Abstract—The exploitation of contextual information can bring several advantages to fusion systems at different levels. Although very promising, this topic is still a scarcely explored. In particular, the inclusion of contextual information in low-level fusion processes has not received much attention in the literature. In this paper we propose a framework for integrating contextual knowledge in a multisensor fusion process in order to improve the estimation of a target's state for tracking. Context will be here encoded in the form of likelihood maps to be fused with the sensors' likelihood functions. The framework presented here allows employing either discounting or pruning strategies for assessing the reliability of sensor observations. In our preliminary experiments, the inclusion of context has provided better accuracy in simulated multisensor tracking scenarios.

Keywords: Tracking, data fusion, contextual information, likelihood masks

I. INTRODUCTION

The exploitation of contextual information can bring several advantages to fusion systems at different levels [1], [2]. The use of combined sensors is deemed to improve the overall performance of a situation awareness system, where many sources of information are fused in order to get more accurate estimates of a target's state. One of the challenges that can be faced is knowing precisely when to discard a sensor observation that is erroneous or non pertinent. A clarification is then needed on how context can be intended, how it can be formalized and represented, and how relevant contexts for particular applications can be discovered and reasoned about [1].

When context is presented in the form of natural language, there are different problems to be addressed: lexical ambiguity, summarization, information extraction, speech processing, and many other. [3]. Up to now, lexical ambiguity is still an open issue, as neither knowledge systems nor machine learning systems efficiently solve uncertainty and ambiguity [4]; for this reason, we decided to avoid the use of the natural language when considering contextual information.

Context is a widely used term, by anyone in everyday communication as well as by philosophers and scientists with many different definitions. The analysis of contextual information is well founded in many and diverse research fields, from Linguistics to Cognitive Psychology to Artificial Intelligence, and can be interpreted as all the information related to the scenario and the entities of interest. In the literature, some consider only position, surroundings, identity and time as “context”, with some extensions to the status and the applications [5]. Nevertheless, there are very general definitions where context is a subset of a physical or conceptual state, which is related to a specific entity, as in [6] where an application-driven definition is reported. Here, context is framed as “any information that can be used to characterize the situation of entities that are considered relevant to the interaction between a user and an application”. More in general, we can say that context is any information that can be used to characterize the situation of entities that are considered relevant to the interaction between the operator and the system. In [7] the authors introduce contexts as abstract mathematical entities in a more general framework which includes context-sensitivity, namely knowledge represented by contextual information systems. More generally, the contextual information has already been incorporated under the umbrella of Bayesian framework and the name of relaxation labeling [8]. In any “representational” scenario, that is in any problematic situation whose goals are disambiguation of referents, correct closure of variables and truth values evaluation (pertaining JDL Fusion levels 2 and 3), two main strands can be recognized as key concepts to start from [9]:

- context is a mere collection of features of the world;
- context is a representation of features of the world.

The first points refers to the state of affairs in the actual world, a set of features of the world, relative to a sentence. Such a concept can be formalized with a tuple of parameters which can be embedded in a semantic model. The second definition concerns a point of view of a situation, or a theory in which a situation can be considered or described in a cognitive fashion. However, in a “real world assessment” the problem is not evaluating variables in a model, getting truth values, but instead building and recognizing variables themselves which are going to play a role in the model. Therefore, two main strands can be recognized [9]:

- context is a collection of objects or properties or features of the world exploited to define or recognize and label simple or complex events or situations.
- Context is a collection of ranging data, sensed in a subset of the world, exploited to build a correct or reliable perception of objects or events.
The first definition can be used, for example, for situation assessment in automotive applications where context consists both in data coming from cameras mounted on vehicles and in other data form sensors measuring steering angle, speed, brakes strength, etc. The joint analysis of on board/off board car context can be used to derive considerations on driver’s behaviour and then to detect possible dangerous complex situations as driver’s sleep. The other definition can be applied to machine perception tasks where properties owned by space-time proximal or causally related objects provide the sensor with sufficient information to discard illusory measurements or, at least, to increase their believes on surrounding world assets.

As it can be imagined, context is a precious source of information in every automatic system for situation awareness. Contextual information can play different roles at different levels, providing significant cues that can range from ancillary data to semantically rich goal-related information. Up to now, its most frequent use is to participate in inferences on the current or future situation, but contextual knowledge can provide indeed a powerful way to semantically bind sensor measurements and real-world observables. This is particularly evident for surveillance systems where the measured position of a target could be checked against the location of sensitive areas with the consequent alarm triggering if a “suspicious” event is detected [10]. The application of information derived from context to JDL Level 1 processing is still under research and only a few attempts have been made, as in [11], where the author proposes to improve tracking performance by selecting a subset of all the available sensors (sensor selection) as reliable sources of measurements for a given target. Updating Kalman filter equations considering also the probability of “validity” of the sensors, the final results was a weighting of their contributions in updating the current target’s state estimate. The paper stresses the importance of taking into consideration contextual information in the fusion process and shows how tracking can benefit from an appropriate weighting of the observations performed by the available sensors. However, the validity domain of a sensor is defined a priori by a human expert and encoded in contextual parameters via fuzzy logic. In general, it is acknowledged that in data fusion problems, multiple measurements of a given observable are filtered to refine an estimate of it. Context can provide essential information by a (detailed) description of the environment, and helps in handling and reducing the ambiguity/uncertainty of detected measurements or events [12] thus being functional in refining filtered estimates.

Unfortunately, in real-world applications it often happens that a sensor detects a sequence of unreliable observations due to partial occlusion of the target, unfavourable weather conditions, reflections, and other problems. In all these cases tracking can be severely disrupted, providing an unreliable estimate of the target’s position and trajectory. For instance, imagine the case of a target moving along a city street and suppose that we want to estimate the vector of Cartesian bidimensional coordinates $x$ that corresponds to its state. If we suppose that the observation $z_t$ at time $t$ is checked against an urban map of the monitored area as falling inside a building, and given the fact that we know that the sensor has no see-through-walls capability, this could be explained as an occasional quirk of the sensor and could be easily filtered out by the tracking algorithm (e.g. Kalman filter, particle filter, etc). This is particularly true if $z_t$ resolves inside a building while both the previous state $x_{t-1}$ and the next measurement $z_{t+1}$ do not. Checking the measurements against a map of the monitored area is a form of contextual knowledge inclusion that could encode an insight on the reliability of the sensor in a specific situation. The knowledge of the sensing capabilities of a source is another form of contextual knowledge that could be exploited conveniently. In the previous example, knowing that the sensor has no see-through-walls capability we can conclude that those measurements may be affected by a form of bias or error, and thus be unreliable. Topographic information is another kind of information that can influence targets’ patterns, as described in [13], where the problem of tracking ground targets is addressed by considering topographic variance as contextual information, and by changing the variance of the process noise at expected turning points. The approach interestingly fuses topographical properties of the monitored area with observations to improve tracking of ground targets.

Section II presents contextual information in a multi-sensor fusion system, exploiting the combination of likelihood masks. Two examples on the effectiveness of the proposed approach in preventing false tracks and filtering noisy measurements, by pruning or discounting, are discussed in Section III.
II. FORMULATION

Contextual information could be further exploited in a multi-sensor system where the same target is observed by multiple sensors. In this case, contextual information could be used to weight the measurements provided by the sensors in the fusion process. That is, measurements from unreliable sensors would be weighted less, thus affecting less the final fused estimate of the target’s position.

Here we propose a way for formalizing and encoding contextual information for target tracking purposes. We describe the multisensor case, where we want to estimate the state (i.e., position) $x$ of a target in the monitored area starting from a set $Z$ of observations coming from multiple sources. Moreover, we describe a framework to handle the exploitation of contextual information, that is represented with likelihood masks to optimize computational costs and memory load and allow a fast application for real-time filtering.

Let us consider a set of $N$ sensors $\{S_1, \ldots, S_N\}$ that produce a set of measurements $Z = \{z_1, \ldots, z_N\}$ regarding an object $o$, in a state $x$ at a certain time in a context $c$. In a Bayesian framework, we want to find the posterior distribution $p(x|Z, c)$ of the state, given a set of observations $Z$ and the context $c$. Applying the Bayes theorem, the posterior can be computed as

$$p(x|Z, c) = \frac{p(x|c)p(Z|x, c)}{p(Z|c)} \quad (1)$$

where $p(x|c)$ is the prior on the state (e.g., the previous known state), $p(Z|x, c)$ is the likelihood of observation $Z$ given the state of the target and the context, and $p(Z|c)$ is a normalizing factor. Supposing the sensors are conditionally independent given the state, we have

$$p(Z|x, c) = p(z_1, z_2, \ldots, z_N|x, c)$$

$$= p(z_1|x, c)p(z_2|x, c) \ldots p(z_N|x, c)$$

$$= \prod_{i=1}^{N} p(z_i|x, c) \quad (2)$$

Then, the posterior probability becomes

$$p(x|Z, c) = \left[ p(x|c) \prod_{i=1}^{N} p(z_i|x, c) \right] [p(Z|c)]^{-1}$$

The interesting term in the equation above is the likelihood $p(z_i|x, c)$ that defines the $i$-th sensor model. This term is a function of all the variables involved and encodes the measurement model of the sensor. The latter can be built by fixing the state $x$ and observing the distribution over $z_i$. Here we will assume the sensor model to be Gaussian and we will investigate how to include context in the likelihood function. When using the model, measurement $z_i$ is observed and the likelihood is a function of $x$ and the context.

A. Likelihood masks

Context will be encoded here in the form of likelihood masks to be fused with the measurement likelihood. Figure 2 shows an example of context likelihood mask generation for urban map data calculated for an electro-optical sensor. From a detailed map of an urban area (a), a full 3D model is constructed, only the buildings are selected (b), a 2D top-view map of the buildings is obtained (c) (detail of the full map shown). The separation between possible and impossible locations is sharply defined: black pixels represent locations where the likelihood of the target’s presence is impossible (coinciding with the buildings’ extent), while white pixels indicate possible locations for the target (in this case the streets). The example mask in Figure 2 (c) could be functional in determining possible and impossible locations for a pedestrian detected by a sensor, while Figure 2(d) depicts the smooth building mask that can be used for tracking a vehicle. The map encodes knowledge about vehicles behaviour, that contrarily to pedestrians are unlikely to be very close to a building; therefore the transition between buildings and streets is blurred. More in general, the mask encodes the capability of the sensor of detecting targets only in open areas, masking the areas covered by buildings. The example in the figure depicts also the possibility of generating different building maps in relation to the type of target.

In general, for tracking purposes, the vast majority of contextual information can be represented by likelihood masks, as every type of knowledge of the environment can be maintained by a likelihood mask, as the detection capabilities of the sensor can be influenced by context in many different ways. The
drawing in Figure 3 depicts possible examples of contexts that can be influential to the final target’s location estimate.

The first row of Figure 3 shows the mask related to the extent of buildings and structures that influence the observation capability of the sensor (as in Figure 2). The mask in the second row represents the field of view of the sensor, excluding the possibility of a detection outside its range. In the third line the weather mask is completely white, as it presents non-disruptive conditions. The fusion of the three contextual masks yields the one shown in the last row following the procedure detailed in Section II-B. The masks here illustrated are merely exemplar and others could be developed depending on the scenario application. Moreover, all contextual masks are in principle dynamic and could be updated as soon as new information is available. This could be done manually by an operator or automatically depending on the type of context: the update of the buildings mask would probably need some kind of human intervention (e.g. a building is collapsed or destroyed), while the field of view map could be recalculated automatically rather easily in the case, for example of sensor relocation. The weather map also could be updated automatically exploiting radar weather maps. However, the updating of likelihood masks is out of the scope of the present paper and it will be investigated in future research.

B. Likelihoods fusion

The masks are here conceived as 2D matrices with values in the [0, 1] interval. The actual fusion of all the ℓ contextual masks is performed by Hadamard product (entrywise multiplication), that is:

\[ C = C_1 \odot C_2 \odot \ldots \odot C_\ell \]  \hspace{1cm} (3)

where \( C \) is the fused mask and the Hadamard product of two matrices \( A \) and \( B \) is defined by \([A \odot B]_{ij} = [A]_{ij}[B]_{ij}\) for all \( 1 \leq i \leq m, 1 \leq j \leq n \).

The masks obtained at the previous step are then used to filter the measurements coming from sensors. The fused mask \( C \) is used in the Bayesian framework in the likelihood function \( p(z|x, c) \) to establish the validity of a sensor observation as follows

\[ p(z_i|x, c) = p(z_i|x)|C|_x \]  \hspace{1cm} (4)

where \( z_i \) is the observation coming from the i-th sensor observing the state \( x \), and \( c \) is the context.

The notation \([C]_x\) indicates that the matrix \( C \) is evaluated at position \( x \). Since the likelihood function \( p(z|x) \) is a function of \( x \), \( p(z|x, c) \) can be expressed by multiplying \( p(z|x) \) by \([C]_x\) for each value \( x \). Thus the observation likelihood is combined with the fused contextual mask to yield the final likelihood.

Figure 4 shows an example of application of (4): the field of view for the two sensors in the same area is shown in (a). In (b) is the observations likelihood for the first sensor, that is filtered in (d) considering urban terrain information and field of view (FOV) of the sensor, while (b) represents the filtering by urban map information only. In Figure 4 (e) is shown the observation likelihood for the second sensor, while in (f) and (g) are respectively shown the likelihood masks, and the fusion of (e) with contextual information. The fusion of the filtered results from the two sensors is shown in (h), where the multiplication of the two masked likelihood functions generates a single observation.

In this example, the observation likelihood for the first sensor is multimodal, meaning that two possible targets have been detected. This may be due to a false alarm given by poor detections (e.g. noise, reflections, etc.). The most and maybe more intuitive representation of contextual knowledge starts from the spatial representation of the scenario that can be translated into visual information. However, as we can see from Figure 4 (b), the urban map information alone could not filter the erroneous detection, as the detections are in an acceptable area, but the integration of contextual knowledge as per (4) overturns the likelihoods of the two detections as it can be seen in the fused likelihood (d), as the observation on the right is not comprised in the field of view of the sensor (FOV). Figure 5 exemplifies the likelihood fusion rationale, and more specifically of equation (4), applied to the multi-sensor case: each measurement is filtered considering the context, that is the fusion of several layers of information. These can be different
III. DISCUSSION

We tested our idea on some preliminary experiments, to demonstrate how tracking process can benefit from filtering the observations of a sensor with contextual information. We simulated three sensors, with different detection errors, that follow a target moving on a straight line for one hundred frames. The purpose of this section is twofold: first, we want to demonstrate how the contextual masks can filter a false track, generated by false detections through time, and then we want to discuss the usage of sharp or smoothed masks to discard faulty observations.

In our experiments we used only the urban terrain of the scenario, that is the information of building location. Additional and more detailed masks, however, can be considered as shown in Figure 4 and as discussed in the previous sections.

A. False tracks

In this first test, we simulated three sensors that produce different observations of a target moving on a straight line. At the same time, we forced the same sensors to detect a false positive, that is a false track; in real world, this situation can be due to a reflective surface, or to multi-path propagations depending on the sensor type. We artificially generated two collections of sensor detections: one set of detections are noisy and scattered about the true target trajectory, and the second set is noisy and scattered about an imaginary track that moves through the building walls. This approach helps avoiding to generate a more complicated measurement process to actually produce such detections.

As we can see from Figure 6, the false track can be discarded considering the planimetry of the scenario, with buildings and roads. The contextual information is used in an online fashion in order to prevent such false track formation (i.e., at the detection level) through applying the contextual likelihood masks to the measurement likelihood as per equation (4).
Figure 6. Example of two tracks of the same object generated by three different sensors with detection issues. The blue line is the target’s true position, while we generated the red line as the false track due to the reflectivity of the building surface.

The detections that fall inside the building walls actually are discarded in the processing. This situation is intended to be an extreme case, where only contextual information can prevent an erroneous behaviour analysis.

B. Noisy observations

This second discussion section is divided in two more parts; in the first one, we produced a toy example using “sharp” contextual masks as the one in Figure 2 (c), that means the building silhouette is perfectly defined and the mask is binary. Here we force the target’s true position to run close to a building, and two sensors out of three to have highly noisy measurements. We add different noise to the measurements, so not every observation falls inside the building area, and not all the sensors produce faulty measurements at the same time. We used the Linear Kalman Filter to track the object trajectory, and we replicated the experiments with a Monte Carlo validation of 10 runs.

As we can see in Figure 7 (a), where the true target position is shown in solid blue line, the three sensors have different noise. In Figure 7 (b) two tracks are shown: in red, the LKF exploits the context-based sensor fusion, while in green is reported the track obtained by the fusion of all the three sensors.

In the second experiment, we used a real-valued mask as in Figure 2 (d), that is “smoother” than the previous one, being the edges not sharp but blurred. Opposite to the pruning strategy [14], where sensor measurements are discarded on the basis of the reliability assigned by contextual information, here we want to explore what happens if we apply a discount strategy. In this case, we want to demonstrate how slightly different contextual information can filter sensor measurement with a diverse impact on the fused result; instead of discarding the observations, we weight them considering the fused likelihood masks value. In particular, the likelihood mask provides the reliability factor to regulate the measurements error covariance matrix as in [15].

The improvement in accuracy is visible in Figure 8, and it is reported in Table I. In the latter, the error statistics are presented, where the lower values are marked in boldface. The contextual information is exploited in the form of a sharp and a smooth map. In the first case, the error on the X axis is lower than the fusion using all the three sensors, but on the Y axis there is a bias, as we can see in and Figure 7 (b), due to the shift toward the centre of the road given by the fusion of measurements that fall on the upper area of the map.

When the smooth map is used to filter the observations, the error is lower, as shown in Figure 8 (b) and in Table I. The fusion with the sharp contextual information mask has an error on x and y axis of 1.68 and 2.78 pixels respectively, while the simple fusion using all the three sensors resulted in 1.79 and 2.29 pixels of error respectively. Using smoothed masks, the contextual information-based fusion scored an error of 1.11 and 1.67 pixels respectively on x and y axis, while fusing all the three sensors produced 1.64 and 2.17 pixels of error in width and height respectively.

The pruning effect of the sharp mask is softened by the use of a smooth mask and by the weighted fusion of the sensors measurements.

The error for the two experiments is shown in Figure 9. The red line represents the predicted trajectory filtered with contextual information, while the green solid line represents the LKF tracker output that uses all the measurements. Figure 9 (a) shows a bias on the Y axis, as discussed previously, due to the inclusion of the upper measurements filtered by the sharp mask only. Contrarily, in Figure 9 (b), where the fusion exploits smooth likelihood masks, the errors on both axes are smaller.

To summarize, in this example the results show that in this case the discounting strategy works better than pruning, but both are effective to discard noisy observations that can lead to error or to false tracks generation.

IV. CONCLUSIONS

We have explored a framework for integrating contextual knowledge in a Bayesian fusion process for target’s state estimation. In particular, we have presented a way of modelling context as a set of masks that can be easily integrated with sensor observation likelihoods for target tracking purposes. Each mask is intended as a way of constraining the detection capability of a sensor according to knowledge relative to

<table>
<thead>
<tr>
<th></th>
<th>Context-based</th>
<th>All sensors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error X</td>
<td>Error Y</td>
</tr>
<tr>
<td>Sharp mask</td>
<td>1.68</td>
<td>2.78</td>
</tr>
<tr>
<td>Smooth mask</td>
<td>1.11</td>
<td>1.67</td>
</tr>
</tbody>
</table>

Table I
ERRORS IN PIXELS FOR DIFFERENT FUSION METHODS. WHEN THE CONTEXTUAL INFORMATION IS EXPLOITED TOGETHER WITH A SMOOTH MAP, THE TRACKER HAS AN ERROR LOWER THAN USING THE FUSION OF ALL SENSORS.
the observed environment. We proposed some examples to demonstrate how the contextual filter works, and we illustrated two simulated cases of false track and noisy observations that generate unreliable measurements, showing that contextual information can effectively prune erroneous observations. Pruning and discounting strategies for discarding or weighting sensors observations have been presented together with a synthetic tracking experiment. Future work will be directed to the implementation in a real-world scenario.

**ACKNOWLEDGMENT**

This work was supported by the European Defence Agency within the DAFNE (Distributed and Adaptive multisensor FusioN Engine) project A-0830-RT-GC.

**REFERENCES**


Figure 9. Tracking error (in pixels) for the context-based fusion (red solid line) and for the full fusion (green line) that exploits all the sensors measurements. On the left (a), the error when using sharp masks is presented, while on the right (b) the results of the usage of a smooth mask are illustrated.


