11th WCRR 2016
World Congress on Railway Research
29th May – 2nd June
Milan - Italy

A fusion method for blurring faces on platforms using belief functions

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

[1] SNCF. Paris. France

{pauline.minary,benjamin.droit}@reseau.sncf.fr

[2] Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), F-62400 Béthune. France

{eric.lefevre,david.mercier,frederic.pichon}@univ-artois.fr

Abstract:

In this paper, we propose a global face detection system based on information fusion. Such a system is necessary to make non-identifiable the faces that are visible on platform videos, as required by French legislation related to public video usage and storage. Our system relies on efficient state-of-the-art face detectors, which find the face positions on images. Specifically, it performs first a pixel-wise combination of their outputs in order to take advantage of the potential complementarities of these detectors, which use different image features and different classification procedures. Then, for each pixel of the image to treat and using the result of the merging, a decision is made whether it should be blurred or not. The combination step is grounded on a now well-established framework for reasoning under uncertainty called the Dempster-Shafer theory of belief functions. In this step, detector outputs are converted into a common representation known as belief function using a calibration procedure, and then are merged using so-called Dempster's rule of combination. Our approach is tested on a classical face detection dataset from the literature, showing good performances.

1. Introduction

Safety is one of the most important challenges for SNCF, the French railway company. Since the 80's, a process called EAS (Driver-Only-Operation) has been developed in order to allow the train driver to watch the railway platform in its entirety so a train can be started without any problem. This is possible using cameras and monitors. For the purpose of checking the proper positioning of cameras, a series of videos is recorded. However, according to the French legislation about respect for private life, a video shall not be retained by the company if it contains identifiable people on it. SNCF would like to keep these videos for different purposes, such as preventive maintenance; thus, the only viable solution is to make people faces unrecognizable. This task can be manually performed frame by frame, but it is a time-consuming, tedious and dull work: it is therefore essential to automate it.

Existing solutions [1], [2], consist in automatic face detectors, which provide for a given image a set of bounding boxes, each associated to a confidence score, and corresponding to the positions of the assumed faces within the image. However, preliminary performance tests indicate that neither of these solutions is sufficient because of the challenging conditions of the application, such as image quality, indoor/outdoor situation, variation of lighting, etc. Yet, these detectors do not always give similar outputs, so it seems interesting to use all the available information and try and combine their outputs in order to increase detection rate and decrease false alarm. As a matter of fact, an interesting approach has recently been proposed (in the context of pedestrian detection) to merge the different boxes returned by detectors, taking into account as well the scores associated with these boxes [3]. In this approach, scores are converted into a common representation using a calibration procedure, and then they are merged using a rule of combination. These steps of calibration and combination are conducted

within a framework for reasoning under uncertainty called the Dempster-Shafer theory of belief function [4], [5]. However, this approach suffers from two main limitations. First, it implies a step of association in order to define which boxes corresponds to which face. This task becomes tricky in a multi objects situation, especially when they are close to each other, which is the case for a crowd waiting for a train. Second, the calibration procedure involves an arguably rather arbitrary parameter, which determines when it should be decided that a box returned by a detector matches or not a face whilst taking into account also the size of this box.

To alleviate these two limitations, we propose an approach positioned at a lower level, meaning pixel-based instead of box-based. Within this scope, a score of a box is associated to every pixel contained in this box. As will be seen, this makes the parameter involved in the calibration step no longer necessary. Besides, the association step is also no longer required as the question we then address is whether a given pixel is part of a face and thus all the scores associated to a given pixel can be simply combined directly. Finally, as it is more critical not to blur faces when there are some than the opposite, a concept of costs is also added in our approach to take into account this aspect. The cost values allow the system to be more or less selective in its choice of blurring. Figure 1 illustrates the different steps of our global system, which will be further detailed in the remainder of this paper.

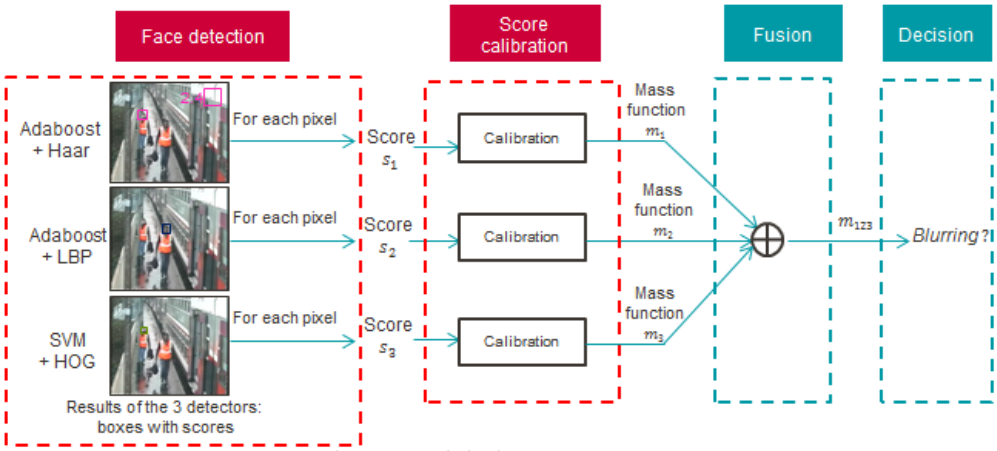

Figure 1: global system steps

This paper is organized as follows. First, selected face detectors as well as the transformations of their outputs into a common representation are discussed in Section 2. Section 3 describes the fusion of the converted scores and the decision step, for each pixel. Evaluation methodology and experimental results are presented in Section 4.

2. Transformation of face detector outputs into a common representation

The first step of our system consists in finding faces in images, thus some face detectors have to be selected; this is described in Section 2.a below. Unfortunately, confidence scores depend on the detectors and have thus to be converted into a common representation so that scores from different detectors may be combined; this is discussed in Section 2.b.

a. Face detectors

A face detector typically relies on two components: a feature extractor, which transforms an image into an alternative form that puts the emphasis on a particular aspect in the image (such as contours of objects), and a classification algorithm, that is a process able to take as input such a form and decide whether it contains a face or not. Let us note that a classification algorithm is able to take such an informed decision thanks to a preliminary so-called training stage performed offline, where the algorithm "learns" how to recognize faces using a database of face and non-face images.

More specifically, a face detector works as follows. Given a test image, it scans across the image at different scales. At each position of the scanning window, image features are extracted and the classification algorithm decides whether the candidate window corresponds to a face. This multi-position multi-scale test generates multiple detections for one face. A post processing is thus required in order to cluster these detections into a single final detection by face. The detector output is then a set of boxes, each supposedly bounding a face.

After a thorough study of the state of the art, three detectors have been selected. The first selected detector is the one proposed by Viola and Jones [1], which is based on a classification algorithm called Adaboost and that uses Haar feature [1] extraction. The second detector is a variant of the previous one: the same classification algorithm is used but with Local Binary Patterns (LBP) [6] feature extraction. The third selected detector is based on a classification algorithm called Support Vector Machine (SVM) and uses Histogram of Oriented Gradients (HOG) features [2].

b. Score calibration

For each box, and thus for each pixel, returned by a detector, it is possible to get a score calculated by the detector. This is valuable information because it provides an indication on how confident the detector is toward each box/pixel. The obtained scores range differs depending on the type of the classification algorithm underlying the detector, but also on the training database. Thus, transposing all of the scores in a common representation, such as a probability distribution, is essential. This step is called score calibration [7]: the goal is to build a function, which for each possible score, returns the probability that the associated box/pixel corresponds to a face.

To build such a function, a database $X = (\{x_1, y_1\}, \dots \{x_n, y_n\})$ composed of n couples $\{x_i, y_i\}$ where x_i are scores and y_i are labels (1 for face, 0 for non-face), is needed [3]. In a box-based approach, a measure based on intersection and union overlapping areas between boxes is usually calculated in order to automatically determine if a box returned by a detector actually corresponds to the ground truth [8]. If the measure exceeds a threshold, the box is regarded as corresponding to the ground truth, so label shall be taken as 1 and 0 otherwise. But this threshold is arbitrarily set. In our pixel-based approach, instead of having *1 detected box = 1 couple (score, label)*, we have *1 pixel of the detected box = 1 couple (score, label).* The measure for attributing label to score becomes: if the pixel of the detected box is contained by a ground truth box the label is 1, otherwise 0. As a consequence, the threshold involved in building the database X in the box-based approach [3] is no longer needed.

Once the database is available, the score calibration can be done. A commonly used method for calibration is called binning [7]. Its principle consists in dividing the score spaces into different bins. For each bin j, the proportion of positive examples k_j over all the examples n_j which fall into this bin is calculated. Given a new score s_k, the bin j which contains this score is found. Then, the probability that the pixel associated to this score belongs to a face is simply $P(1|s_k) = \frac{k_j}{n_j}$, and that it does not $P(0|s_k) = 1 - P(1|s_k) = 1 - \frac{k_j}{n_j}$.

However, this method presents the disadvantage of not taking account of the uncertainty produced by the database. Indeed, some score values are less present than others in this database; their associated probability is thus less precise because there is less information available. In order to manage these uncertainties, a binning calibration method based on belief function theory has been proposed in [7], which we briefly recall now.

Let $\Omega = \{0,1\}$, called the frame of discernment, be the finite set of possible answers to the problem, with in our case 0 corresponding to non-face, and 1 to face. A mass function over Ω is a function $m: 2^{\Omega} \rightarrow [0,1]$ which verifies $\sum_{A \subseteq \Omega} m(A) = 1$, which means in our case $m(\{0\}) + m(\{1\}) + m(\{0,1\}) = 1$. Mass $m(\{1\})$ represents the belief committed exactly to the hypothesis that the pixel belongs to a face, $m(\{0\})$

that it does not, and $m(\{0,1\})$ represents the amount of ignorance. This latter mass highlights the difference with the probability framework. In the belief function framework, the score calibration can be done using Dempster's model, which gives the following mass function [7]:

$$m(\{1\}|s_k) = \frac{k_j}{n_j + 1}, \qquad m(\{0\}|s_k) = \frac{n_j - k_j}{n_j + 1}, \qquad m(\{0,1\}|s_k) = \frac{1}{n_j + 1}.$$

We note that there may be some pixels in an image for which some or even all of the detectors do not provide a score. When a detector does not provide a score for a pixel, the calibration procedure cannot be used; however, in this case, it is safe to assume that the detector is almost certain that the pixel does not belong to a face, which can be modelled by the following mass function:

$$m(\{1\}|no\ score) = 0, \qquad m(\{0\}|no\ score) = 0.99, \qquad m(\{0,1\}|no\ score) = 0.01\,.$$

3. Fusion and decision making processes to blur or not pixels

Once the score calibration is done, the fusion step can be performed. For each pixel, the obtained mass functions are merged and result in a final mass function; this is presented in Section 3.a. Using this final mass function and a cost mechanism, it is decided if the pixel belongs to a face, and thus has to be blurred, or not, as described in Section 3.b.

a. Fusion

As three detectors are used, three mass functions have to be merged for each pixel. The basic operation for combining two mass functions m_1 and m_2 induced by independent sources of information is Dempster's rule of combination. It is defined by:

$$m_{1,2}(A) = m_1 \oplus m_2(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \neq \emptyset,$$

where $K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ represents the degree of conflict between these two mass functions and where $m_{1,2}$ represents the mass function resulting from the combination of m_1 and m_2. This rule is commutative and associative, which means $m_{1,2}(A) = m_{2,1}(A)$ and $m_{1,2} \oplus m_3\,(A) = m_1 \oplus \mathrm{m}_{2,3}(A)$.

b. Decision

In order to decide whether to blur a pixel, a concept of costs [9] can be introduced to take into consideration the difference of impact on the final result depending on the decision.

Let $D = \{d_0, d_1\}$ be the set of decisions that can be made, with d_1 corresponding to blurring and d_0 the opposite. A cost $c(d_i, \gamma)$, with $i, \gamma \in \{0,1\}$ represents the cost of making decision d_i when γ is the correct state. The cost $c(d_0, 1)$ has to be higher than $c(d_1, 0)$ because it is more serious to consider a face pixel as non-face than the opposite, because the purpose is to minimize the number of non-blurred faces. The costs $c(d_0, 0)$ and $c(d_1, 1)$ are equal to 0.

The risks are defined by $R_0 = c(d_0, 1) m_{1,2,3}(\{1\})$ and $R_1 = c(d_1, 0) m_{1,2,3}(\{0\})$ with $m_{1,2,3}$ the mass function resulting from the combination of m_1, m_2 and m_3. Thus, for each pixel, the risks are calculated and the final decision is the one, which corresponds to the lowest risk. The costs can then be adapted in order to influence the decision and obtain more or less blurred pixels on the final image.

4. Experimental results

In this section, the obtained results with the proposed approach on a literature database are presented.

a. Database

A large image database with annotations is required for several reasons. Annotations give the positions (ground truth) of all the faces in the images. As explained, classification algorithm need examples to be trained, but it is also required for building a transformation function during the step of score calibration (see Section 3). Finally, by comparing the system outputs and the ground truth, the performances can be measured.

The database that we used is a literature database for face detection called Face Detection Data Set and Benchmark (FDDB) [10]. This data set contains the annotations for 5171 faces in a set of 2845 images, which represent various situations and include occlusions, difficult poses, and low resolution faces. Nevertheless, the aim is to apply the system on station platforms videos, and the database examples must be the most representative as possible regarding the considered application. Thus our own database, with annotations, is currently being created to complete the previous one.

b. Evaluation methodology

The goal of the evaluation consists in confronting the system results and the ground truth and to determine if the system concealed the faces correctly. Results of a detection system can be expressed by the recall and precision rates, often used to quantify performance [11]. When a blurred pixel belongs to a ground truth bounding box, i.e. is correctly classified as face, it is a True Positive (TP). When it actually does not belong to a face, it is called a False Positive (FP). A face pixel classified as non-face is a False Negative (FN). The rates are then defined by:

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP}.$$

Recall indicates the number of pixels, which the system correctly blurred, compared to what it should have blurred. Precision gives information about the amount of false positives: if the precision is low, a lot of non-face pixels are blurred.

c. Results and comments

The test database includes 200 images of FDDB, which contain 358 faces. By fixing $c(d_1, 0)$ at 1 and varying $c(d_0, 1)$ from 1 to infinity, different results of recall and precision rates can be obtained. Figure 2 illustrates these rates for the three different detectors and for our approach. As we can see, for a given precision rate, our approach has the highest recall rate most of the time.

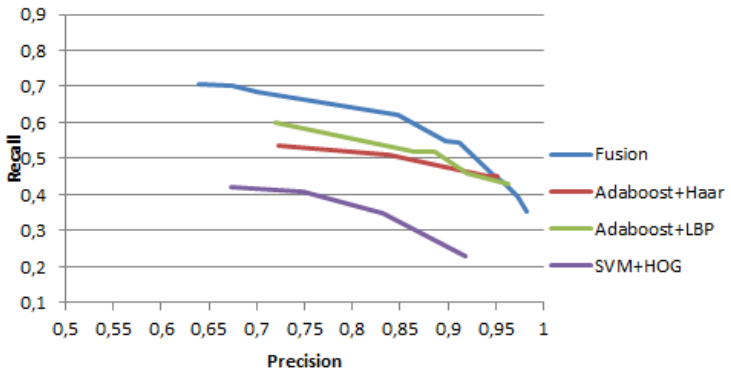

Figure 2: results of precision and recall rates

In terms of faces, and for fixed costs, if we consider that a face is correctly burred when 30% of their pixels are, we obtain 342 correctly blurred faces over 358, meaning a recall of 95.5%. This recall is respectively equal to 82.1% and 52.8% if a face is regarded as correctly blurred when 50% and 70% of their pixel are.

5. Conclusion

In this paper, we proposed a pixel-based approach, which merges several sources of information and uses the theory of belief functions in order to manage uncertainties. Some tests on a literature database show that an information fusion allows getting better performances than existing face detectors taken alone.

In terms of outlook, several improvements are envisioned, such as extending the calibration procedure to learn the mass function associated to pixels having no score. In addition, another face detector, based on colour skin detection, will further be added to the global system. Moreover, a SNCF database is created in order to have face examples closer to the application. Finally, this paper only presented work on still images; however the system inputs are videos. Thus, it would be interesting to take into account for each image the information contained by the previous and next frames, in particular the previous and next face positions.

This research project will allow SNCF to obtain a robust system for automatically blurring faces on videos, and leads to a more efficient use of video-protection cameras filming railway platforms and stations. Afterwards it can also be greatly useful for developing others applications which can improve the users' safety.

References

[1] P. Viola, and M. J. Jones, Robust real-time face detection. *International journal of computer vision*, 57(2), 137-154, 2004.

[2] 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*, pp. 130-136, 1997.

[3] Ph. Xu, F. Davoine and T. Denœux. Evidential Combination of Pedestrian Detectors. *In Proceedings of the 25th British Machine Vision Conference (BMVC)*, Nottingham, UK, September 1-5, 2014.

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

[5] Ph. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191-243, 1994.

[6] A. Hadid, M. Pietikäinen and T. Ahonen. A discriminative feature space for detecting and recognizing faces*. In Proceedings of the Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. II-797, 2004.

[7] P. Xu, F. Davoine, H. Zha and T. Denoeux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning*. 2015. DOI:10.1016/j.ijar.2015.05.002.

[8] M. Everingham, L. Van Gool, C. KI Williams, J. Winn and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[9] T. Denoeux. Analysis of evidence-theoric decision rules for pattern classification. *Pattern Recognition*, 30(7):1095-1107, 1997.

[10] V. Jain and E. Learned-Miller. FDDB: A Benchmark for Face Detection in Unconstrained Settings. Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst. 2010.

[11] C. D Manning, P. Raghavan and H. Schütze, *Introduction to information retrieval*, Vol. 1, p. 496, Cambridge, Cambridge university press, 2008.