PARTICLE FILTERING WITH A SOFT DETECTION BASED NEAR-OPTIMAL IMPORTANCE FUNCTION FOR VISUAL TRACKING

M. Oulad Ameziane[1,2], *C. Garnier*[1], *Y. Delignon*[1], *E. Duflos*[2] *and F. Septier*[1]

[1] Institut Mines-Telecom / Telecom Lille / CRIStAL UMR CNRS 9189, France
[2] Ecole Centrale de Lille / CRIStAL UMR CNRS 9189, France

ABSTRACT

Particle filters are currently widely used for visual tracking. In order to improve their performance, we propose to enrich the observation model with soft detection information and to derive a near-optimal proposal to efficiently propagate particles in the state space. This information reflecting probabilities about the object location is more reliable than the usual binary output which can yield false or missed detections. Moreover, our proposal not only incorporates the observations as in previous works, but relies on a close approximation of the optimal importance function. The resulting PF achieves high tracking accuracy and has the advantage of coping with unpredictable and abrupt movements.

Index Terms— Visual tracking, Monte-Carlo methods, particle filtering, optimal importance function, soft detection.

1. INTRODUCTION

Visual tracking is one of the most fundamental tasks in many vision applications, such as intelligent surveillance, traffic monitoring, human-computer interaction... Among the many tracking methods [1–3], particle filters (PFs) are currently widely used. Their efficiency strongly depends on the importance function (also called the proposal) which explores the state space. The simplest proposal is the prior density related to the dynamic model. But the object displacement between two video frames can be difficult to predict in case of abrupt movements due to dynamics variations, low frame rate videos and switching between cameras.

As demonstrated in [4], the optimal importance function in the sense of weight variance minimisation takes into account the current observations. In a few cases, a Gaussian approximation can be derived by linearisation [5]. But in most cases, computing the optimal proposal is impossible because the analytic form is unavailable or calculations are prohibitive. The challenge in visual tracking is then to select the most relevant observations and to make the best use of them to efficiently propagate the particles in the state space.

Two main strategies have been proposed. Implicit approaches use the prior proposal and add a step to guide the particles from the observations. The auxiliary PF [6–8] preselects particles before their propagation and improves performance when the state noise is small. Hybrid PFs include an optimization stage based on mean shift [9, 10] or heuristics [11, 12]. The drawback is that they can leave the theoretical framework of PFs, since each particle move can alter the filtering density. Explicit approaches are more direct and build the importance function from the observations. It is based on a Gaussian mixture model (GMM) between the prior density and a density centred on specific points, which can be the locations of high motion activity [13] or the centroids of detected silhouettes [14, 15]. Several detectors can be combined to overcome their unreliability [16].

In this paper, we propose to enrich the observation model with soft detection information and to derive a near-optimal proposal. This intermediate information, obtained in detectors before hard decision, reflects probabilities about the object location. It is more reliable than the usual final binary output which can yield detection errors. Soft detection information has already been used in PFs for particle weighting [17], but according to our knowledge, it has never been exploited for particle drawing. Moreover, our proposal not only incorporates the current observations as in previous works [13–16], but it relies on a close approximation of the optimal importance function [4]. The resulting PF achieves high tracking accuracy and successfully handles the uncertainty of the dynamic model encountered in real-world situations.

This paper is organised as follows. Section 2 presents the general formulation of the visual tracking problem. In Section 3, our soft detection based near-optimal proposal is described. Section 4 presents performance results obtained on public datasets. Conclusions are finally drawn in section 5.

2. VISUAL TRACKING PROBLEM FORMULATION

This paper deals with single object tracking along a sequence of images. The aim is to estimate the object dynamic state x_k from a sequence of observations $y_{1:k} = (y_1, ..., y_k)$. In the Bayesian framework, the distribution of interest is the posterior $p(x_k|y_{1:k})$, also called the filtering density. This density is recursively expressed using the Bayes rule:

$$p(x_k|y_{1:k}) \propto p(y_k|x_k) \cdot \int p(x_k|x_{k-1}) \cdot p(x_{k-1}|y_{1:k-1}) dx_{k-1}$$

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

where the prior density $p(x_k|x_{k-1})$ represents the dynamic evolution of the state x_k given the previous state x_{k-1}, and the observation likelihood $p(y_k|x_k)$ measures the matching accuracy of the observation y_k given the state x_k.

2.1. Dynamic model

The object is represented by a bounding window. The state vector is defined as $x_k = \{c_k, s_k\}$ with $c_k = \{c_k^x, c_k^y\}$ the position of the top left corner and $s_k = \{s_k^x, s_k^y\}$ the size of the window. To address any type of movement, we consider a dynamic model with little information. As in most works [13–16], we assume that the components of x_k evolve as mutually independent Gaussian random walks: $x_k|x_{k-1} \sim \mathcal{N}(x_{k-1}, \Sigma)$ where $\Sigma = diag(\sigma_c^2, \sigma_c^2, \sigma_s^2, \sigma_s^2)$ is the covariance matrix which defines the uncertainty region around the previous state. In real scenarios, the object can perform large amplitude changes in position while the size evolves smoothly. Therefore the position variance σ_c^2 is much larger than the size variance σ_s^2.

2.2. Observation model

The observation model includes the usual colour information and is enriched with soft detection information extracted from each image I_k.

The colour information is expressed as a set of RGB histograms: $y_k^H = hist(I_k \cdot \mathbb{1}_{R(x_k)})$ with $R(x_k)$ the region defined by x_k. As in [18], the region $R(x_k)$ is divided into multiple subregions to take into account the colour spatial distribution. A histogram is then computed for each colour and each subregion.

The soft detection information is provided via a motion detector [19] able to detect any kind of object. The principle is to model the background and foreground homogeneous regions by an adaptive GMM in a spatio-colorimetric feature space. Then the pixel classification is based on maximum likelihood and provides a binary mask called the hard detection map. Here, we exploit a richer information that is available in the algorithm before classification. This is the probability map (or soft detection map): $y_k^D = [P_{i,j}]$ where $P_{i,j}$ is the probability that the pixel located at the position (i, j) belongs to the foreground. This type of map is accessible in any visual tracking system because a detector is always required to automatically detect the presence of an object of interest. Figure 1 shows an example of soft and hard detection maps.

Both of these information are fused in the conventional way by assuming that they are conditionally independent given the state [13, 20]. Then the overall likelihood is:

$$p(y_k|x_k) = p(y_k^H|x_k) \cdot p(y_k^D|x_k) = L_H \cdot L_D \quad (1)$$

where L_H is the usual colour likelihood [21] defined from the Bhattacharyya distance D_B between the N_b bin reference

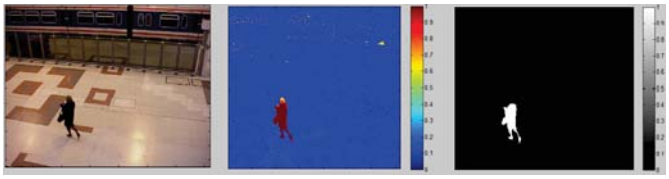

Fig. 1. Detection maps (left to right: original image, soft detection map, hard detection map)

histograms H_k^{ref} and candidate histograms y_k^H for the 3 RGB channels and the S subregions of $R(x_k)$:

$$L_H \propto \exp\left(-\lambda \sum_{p=1}^{3}\sum_{r=1}^{S} D_B^2\left(y_k^H(p,r), H_k^{ref}(p,r)\right)\right) \quad (2)$$

And L_D is the detection likelihood defined from the soft detection map $y_k^D = [P_{i,j}]$ as follows:

$$L_D \propto \exp\left(\lambda_1 . \sum_{(i,j)\in R(x_k)} P_{i,j} - \lambda_2 \cdot N(s_k)\right) \quad (3)$$

with $N(s_k)$ the number of pixels inside the region $R(x_k)$ with size s_k. This formulation avoids that larger regions are systematically encouraged. We can note that this expression is similar to that proposed in [22], except that it contains the probability $P_{i,j}$ instead of a normalized distance from the background. λ, λ_1 and λ_2 are weighting coefficients whose values are empirically chosen.

3. PF WITH THE NEAR OPTIMAL PROPOSAL

Our approach is based on particle filtering which consists in recursively approximating the filtering density $p(x_k|y_{1:k})$ with a set of N_p weighted particles $\{x_k^{(i)}, w_k^{(i)}\}_{i=1}^{N_p}$ evolving in the state space: $p(x_k|y_{1:k}) \approx \sum_{i=1}^{N_p} w_k^{(i)} \cdot \delta(x_k - x_k^{(i)})$.

3.1. General framework

To obtain the set of particles at time k from the previous particles $\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^{N_p}$, we draw the samples $x_k^{(i)}$ using a generic importance function (also called a proposal) $q(x_k^{(i)}|x_{k-1}^{(i)}, y_k)$, and then we update the associated weights according to the recursive expression:

$$w_k^{(i)} \propto w_{k-1}^{(i)} \cdot \frac{p(y_k|x_k^{(i)}) \cdot p(x_k^{(i)}|x_{k-1}^{(i)})}{q(x_k^{(i)}|x_{k-1}^{(i)}, y_k)} \quad (4)$$

The choice of the proposal is essential to perform an efficient exploration of the state space and to ensure a high level of tracking performance. The optimal proposal (in the sense

of weight variance minimisation) takes into account the current observations [4] and is expressed as:

$$p(x_k|x_{k-1}, y_k) = \frac{p(y_k|x_k) \cdot p(x_k|x_{k-1})}{\int p(y_k|x_k) \cdot p(x_k|x_{k-1}) \cdot dx_k} \quad (5)$$

As shown in section 2, the prior density is simple but the likelihood is complex. So the analytic evaluation of the optimal proposal is intractable. Moreover, a pointwise evaluation is computationally too expensive.

3.2. Proposed near-optimal importance function

To efficiently explore the state space, our objective is to find a compromise between computational complexity and importance function optimality. Our approach relies on an approximation of the optimal proposal (5) from an approximation of the likelihood $p(y_k|c_k, s_k)$ (1), which is the most expensive component in terms of computations.

According to the assumptions of subsection 2.1, the position c_k and the size s_k of the object evolve independently, and even in case of abrupt changes in position, the size varies smoothly. Then the likelihood depends much more on the position c_k than on the size s_k of the region $R(x_k)$. To significantly reduce the computational cost of the denominator in expression (5), the likelihood is evaluated for a unique value of s_k: $\tilde{s}_k = E[s_k|\hat{s}_{k-1}]$ with $\hat{s}_{k-1}$ the estimated size at time $k-1$. For a Gaussian prior density, $\tilde{s}_k = \hat{s}_{k-1}$.

With the same purpose, since the likelihood based on soft detection information is computationally much less expensive than the usual colour likelihood L_H, only the former $L_D = p(y_k^D|c_k, s_k)$ is considered in the optimal proposal expression (5). Indeed, according to expressions (2) and (3), for a given value of x_k, L_D requires $N(s_k)$ additions, while L_H needs to calculate $3S$ histograms for the region $R(x_k)$ and all the distances to the reference histograms, which represents $3\left(N(s_k) + 3N_bS\right)$ operations. However L_H is still used in the calculation of particle weights.

The optimal proposal is then approximated by:

$$\hat{p}(x_k|x_{k-1}, y_k) = \frac{p(y_k^D|c_k, \tilde{s}_k) \cdot p(c_k|c_{k-1}) \cdot p(s_k|s_{k-1})}{\hat{p}(y_k^D|c_{k-1}, s_{k-1})}$$

with:
$\hat{p}(y_k^D|c_{k-1}, s_{k-1}) = \int\int p(y_k^D|c_k, \tilde{s}_k) \cdot p(c_k|c_{k-1}) \cdot p(s_k|s_{k-1}) \cdot dc_k \cdot ds_k = \int p(y_k^D|c_k, \tilde{s}_k) \cdot p(c_k|c_{k-1}) \cdot dc_k$

Finally, the near-optimal importance function can be written as the product of two densities: a near-optimal proposal for c_k and the prior proposal for s_k:

$$\hat{p}(x_k|x_{k-1}, y_k) = p(c_k|c_{k-1}, \tilde{s_k}, y_k^D) \cdot p(s_k|s_{k-1}) \quad (6)$$

with:

$$p(c_k|c_{k-1}, \tilde{s}_k, y_k^D) = \frac{p(y_k^D|c_k, \tilde{s_k}) \cdot p(c_k|c_{k-1})}{p(y_k^D|c_{k-1}, \tilde{s}_k)} \quad (7)$$

In order to simplify the sampling of c_k, $p(c_k|c_{k-1}, \tilde{s_k}, y_k^D)$ is considered as a discrete distribution with a finite support on c_k, therefore $p(y_k^D|c_{k-1}, \tilde{s}_k) = \sum_{c_k} p(y_k^D|c_k, \tilde{s}_k) \cdot p(c_k|c_{k-1})$.

3.3. Tracking algorithm

Our tracking algorithm relies on an implementation of the PF using the near-optimal proposal. According to (6), we draw the particles in two steps: the position is drawn from the discrete distribution defined by (7), and the size is drawn from the Gaussian prior density. By replacing the importance function by the near-optimal proposal (6) in the weight expression (4), we obtain the following expression to update the weights:

$$w_k^{(i)} = w_{k-1}^{(i)} \cdot \frac{p(y_k|c_k^{(i)}, s_k^{(i)}) \cdot p(y_k^D|c_{k-1}^{(i)}, \tilde{s_k})}{p(y_k^D|c_k^{(i)}, \tilde{s_k})} \quad (8)$$

4. EXPERIMENTAL RESULTS

To validate our approach, simulations have been carried out on several video sequences: "Walking" and "Running" extracted from PETS'06 and BEHAVE datasets and "Side" and "Front" from our own dataset. As shown in Figure 2, they represent a person walking or running in different directions relative to the camera (diagonal, sideways, front) and correspond to scenarios with different position and size variations. In addition, to simulate video streams with a variable frame rate, the sequences are downsampled with a rate DS. We compare several PFs using different proposals:
- A Gaussian prior proposal for the **conventional PF** (PF) [4] and the **auxiliary PF** (APF) [6] which also includes a preselection of particles.
- A proposal based on hard detection information for the **boosted PF** (BPF) [14]. This proposal is a GMM between the prior density and a density centred on the detected object. The weight of the detection based density is denoted α.
- A proposal based on soft detection information (6) for our **near optimal filter** (NOPF).

For each experimentation, the initialisation is manual and the parameters are : $S = 4$ bands, $N_b = 10$ bins for each RGB channel, $\lambda = 3$, $\lambda_1 = 4, 55.10^{-4}$, $\lambda_2 = 5, 5.10^{-5}$. The number of particles and the covariance matrix are chosen to fit the sequences. For "Walking" and "Running", $N_p = 100$ and for "Front" and "Side" of larger size, $N_p = 200$. For all the sequences except "Front", $\Sigma = diag(1600, 1600, 2, 2)$ to define a large exploration area around the previous position. For the "Front" sequence, $\Sigma = diag(200, 200, 50, 50)$ to take into account lesser changes in position and larger variations of the size.

Figure 2 shows some tracking results for a downsampling rate DS $= 10$. In the "Walking" sequence, the PF and BPF

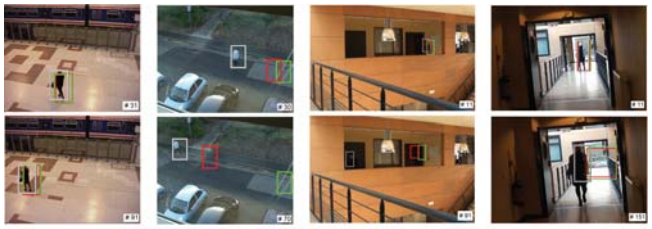

Fig. 2. Tracking results obtained with PF (green), BPF with $\alpha = 2/3$ (red) and NOPF (black) for DS $= 10$ (left to right: "Walking", "Running", "Side" and "Front" sequences).

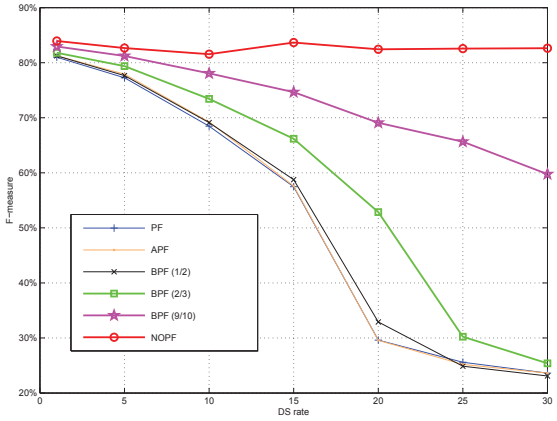

Fig. 3. Average F-measure versus the DS rate for the "Walking" sequence.

can track the object but with less accuracy than the NOPF. The "Running" and "Side" sequences present more challenging scenarios : a fast moving target with considerable changes in shape and a dark coloured target passing through dark areas. The PF and BPF get trapped due to the high speed and the similarity of the target with the dark background area. They also lose the target near the end of the "Front" sequence because they are misled by illumination variations. In all these situations, the NOPF succeeds in accurately tracking the object through better exploration of the state space.

Figure 3 represents the average F-measure versus the DS rate for the "Walking" sequence. This indicator [23] combines the precision and the recall between the ground truth data and the estimated data. Even with a high position variance σ_c^2, the performance of the PF and APF algorithms quickly decrease as DS increases. The prior proposal does not properly explore the state space in case of abrupt movements. In the APF, the pre-selection of the particles according to the observations does not bring any improvement because the state noise is too large. The BPF provides better tracking performance specially when giving more weight to the detection component ($\alpha = 2/3$ and $9/10$). These results highlight

	DS rate	PF/APF	BPF (0.5)	BPF (0.66)	BPF (0.9)	NOPF
"Running"	2	34%	55%	66%	79%	100%
	5	27%	37%	55%	78%	100%
	10	24%	31%	36%	65%	94%
"Side"	1	34%	41%	49%	50%	87%
	5	26%	34%	40%	44%	79%
	10	27%	33%	32%	41%	68%

Table 1. Success rate versus the DS rate for "Running" and "Side" sequences.

DS rate	PF/APF	BPF (0.5)	BPF (0.66)	BPF (0.9)	NOPF
1	38%	38%	37%	38%	56%
10	36%	36%	37%	38%	61%

Table 2. Average F-measure versus the DS rate for the "Front" sequence.

the benefits of introducing detection information in the proposal. The NOPF offers a very great tracking accuracy and robustness against unpredictable movements. The average F-measure remains constant and higher than 80% up to DS $= 30$. Our proposal using soft detection results closely approximates the optimal importance function while the GMM based on hard detection is a rough approximation.

Table 1 summarises the success rate obtained with the different trackers over "Running" and "Side" sequences. In each frame, the tracking result is considered as a success if the F-measure is higher than 50%. Compared with the previous video, on "Running", the decrease in performance occurs for lower DS rates with the PF and BPF because of the target high speed. The results confirm the previous analysis: the interest in using detection information in the proposal and the best performance of the NOPF with a success rate close to 100%. For "Side", the success rate is already low without downsampling. The hard detection results are not precise enough to avoid the local traps in the background areas similar to the object. Due to more reliable soft detection information and close approximation of the optimal proposal, the NOPF outperforms the PF and BPF.

The "Front" sequence is interesting to validate our approach in the case of larger variations of the target size. Table 2 provides the average F-measure obtained with the different PFs. The good performance obtained by the NOPF shows that our proposal, which is the product of a near-optimal proposal for the position and the prior proposal for the size, still efficiently guides the particles when the object size varies more significantly.

Concerning the computational complexity, our algorithm introduces additional operations for the calculation of the discrete distribution (7). Its evaluation involves the multiplication of the detection likelihood by the Gaussian prior density for several positions c_k within the 99.7% confidence interval around the previous sample $c_{k-1}^{(i)}$. The asymptotic complexity is still in $O(N_p)$ as for a conventional PF but the number of operations is approximatively multiplied by $36\sigma_c^2$. In return, the NOPF can perform a successful tracking with a very small number of particles.

5. CONCLUSION

In this paper, we have proposed a PF with a soft detection based near-optimal importance function for single object tracking in videos. Soft detection information is more reliable than usual binary information which can yield detection errors and allows us to design an efficient proposal. Our proposal relies on a close approximation of the optimal importance function and ensures that particles are exclusively drawn in the most likely areas of the state space. Experimental results highlight the benefits of exploiting soft detection information in the proposal, the relevance of the hypotheses retained in the approximation and the robustness of the tracking algorithm against abrupt movements.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, pp. 1–45, December 2006.

[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, pp. 3823–3831, November 2011.

[3] L. Mihaylova, A. Carmi, F. Septier, A. Gning, S. K. Pang, and S. Godsill, "Overview of Bayesian sequential Monte Carlo methods for group and extended object tracking," *Digital Signal Processing*, vol. 25, pp. 1–16, 2014.

[4] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, March 2000.

[5] E. Arnaud and E. Memin, "Partial linear gaussian models for tracking in image sequences using sequential monte carlo methods," *International Journal of Computer Vision*, vol. 74, pp. 75–102, January 2007.

[6] M.K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *Journal of the American Statistical Association*, vol. 94, pp. 590–599, June 1999.

[7] D.Y. Kim, E. Yang, M. Jeon, and V. Shin, "Robust auxiliary particle filter with an adaptive appearance model for visual tracking," in *Asian Conference on Computer Vision (ACCV)*, 2010, pp. 718–731.

[8] F. Desbouvries, Y. Petetin, and E. Monfrini, "A non asymptotical analysis of the optimal sir algorithm vs. the fully adapted auxiliary particle filter," in *Statistical Signal Processing Workshop (SSP), IEEE*, June 2011, pp. 213–216.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, May 2002.

[10] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, March 2009, pp. 221–224.

[11] J.J. Pantrigo, A. Sanchez, A.S. Montemayor, and A. Duarte, "Multi-dimensional visual tracking using scatter search particle filter," *Pattern Recognition*, vol. 29, pp. 1160–1174, June 2008.

[12] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu, "Sequential particle swarm optimization for visual tracking," in *Computer Vision and Pattern Recognition,(CVPR)*. IEEE, 2008, pp. 1–8.

[13] P. Perez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proceedings of the IEEE*, vol. 92, pp. 495–513, February 2004.

[14] K. Okuma, A. Taleghani, N. De Freitas, J.J. Little, and D.G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *European Conference on Computer Vision (ECCV)*, 2004, pp. 28–39.

[15] W. Lu, K. Okuma, and J.J. Little, "Tracking and recognizing actions of multiple hockey players using the boosted particle filter," *Image and Vision Computing*, vol. 27, pp. 189–205, January 2009.

[16] I. Zuriarrain, A.A. Mekonnen, F. Lerasle, and A. Nestor, "Tracking-by-detection of multiple persons by a resample-move particle filter," *Machine Vision and Applications*, vol. 24, pp. 1751–1765, November 2013.

[17] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1820–1833, September 2011.

[18] D.N. Truong Cong, F. Septier, C. Garnier, L. Khoudour, and Y. Delignon, "Robust visual tracking via MCMC-based particle filtering," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2012, pp. 1493–1496.

[19] D.N. Truong Cong, L. Khoudour, C. Achard, and A. Flancquart, "Adaptive model for object detection in noisy and fast-varying environment," in *International Conference on Image Analysis and Processing (ICIAP)*. Lecture Notes in Computer Science, 2011.

[20] E. Erdem, S. Dubuisson, and I. Bloch, "Visual tracking by fusing multiple cues with context-sensitive reliabilities," *Pattern Recognition*, vol. 45, no. 5, pp. 1948–1959, 2012.

[21] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *European Conference on Computer Vision (ECCV)*, pp. 661–675. 2002.

[22] J. Yao and J.M. Odobez, "Multi-person bayesian tracking with multiple cameras," in *Multi-Camera Networks: Principles and Applications*, pp. 363–387. Academic Press, 2009.

[23] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA Broadcast News Workshop*, 1999, pp. 249–252.