Soft Data Fusion of Categorical Crowdsourced Data and its Application to Urban Situation Assessment

Barry Park, Anders Johannson
Systems Centre, Department of Civil Engineering
University of Bristol
Bristol, UK
{cebsp, a.johansson}@bristol.ac.uk

David Nicholson
Advanced Technology Centre
BAE Systems
Bristol, UK
david.nicholson2@baesystems.com

Abstract—Conventional “hard” sensing in urban spaces is challenged by the complexity of the environment, creating gaps in situation assessment and possible confusion due to data association errors. Crowdsourced “soft” reports from human observers may remedy this problem but require techniques for fusing hard and soft data. This paper describes an experimental crowdsourcing system to evaluate the potential improvement in situation assessment resulting from the fusion of hard and soft data. The paper then applies a new combination of Bayesian inference algorithms, Particle Filtering and Softmax learning, to a canonical test problem: tracking a single moving object moving along a road network. The fusion of soft reports with intermittent hard data is shown to yield a marked improvement in situation assessment performance. Key to achieving such gains in practice will be appropriate incentives to reward trustworthy reporters along with methods to reduce sensitivity to any remaining untrustworthy reports.

Index Terms—crowdsourcing; tracking; fusion; uncertainty; trust

I. INTRODUCTION

Reliable and accurate situation assessment (SA) underpins decision-making in a wide range of applications. Typically SA is derived from “hard” physical sensors, such as radar, and contextual sources of information such as geographical databases [1].

Consider an urban scene in which a decision-maker requires an accurate and complete picture of the whereabouts of one or more moving objects of interest. This could pose a difficult computational challenge for an information fusion process because the hard sensors offer only sparse coverage of the objects due to limited fields of view and obscuration.

Crowdsourcing is an emerging approach for mitigating difficult computational problems by utilizing widely available human resource [2]. In the case of urban SA, human “sensors” could be tasked to provide “soft” observations about an object that extend and complement hard data from physical sensors. They extend the hard data by plugging gaps in coverage and complement the physical attributes of objects measured by hard sensors by providing soft relational measurements [3].

This paper describes a key aspect of our work in the development of a crowdsourcing platform for investigating urban SA for applications such as defence, policing, emergency response, and future smart cities. It focuses on the problem of how to fuse hard and soft observations within a rigorous and robust mathematical framework for SA. The goal was to quantify under what circumstances and to what extent the soft reports are able to improve SA performance. The background to this work is: (a) a growth in the theory and application of crowdsourcing and our ongoing effort to harness this for improving urban SA, and (b) the need to integrate existing mathematical approaches for SA, particularly hard/soft fusion, in a common framework.

This background is expanded in Section II, and demonstrates examples of applying the crowdsourcing paradigm to urban SA. It is followed in Section III by a description of a simulation-based example designed to evaluate SA performance with and without soft data. Section IV describes a particularly versatile Bayesian inference framework for hard/soft data fusion. A number of simulation-based experiments were performed to evaluate this framework and the results are presented and discussed in Section V, followed by conclusions in Section VI.

II. BACKGROUND

A. Crowdsourcing

Some problems remain immensely challenging to solve in sufficient time with the required level of accuracy due to their computational demands or the limitations on data collected by conventional means. Crowdsourcing is now being increasingly utilized as an effective means of overcoming these problems [4].

The idea behind crowdsourcing is to outsource a problem to a human task force and request data (e.g. observations or labels) that facilitates in its solution. The data is typically returned in the form of a soft report or preference. Incentives are awarded to stimulate timely and truthful reporting. However, the crowdsourced data is still likely to be of variable quality and must be carefully filtered and aggregated to reduce bias and uncertainty.

In the case of urban SA, human observers could be tasked to report on vehicle positions relative to certain landmarks. These tasks might be undertaken in a live situation, making use of smartphone devices to input data. Alternatively, they could be performed offline by tasking participants to report on video playback of a scene from their laptop or desktop computers at home.
This paper is concerned with the data fusion aspects of a crowdsourcing workflow for generating urban SA. Two of the main considerations in this regard are the heterogeneity of the data both in terms of its type (hard and soft) and its level of trust and uncertainty. Only the type heterogeneity issue will be considered here.

B. Crowdsourcing and Urban SA

There are many examples of crowdsourcing approaches being used to generate large amounts of both hard and soft data from human reporters, for applications closely related to SA.

Hard data collected by crowdsourced reporters has been aggregated to improve estimates in radiation levels [5], while DARPA have carried out experiments that have shown crowdsourcing can be used to locate objects across the mainland US [6], and people globally [7], in a matter of hours.

SA has been improved in disaster relief through soft crowdsourced data. Ushaidi, initially developed to map reports of violence in Kenya in 2008, provides a platform for users to submit soft text-based data that can used for crisis mapping [8]. A significant bottleneck is the need for volunteers to label and aggregate incoming data due to its subjective nature.

This paper provides a first proof of concept for the fusion of soft, categorical, crowdsourced data with conventional hard data sources applied to urban SA.

C. Hard/Soft Data Fusion

When data is generated by physical sensors and by crowdsourcing, its combination will likely require a hard/soft data fusion framework. Previous work, although not directly framed in the context of crowdsourcing, has focused on three main theoretical frameworks: Belief Theory, Random Set Theory, and Bayesian Probability Theory (BPT). In this paper a BPT approach is used.

BPT is an axiomatic framework for data fusion and is the basis for established techniques such as object filters for object tracking and Bayesian Networks for SA. More recently BPT-based machine learning has been applied to hard/soft data fusion test cases in the military [9] and bioinformatics [10] application domains. BPT can be used for flexible nonparametric modeling of soft data and combined with decision theory to allocate human effort in crowdsourcing [11].

D. Trust

This paper assumes that reporters will give noisy but trustworthy responses. Here trustworthiness can be seen as a measure of the probability that a user has intentionally misreported data. Live crowdsourcing systems must be robust to both noise in reporter responses and untrustworthy reports. Here we present a short summary of approaches that aim to compensate for sources of untrustworthy data. Future work will be applying several trust based fusion processes to crowdsourced data for urban SA.

1) Validation and Gold Standard approaches: Validation of data, using a trusted third party involves comparing reported labels or data against a known ground truth. In urban SA, this could involve sending a trained observer to a particular location to confirm reports from that region. This is clearly an expensive operation when there are a large number of reports and regions that require attention.

Similarly, Gold standard data approaches use pre-labelled trusted data to test users quality. An example of this is reCAPTCHA [12], which uses labelled data to ensure a website user is a person, and not an automated bot. More generally, when a report disagrees with the gold-standard, this disagreement lowers the trust in a particular reporter, and depending on which set of rules are used, can lead to all subsequent reports from that reporter being ignored. In the case of urban SA, there is often no source of pre-defined gold standard data that could be used to test a reporters response in this manner.

2) Majority Voting and Consensus: Majority voting, and voting consensus, is a heuristic approach to trust that is commonly applied where there are repeated labels of the same object. In this case, when a a certain level of agreement between reports is achieved, then the agreed label is treated as a gold-standard response [13]. It is difficult to apply this approach to reports that contain continuous variables, as there will rarely be complete agreement between reports. Also, any reported uncertainty tends to be ignored. Clustering of continuous-value reports can be used to estimate consensus, with higher density clusters pointing towards trustworthy estimates [14]. However, finding a correct clustering assignment can be challenging when there is not a clear pattern to distinguish between trustworthy and untrustworthy reports.

3) Supervised Bayesian Models: Supervised learning approaches, which combine labels from multiple reporters, over multiple tasks, have been implemented for crowdsourcing systems, with the aim of estimating the true labels for particular tasks, while uncovering the accuracy or trustworthiness of individual reporters. [15] have applied this approach to numerous crowdsourcing tasks, while [16], have further extended this approach to cope with more complicated Bayesian models, where annotator accuracy can change with time. These approaches can give robust measures of reporter trustworthiness, so long as a complete specification of model parameters and probabilistic correlations are possible.

4) Mechanism Design: A further approach, Mechanism design (MD), aims to negate the impact of untrustworthy data at the source, by encouraging the reporting of trustworthy and accurate data. MD revolves around two key components. The first is the exchange of resources, often money, which can be used as an incentive for truthful reporting. The second component is rational behaviour of the agents or people in the system. Under these conditions, MD approaches carefully construct particular ways of exchanging information and resources, that while agents behave rationally and look to take the best actions for themselves, gives an overall system-wide desirable outcome. An example of MD applied to crowdsourc-
ing can be seen in the DARPA Network Challenge, where the winning team used MD to encourage people to both report observations, and to refer the task on to their social network [17]. It is worth noting that some users tried to game the system, so MD approaches still need to be combined with robust filtering and fusion processes when used in the wild.

III. URBAN SA TEST CASE

The data fusion component of the system has been evaluated on a test case involving a single mobile sensor and simulated observation data (hard and soft). The aim of the test case was to prototype a novel combination of hard/soft data fusion algorithms and to motivate a future crowdsourcing experiment by demonstrating proof of concept.

In the test case a constant speed object moves along a network of roads. The roads are treated as lines with zero width. For tracking purposes, there was a known upper bound on the objects velocity of twice its constant value. Observation data was generated in the form of multiple simulated hard and soft reports. For evaluation purposes, four experimental set-ups were considered:

1) **Good hard sensor coverage**: Three hard sensors were dispersed along the roads. They were placed so that at least one sensor was able to detect the object at any time, i.e. the object was always within the maximum range of one sensor.

2) **Poor hard sensor coverage**: Only a single hard sensor was available. It detects the object at the start of the simulation but then the object moves out of range of the sensor.

3) **Good hard sensor coverage with periodic soft reports**: As for the first setup with the addition of soft reports.

4) **Poor hard sensor coverage with periodic soft reports**: As for the second setup with the addition of soft reports.

The simulated hard sensor reports are real-valued bearing-only observations of the object of interest. The observations are perturbed by zero-mean Gaussian random noise. The hard sensor is assumed to have a maximum range \( R_h \), with the probability of detection \( PD \) given by

\[
PD = 1 - r/R_h
\]

where \( r \) is the absolute distance between object and sensor. For all the experiments, \( R_h \) was set to 200m. No false detections (clutter) were generated for the purpose of this evaluation.

The simulated soft sensor reports are discrete multi-categorical values for the range and bearing of the object with respect to a reference landmark. Each sensor provides one of 17 mutually exclusive and exhaustive soft location labels from the 16 range-bearing categories \{'near to', 'far from'} \times \{'north', 'north-east',..., 'north-west'} \} and the additional range-only category \{'next to'}\). This work uses a model of noisy human responses without bias or deceit, and thus presents an initial baseline for fusion of crowdsourced data for situation assessment.

The observation data, along with (assumed known) geographic information about the road layout and the reference landmarks are then passed to the data fusion algorithms described in the next section.

IV. FUSION APPROACH

The BPT framework from Section II forms the basis of the data fusion approach. The desired output is an estimate for the posterior probability density function \( p(X_k|H_{1:k}, S_{1:k}) \) where:

\( k \) is a discrete time index; \( X_k \) is the continuous random state vector of the object at time \( k \); and \( H_{1:k} \) and \( S_{1:k} \) are the respective sequences of hard and soft data up to time \( k \).

A. Prediction

In the urban environments of interest here, objects are usually constrained to move along road networks. A Bayesian estimator must be able to deal with the multi-modality of the probability distributions created by road intersections as well...
as the dynamic motion uncertainties arising from different road characteristics.

Here the multi-modality issue was handled with a Particle Filter (PF) [18]. The PF prediction was implemented using a constant velocity motion model with additive Gaussian noise. A relatively high variance in the noise element was used. This was an intentionally sub-optimal solution designed to challenge the fusion process when there were gaps between observations, even for the simplest test case of a constant velocity object on a straight road.

### B. Hard Fusion

The hard fusion process is a conventional PF update for a noisy bearing-only sensor observation, followed by the application of road constraints.

### C. Soft Fusion

The soft fusion update at time $k$ is given by Bayes’ Rule:

$$p(X_k | H_{1:k}, S_{1:k}) \propto p(S_k | X_k)p(X_k | H_{1:k}, S_{1:k-1}) \quad (2)$$

A soft report $S_k = j$ is an observed category for $X_k$, where $j \in \{1, \ldots, m\}$ and $m = 17$ for the test case since there are 2 x 8 soft range-bearing categories and 1 soft range-only category.

The first key issue is how to specify the likelihood model $p(S_k | X_k)$. This is a function that maps the continuous state space variable $X_k = x$ to discrete probabilities on the soft observation space. Following [18], a natural choice for ‘continuous-to-discrete’ mapping is the softmax function:

$$p(S_k | X_k) = \frac{\exp(w_j^T x + b_j)}{\sum_{h=1}^m \exp(w_h^T x + b_h)} \quad (3)$$

where $w_j, w_h \in \mathbb{R}^n$ and $b_j, b_h \in \mathbb{R}^1$ are vector weights and scalar biases for classes $j, h \in \{1, \ldots, m\}$, respectively.

The softmax parameters are learned from labeled training data using convex optimization procedures based on maximum likelihood (or MAP) estimation.

Training can be an expensive process, both in term of data volume and processing time, so constraints on the softmax model were enforced to ease the problem. These amounted to an assumption that given a state estimate, its bearing observation (relative to a landmark) is perfectly known, i.e. the soft reporters are able to locate the object in the correct quadrant of a circle around the landmark without error. Training therefore reduces to estimating the mapping between the state and the soft range observations.

The soft sensor model used in these experiments to train and test the softmax function had the following range categories: ‘next to’ < 100m, 100m < ‘near to’ < 200m, and 200m < ‘far from’ < 300m. A soft sensor with no noise will assign the ‘correct’ categorical label every time.

Noise was introduced to this model by perturbing the range measured by the soft sensor, between object and landmark, by zero mean Gaussian noise, with variance $\sigma^2$, leading to possible mis-assignment in the soft range category. In later experiments, this data will be created by real crowdsourced human observers.

To quantify the training data requirement, mean training curves for varying soft range noise levels were produced. Based on 50 simulation runs, these curves show the probability of correctly predicting unseen observation labels. Fig. 1a shows an example of the mean and variance in the classification probability across the 50 simulations when $\sigma = 20$. It can be seen that on average, with 10 training points per range category, the softmax model correctly predicts the correct range-bearing category approximately 90% of the time.

Fig. 1b shows the impact on the mean classification performance with increasing noise.

An example of a learned softmax function is shown in Fig. 1c. This displays the probability contours of a learned softmax likelihood function along with the 10 samples of training data in each range category that were used to train it. This model was trained using $\sigma = 0$. Notice the ‘fuzzy’ boundaries between the categories. The degree of ‘fuzziness’...
is controlled by the softmax weights. Also of note are the square edges between the different range categories. This means that even with infinite training data, and a noise free soft sensor model, the softmax function will not always correctly categorise the soft range. This is a limitation of the current approach, however, with a small amount of injected noise in the range data, or limited amounts of training data, the impact of this is minimal.

The second main issue is how to evaluate (2) given a softmax likelihood function (3) and a prior \( p(X_k|\cdot) \). The simplest solution is likelihood weighted importance sampling (LWIS) [18], for which the new particle weights following a soft update at time \( k \) can be determined, up to a normalization constant:

\[
\tilde{w}_k^i \propto w_k^i \times p(S_k|x_k^i) \tag{4}
\]

This approach, while computationally convenient, is susceptible to inconsistent estimates in situations that drive many particle weights to near zero values. A more involved but robust approach, Variational Bayesian Importance Sampling (VBIS), has been developed in [19].

Both LWIS and VBIS were applied in this investigation and found to perform very similarly. The results in the next section are therefore based on the simpler LWIS method.

V. RESULTS AND DISCUSSION

This section contains a selection of results for the test case outlined in Section III. First, an illustration of poor situation assessment is shown in Fig. 2a. This has arisen because the object has turned right at a road intersection, moving out of range of two hard sensors. The particle-based state estimate has propagated particles consistent with the possible multiple motion hypotheses given the known road layout. The consequence is the mean object position estimate is significantly displaced from the true object position.

A soft report is now received stating that the object is “near to” and “south east” of the upper landmark, as show in Fig. 2b. Probability contours are displayed for the learned softmax likelihood function. The soft report is fused with the current particle-based estimate and the impact is clear: an improved mean object position estimate and a reduction in the spread of particles (variance).

Detailed statistical performance analysis results are now presented for different combinations of hard and soft sensor updates. The results were obtained for the simple test scenario shown in Fig. 2c. A single object moves along a road next to which are situated 3 hard sensors. At any time step at least one sensor has the object within its range. A single landmark is used as the anchor point for periodic soft reports.

First, it is instructive to look at the estimate of one of the object’s position variables for a representative run of the simulator. This is shown in Fig. 3a for a situation in which there is poor hard sensor coverage with no soft reports. In this scenario, the target is within range of a hard sensor until time step \( k = 100 \), where the object moves beyond its range. There
is a significant drift in estimated position relative to the true object position in this case.

The estimates for a situation in which there is again poor hard sensor coverage but periodic soft reports every 50 time steps are shown in Fig. 3b. The soft reports provide sufficient information to offset much of the drift that was evident in the previous case.

The statistical performance analysis was based on 20 Monte Carlo runs of the simulator, after which the root-mean-square position estimate of the object was calculated at each time step. First, the results for poor hard sensor coverage, with periodic soft reports were generated. This is shown in Fig. 3c, where the soft range noise was set to $\sigma \in \{0, 20, 40\}$. Each softmax function was trained on 10 data points per range category, leading to a correct classification probability of approximately 95%, 90% and 80% respectively. The main impact of the noise can be seen when the object approaches the boundary between two soft range categories, which in this case occurred between the time steps $150-200$, and $250-300$. The noise has little impact when the object is located near the centre of a soft range-bearing region, which is shown for time steps $k > 300$.

Using the softmax model with $\sigma = 0$, Fig. 3d shows the impact of soft reports for various hard and soft sensor combinations. Bounds on performance are provided by the poor and good hard sensor coverage cases. In the case of good sensor coverage, the soft reports have little impact on performance. However, soft sensor reports clearly improve situation assessment performance when hard sensor coverage is poor.

VI. CONCLUSIONS

Crowdsourcing applications raise interesting new challenges for defence researchers. The crowdsourced data from human reporters is likely to be of a soft type, of variable quality, and may not always be trustworthy.

This paper has focused on the soft aspect of crowdsourced data and shown how it can be fused with hard data to provide SA performance benefit when the coverage provided by hard sensors is poor.

This paper has only provided a proof-of-concept that fusing soft reports from human sensors can improve urban SA. Its greatest limitation in this regard was the use of synthetic soft data. The next step is therefore to conduct a trial with real soft data generated by human observers recruited from Amazon Mechanical Turk. This expands the scope of the data fusion problem considerably.

ACKNOWLEDGMENT

This work was jointly supported by the EPSRC funded Industrial Doctorate Centre in Systems (Grant EP/G037353/1) and BAE Systems. Thanks are due to Nisar Ahmed (Cornell) for providing software and advice on the VBIS soft fusion algorithm, and to Steve Reece (Oxford) and Gopal Ramchurn (Southampton) for useful comments.

REFERENCES