encoded by the discrete information $X_i = 1$. As the positions of the faces on these L images are available, its true label Y_i is available. Thus, using this training set and Eq. (13), the MF representing the uncertainty with respect to the presence of a face on pixel $p_{x,y}$ when this pixel is classified as skin can be calculated. In addition, given a pixel classified as non skin, the whole process can be applied to define a MF representing the uncertainty with respect to the presence of a face on pixel $p_{x,y}$.

The same experiment as in Section 5 was performed, including the four face detectors, the two different datasets, and the decision strategy. The repartition of

the images for the calibration training and the tests was also the same. The fifth source, *i.e.,* the skin detector which gives information on pixels, was simply added to the global system. Figure 8 compare the results of the pixel-based approach proposed in Section 5 and the new system now relying on a combination of the outputs of five detectors instead of four. The color detector is only represented by one point in Figure 8 because all the pixels considered as skin have the same MF, likewise for the pixels indicating non skin. Thus, as for the deep neural network detector, increasing the cost c(0,1) does not gradually increase the number of blurred pixels. Actually, at some value of cost c(0,1), which is not represented in Figure 8, a second point for the color detector is obtained but it corresponds to a useless point where all the pixels are blurred by the color detector.

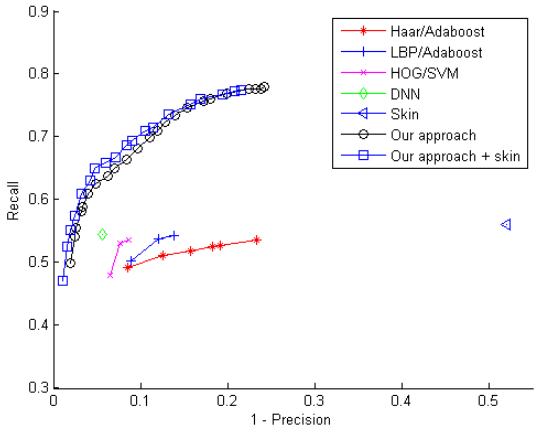


Fig. 8: Integration of skin color information to the proposed approach on FDDB.

As it can be noticed, the addition of the skin color information improves the global combination although the performance of skin detection is not that good. Finally, we conducted the experiment on the SNCF dataset and the results are shown in Figure 9. It also improves the overall performances.

7 Conclusion

In this paper, a pixel-based face blurring system relying on evidential calibration and fusion of several detector outputs was proposed. This pixel-based approach brings several advantages over a previous box-based proposal. First, an overlap threshold is no longer necessary, as well as a clustering step. Furthermore, it enables to integrate all kind of detectors, either returning discrete or continuous information, as well as pixel-based or box-based information. In particular, in the considered blurring problem, it allows us to model and to integrate to the

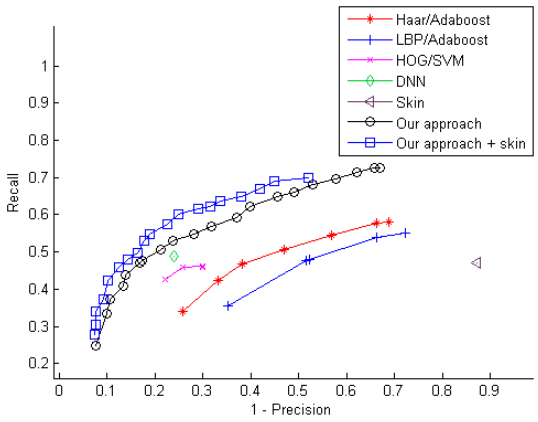


Fig. 9: Integration of skin color information to the proposed approach on SNCF dataset.

fusion process the information of score absence for each detector, *i.e.*, a MF is defined for pixels which are not contained by any of the boxes returned by the detector. The proposed system also shown better performances than the box-based approach, either on a literature dataset or on a more challenging one. We also illustrated the ability of natively integrating a detector giving pixel-based outputs by adding a skin color detector to the global system; this latter addition further improved the overall performances.

The proposed approach can be applied with other detectors, which may return discrete or continuous information, and can be based on boxes or pixels. One perspective consists in replacing one of the face detectors, or to add one to the global system.

Another perspective is to make use of the spatio-temporal context of a given pixel. It is reasonable to consider that a pixel is more likely to be blurred if its neighbours have been blurred. Similarly, if videos have to be handled instead of still images, one could use the fact that a pixel is more likely to be blurred if it has been blurred on a previous image. Taking advantage of such contextual information is an important field of application of the Markov random field theory [19, 20, 22], which has been extended to the evidential framework in [31, 12], and it could be an inspiration to extend our approach or, alternatively, a detector using this theory could be added to the global system.

References

1. P. Aarabi, J. C. L. Lam, and A. Keshavarz. Face detection using information fusion. In *Proceedings of the 10th International Conference on Information Fusion*, pages 1–8, Quebec, Canada, July 2007.

2. M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.
3. J. Brand and J. S. Mason. A comparative assessment of three approaches to pixel-level human skin-detection. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 1, pages 1056–1059, Barcelona, Spain, Sept. 2000.
4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, San Diego, California, June 2005.
5. A.P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, 37(2):355–374, 1966.
6. T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.
7. T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.
8. 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, Sept. 2009.
9. 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.
10. S. S. Farfade, M. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the International Conference on Multimedia Retrieval*, Shanghai, China, June 2015.
11. F. Faux and F. Luthon. Theory of evidence for face detection and tracking. *International Journal of Approximate Reasoning*, 53(5):728–746, 2012.
12. L. Fouque, A. Appriou, and W. Pieczynski. An evidential Markovian model for data fusion and unsupervised image classification. In *Proceedings of the Third International Conference on Information Fusion*, volume 1, pages TUB4–25, Paris, France, July 2000.
13. A. Hadid, M. Pietikäinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–797, Washington, DC, June 2004.
14. V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, Univ. Massachusetts, 2010.
15. M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.
16. I. A. Kalinovskii and V. G. Spitsyn. Compact convolutional neural network cascade for face detection. In *Proceedings of the 10th Annual International Scientific Conference on Parallel Computing Technologies (PCT)*, volume 1576, pages 375–387, Arkhangelsk, Russia, March 2016.
17. O. Kanjanatarakul, T. Denœux, and S. Sriboonchitta. Prediction of future observations using belief functions: A likelihood-based approach. *International Journal of Approximate Reasoning*, 72:71–94, 2016.
18. O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Forecasting using belief functions: an application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.

19. S. H. Khatoonabadi and I. V. Bajic. Video object tracking in the compressed domain using spatio-temporal Markov random fields. *IEEE Transactions on Image Processing*, 22(1):300–313, 2013.
20. R. Kindermann and L. Snell. *Markov random fields and their applications*. American Mathematical Society, 1980.
21. D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
22. S. Z. Li. Markov random field models in computer vision. In *Proceedings of the European Conference on Computer Vision*, pages 361–370, Stockholm, Sweden, May 1994.
23. G. C. Luh. Face detection using combination of skin color pixel detection and viola-jones face detector. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, volume 1, pages 364–370, Lanzhou, China, July 2014.
24. H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC press, 2006.
25. E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–136, San Juan, Puerto Rico, June 1997.
26. J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large-Margin Classifiers*, 10(3):61–74, 1999.
27. H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
28. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.
29. P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
30. M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen. Using the skin locus to cope with changing illumination conditions in color-based face tracking. In *Proceedings of the Nordic Signal Processing Symposium*, volume 38, pages 383–386, Kolmarden, Sweden, June 2000.
31. H. Soubaras. On evidential Markov chains. In *Foundations of Reasoning under Uncertainty*, pages 247–264. Springer, 2010.
32. Z. S. Tabatabaie, R. W. Rahmat, N. I. B. Udzir, and E. Kheirkhah. A hybrid face detection system using combination of appearance-based and feature-based methods. *International Journal of Computer Science and Network Security*, 9(5):181–185, 2009.
33. J-C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Proceedings of the fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 54–61, Grenoble, France, March 2000.
34. P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
35. P. Xu, F. Davoine, and T. Denœux. Evidential combination of pedestrian detectors. In *Proceedings of the 25th British Machine Vision Conference (BMVC)*, pages 1–14, Nottingham, England, Sept. 2014.
36. P. Xu, F. Davoine, H. Zha, and T. Denœux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning*, 72:55–70, 2016.

37. S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3676–3684, Santiago, Chile, Dec. 2015.
38. B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the International Conference on Machine Learning*, pages 609–616, Williamstown, Massachusetts, June 2001.
39. B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 694–699, Edmonton, Canada, July 2002.
40. C. Zhang and Z. Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1036–1041, Steamboat Springs, Colorado, March 2014.