An Evidential Framework for Pedestrian Detection in High-Density Crowds

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

Abstract

This paper addresses the problem of pedestrian detection in high-density crowd images, characterized by strong homogeneity and clutter. We propose an evidential fusion algorithm which is able to exploit multiple detectors based on different gradient, texture and orientation descriptors. The evidential framework allows us to model the spatial imprecision arising from each of the detectors. A first result of our study is that the fusion results underline clearly the good complementarity among the four descriptors we considered for this specific context. Moreover, the proposed algorithm outperforms a fusion solution based on Multiple Kernel Learning on difficult high-density crowd images acquired at Makkah at the height of the Muslim pilgrimage.

1. 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 security risks linked to this context. Although a significant effort has been devoted in the last decade to pedestrian detection, the advances proposed in the literature are not always transferable to crowd detections for multiple reasons, among which we can recall the absence of background, the heavy occlusion of body parts, the high visual homogeneity and the small size of the targets. Under these circumstances, it is essential to rely on multiple visual descriptors which are able to provide augmenting views of the data, and which may thus represent the foundation for independent detectors. Then, a global decision has to be performed 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.

The objective of our work is twofold. We aim to highlight, among some descriptors which are widely used in pedestrian detection, those which are the best suited for detection in high-density crowds. Secondly, we propose an evidential fusion solution adapted for this decision context which is particularly difficult for individual detectors. The experiments show that the proposed method performs better than a state of the art statistical learning fusion framework.

2. Related work

For the detection task in high-density crowds, it is not immediately clear which descriptors are the most adapted or discriminative, and which fusion strategy is the most effective. For example, one of the most effective preprocessing steps in pedestrian detection is the background subtraction. Beside removing (potentially significant) parts of the scene which do not contain objects of interest, background subtraction allows for the use of descriptors built upon the blobs associated to the foreground, such as their skeleton or the shape of the foreground connected components [13]. Unfortunately, in a cluttered scene this approach is ineffective due to the limited presence of background, and detectors adapted for crowded areas need to rely exclusively on appearance cues.

Detectors Among the appearance cues, the simplest descriptors rely on a local color histogram, which may be associated to skin, hair or clothes. However, this approach is limited by multiple factors: the object resolution needs to be relatively high, the color spaces are not discriminative enough for difficult tasks, and lastly many surveillance cameras provide gray level data. Likewise, common face detectors such as Viola-Jones [21] are unsuited, since pedestrian faces are not detailed enough.

Among the descriptors related to the image gradient, the Histogram of Oriented Gradients (HOG) descriptor [5] is very popular and has exhibited in various contexts an excellent performance when used either in conjunction with

a linear SVM, or with a histogram intersection kernel (HIK) [24]. More generally, the contour related to the specific shape of the head and shoulders is indeed highly discriminative, but it may fade away due to clutter. Supervised learning may be used in order to enhance the local edge map according to a training contour dataset [23], but it is also advisable to rely on descriptors aimed at other features than shape.

Such a feature which has been often used for detection in crowded scenes is the Local Binary Pattern (LBP) operator [15]. The traditional use of LBP is in texture classification, but due to its local sampling strategy it exhibits a reasonable robustness to occlusion as well. Some alternative solutions are the covariance matrix based descriptors [10], but the main advantages of LBP are its compactness and low computational cost. Also related to texture representation, Gabor filter banks have been used for head detection [11] to encode the local frequency and orientation.

Fusion Regarding the fusion of detectors based on different features, some fundamentally different approaches are popular in the literature for the pedestrian detection task. In order to avoid fusing different detectors altogether, a radical choice is to simply concatenate all the available representations before classification. The fundamental limitations of this solution are due to the suboptimal projection of all the features into a common discriminative space, and to the risk of overfitting.

Another straightforward, popular 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 [10]. Although this strategy is beneficial for real-time performance, the decision fusion does not benefit systematically from all the available information, and this fact is detrimental in a difficult setting such as high-density detection, where individual classifiers exhibit rather modest performance.

In order to benefit simultaneously from all the available features, AdaBoost is well positioned in terms of computation requirements and it remains very popular on architectures with limited power and/or real-time constraints, but it has been shown that nonlinear classification exploits better the available data [8, 3] (with a significant increase in computation cost). In this latter case, the fusion of the classifiers may be performed 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 [7] in order to cope better with individual missed detections. 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 [9]. The problem scales very well with the number of individual classifiers, and efficient implementations can cope with millions of data sources. However, the main limitation of MKL, as with nonlinear SVMs in general and with neural networks, is the difficulty to interpret the final decision function, and to account for spatial imprecision.

Belief function theory is often presented as a general framework that allows us to model both the uncertainty as using probability theory and imprecision as using fuzzy sets or possibility theory. From the seminal work of Dempster and Shafer [18], the Transferable Belief Model proposed by Smets [19] makes the theory very popular and numerous applications have been carried out in very various domains [6], including (video-)surveillance problems such as target recognition [17] or multi-sensor system [2]. Combination of detectors has been proposed by [25] but in a very different context: multi-class problem but with much higher target resolution, whereas in this work we focus on spatial features of the considered descriptors.

3. Proposed fusion method

In this study, we propose to use belief function framework [18] to perform fusion between different SVM based pedestrian detectors which provide a probabilistic output. In our setting, the crowd exhibits high density characterized by strong occlusions, rendering the head of each pedestrian barely visible. Besides, due to the specific geometry of such 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 increase the spatial imprecision of the detection.

Since belief function theory has been proposed to handle both uncertainty and imprecision, we naturally look to this framework. Details about it will be provided after the presentation of the descriptors we considered.

3.1. Descriptors

HOG features The HOG, introduced in [5], grasps the shape of interest by histograms of the distribution of local intensity gradients or edge directions. For each pixel, HOG is computed considering a larger window around it, which is in turn divided into smaller cells, and for each cell a local histogram of gradient directions is accumulated. Cells can be then grouped into blocks, and the histograms contained in each block are normalized to take into account changes in illumination and contrast. The resulting histograms are lastly concatenated into a single vector.

LBP features The LBP [15] is a powerful texture descriptor. For each pixel, the 3×3 neighborhood is thresholded with respect to the pixel value, and the resulting string reading the neighbors' values clock-wise is interpreted as a binary number and used as a label. Then, a histogram is de-

rived from the labeled image.

LBP has been enhanced in [16] to support neighborhoods of different sizes, and to consider uniform patterns, i.e. patterns which have at most two transitions from 0 to 1 or vice-versa in a circular fashion. In fact, it has been observed that uniform patterns represent well local primitives, such as corners or edges, and hold most of the information related to texture. This has the advantage to reduce considerably the size of the histogram, since all the non-uniform patterns fall in the same bin. Formally, the LBP operator at (x_c, y_c) location having a gray-scale value of g_c is defined as:

$$LBP_{p,r}(x_c, y_c) = \sum_{p=0}^{p-1} s(g_p - g_c)2^p, \qquad (1)$$

where p is the number of points interpolated on a r radius. The notation $LBP^{u2}_{p,r}$ means that uniform patterns are considered.

Following the idea of [1], we subdivide the image in small regions from which histograms are extracted and then concatenated, in order to enhance the locality of the LBP.

Gabor features Gabor filter banks are widely used in object recognition and texture segmentation, for capturing global information thanks to the flexibility in the choice of spatial frequencies and orientations. An input image $I(x, y)$ is convolved with Gabor filters $G(x, y)$, i.e. gaussian functions modulated by an oriented complex sinusoidal signal, at different scales and orientations:

$$G(x, y) = e^{-\frac{x'^2+\gamma y'^2}{2\sigma^2}} \cdot \cos\left(2\pi f x' + \phi\right), \qquad (2)$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$, σ regulates the ellipticity of the kernel, f is the frequency of the spatial wave and θ determines the rotation angle in the $[0, 180°)$ domain.

To build the final feature vector, instead of just concatenating the raw responses of every filter in the bank, we subdivide the window around each pixel in several blocks, and then we compute their first and second order statistics.

Daisy features Daisy is a more recent descriptor which has gained popularity particularly in the field of stereo matching [20]. For a given input image, H orientation maps are firstly computed and then convolved several times with Gaussian kernels of different Σ on Q concentric layers with T circles centered on the considered location. Then, histograms of orientations are derived from the central location and from all the Gaussian in every layer, and finally concatenated. Normalization can be applied for every histogram independently or for the whole vector.

The amount of Gaussian smoothing is proportional to the radius of the circle, giving rise to larger Gaussian kernels in the outers rings. For this reason, we consider that Daisy is well suited for our application, as we benefit from a finer description in the center of the head and a coarser description moving away from it. Gaussian smoothing, together with the sampling overlap, naturally enforce spatial consistency. To our knowledge, this descriptor has not been previously used for head detection in crowds.

3.2. Evidential fusion

To handle both uncertainty and imprecision, belief functions are defined on a larger hypothesis set than 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*, $2^\Theta = \left\{\emptyset, H, \bar{H}, \left\{H, \bar{H}\right\}\right\}$.

There are five basic belief functions that are in one-to-one relationship so that the definition of any of them is sufficient to define the other ones. Classically, *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 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. Pl and Bel functions may also be interpreted as upper and lower probabilities [18] 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).

In our case we aim at using bba to model the spatial imprecision due to the close resolutions of object (head) and descriptor respectively. Now, applied in the spatial domain, the work of [4] allows us to define a bba taking into account the spatial imprecision. Specifically, let us consider a structuring element b (related to spatial neighborhood of interest) and an initial bba having only two focal elements simultaneously on b. Then, erosion and dilation (respectively opening and closing) morphological operators may be applied to the initial mass of the considered focal element in order to get Pl and Bel values for the final bba. Indeed, duality property between erosion and dilation (or between opening and closing) allows us to get a well-defined bba. In our case, initial bbas are bayesian (i.e. bbas having only singleton focal elements) defined from binary classifier output (scores for each Θ class) so that initial bbas have only two focal elements. Then, in every pixel s of the image, the bba $m_{i,s}$ associated to descriptor i is simply defined by

$$\begin{aligned} m^{i,s}(A) &= \gamma_a\left(m_0^{i,s}(A)\right), \forall A \in \left\{H, \bar{H}, \right\}, \\ m^{i,s}(\left\{H, \bar{H}\right\}) &= 1 - m^{i,s}(H) - m^{i,s}(\bar{H}), \qquad (3) \\ m^{i,s}(\emptyset) &= 0, \end{aligned}$$

where γ_a is the opening operator of parameter a. We consider opening rather than erosion since, as already pointed in [4], obtained results are much better due to the filtering property of this morphological operator. Now, conversely to [4], we propose the use of a structuring element especially crafted for the spatial descriptors, namely a spatial Gaussian structuring element fitted in a window of radius a.

At the end of bba allocation, considering N descriptors (for this study, $N = 4$), in every pixel s, N bbas are defined that represent the soft output of each of the N binary classifiers. According to bba i, the uncertainty of a head presence in s ranges between $Bel^{i,s}(H) = m^{i,s}(H)$ and $Pl^{i,s}(H) = m^{i,s}(H) + m^{i,s}(\Theta)$ so that $m^{i,s}(\Theta)$ represents the imprecision on the uncertainty value provided by i^{th} descriptor in s. In our model, uncertainty comes from the binary classifier score whereas imprecision comes from spatial heterogeneity of uncertainty values within the considered structuring element.

Defining the bbas associated to the N descriptors allows for combining them. As the descriptors are considered as *cognitively* independent, the orthogonal sum or its unnormalized version, the conjunctive combination rule [19], are well-suited for this task. For two sources m_1 and m_2, conjunctive rule writes $\forall A \in 2^\Theta, m_{1\textcircled{2}}(A) = \sum_{(B,C)\in 2^\Theta\times 2^\Theta,\ B\cap C=A} m_1(B)\, m_2(C)$. When $N > 2$, associativity property of the conjunctive rule may be used. However, in our case where $|\Theta| = 2$, the analytical result may be easily derived:

$$\forall A \in \left\{H, \bar{H}\right\},$$

$$\begin{aligned}
m^s_{\textcircled{1}^N}(A) &= \sum_{\substack{(B_1,\ldots,B_N)\in\{A,\Theta\}^N,\\ \exists j\in[1,N] s.t. B_j=A}} \prod_{j=1}^{N} m^{j,s}(B_j), \\
m^s_{\textcircled{1}^N}\left(\left\{H,\bar{H}\right\}\right) &= \prod_{j=1}^{N} m^{j,s}\left(\left\{H,\bar{H}\right\}\right), \qquad (4)\\
m^s_{\textcircled{1}^N}(\emptyset) &= 1 - m^s_{\textcircled{1}^N}(H) - m^s_{\textcircled{1}^N}\left(\bar{H}\right) - \\
&\quad m^s_{\textcircled{1}^N}\left(\left\{H,\bar{H}\right\}\right).
\end{aligned}$$

Finally, in every pixel, the decision is taken from $m^s_{\textcircled{1}^N}$. 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 [19]. However, with only two hypotheses in Θ, previous criteria boil down to the same decision: $\hat{A} = argmax_{A\in\left\{H,\bar{H}\right\}} m^s_{\textcircled{1}^N}(A)$.

To illustrate the interest of modeling imprecision in addition to uncertainty, let us consider the following toy example with a pixel belonging to a head and four sources available to detect it. Three of them provide a probability of H equal to .6 ($P_{i\in\{1,2,3\}}(\bar{H}) = .4$); however punctual noise present in the fourth source leads to $P_4(\bar{H}) = .9$ (and $P_4(H) = .1$) so that decision based on probability product leads to the wrong label, $\bar{H}$. Now, using the proposed evidential approach, neighborhood information introduced during bba allocation allows for the discounting of the unreliable source (according to the pixel spatial neighborhood). Table 3.2 shows that it leads to the right decision, H.

Table 1. Probability of H in s neighborhood; s is the central pixel; Probability of $\bar{H}$ is the complement with respect to 1.

descriptors 1 to 3			descriptor 4		
	.7			.5	
.5	.6	.5	.5	.1	.5
	.5			.5	

Table 2. Mass allocation, combination and decision (in bold) in case of Table 3.2 probability maps; for example simplicity, erosion with a flat 4-connectivity structuring element is used; for comparison, probability product is shown.

hypothesis, pixel s	H	$\bar{H}$	$\left\{H,\bar{H}\right\}$	$\emptyset$
$m^{s,1} = m^{s,2} = m^{s,3}$	.5	.3	.2	0.
$m^{s,4}$	.1	.5	.4	0.
$m^s_{\textcircled{1}^4}$	**.17**	.11	.0.004	.18
$\prod_{j=1}^{4} p^{s,j}$	.02	**.06**	/	/

4. Experimental results

4.1. SVM settings

For the classification step we rely on SVM, with kernels adapted to each descriptor; cross-validation is performed to set the parameters. Considering that head sizes in our dataset span between 8 and 12 pixels, we compute HOG descriptors in 24×24 windows, in order to include information about the immediate surrounding of the actual head while at the same time avoiding other targets. A L2-hys normalization is applied for each block. For learning, we rely on the HIK.

A $LBP^{u2}_{1,8}$ is used over 12×12 windows subdivided into four 6×6 blocks. The choice of the window size is sensitive, as larger windows result in wide detections, overflowing the actual heads. Stride between blocks has also been tested but it does not provide consistent improvement. Following the example of [1] which employs on a χ^2 distance as a dissimilarity measure, we rely on a χ^2 kernel function which has been shown to be positive definite and suited for data generated from histograms [12].

We use a Gabor filter bank of 5 scales and 4 orientations; a high number of scales is essential to obtain good results, while increasing the number of orientations does not provide an effective gain in performance. For each Gabor filter response image, we compute and concatenate mean and

standard deviation over 4×4 blocks on 16×16 windows. Then, a RBF kernel is considered for learning.

For the Daisy descriptor, we use a radius $R = 8$ from the center to the outer ring, with $Q = 3$ number of layers and $T = 8$ histograms of $H = 8$ bins at each layer. As for the HOG, the HIK is employed for SVM classification.

4.2. Results

We tested our proposed fusion method on high-density crowd images acquired at Makkah during Hajj.

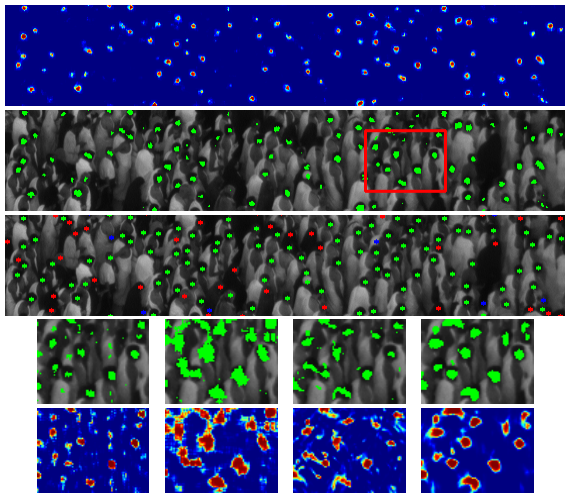

Figure 1. Example of classification results. First row: colormap of the fusion output. Second row: detections after thresholding at $precision = 0.92$ and $recall = 0.72$. Third row: detection decision after NMS. Fourth and fifth rows: detections and colormaps respectively for the employed descriptors, namely HOG, LBP, Gabor and Daisy.

Figure 1 shows an example of the classification result, namely a colormap of the final $BetP(H)$ map (for $|\Theta| = 2$, $BetP(A) = m(A) + \frac{m(\Theta)}{2}, \forall A \in \Theta$), and a thresholding result to visualize easier the inferred head locations. For the highlighted box, each independent source is presented. We notice that each source has a specific behavior, which underlines their complementarity. HOG and Gabor provide a localized detection, but more small false positives are present. On the contrary, LBP and Daisy provide larger and rougher results. In particular Daisy provides very smooth detections, due to the Gaussian based spatial sampling.

SVM performs a dense classification across the test area, and decision with fusion is taken at pixel level. To be able to quantitatively assess the results with object level statistics based on detection level, we perform a non maxima suppression (NMS) on the output probability map [14]. Setting $r = 2$ as head radius, the minimum distance between two maxima is fixed to $2r+1$, to avoid overlapping between different detections. Fig. 1 shows an example of the detections after NMS, highlighting in green True Positives (TPs), in red False Negatives (FNs) and in blue False Positives (FPs). Although the crowded scene is difficult, most of the heads are correctly detected, even if a clear problem remains with regards to the detection of dark veils, which are not well handled by the initial sources. Nevertheless, the number of FPs is consistently reduced.

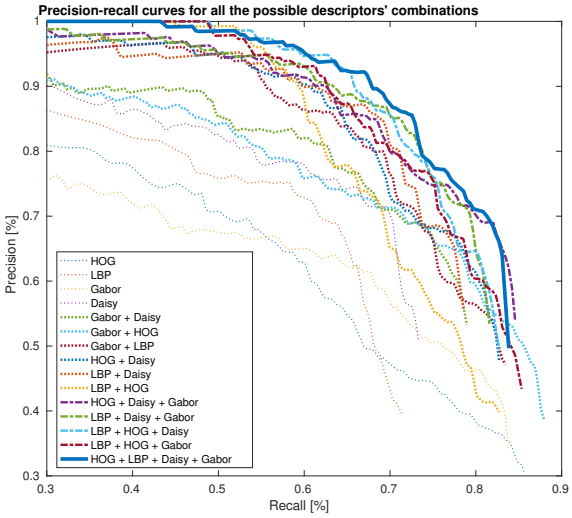

Figure 2. Precision-recall curves for the fusion of all the possible descriptors' combinations. Initial sources are shown as reference. The proposed fusion approach outperforms other approaches.

Figure 2 shows the precision-recall (PR) curves for the four descriptors, namely the *sources* of the proposed evidential approach, and the fusion results achieved combining any subset of sources. The plot underlines the tendency of HOG and Gabor descriptors to detect more FPs, which translates into having a lower overall precision, while LBP and Daisy reach higher levels of precision but have a lower recall, as they miss more heads. Likewise, we note that the combination HOG+Gabor provides the best recall, and in the same way LBP+Daisy consistently improves the precision values. This plot shows that each descriptor contributes to the overall results, which highlights again their complementary relationship.

Figure 3 shows the PR curve of the proposed fusion method compared to the MKL implementation provided by [22] and the 'naive' product of probabilities. The curves related to the four initial descriptors are left as a reference. The proposed fusion method provides higher recall and precision values with respect to product and MKL. Compared to the linear kernel combination computed by MKL, the fusion algorithm takes advantage at a local scale of the information provided by the independent detectors. The good performance of the product can be explained by the relatively consistent behavior of the classifiers and by the presence of only two singleton hypotheses. However, as shown

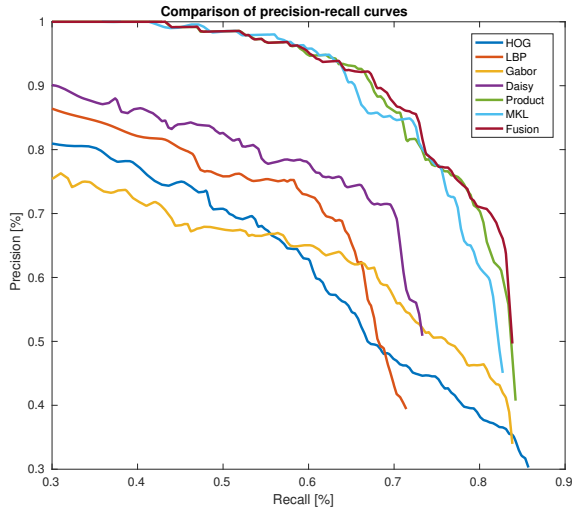

Figure 3. Precision-recall curves for our fusion approach, MKL and product of probabilities. Initial sources are shown as reference. The proposed fusion provides overall better precision and recall values.

in the toy example, this approach would not extend well to less reliable detectors.

5. Conclusion

In this paper we proposed an evidential framework which addresses pedestrian detection in high-density crowds. The study of the combinations of the different classifiers which were considered proves that the underlying descriptors complement each other, and that the fusion algorithm is able to exploit their synergy and to take into account spatial imprecision. Finally, since our solution does not perform as a "black box" algorithm, our work opens up some interesting avenues for refining the selection, training and use of the individual detectors which were considered in this work.

Acknowledgments

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

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. *ECCV*, pages 469–481, 2004.

[2] C. André, S. Le Hégarat-Mascle, and R. Reynaud. Evidential framework for data fusion in a multi-sensor surveillance system. *Engineering Applications of Artificial Intelligence*, 43:166–180, 2015.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten Years of Pedestrian Detection, What Have We Learned? In *ECCV Workshops*, pages 613–627, 2014.

[4] I. Bloch. Defining belief functions using mathematical morphology–application to image fusion under imprecision. *Int. journal of approximate reasoning*, 48(2):437–465, 2008.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893. IEEE, 2005.

[6] T. Denœux. 40 years of dempster–shafer theory. *International Journal of Approximate Reasoning*, (79):1–6, 2016.

[7] 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*, pages 325–330, 2012.

[8] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE trans. on pattern analysis and machine intelligence*, 31(12):2179–2195, 2009.

[9] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011.

[10] R. Hu, R. Wang, S. Shan, and X. Chen. Robust head-shoulder detection using a two-stage cascade framework. In *ICPR*, pages 2796–2801, 2014.

[11] M. Li, S. Bao, W. Dong, Y. Wang, and Z. Su. Head-shoulder based gender recognition. In *ICIP*, pages 2753–2756, 2013.

[12] P. Li, G. Samorodnitsk, and J. Hopcroft. Sign cauchy projections and chi-square kernel. In *Advances in Neural Information Processing Systems*, pages 2571–2579, 2013.

[13] D. Merad, K. E. Aziz, and N. Thome. Fast people counting using head detection from skeleton graph. In *AVSS*, pages 151–156, 2010.

[14] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *ICPR*, pages 850–855, 2006.

[15] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[16] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *ECCV*, pages 404–420. Springer, 2000.

[17] B. Ristic and P. Smets. Target identification using belief functions and implication rules. *IEEE transactions on Aerospace and Electronic Systems*, 41(3):1097–1103, 2005.

[18] G. Shafer et al. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.

[19] P. Smets and R. Kennes. The transferable belief model. *Artificial intelligence*, 66(2):191–234, 1994.

[20] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE TPAMI*, 32(5):815–830, 2010.

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001.

[22] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, and M. Varma. Multiple kernel learning and the SMO algorithm. In *NIPS*, December 2010.

[23] S. Wang, J. Zhang, and Z. Miao. A new edge feature for head-shoulder detection. In *ICIP*, pages 2822–2826, 2013.

[24] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

[25] P. Xu, F. Davoine, and T. Denœux. Evidential combination of pedestrian detectors. In *British Machine Vision Conference*, 2014.