A novel approach for multi-object tracking using evidential representation for objects

Wafa Rekik, Sylvie Le Hégarat-Mascle, Emanuel Aldea

A novel approach for multi-object tracking using evidential representation for objects

Wafa Rekik
University of Sfax, Tunisia
Advanced Technologies for Medicine & Signals
Email: wafa_rk@yahoo.fr

Sylvie Le Hégarat-Mascle, Emanuel Aldea
SATIE, Université Paris-Sud
Université Paris-Saclay, France
Email: {sylvie.le-hegarat, emanuel.aldea}@u-psud.fr

***Abstract*—Despite many proposed solutions, multi-object tracking remains a challenging problem in complex situations involving partial occlusions and non-uniform and abrupt illumination changes. Considering modular systems, the tracking performance strongly depends on the consistency of the different blocks relatively to error features. In this work, using the Belief Function framework, we take into account the reliability and the imprecision of the object detection and location to characterize objects and to derive a reliable descriptor. Since this latter is then estimated only on safe object subparts, even in case of crosses between objects, we use a distance between descriptor robust to partial occlusion, namely the recently proposed Bin-Ratio-Distance. Results obtained on various actual sequences underline the interest of the proposed algorithm by outperforming the tested alternative approaches.**

I. Introduction

During the last two decades, tracking has become an expanding axis of research in computer vision, e.g. see [33], [44], opening original fields of applications. Basically, tracking aims at providing the trajectory of each moving object (either in the image or in the 3D scene). For instance, in medical imaging [28], it allows for organ or cancer monitoring (diagnosis, characterization and response to therapy) as well as respiratory motion tracking during an operation [24]. In video-surveillance, it allows us to study intruder behavior and to infer the level of danger [17]. However, in complex scenes, the number of objects to track may be highly variable. Thus, one of the main challenges of tracking using a monocular camera, is the robustness to acquisition conditions (e.g., luminosity variation and/or cluttered or changing background in outdoor scenes) and to occlusion. Indeed, in most tracking algorithms, two processes interact: detection and recognition. Now, the master process depends on the considered algorithm: e.g., tracking by detection versus detection by tracking, e.g. [4]. Basically, in the first case, for each frame, the objects are detected and then associated (using a data association algorithm) to the previously tracked objects that are the closest, in terms of appearance and/or spatial distance and/or any feature characterizing the objects; whereas in the second case, the already tracked objects are searched across the image.

Now, in cluttered scenes presenting severe and long-time occlusions (i.e., people fully or almost fully occluded for long periods of times), both detection and recognition processes are impossible in some individual frames. Then, motion models have been proposed (e.g., [8], [18], [21]) to interpolate tracklets during occlusions, and make the spatial distance still relevant. In order to improve the relevance of the appearance model, the used measure of distance (on which data association will be based) between two objects/people observed at different instants must be robust to partial occlusion and to background changes. The contribution of this paper consists in extending the work of [35] to construct tracklets robust to object crossing (i.e., limited-time occlusions). We use a distance robust to missing subparts of the objects and we take into account the local uncertainty of the object subparts. Section II-A1 presents the uncertainty computation using belief function framework. Our results show that both the evidential functions characterizing the object and the used distance robust to the missing parts in the object are required to achieve a good performance even in complex outdoor scenes.

A. Related work

The representation of objects is a part of the definition of a tracking algorithm, since it specifies the object characteristics used to distinguish an object from the background or from the other objects. This representation is highly linked to the application, and also to data characteristics/properties. For instance, when objects correspond to only few image pixels, the centroid of the object is a common representation. Small and rigid objects are often represented by simple geometric shapes such as rectangles or ellipses [7], whereas rigid and complex objects are represented by their contour [43] or silhouette (region within the contour). Finally, most sophisticated geometric representations deal with articulated objects (set of sub parts collected by junctions) as sets of geometric shapes (cylinder or ellipse for each sub part) or skeletons [1]. This work extends one of our previous studies namely [35] where an evidential approach has been proposed to construct an object from fragmentary detections (i.e. subsets of pixels that are assumed to be a subpart of an object). Fragmentary detections correspond to an intermediate level between pixel level and object level, used to obtain a lighter representation of the data in the perspective of higher-level processing. For instance, this is a classic strategy adopted by parsing algorithms based on superpixels, e.g. [38]. However, object fragments are non-dense, as opposed/contrarily to superpixels. Besides, in [35], due to computation constraints, fragments

are rectangular windows (2D-tiles), thus not respecting the boundary adherence property. Then, from detection fragments that are uncertain since they may correspond either to false positives/false detections that we denote as false alarms in the rest of the paper or to actual subparts of the objects of interest, objects are reconstructed taking advantage of the accumulation through time. Such an approach provides a geometric representation of the objects in terms of subsets of uncertain fragments (rectangles in our case), where each uncertainty value is also associated to an imprecision.

Beside shape representation, object characterization may benefit from appearance features. Among proposed object descriptors, the most popular one is the histogram [43] representing the probability distribution of an object characteristics (colors, texture, gradient, shape...), generally directly computed from object spatial representation. Then, the problem of robustness relatively to illumination changes has been addressed either using value preprocessing (e.g., change of color space, normalisation etc., [32]) or considering more robust features such as orientation of gradient, e.g. SIFT [26] or Histogram of Oriented Gradient [9]. Finally, the active appearance model [13] represents the shape of an object as well as its appearance. However, it requires a learning step. The multi view appearance model [6], [29] allows coding an object in different views based on approaches such as the principle component analysis and the independent component analysis. However, multiple views of the object are simultaneously required.

Considering multi-object tracking, the next problem to handle is the data association, i.e. the association between each current object (at instant t) and one of the previously observed objects. For this, having defined the/an object descriptor, the association is generally derived considering a measure of similarity or dissimilarity between two descriptors. Hu et al. [19] propose a histogram distance that they claim robust to missing subpart of histograms, as it may occur in case of occlusion. However, even if this distance is robust to occlusion, it is sensitive to change in the background (leading to actual changes in the histograms) so that histograms should be carefully estimated by removing or at least discounting the background contribution.

In our case, having at our disposal not only a geometric description of the objects but also uncertainty values associated to subparts of this geometric representation, we aim at using this information to derive more robust appearance descriptors of the objects. The idea of weighting the contribution of the objects pixels is rather intuitive, but the difference with [7] for instance is that in this work uncertainty distribution is object specific (rather than generic to any object).The underlying justification is that, even if some background pixels may be included in the object geometric description, they are associated to lower uncertainty values than pixels that actually belong to the object.

Regardless of the chosen descriptor, since objects are moving their spatial representation is imprecise and since false alarms cannot be excluded from the detection process, some objects may be uncertain. To manage imprecise and uncertain data (objects characteristics, set of objects, false alarms...), belief function theory is a suitable framework [39]. It has been effective in developing solutions for several problems such as data association and filtering [34], [35]. To reduce the uncertainty of objects, an evidential filtering approach has been proposed in [34]. It aims at estimating the set of objects of interest called discernment frame i.e. discarding false alarms [34]. Assuming that an object reliability increases with its size and temporal persistence, evidential rules have been used to rank objects according to their reliability.

In this work, we show that belief function theory, especially the evidential representation is effective for recognizing imprecise/uncertain objects through time and tracking them.

B. Basics in Belief Function Theory

In this section, we will not present the Belief Function Theory, since the reader can refer to [37], [39]–[42], we only introduce the functions and operators we employ in the following sections of this paper.

- A discernment frame, noted Ω, is a set of mutually exclusive hypotheses and has as powerset 2^Ω that is the set of Ω subsets.
- A basic belief assignment (bba) on Ω is defined through its mass function, noted m, that maps 2^Ω to $[0,1]$ such that $\sum_{A\in 2^\Omega} m^\Omega(A) = 1$, where $m^\Omega(A)$ is the elementary part of belief on A that cannot be given to any more specific hypothesis $B \subsetneq A$. Ω subsets with non null mass are m^Ω focal elements.
- The pignistic probability $BetP^\Omega$ is a mapping from Ω to $[0,1]$ that has the properties of a probability function, and is often used to take decision in Ω. It can be derived from m^Ω assuming equiprobability in case of compound hypotheses (disjunctions of Ω elements):

$$\forall H \in \Omega, BetP^\Omega(H) = \sum_{A\in 2^\Omega, H\in A} \frac{m^\Omega(A)}{|A|\,(1-m(\varnothing))}. \tag{1}$$

- The global discounting by a factor α modifies a bba m to give a less committed bba m^α:

$$\begin{cases} m^\alpha(A) = \alpha m^\Omega(A) & \forall A \subsetneq \Omega, \\ m^\alpha(\Omega) = \alpha m^\Omega(\Omega) + 1 - \alpha. \end{cases} . \tag{2}$$

- The conjunctive combination rule allows to combine several bbas to give a more committed bba. In case of two bbas, m_1^Ω and m_2^Ω, its normalized version (also called orthogonal sum) is defined by:

$$\begin{cases} K = \sum_{\substack{(B,C)\in 2^\Omega\times 2^\Omega, \\ B\cap C=\varnothing}} m_1^\Omega(B) m_2^\Omega(C), \\ m_{1\oplus 2}^\Omega(A) = \frac{1}{1-K} \sum_{B\cap C=A} m_1^\Omega(B) m_2^\Omega(C) \quad \forall A \neq \varnothing, \\ m_{1\oplus 2}^\Omega(\varnothing) = 0. \end{cases} . \tag{3}$$

- The disjunctive combination rule allows to combine several bbas to give a less committed bba. In case of two bbas, m_1^Ω and m_2^Ω,

$$\forall A \in 2^\Omega, m_{1\cup 2}^\Omega(A) = \sum_{\substack{(B,C)\in 2^\Omega\times 2^\Omega \\ B\cup C=A}} m_1^\Omega(B) m_2^\Omega(C). \quad (4)$$

II. Multi-object tracking system

A. Object representation

The key point of our contribution lies in the use of an object representation that is both imprecise and uncertain. As pointed in Section I-B, belief functions allow for handling intervals of uncertainty (corresponding to values between instanciations of plausibility and belief functions) rather than uncertainty scalar values, thanks to the handling of a much larger number of hypotheses, namely $2^{|\Omega|}$ for a discernment frame Ω having $|\Omega|$ elements. In this work, we propose different bbas to characterize the objects. Specifically, an object is characterized by (i) the uncertainty about its existence, (ii) its uncertain and imprecise location in the image and (iii) its descriptor in terms of image features.

1) object uncertainty: Due to the presence of false alarms in the set of object fragments, we define, per object i, a bba m^{ω_i} representing our belief in the fact that a fragment or a group of fragments (gathered because of their spatial overlapping, e.g. using connected component labeling) either corresponds to a false alarm or to an actual object. This bba has for specific discernment frame $\omega_i = \left\{O_i, \bar{O}_i\right\}$ with O_i denoting the hypothesis 'the i^{th} fragment group is an actual object'. Based on the interpretation of the detection phenomenon, each bba m^{ω_i} has two focal elements, namely O_i and ω_i (*simple* bba).

2) object location: To represent object location, we consider a discernment frame Ω that is the image lattice: each pixel corresponds to a possible location of the object in the image at the considered instant. The hypotheses in 2^Ω are thus the subsets of image pixels. Note that such an approach dealing with Ω subsets is completely different from the evidential occupancy grids [31] that extend the classic occupancy grids originally proposed by [14] for the robot perception and which are a tessellation of spatial information. Indeed, in evidential occupancy grids 1D bbas are attached to each tessel dealing with the belief in its occupancy (or not), whereas in our case we handle 2D bbas.

For standard image sizes of 256×256 pixels, the cardinality of 2^Ω is $2^{2^{16}}$ so that considering binary coding of every hypotheses (singletons and compound) is untractable. Besides, some pixel subsets are much more likely to represent object location, e.g. connected components. Then, we adopt a sparse representation of Ω: instead of enumerating all its possible elements, we only focus on the focal elements of the considered bba. Following the works [3], [35], we represent (non-unequivocally) a focal element by a subset of rectangular subsets of pixels (windows or boxes). However, with such a representation, the bit-wise operators cannot be used (as usually when handling the binary representation of all Ω elements) to derive the basic set operators (inclusion, intersection, union) that have thus to be redefined in an efficient way. Specifically, writing a focal element A given by the set of rectangular boxes $\mathcal{S}_A$ defined by their column and line intervals, as:

$$A = \left\{ \left[a_{0,0}^i, a_{0,1}^i\right] \times \left[a_{1,0}^i, a_{1,1}^i\right], i \in \{1 \dots |\mathcal{S}_A|\} \right\}, \quad (5)$$

we propose to define operators between A and B focal elements from their representations (Eq. (5)) as follows:

- Union $A \cup B$ can be very simply achieved by adding the subset of boxes representing A to the subset of boxes representing B:

$$A \cup B = \left\{ \left[a_{0,0}^i, a_{0,1}^i\right] \times \left[a_{1,0}^i, a_{1,1}^i\right], i \in \{1 \dots |\mathcal{S}_A|\}, \left[b_{0,0}^j, b_{0,1}^j\right] \times \left[b_{1,0}^j, b_{1,1}^j\right], j \in \{1 \dots |\mathcal{S}_B|\} \right\}. \quad (6)$$

- Intersection $A \cap B$ is the set of non-empty intersections between any pair of boxes extracted from each subset representing A or B, respectively:

$$A \cap B = \left\{ \left[max\left\{a_{0,0}^i, b_{0,0}^j\right\}, min\left\{a_{0,1}^i, b_{0,1}^j\right\}\right] \times \left[max\left\{a_{1,0}^i, b_{1,0}^j\right\}, min\left\{a_{1,1}^i, b_{1,1}^j\right\}\right], i \in \{1 \dots |\mathcal{S}_A|\}, j \in \{1 \dots |\mathcal{S}_B|\} \right\}. \quad (7)$$

- Inclusion is defined from intersection (Eq. (7)): $A \subseteq B$ if and only if $A \cap B = A$ and equality comes from inclusion: $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

The representation given by Eq. (5) is thus effective provided that we control its size. Indeed, focal element operations (Eq. (6) and Eq. (7)) increase the cardinality of $\mathcal{S}_A$ (where A is a resulting focal element). Therefore, to avoid excessive memory load, we regularly simplify the representation of focal elements by computing a sub-paving of their geometric extension without overlapping boxes. Such a process may be achieved in a similar way to the *regularization* process in Interval Analysis [20].

In conclusion, for object location representation, we have as many bbas as considered objects but these bbas are all defined on the same discernment frame Ω. To distinguish them (when ambiguous) the object index is written as a subscript, e.g. m_i^Ω.

3) object descriptor: The third item characterizing an object is its descriptor in terms of image features, that will allows us to recognize it through time. In this application, we focus on classic descriptors that correspond to color and oriented gradient histograms computed on different subparts of the object. Indeed handling histograms that are global to the objects does not allow us to catch the spatial information intrinsec to an object. Then, for every object, we divide it in four subparts corresponding to the intersection between its spatial location (provided by its bba on Ω presented in the previous paragraph) and the four quarters of its bounding box. The object descriptor will then be the concatenation of the histograms computed on each of these four object subparts.

Besides, we aim at taking into account the fact that the object spatial inprints are imprecise. This imprecision is mainly due to the facts that (i) objects are moving and (ii) their

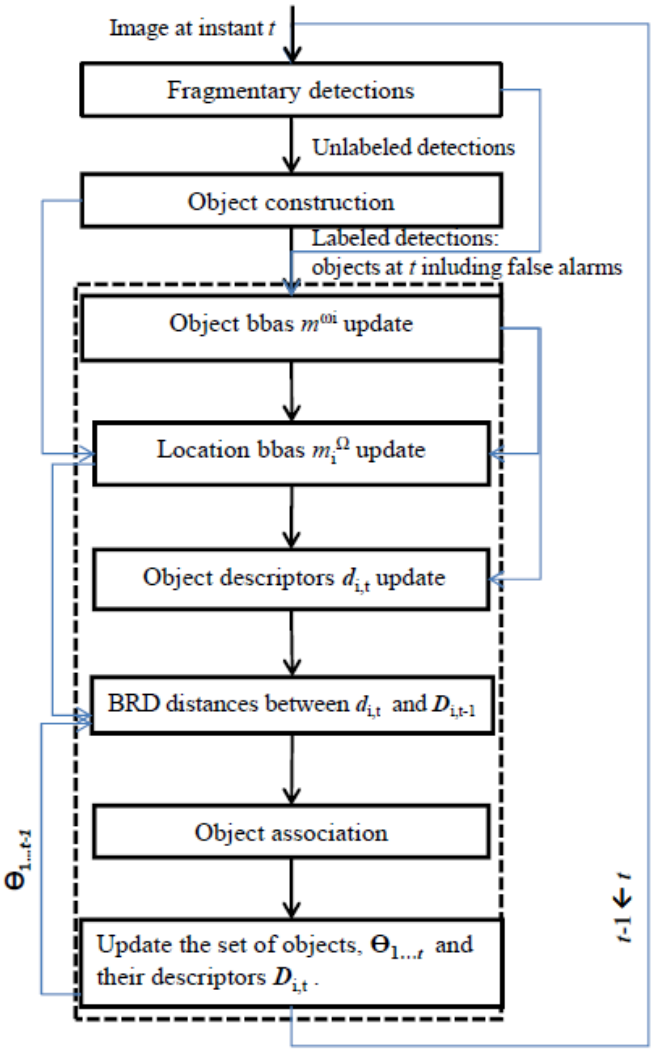


Fig. 1. Diagram of the multi-object tracking system.

detection considering only one instant is incomplete (sparse fragments). Now, the bbas representing object location provide us an estimation of this spatial inprint imprecision. In this study, we compute the $BetP$ value for every object pixel s. Denoting $\mathcal{E}_{m^\Omega}$ the set of m^Ω focal elements, $|A|$ the number of pixels in focal element A,

$$BetP^{\Omega}(s) = \sum_{A \in \mathcal{E}_{m^\Omega}, s \subseteq A} \frac{1}{|A|} m^{\Omega}(A). \quad (8)$$

Then, the $BetP(s)$ values are used to weigh the contribution of the pixels s in the computation of the histograms. Finally, since we consider either color (or gray level) values or orientations of gradient, we have two descriptors per object.

B. System architecture

Let us now present the global tracking system and its key modules. Being modular, the proposed system involves several blocks addressing different sub-problems that can be processed separately. However, as illustrated in Figure 1, blocks present numerous interactions.

The global algorithm is carried as follows. From fragmentary detections at instant t and object construction [35], evidential representations are derived for each object dealing with its actual existence (Section II-A1) and its location (Section II-A). Then, the descriptor (histogram-based) is computed for each object (Section II-A3). Finally, based on the chosen distance (Section II-B6), the data association step identifies the actual objects (at instant t) in the set of the objects previously detected (until instant $t-1$).

1) Fragmentary detection block: At a given instant t, the detection block provides new fragments (corresponding to potential subparts of the objects). In our case, we use an a-contrario detection algorithm [2] working on the difference image between the current image and the background image automatically estimated using codebook and sigma-delta filter. However, any detection algorithm favouring the false negatives relatively to the false positives can be considered (since false negatives will be corrected by the fragment accumulation through time).

2) object construction block: This block corresponds to the algorithm proposed in [35]. To be able to construct objects by fragment accumulation when objects are moving, it is necessary to track the object under construction, i.e. to estimate tracklets. These tracklets are valid when objects do not interact, as objects that are two close may be merged by the construction process. In this work, object construction is controlled by the bbas m^{ω_i} about the existence of the object (cf. Section II-A1). Specifically, when two objects present a pignistic probability (Eq. (1)) greater than a given threshold (0.8 in our case), they cannot be merged and if they intersect each other (typically during a crossing situation) their constructions are stopped.

3) m^{ω_i} update block: As said in Section II-A1, m^{ω_i} is a *simple* bba having O_i focal element besides ω_i. These bbas result from the conjunctive combination (cf. Eq. (3)) of *simple* bbas derived from fragment detection.

Specifically, let us assume that n object bbas have already been defined: $\{m^{\omega_i}, i \in \{1 \dots n\}\}$. A new fragment detection leads to a *simple* bba m_0, with two focal elements F and $\{F, \bar{F}\}$ representing the belief that the fragment corresponds to an actual object rather than a false alarm and the ignorance about this question, respectively. The value $m_0(F)$ depends on the size (in number of pixels) of the detection: the larger the detection, the higher $m_0(F)$ is. The fragment is then interpreted as either belonging to a new object or belonging to an already existing object. In the first case, a new object bba $m^{\omega_{n+1}}$ is initialized with m_0. In the second case, i being the object to which the fragment belongs, m^{ω_i} is updated by $m^{\omega_i} \oplus m_0$.

Finally, before next instant, every m^{ω_i} is discounted (cf. Eq. (2)) by a factor α to take into account the aging of the objects.

4) m^{Ω} update block: For every given object i, m_i^{Ω} models the imprecise and uncertain locations of possible detections

corresponding to object i. Like in [35], it is constructed in two steps.

First, when a detection is associated to object i, m_i^Ω is disjunctively combined (Eq. (4)) with the detection bba m_0 to expand the set of possible locations of future detections. The second step takes into account the object displacement (in time and space) in the video sequence, by invalidating or discounting some possible locations and reinforcing some other locations: locations that are temporally distant from recent apparitions of the object are discounted, close locations are reinforced.

Besides, to manage objects separation (for instance an object that initially corresponds to a group of persons), a spatial conditioning has been proposed in [35]. It allows us to discard locations that are not in the main connected component of the object, assuming that an object location can be represented by its main connected component. We refer the reader to [35] for further details about the management of the object location bba, in particular during object construction.

In this application, as stated previously, we stop the update of m_i^Ω when two sufficiently certain objects cross each other. Specifically, using m^{ω_i} and m^{ω_j} to derive the pignistic probabilities of objects i and j, m_i^Ω and m_j^Ω are not updated if (i) the two objects intersect each other (after fragment association) and (ii) their pignistic probabilities exceed a given threshold.

5) object descriptor update block: Having updated the location bbas m_i^Ω of the different objects i present at time t, their descriptor can be updated as follows. Let $\boldsymbol{D}_{i,t-1}$ denote the descriptor of object i at $t-1$ and $\boldsymbol{d}_{i,t}$ the instantaneous descriptor of object i computed only considering image at t (and m_i^Ω), $\boldsymbol{D}_{i,t} \leftarrow \beta \boldsymbol{D}_{i,t-1} + \gamma \boldsymbol{d}_{i,t}$ with $\beta + \gamma = 1$, $(\beta, \gamma) \in [0,1]^2$. When $\beta = 0$, there is no filtering of instantaneous descriptors and when $\beta = 1$, there is no update of the descriptor.

6) object comparison: Several histogram distances have been proposed to compare descriptors, such as the bin to bin distances and the cross-bin distances. The first ones are commonly used due to their simplicity to implement and relative efficiency. Among them, we can cite: the Mahalanobis [27] and the Euclidean [10] distances that process the difference of bin values using either the L_1 [27] or the L_2 [10] norm; the intersection distance [15] that considers the minimal value between bins; the Bhattacharyya distance [12] that considers the product of corresponding bins and the χ^2 distance [23]. The cross bin distances [16], [25] propose a cross bin comparison between two histograms that aims at improving the similarity measure between histograms. However, since these distances do not take into consideration the correlation between bins, disappointing results occur sometimes, e.g. two histograms of completely different objects may present a low distance.

In our case, the main feature of geometric estimation of the objects is underestimation rather than overestimation. This means that some subparts of an object may still miss (at a given time) but only few pixel background are included. Then, we focus on the Bin Ratio Distance (BRD), recently proposed [19], that allows us to consider partial matching by relying on the correlations between histogram bins:

$$d_{l_1-BRD}(h_1, h_2) = \|h_1 - h_2\|_1 - \|h_1 - h_2\|_2^2 \sum_{i=1}^{N} \frac{|h_{1i} - h_{2i}|\, h_{1i} h_{2i}}{(h_{1i} + h_{2i})^2}, \quad (9)$$

where h_1 and h_2 denote the considered N bin histograms, h_{1i} and h_{2i} are the i^{th} bins of histograms h_1 and h_2, $\|.\|_1$ and $\|.\|_2$ denote respectively the l_1 and the l_2 norms. Using the BRD (Eq. 9), it is thus possible to recognize an object that has been only partially detected (e.g., detecting only the torso or the feet of a pedestrian) and associate it correctly to the complete (totally detected) corresponding object.

7) data association: Finally, data association is performed based on all the distances between couples of objects: one belonging to the set of objects present at t and one belonging to the set of objects previously handled. These two sets of objects are respectively denoted Θ_t and $\Theta_{1\ldots t-1}$.

The problem of data association in multi-object tracking has been widely addressed for a long time (e.g. [5]). In the case of an additive cost function with positive elementary costs (distances between pairs of histograms in our case), a variant of the Hungarian algorithm [22], [30] enables the efficient search of the optimal solution (global minimum). This problem has also been widely addressed using the belief function framework, from [36] to [11], allowing to model different levels of ignorance about the cost of a given elementary association.

In the proposed tracking system, results have shown that the computed distances are rather reliable and precise so that the evidential data association [11] does not improve the results obtained considering classic solution [22].

More specifically, a non-association cost is specified. When an object from Θ_t is non-associated, a new label is assigned to it. It corresponds either to a group of objects (with no correspondent in the set $\Theta_{1\ldots t-1}$) or to a novel object just appearing in the scene.

III. Results

To evaluate the proposed method, we tested it on different sequences of real data. Figures 2 and 3 illustrate the kind of results obtained in three sequences. Figures 2 and 3e-3h correspond to outdoor scenes that are affected by fast changes in illumination, which may adversely affect the object descriptors $\boldsymbol{d_{i,t}}$. In the chosen video sequences, occlusions occur at several instants (Fig. 2). Specifically, the objects (persons) move in such a way as to cross each other and change the direction of their trajectories so that they present different orientations and poses over time. The sequence in Fig. 3a-3d is an indoor sequence with only two objects , and although it might appear as simpler, the number of crosses is very high. Finally, the sequences in Fig. 3 are in color whereas the sequence in Fig. 2 is in gray levels.

A. Qualitative validation

In Fig. 2 and Fig. 3, we present some results of the proposed tracking algorithm. The different labels are highlighted by different colors in the images whereas black represents the false alarms or objects whose construction is not still achieved.

Qualitatively, the proposed algorithm copes with occlusions and object crossing which were the main challenges of multi-objects tracking:

- Using the object bba m^{ω_i}, the labels of objects are protected again adverse effects specific to occlusion, particularly spatial overlap with other objects (Fig. 2c-2d, Fig.3b-3c and Fig.3f).
- The proposed object descriptor allows for recognizing objects in challenging situations. As an example, by using as characteristics the oriented gradient and the gray level intensities, the brown object maintains its label after crossing two other objects (the orange and the green objects) (Fig. 2e-2f) despite of the abrupt illumination change (moving from the shadow to a sunny place). It is also the case of the objects in Fig. 3d and Fig. 3g being identified in the HSV (Hue, Saturation, Value) color space.
- In Fig. 2g corresponding to instant $t = 170$, the object to the right of the scene is divided into two parts: the top and the bottom. The top has been merged with two other objects just to the left and form the new white object. However, the bottom remains separated and the object recovers its true label (green) thanks to the use of the BRD distance, despite its partial detection.
- In Fig. 2h, at instant $t = 177$, the white object (Fig. 2g) is lost due to the crossing. As it had partially merged with the cyan object, the white object was destructed by the spatial conditioning process of object (re-)construction [35] that only keeps the principle component. At instant $t = 177$, its reconstruction being not achieved, it appears in black. However, at instant $t = 190$ when reconstruction is completed, the object recovers its orange (true) label (Fig. 2i).

B. Quantitative evaluation

In this section we aim at providing a quantitative comparison of the proposed approach with two alternative ones. The first one uses a representation of objects by bounding boxes rather than the spatial evidential representation (m_i^{Ω}). The second one uses the Euclidean distance between histograms instead of the BRD distance.

The considered quantitative criterion is the recall ($=\frac{number\ of\ correct\ correspondences}{number\ of\ actual\ correspondences}$) [44] that is evaluated knowing the ground truth (GT) of the sequence. We only plot the recall since, on results, the precision ($=\frac{number\ of\ correct\ correspondences}{number\ of\ performed\ correspondences}$) is always equal to 1 with our tracking algorithm and very close to 1 for the two considered alternative approaches. This is due to the fact that, because some objects are under construction and false alarms are already detected as such thanks to m_i^{Ω} bbas, the set of associated objects is either equal or a subset of the set of actual (Ground Truth) objects.

Figure 4 shows the recall over time for the three approaches in the case of the sequence presented in Fig. 2. The quantitative evaluation is performed on the sequence in Fig. 2, due to the availability of the ground truth annotation. However, the behavior of the algorithm is consistent across all the sequences that we used in the experimental setup. We notice that the proposed algorithm allows for a significantly higher performance. The fluctuations between $t = 100$ and $t = 170$ are mainly due to occlusions (Fig. 2h) which trigger the destruction of some objects (Fig. 2h) and thus some missed associations (relatively to the actual correspondences) until the destroyed objects are completely reconstructed. However, contrarily to the alternative approaches showing decreasing curves with time which means that errors in association cannot be recovered, when objects are reconstructed again (from $t = 170$), the proposed algorithm recovers their labels and recall value reaches 1 again.

For a deeper analysis of the alternative approaches, we proposes an evaluation during sub parts of the sequence. We consider a temporal window of 35 instants, so that at the beginning of each 35s-window, the objects labels are initialized like in the Ground Truth. This allows us to illustrate the frequency of 'stall' of the different approaches. Figure 5 shows that the approach using Euclidean distance regularly fails to track objects: the recall, equal to 1 at each reinitialisation of the labels, systematically decreases to a very low value. The alternative approach using a bounding box, is more robust since when correctly reinitialized after reconstruction, it is able to follow the objects and even resist to punctual association error (e.g. at instant $t = 280$).

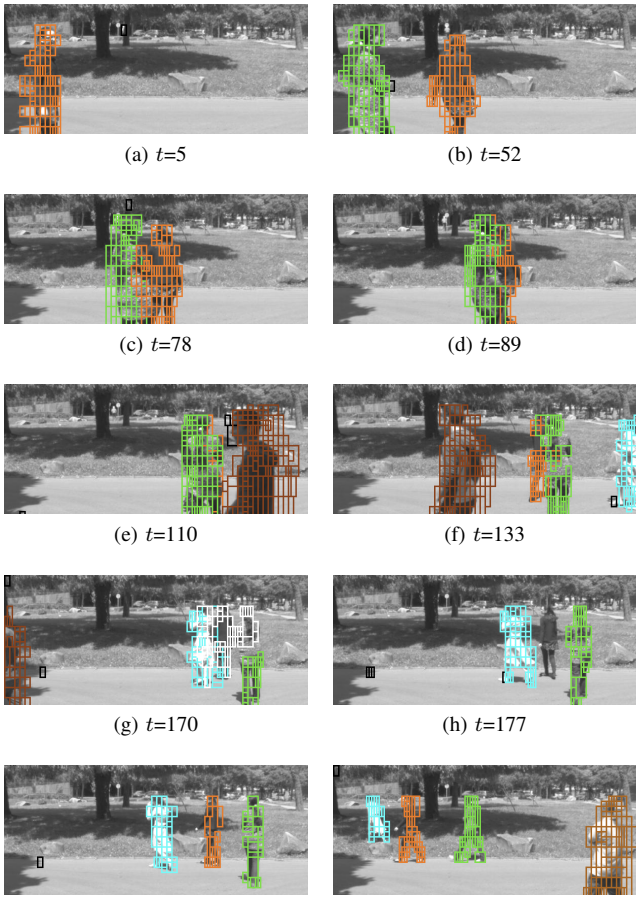
IV. Conclusion

Tracking is an important field dealing with the automatic estimation of objects trajectories. One of its challenges is occlusion due to object interactions. This paper extends our previous work about the construction of objects from fragmentary detections from which only geometrical information (size, location) are provided and thus insufficient to track objects, specifically after long occlusions. To cope with occlusions, richer information for objects characterization such as color, texture, gradient are required. Thus, we propose to convert the imprecise and uncertain geometrical representation of an object into a descriptor in terms of image features based on the uncertainty of a pixel to belong to an object. Using an adequate distance (Bin Ratio Distance) between descriptors, experimental results show the robustness of our tracking algorithm to objects occlusion and crossing.

References

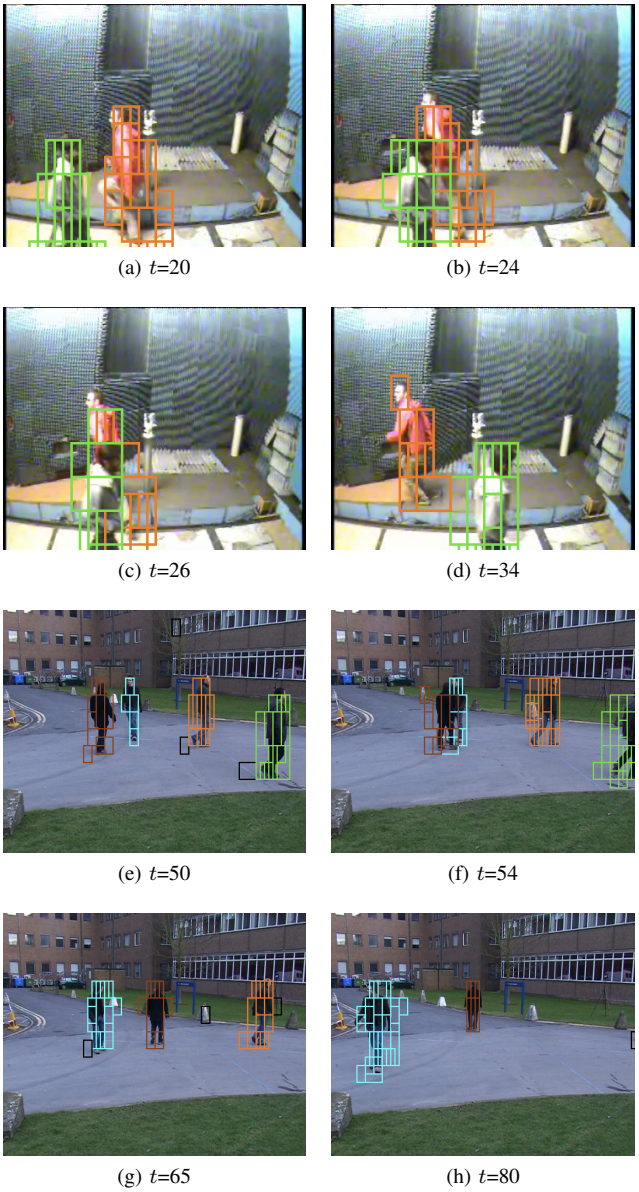
[1] Anjum Ali and JK Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and recognition of events in video, 2001. Proceedings.*, pages 28–35. IEEE, 2001.

[2] Moez Ammar, Sylvie Le Hégarat-Mascle, Roger Reynaud, and Amandine Robin. An a-contrario approach for object detection in video sequence. *International Journal of Pure and Applied Mathematics*, 89:173–201, 2013.



(a) t=5 (b) t=52 (c) t=78 (d) t=89 (e) t=110 (f) t=133 (g) t=170 (h) t=177 (i) t=190 (j) t=281

Fig. 2. Examples of tracking results on a gray level sequence. False alarms and objects still under construction are represented in black. Tracked objects are represented in colors.



(a) t=20 (b) t=24 (c) t=26 (d) t=34 (e) t=50 (f) t=54 (g) t=65 (h) t=80

Fig. 3. Examples of tracking results on two color sequences. False alarms and objects still under construction are represented in black. Tracked objects are represented in colors.

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

[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[5] Yaakov Bar-Shalom. *Tracking and data association*. Academic Press Professional, Inc., 1987.

[6] Michael J Black and Allan D Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[7] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[8] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 25(5):564–577, 2003.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[10] Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.

[11] Thierry Denoeux, Nicole El Zoghby, Véronique Cherfaoui, and Antoine Jouglet. Optimal object association in the dempster–shafer framework. *IEEE transactions on cybernetics*, 44(12):2521–2531, 2014.

[12] Abdelhamid Djouadi, Oe. Snorrason, and FD Garber. The quality of training sample estimates of the bhattacharyya coefficient. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):92–97, 1990.

[13] Gareth J Edwards, Christopher J Taylor, and Timothy F Cootes. Interpreting face images using active appearance models. In *Third IEEE International Conference on Automatic Face and Gesture Recognition.*, pages 300–305. IEEE, 1998.

[14] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[15] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8(Apr):725–760, 2007.

[16] James Hafner, Harpreet S. Sawhney, William Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE transactions on pattern analysis and machine intelligence*, 17(7):729–736, 1995.

[17] Weizhe Hong, Ann Kennedy, Xavier P Burgos-Artizzu, Moriel Zelikowsky, Santiago G Navonne, Pietro Perona, and David J Anderson. Automated measurement of mouse social behaviors using depth sensing,

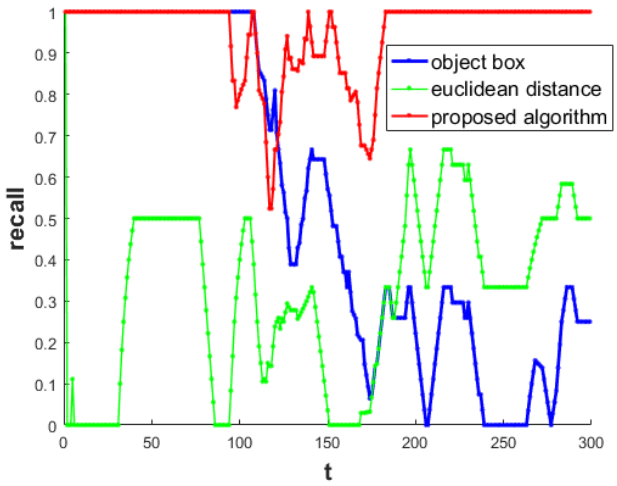


Fig. 4. Quantitative performance of the proposed method and alternative ones in terms of recall versus time.

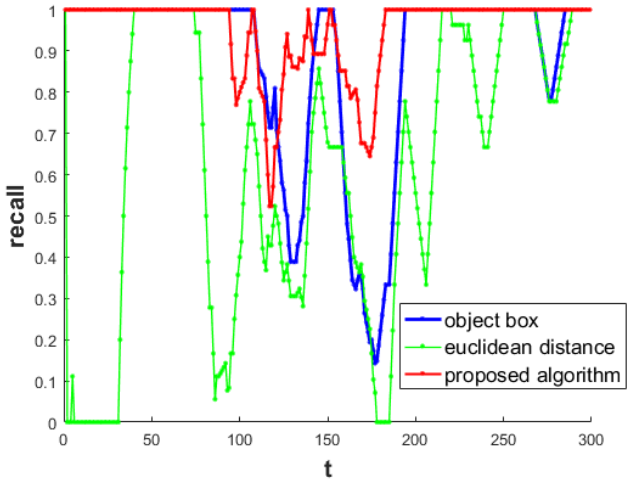


Fig. 5. Quantitative performance of the proposed method and alternative ones in terms of recall versus time with a periodic ($35s$) label reinitialization.

video tracking, and machine learning. *Proceedings of the National Academy of Sciences*, 112(38):E5351–E5360, 2015.

[18] Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical Symposium East*, pages 319–331. International Society for Optics and Photonics, 1981.

[19] Weiming Hu, Nianhua Xie, Ruiguang Hu, Haibin Ling, Qiang Chen, Shuicheng Yan, and Stephen Maybank. Bin ratio-based histogram distances and their application to image classification. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2338–2352, 2014.

[20] Luc Jaulin and Eric Walter. Set inversion via interval analysis for nonlinear bounded-error estimation. *Automatica*, 29(4):1053–1064, 1993.

[21] Allan D Jepson, David J Fleet, and Thomas F El-Maraghi. Robust online appearance models for visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 25(10):1296–1311, 2003.

[22] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[23] G Laplace. Distance du khi 2 et algorithmes de classification hierarchique. *Dialektikê. Cahiers de Typologie Analytique Coarraze*, pages 22–37, 1975.

[24] Shi H Lim, Ehsan Golkar, and Ashrani A Abd Rahni. Respiratory motion tracking using the kinect camera. In *IEEE Conference on Biomedical Engineering and Sciences (IECBES)*, pages 797–800. IEEE, 2014.

[25] Haibin Ling and Kazunori Okada. Diffusion distance for histogram comparison. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 1, pages 246–253. IEEE, 2006.

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[27] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[28] Hengameh Mirzaalian, Tim K Lee, and Ghassan Hamarneh. Skin lesion tracking using structured graphical models. *Medical image analysis*, 27:84–92, 2016.

[29] Baback Moghaddam and Alex Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):696–710, 1997.

[30] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[31] Daniel Pagac, Eduardo Mario Nebot, and Hugh Durrant-Whyte. An evidential approach to map-building for autonomous vehicles. *IEEE Transactions on Robotics and Automation*, 14(4):623–629, 1998.

[32] George Paschos. Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE Transactions on Image Processing*, 10(6):932–937, 2001.

[33] F. Porikli and A. Yilmaz. Object detection and tracking. In *Video Analytics for Business Intelligence*, pages 3–41. 2012.

[34] Wafa Rekik, Sylvie Le Hégarat-Mascle, Roger Reynaud, Abdelaziz Kallel, and Ahmed Ben Hamida. Dynamic estimation of the discernment frame in belief function theory: Application to object detection. *Information Sciences*, 306:132–149, 2015.

[35] Wafa Rekik, Sylvie Le Hégarat-Mascle, Roger Reynaud, Abdelaziz Kallel, and Ahmed Ben Hamida. Dynamic object construction using belief function theory. *Information Sciences*, 345:129–142, 2016.

[36] Cyril Royère, Dominique Gruyer, and Véronique Cherfaoui. Data association with believe theory. In *Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on*, volume 1, pages TUD2–3. IEEE, 2000.

[37] Glenn Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[38] Guang Shu, Afshin Dehghan, and Mubarak Shah. Improving an object detector and extracting regions using superpixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3721–3727, 2013.

[39] Philippe Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[40] Philippe Smets. Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9(1):1–35, 1993.

[41] Philippe Smets. Decision making in a context where uncertainty is represented by belief functions. *Studies in Fuzziness and Soft Computing*, 88:17–61, 2002.

[42] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66(2):191–234, 1994.

[43] A. Yilmaz, Li. Xin, and M.Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1531–1536, November 2004.

[44] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM computing surveys (CSUR)*, 38(4):13, 2006.