OBJECT RECONSTRUCTION IN AN IMAGE BASED ON BELIEF FUNCTION REPRESENTATION

W. Rekik\textsuperscript{1,2}, S. Le Hégarat-Mascle\textsuperscript{1}, C. André\textsuperscript{1}, A. Kallel\textsuperscript{2}, R. Reynaud\textsuperscript{1}, A. Ben Hamida\textsuperscript{2}

\textsuperscript{1} IEF, University Paris-Sud, 91405 Orsay Cedex (France)
\textsuperscript{2} ATMS-ENIS, University of Sfax, (Tunisia)

ABSTRACT

This study focuses on the problem of object reconstruction through several frames of a video sequence. Elementary detections on which this reconstruction is based are assumed to be fragments of the objects. Belief function framework allows then the modelling of the uncertain and imprecise location of these object fragments within the image. We show that the two competing mechanisms for object reconstruction, namely the data accumulation and their temporal removal or weighting, can be implemented using belief function operators. Results illustrate the robustness of the proposed approach to object partial occlusion and crossing.

Index Terms— Object detection, belief function theory, data association

1. INTRODUCTION

Tracking is an important field in computer vision used in several applications such as automated surveillance, human behavioural analysis and traffic monitoring [1, 2]. Fundamentally, it aims at estimating automatically the trajectory of each moving object (either in the image or in the 3D scene). Then, classical sub-problems are the object detection and the assignment of consistent labels, i.e. giving each object a unique label. Now, the developed methods depend on the object representation [1], varying from centroids [3] or set of points [4] in the case of objects that occupy small regions in an image, primitive geometric shapes like box or ellipse [5] in the case of simple rigid objects, to silhouette or contour [6] in the case of non-rigid objects. Obviously the object features (e.g. colour, texture, optical flow, edge) used for their tracking depend on their representation. Then, the ability of a tracker to overcome the main challenges of the tracking problem, such as multiple object tracking with object merging, splitting and crossing, depends on (i) the object representation, from which some discriminative features are extracted, and (ii) the object detection on which depends the precision and reliability of the extracted features.

Besides, in video processing (conversely to radar application), the detections often correspond to fragments of the physical objects [7]. This fragmentary aspect may be due to some occultation phenomena: e.g., by performing detection based on background subtraction, not only self occlusion and interobject occlusion may occur but also occlusion by the background (when this latter is highly similar to parts of the object). On one hand, accumulating the detections through several images (of a video sequence) should allow us to achieve the entire object detection. On the other hand, in the case of multi-object scene, the reconstruction of the objects from the sets of their unlabelled detections may be not trivial (e.g. objects are moving etc.), so that a key subproblem is the association of the fragments corresponding to the same object.

This work deals with the object reconstruction from detections of object fragments in a video sequence. A difference with regards to a complete tracking problem is that we do not consider any a priori knowledge nor any constraint on the initialisation of a new object nor on its labelling regarding the previous ones. This work already performs a monitoring of the scene in terms of objects or intruders present. Taking into account the incomplete and imprecise nature of detections, we propose an object representation using belief function (BF) framework [8, 9]. Besides allowing the management of both uncertainty and imprecision, BF theory provides a well-established framework with numerous operators, explaining its success for several image processing problems from medical imaging [10], remote sensing [11], videosurveillance [12] to data association [13]. The remainder of the paper is as follows: Section 2 briefly presents the used image processing algorithm providing the elementary detections, Section 3 describes the object representation in BF framework (assuming that the reader knows the BF theory main outlines), Section 4 provides some examples of results in order to analyse the algorithm performance and Section 5 gives the perspectives of this work.

2. FRAGMENTARY DETECTIONS

The approaches for object detection are numerous. For instance, [1] enumerates four categories: point detectors, background modelling, segmentation and supervised classifiers. Now, in the perspective of object reconstruction using several images, false negatives seem preferable to false positives, so
that we focus on methods using a criterion based on the Number of False Alarms (\(NFA\)).

A-contrario approaches have been proposed for image processing by Desolneux [14]. Applied to several problems, such as structure detection [15], change detection either in remote sensing images [16] or in video scene [17], they detect very improbable realizations of an unstructured model, so-called ‘naïve model’. Modelling the absence of the researched objects, i.e. the noise represented by the naïve model, allows us to overcome object modelling. Besides, these approaches have often been presented as free from parameter fitting. Indeed, they allow the conversion of a detection problem implying threshold value(s) into an optimisation problem [18].

In [18], the a-contrario principle was applied to detect the objects not present in a scene model that may be either a a-priori scene (e.g. a blank scene except the road) or, in a more classical way, a background image. Two \(NFA\) criteria were used successively. The first one allows detecting the subset of pixels (i.e. the image subdomain) presenting the most consistent grey-level values relatively to the background, according to a naïve model that is a Gaussian noise. It provides a binary image \(B\) of the pixels likely to belong to the researched objects. The second \(NFA\) criterion performs a measure of significance over a small subset of connected pixels representing pixel neighbourhood (‘window’ level). It allows detecting the subset of ‘windows’ presenting the most significant number of ‘true’ pixels (value equal to 1) relatively to \(B\), using as naïve model a binomial distribution representing the absence of clusters in \(B\). In this work, conversely to [18] where the background image was static (e.g. first image of the sequence), we use several backgrounds. In a trivial way, during the minimisation of the first \(NFA\) criterion, in each pixel, the background can be chosen in order to minimise the quadratic error. The selection and the updating of the used background image is done by coupling the \(\Sigma-\Delta\) filter [19] and the codebook model [20]. Examples of detection results are shown on Fig. 4, where the detections at time \(t\) are drawn with bold lines (not paying attention to colour). In summary, the detection algorithm minimises the Number of False Alarms (relatively to a naïve model) to provide a non-dense set of detections (fragments not overlapping). Rather than using a more or less ad-hoc criterion to gather the detected fragments (e.g. defining a bounding box), they are used to provide a credal representation of the objects under construction, so that the problem of object detection is changed into a problem of data (the fragments) fusion.

3. CREDALE REPRESENTATION OF OBJECTS

Let us first define the discernment frame, \(\Omega\), that is a set of mutually exclusive hypotheses. Four main BFs, namely the mass, the plausibility, the belief and the commonality, are defined on \(2^\Omega\), the set of the disjunctions of \(\Omega\) elements, to \([0,1]\).

These four BFs are related by one to one relationships, so that they may be represented by a generic notation: the basic belief assignment \(bba\). We note also that the mass function is such that the sum of its values equals 1. In the perspective of object reconstruction based on the association of some fragmentary detections, we propose to define a bba for each object under reconstruction, which represents the belief on the location in the image of the object fragments. Then, the discernment frame is the image support containing the potential locations of a given fragment (identified by its centroid for instance). Then, an object is characterised (or represented) in terms of beliefs on the location of its fragments. Now, as new fragments of a given object are detected, the object bba should be updated. Two mechanisms compete. The first one aims at ‘adding’ the new detections to the object in order to complete it, whereas the second one aims at ‘removing’ from the object the ‘obsolete’ detections in order to specify the object at current time. Precisely, let us denote by \(m_{o_t}\) the mass function representing the object \(o_t\) at \(t\). For each new detection \(d_j\), a categorical \(bba\) is constructed, i.e. a bba with only one focal element \(A\) (i.e. \(A \in 2^\Omega \mid m(A) > 0\)) such that \(A\) is simply the disjunction of the pixels in \(d_j\) area. According to the first mechanism, the reconstructed object is the disjunction of all the fragmentary detections associated to it (we discuss further the association criterion). Then, \(m_{o_t}\) is updated using the disjunctive combination rule [21], noted \(\ominus\):

\[
m_{o_t} \leftarrow m_{o_t} \ominus m_{d_j}.
\]

Similarly, if during their reconstruction, two objects under construction, \(o_i\) and \(o_j\), are identified as subparts of the same object (e.g. legs and torso of a person), they are simply fused using the disjunctive rule: \(m_{o_t} \leftarrow m_{o_t} \ominus m_{o_i \cap o_j}\).

The second mechanism takes into account the eventual object displacements and/or disappearance from the scene, so that the future fragments are more likely to be located in the most recent detections. Only considering the \(\Delta t\) last instants for the object reconstruction, boils down to conditioning the object bba on the disjunction of the detections during the last \(\Delta t\) instants. Let \(m_{\Delta t}\) be the categorical \(bba\) allowing this temporal conditioning.

Besides, reinforcing the belief that future fragments occur close to the detections at previous time \((t-1)\) can be achieved by the conjunctive combination of \(m_{o_t}\) with a bba defined as follows. Let \(m_{d_{o_{t-1}}}\) be the categorical bba with focal element equals to the disjunction of the detections at \(t-1\) (associated to \(o_t\) and \(m_{d_{o_{t-1}}}^\alpha\) be the result of its discounting by a factor \(\alpha\), so that \(m_{d_{o_{t-1}}}^\alpha\) is a simple \(bba\) (i.e. a bba having only two focal elements with \(\Omega\) among them). Then,

\[
m_{o_t} \leftarrow m_{o_t} \ominus m_{d_{o_{t-1}}}^\alpha \ominus m_{\Delta t}.
\]

If the temporal conditioning (represented by \(m_{\Delta t}\)) allows us to manage an object disappearance, the spatial conditioning aims at managing the object splitting, as in the case of a group of persons such that at \(t\) each one takes a different
way. In our case, it allows separating two objects mistakenly fused during their crossing. For this, we assume that the disjunction of the focal elements (of a bba, $m_{n_o}$, representing one object) is 1-connected component. This disjunction is noted $A$. Implementing this hypothesis simply boils down to a conditioning on the principal connected component of $A$. In practice, we consider as supplementary hypothesis that the actual objects in the scene present a symmetry with respect to the column axis, noted $X$, (that is true for human ‘objects’ imaged on a vertical plane). Then, the object fragmentation is assumed to be mainly according to the row axis, and the main connected component for conditioning is estimated after projection on $X$: Let denote by $A^1X$ the projection of $A$ on $X$, and by $C^1X$ the greater connected component of $A^1X$. The spatial conditioning is then performed on the disjunction of the focal elements (of $m_{n_o}$) such that their projection on $X$ belongs to $C^1X$.

In the presentation of the bba updating, we assume that the association between detections at $t$ and objects already under construction (derived from previous detections) are known. A main interest of the proposed representation is the ability to define a simple and efficient data association. Let us remind that, for each object, the bba gives the belief about the imprecise locations of the next detections. Then, we base the cost of an elementary association on the conflict generated by the conjunctive combination of the bbas of the associated elements. Using BF framework, the conflict between two bbas, $m_1$ and $m_2$, is measured by the empty set mass resulting from their combination: $m_{1@2}(\emptyset) = \sum_{(A,B)\in 2^n X 2^n \cap A^1 B=\emptyset} m_1(A) m_2(B)$. In the case of associations of type $1 - 1$ represented by a permutation vector $A = [a_1...a_n]$ of size $n$, [22] showed that, based on BF representation of the elements to associate, the plausibility of the association writes $\prod_{j=1}^{n} (1 - m_{1@2}(a_j(\emptyset)))$. Then, the maximisation of the plausibility boils down to a well-known problem of minimisation of the sum of positive costs defined as

$$c_0(a_j = i) = -\log (1 - m_{j@1}(\emptyset)), \quad (3)$$

for which efficient solutions exist (e.g. [23]).

In our case, let us also remind that, at a given instant $t$, a given object is detected in the form of multiple fragments. A multiple association $(1 - N)$ seems then required to allow the association of more than one detection to the same object. However, since such approaches are sometimes difficult to control unless involve several threshold parameters difficult to adjust, we rather propose to perform iterative associations of type $1 - 1$. The first one is between objects (under construction) and fragments detected at $t$. At the end of this first step, for each object, only one fragment (at most, since non-associations are also allowed) may be associated, the other fragments being interpreted as potential new objects. Then, we propose to iterate $n_{it}$ times the association between objects. As in the case of the association between current objects and new detections, the potential (non-associations being still possible) associations between objects are researched by global minimisation of elementary costs, defined from the bbas of the associated elements. However, two specificities appear. Firstly, the elementary cost involves a weighting coefficient:

$$c_1(a_j = i) = -\log \sum_{A\cap B=\emptyset} m_j(A) m_i(B) \frac{|A_1X \cap B_1X|}{|A_1X| |B_1X|} \quad (4)$$

where $C_1X (C \in \{A, B\})$ is the projection of $C \subseteq \Omega$ on $X$.

The implicit assumption for this ad-hoc weighting is that a bba with large focal elements corresponds to an object already well-reconstructed and that completing it with a new detection (or set of detections) should be performed carefully, i.e. taking into account the relative size of the intersection between the focal elements. Besides, like for the spatial conditioning, focal elements are considered after projection on $X$ axis because of anisotropic fragmentation. Secondly, in such association internal to a set of objects (conversely to the case where the graph is bipartite), the research of the optimal solution implies a modification of the classical hungarian algorithm in order to keep the symmetry of the cost matrix during its transformations into equivalent matrices. In this case, the convergence time is no longer guaranteed. However, experimentally, we noted that the algorithm rapidly converges in the majority of cases. Five main steps are then implemented for object reconstruction: (i) one-to-one association between fragments at $t$ and objects under reconstruction based on Eq. 3 costs; (ii) an iterative $1 - 1$ association between partially reconstructed objects using Eq. 4 costs; (iii) the bbas updating by combination of the associated elements (Eq. 1); (iv) temporal conditioning and refinement using Eq. 2 and (v) spatial conditioning.

4. EXAMPLES OF RESULTS

To evaluate the proposed method, we tested it on a sequence of real data acquired at 25 images per second. The presented results correspond to a subpart of the sequence (about fifteen seconds in length) that has been chosen as representative of the entire sequence, and for which the ground truth was manually defined: six persons and one car move in the scene and cross each other, so that until six objects of interest may be simultaneously present.

The first line of Fig. 1 shows an example of the proposed method. For comparison, two variants either in terms of the detection-object association or of the object-object association (either only involving $c_0$ criterion or only involving $c_1$ criterion) are shown on Fig. 1 second and third lines, respectively. Qualitatively, we see that the method globally succeeds in reconstructing the objects from multiple fragments: detect-
Fig. 1. Example of object reconstruction at 2 instants (respectively 1st and 2nd columns), and considering 3 different association criteria for the detection-object association and the object-object one: 1st line: \((c_0, c_1)\), 2nd line \((c_1, c_1)\), 3rd line \((c_0, c_0)\). At a given instant \(t\), a same colour notifies that the detections are associated to the same object, with the detections at \(t\) in bold lines.

Quantitatively, given a ground truth (GT), the method has been evaluated in terms of precision \((= \frac{n_{tp}}{n_{tp} + n_{fp}})\) and recall \((= \frac{n_{tp}}{n_{tp} + n_{fn}})\) measures, where \(n_{tp}, n_{fp}\) and \(n_{fn}\) denote the number of ‘true positives’ \((tp)\), ‘false positives’ \((fp)\) and ‘false negatives’ \((fn)\), respectively. Fig. 2 illustrates their definition in a rather complex case. Fig. 2a shows the GT of the detection labelling (i.e. unique colour for a given object) corresponding to Fig. 1c. Using a bipartite graph representation, Fig. 2b shows the relationships (graph vertices) given by the detection labelling either in the GT or in the result (graph nodes). When two nodes are linked by only one vertex, it is a \(tp\), whereas when several vertices are present, one is counted as a \(tp\) and the other(s) either as \(fp\) (two or more labels in the result for a unique label in the GT) or \(fn\) (one label in the result for several labels in the GT).

Fig. 3 shows the ‘precision’ \((y\)-axis) versus the ‘recall’ \((x\)-axis) values computed either in terms of labels (Fig. 3a) or in terms of fragments (Fig. 3b). The different curves represent the results of the proposed method for different combinations of the association criteria \((c_0\) or \(c_1\)) for the detection-object association and the object-object one. The points on a given curve correspond to different ‘delays in decision’, \(\delta t\) varying in \([0, 4]\), introduced as follows: the fragmentary detections at \(t\) are labelled as in the result at \(t + \delta t\), so that some fragments not associated at \(t\) may be associated at \(t + \delta t\). Besides the fact that the results are very promising, we note that, as expected, when \(\delta t\) increases, both \(n_{fp}\) decreases and \(n_{fn}\) increases so that, both ‘precision’ increases and ‘recall’ slightly decreases.

5. PERSPECTIVES

The proposed method is strongly linked to the tracking problem. It already performs a part of the data association (between fragmentary detections) using only location features but without handling the label continuity versus the object crossing, merging and splitting, or strong occultation. Therefore, future work will deal with the label managing using a hierarchical approach such that the first level is the object reconstruction described in this work and the second level is a more classical tracking based on richer features (e.g. colour, texture, speed) extracted from the reconstructed objects.
6. REFERENCES


