Augmenting Bayes filters with the Relevance Vector Machine for time-varying context-dependent observation distribution

Alexandre Ravet, Simon Lacroix, Gautier Hattenberger

Augmenting Bayes filters with the Relevance Vector Machine for time-varying context-dependent observation distribution

Alexandre Ravet[1,2], Simon Lacroix[1,3] and Gautier Hattenberger[4]

***Abstract*— Bayesian filtering often relies on a reduced system state relating to robot internal variables only. The exogenous variables and their effects on the measurement process are then encompassed within a global observation noise model. Even if Bayes filters proved to be robust to such approximations, special care has to be taken to handle some of these exogenous effects, usually by introducing complex observation distributions or rejection rules. No matter how complex these models are, they often fail in dealing with contextual incidence which can hardly be explicitly encoded. This article shows how contextual information can be introduced within the Bayesian filtering framework by coupling a filter with classification and regression probabilistic models. The classification model provides an efficient context-dependent measurement selection mechanism and is specifically trained with respect to the filter estimation performance. This first component is enhanced by the introduction of context-dependent observation noise provided by the regression model. The performance of this is approach is evaluated and compared with other methods in the context of altitude estimation for a UAV.**

I. Introduction

The issue of understanding and exploiting measurements provided by different sensors is of major importance in state estimation. It is commonly accepted that the more accurate the observation model, the better. However, the physical measurement process is often too complex to be fully understood and exactly modeled. In practice, one has to deal with a whole range of performance alterations going from the unavoidable average observation noise to completely unreliable measure values. Instead of trying to explain the whole measurement set with an accurate observation model, a common way to cope with unmodelled observation alterations is to select the measurements which are known to be positively informative for the estimation process – in other terms to reject outliers. This yields robustness of the estimation process with respect to unmodeled effects on the observations, but at the risk of under-exploiting information. It is desirable to define an estimation scheme that goes beyond outlier rejection, introducing *context information* for a time varying modeling of measurement contribution.

a) Related work: Most approaches for solving the estimation problem rely on the framework of Bayesian filtering which proved to be robust and highly reliable. Any implementation of a Bayes filter is based on two key components, the state transition model, and the observation model. The optimal observation model should describe how sensor measurements are generated in the current system state. In the Bayesian filtering framework, this model is described as a conditional distribution $p(z_t|x_t)$ where z_t is the set of measurements at time t and x_t is the state of the system. As the underlying physics explaining the process of measurement formation is often too complex to be perfectly described by the observation model, contextual information (*i.e.* information about exogenous effects required to explain the formation of measurements) is usually not included in the state x_t. Such information should also account for effects stemming from the unpredictability of dynamic environments, resulting for example in unexpected occlusions. All the unmodeled effects are usually represented through the observation noise, by making the assumption that they are similar to random effects. This is a strong assumption that often has to be compensated, for example through the introduction of outlier rejection methods [1], or through the derivation of complex observation pseudo-densities modeling sensor specific behaviors, such as the well known beam model described in [2]. These approaches suffer from over-simplification as they rely on a subjective understanding of the sensor characteristics, and also remain self-contained systems that implicitly accommodate all the context influence in a single static model. A direct consequence of this implicitness is that these systems are quite likely to diverge [3].

As deriving an accurate model by hand is too complex, learning algorithms have been applied in order to build, at first, parametric models [4], and more recently non-parametric models [5]. By learning an observation model without introducing any prior knowledge over the sensor behavior, the latter approach proved to be very efficient, and has been extended to a fully state-dependent observation model through the introduction of heteroscedastic noise (meaning the observation noise is state-dependent), whereas usual models make the assumption of homoscedastic (constant) noise. The resulting model is of the form $z_t = f(x_t) + \epsilon(x_t)$ where f and ϵ are represented by Gaussian processes (GP). Still, all these attempts to learn a precise model rely on a system oriented reduced state, and are consequently unable to model contextual influence over the measurement process. Moreover the global performance of the filter can strongly depend on the algorithm that utilizes the model [2].

b) Approach: This paper aims at showing that the joint set of sensor measurement values z_t, potentially extended with any relevant contextual information i_t (*i.e.* other sensor measurements, or other internal data), provides a succinct

[1]CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France
[2]Univ de Toulouse, INSA, LAAS, F-31400 Toulouse, France
[3]Univ de Toulouse, LAAS, F-31400 Toulouse, France
[4]Ecole Nationale de l'Aviation Civile (ENAC), 7 Avenue Edouard Belin, BP-54005, Toulouse, France

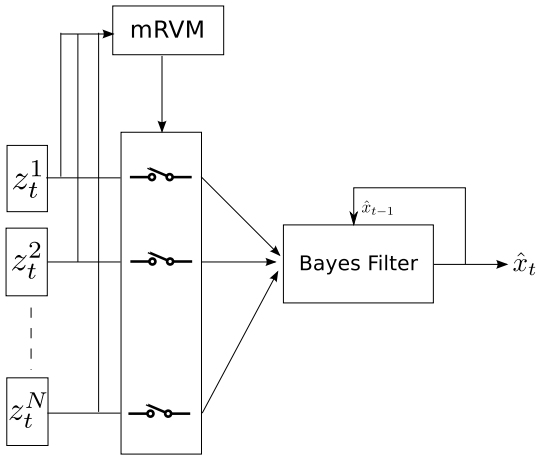

Fig. 1. Principle of mRVM classification based measurement selection for Bayesian filtering (contextual information i_t is not represented for clarity)

yet rich representation of the perception context that can be used as the input of a context-dependent observation model. Intuitively, we can see this joint set of measurement values as the minimum information relating the influence of the current context over sensor performance. As shown in [5], [6], one intuitive way to learn an observation model is to apply a direct regression method. However, we rely on a different approach for the two following reasons:

- In our case the contextual information is not part of the system state. It is an additional set of variables $c_t = \{z_t, i_t\}$. Therefore learning a direct mapping from the context information to the measurement is irrelevant, as the output of the system is contained in the input.
- It is legitimately more appealing to optimize the observation model with respect to the ultimate system output performance (the state estimate accuracy) rather than learning a direct observation model which can be seen as an intermediate goal (explaining at best the mapping between state and context to measurements), but which does not provide any guarantee in term of global performance.

The core component of our model is used to encode a *high-level knowledge* over measurement validity which, conversely to other approaches, is learned with respect to the only goal of helping the system to provide the best estimate. Learning such knowledge implies to take into account the whole prediction-update process in the optimization and requires to examine each filter algorithm separately. A simple way to achieve this with classical learning methods, and without any consideration regarding the type of filter, is to view the observation function as a selector over measurement values. The function is then context-dependent in the sense that the actual subset of measurement values used for the estimation process change over time. Practically, this is done by running at first a group of different filters, all of these filters using a distinct yet combinatorially exhaustive subset of measurement values from a training set $D = \langle X, C \rangle$, where X is a vector of ground truth states and C a vector containing measurements $c_t = \{z_t, i_t\}$ made at state x_t. We then apply the Relevance Vector Machine (RVM) classification technique in order to learn a discrete mapping from the context space (c_t) to the most accurate filter at time t. This provides a context-dependent observation function based on a classifier optimized with respect to the global performance of the system (measured w.r.t X). This work is thus closely related to the bank of filter approach [4] [3], if we consider a bank of filters differing only in the subset of measurement values they are using. However, we enhance this first classification model by learning a context-dependent (heteroscedastic) observation noise model for each sensor with RVM regression. The classification and regression models are then combined to obtain what can be seen as a fully context-dependent observation model.

c) Outline: The next section provide background details of RVM. We then present the specific implementation for learning context-dependent observation function (Section III), and observation noise (Section IV). Finally, we provide experimental results and comparison to existing approaches.

II. Background on Sparse Bayesian Learning

In the past decade, machine learning has seen the emergence of sparse Bayesian methods, a specialisation of which being known as the Relevance Vector Machine (RVM) [7]. In this work, RVM is chosen for the regression and classification task for two main reasons. First, as a nonparametric model, RVM introduces no assumption over the functional form underlying the mapping from the context to the appropriate measurement subset, or from the context to the current observation noise. It then provides a generic approach for different robot configurations. Secondly, RVM training naturally provides sparsity thanks to the *automatic relevance determination* mechanism. It is a considerable feature in the context of real-time Bayesian filtering, since inference over the model is fast. This is in contrast with GP, widely used for enhancing Bayes filters and for the task of heteroscedastic regression [5], [8], [9], [10], which require to turn to more complex sparse GP techniques when the system is intended to be used in real time.

A. RVM principle

Based on the probabilistic Bayesian framework, RVM provides a powerful solution to the problem of supervised learning, *i.e.* learning a mapping model between a set of input vectors $\{x_n\}_{n=1}^N$ and corresponding target values $\{t_n\}_{n=1}^N$ considered as noisy outputs of an underlying *noise free* function $y(x_n)$. RVM can be seen as the equivalent probabilistic treatment of the Support Vector Machine (SVM), whilst avoiding most of its limitations. The shared principal characteristic of these methods is that they both base their predictions upon a function $y(x)$ defined as a weighted sum of basis functions given by kernels K, with one kernel defined for each input of the training set:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0 \tag{1}$$

where w_0 is a bias parameter and $w = \{w_1, \ldots, w_N\}$ are the weights. The most appealing aspect of the RVM is that the obtained prediction model is sparse (most of w_i parameters are close to zero), while still providing excellent generalisation performance for new input values as long as they remain in the implicit boundaries defined by the exploited training set.

A widely used kernel function is the Gaussian kernel defined as

$$K(x_m, x_n) = exp(-\frac{1}{r^2} \parallel x_m - x_n \parallel^2) \qquad (2)$$

where r is called the width parameter and can be easily tuned during the learning process. By experience this function provides good results for our application. However the use of a different kernel function may improve drastically the performance of the learned model, for example if we prefer to use a different width parameter along the different input dimensions.

Training and inference over the model are done within the Bayesian framework. By lack of space we will not derive equations for training and predicting, but all details can be found in [7]. The RVM classification method originally introduced by Tipping is however designed for binary classification, and requires adaptations for the multiclass case. This is why we prefer to rely on the multiclass Relevance Vector Machine (mRVM) introduced in [11] which naturally handles the multiclass setting and achieve better sparsification.

III. Learning context-dependent observation function

In many cases deriving the noise free component f of the observation model $z_t = f(x_t) + \epsilon$ is not the most problematic task, as it can be obtained directly through deterministic physics rules. In next sections we assume that these functions are known, and then alleviate the rest of the system which can focus on capturing the remaining unmodeled aspects of the observation model. We now detail the procedure for learning the context-dependent selection scheme and give an illustration of its application on simulated data.

A. Goal-oriented learning

The complete observation vector z_t is composed of N measurements $\{z_t^n\}$. Based on pre-defined noise free component of the observation model for each of these measurements $\{z_t^n\}$, we are able to derive the combinatorially exhaustive set of possible observation functions. As these functions correspond to different contexts, we note them f_c where $c \in \{1, ..., C\}$ with C the total amount of distinct functions. We then train a RVM model such that, at each filter iteration, the model selects the most appropriate subset of measurements $\{z_t^n\}_{n \in c}$ in the complete measurement vector z_t according to the current context, *i.e.* we use the corresponding observation function $f_c(x_t)$:

$$\left[\{z_t^n\}_{n \in c} \subseteq z_t\right] = f_c(x_t)$$

1) Defining the training data: The definition of the training data is a crucial step in this approach. The idea is to form an alternative training set based on the original data $D = \langle X, C \rangle$ in order to explicitly introduce the requirements in terms of global system performance. Thus we run C Bayes filter, each of which based on a different function f_c, and save the corresponding state estimates and uncertainty. Then, each sample of D is associated with an activation vector A_t – *e.g.* $A_t = \{0, 0, 1, 0, ..., 0\}^T$, which means that the best estimate output at time t was provided by the filter based on the observation function $f_{c=3}$. It should be noted that the definition of the 'best' output requires the evaluation of a performance metric which represents the expected requirements to the system behavior. In this paper, we evaluate the output distribution of each filter at the corresponding ground truth value x_t (given by $\mathcal{N}(x_t|\hat{x}_t^c, \sigma^c)$ where $\hat{x}_t^c$ and σ^c are the output mean and variance of the filter based the observation function f_c). Choosing the highest value among all filters is then equivalent to selecting the filter with the lowest estimation error and uncertainty. However, other requirements may be easily introduced through the evaluation of different metrics or the addition of penalty terms.

A new training set $D' = \langle C, A \rangle$ is formed, and is readily used for training a mRVM classifier with C class labels. This classifier is then used at runtime to select the most appropriate subset of measurements according to the context $c_t = \{z_t, i_t\}$. This process is illustrated in Fig.1.

The training method of the selection model can be seen as a case of discriminative training, analogous to the one suggested in [12]. This is due to the optimization of the RVM model done with respect to the ultimate filter performance, while the observation model is usually optimized through generative training, *i.e* in order to maximize the observation likelihood $p(z_t|x_t)$ over the training set. A strong benefit of this specific training approach is the subsequent model implicit capability in compensating for mis-modeled aspects of the real system. Discriminative training has then been chosen here as it serves the common purpose of compensating for Bayes filter model inaccuracies. Note that, alternatively, the RVM model could however be trained generatively.

B. Illustration

We demonstrate here the capabilities of the selection approach on a simulated dataset. For this test case the original training set D contains 6000 samples of the simulated altitude ground truth of a UAV and the corresponding altitude measurements provided by three sensors (Fig.2) presenting some realistic characteristics, such as maximum range threshold and outliers occurrences – section V deals with real data. From these 3 sensors, seven observation functions f_c are built in order to cover the exhaustive measurement combinations: $\{[z^1], [z^2], [z^3], [z^1 z^2], [z^1 z^3], [z^2 z^3], [z^1 z^2 z^3]\}$. These observation functions are used within a simple Kalman filter with constant velocity transition model. The examination of the function activation frequencies provided by the classifier

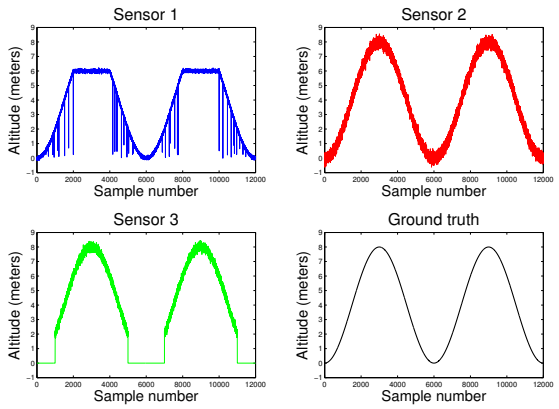

Fig. 2. Altitude measurement provided by three sensors. Sensor 1 reproduces typical ultrasonic measures, low observation noise, outliers occurrences and maximum range threshold. Sensor 2 permanently provides measures with high observation noise. Sensor 3 does not provide relevant measures below 2 meters.

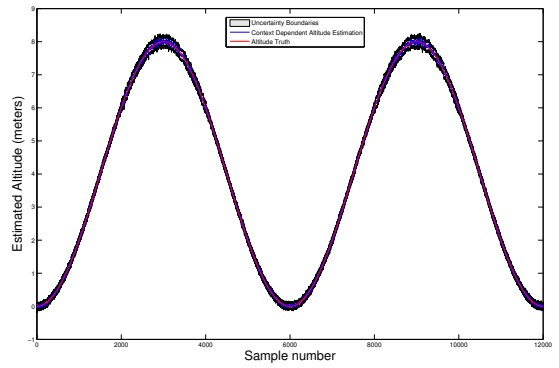

Fig. 3. Final estimation and associated uncertainty on validation set.

on the training set then provides useful insights on the relevant sensor combinations: some functions may appear to be rarely used and therefore can be removed from the initial set before re-training a refined classifier. This process can simplify the classification problem and directly reduce the computational cost of the model at runtime. Once the classifier is trained, it is evaluated on a validation data set.

The final estimate value is shown Fig.3. Clearly, the classifier recognizes all different contexts which for sensor measurements could alter the estimation, while combining the maximum of reliable data so as to provide a low estimation uncertainty. In practice, this means that the classifier behavior consists in activating as often as possible the observation functions corresponding to combined measurements. The effect of measurement rejection over the final estimate uncertainty are well illustrated for the highest altitudes where sensor 1 does not provide any useful information and hence is not used. As a consequence the estimation uncertainty is slightly increased.

IV. Learning context-dependent observation noise model

In practice, the core component described in the previous section is a powerful mechanism which is also capable of compensating further inaccuracies made in the estimation model. This is a simple consequence of the fact that the selection function has been trained to choose the appropriate measurement combination such as obtaining the best estimate output, no matter how well tuned the observation and transition models are. As explained in section III, the parameters of the observation function f are not the most complex to obtain, but the observation noise can be difficult to estimate. Moreover, introducing an accurate and realistic noise model in the filter helps in tempering the rejection scheme operated by the classifier. Indeed, and as shown in [5], the use of heteroscedastic observation noise models can greatly improve the estimation precision while providing a more realistic estimation uncertainty. When combined with the selection scheme described in the previous section, the use of such noise models results in a better exploitation of each filter by extending their domain of 'reliability'.

An elegant approach to learn input dependent noise models in the context of regression was introduced in [10]. A key idea of this method is to introduce a second regression model dedicated to modeling the empirical 'observation' standard deviation for each sample. In a similar approach we form again N new dataset $D''^i = \langle C, S^i \rangle$ where $S_t^i = log[\sqrt{(z_t^i - f(x_t))^2}]$. This value provides an empirical estimation of (the logarithm of) the observation noise level for the measurement z^i at time t. N RVM regression models are then trained on these different training sets.

Used along with the different functions f_c these models notably prevents the different filters from diverging during the performance evaluation step preceding the classifier training. Note that the RVM model provides Gaussian prediction distribution over the noise level. In a fully Bayesian approach, evaluation of the final observation distribution $p(z_t|x_t)$ would require to marginalize over (integrate) the noise distribution. However, this integral is analytically intractable and then requires to turn to approximation methods. For computational efficiency, the integral is then replace by the most likely approximation, meaning that the noise level is approximated by a point estimate corresponding to the maximum of the prediction distribution (its mean). At runtime, computation time can be saved by only predicting noise values for the subset of measurements required by the active function f_c. These functions associated with the classification and regression models then define a complete context-dependent observation model whose performance is illustrated in the next section.

V. Experiments

Experiments on real data are performed in the context of altitude estimation for a quadrotor UAV. The paparazzi platform [13] was used to collect sensor measurements provided by an ultrasonic sensor and a barometer, along with other UAV internal data. The altitude ground truth is provided

by a motion capture system and its data is synchronized at $50Hz$ with the UAV telemetry. All data is collected by manually flying the UAV such as covering the practicable space with different dynamic behaviors (smooth and more aggressive motions). The different datasets all contain around 6000 samples.

We aim at evaluating the benefits of introducing context-dependency over the estimation process. The proposed approach is then compared with the GP-PF algorithm defined in [5]. This state-of-the-art algorithm is based on a GP modeling the whole observation model $p(z_t|x_t)$. Both filters instantiations are trained on a common dataset similar to the one shown in Fig.4. For all datasets, the ultrasonic sensor measurements show frequent occurrences of strong outliers that we intuitively know to be caused by perturbations originating from the actuation of the 4 motors. We however never made deeper investigations on the nature of the perturbations and thus train the RVM-based filter by learning observation noise and measurement selection for a total of 3 measurement combinations: one filter based on the ultrasonic measurement, an other one based on the barometer, and a last one based on both ultrasonic and barometer sensors. The context data c_t contains both sensor measurements and the motor thrust command which, according to our intuition about ultrasonic sensor perturbation, may provide useful information for the selection task. The GP-PF is trained on the raw dataset containing the ground truth and the measurements. A Kalman filter with 3 sigma rejection scheme on all measurements is also evaluated on the common validation set. All tested filters share a common constant velocity prediction model with identical parameters.

The final estimation error for all filters is shown in Fig.5. As can be seen, the GP-PF diverged in many situations. Analysis of the trained observation GP shows that it failed to model the occurrence of outliers in the ultrasonic measurements. For some cases outliers are ignored, while for frequent occurrences phases the observation model tends to converge to the outlier values. Less importantly, one can notice that fluctuations of the barometer measurements when the motors are turned on (noticeable on the thrust command from Fig.4) are also not generalizing well on the validation set. The Kalman filter provides a lower estimation error thanks to the rejection scheme, but still shows some cases of strong divergences, especially for the high dynamic parts of the flight. Finally the RVM-based filter shows a noticeable improvement in both terms of error amplitude and frequency.

Clearly, both Kalman filter and GP-PF reveals the need for a context-dependent knowledge during the estimation process. While the Kalman filter manages to get rid of most outliers, the rejection mechanism can not handle some specific configurations. For example this is the case for the part of the dataset contained between samples 3000 and 4000 corresponding to the highest observed vertical speed variations. Here, ultrasonic sensor outliers and barometer measurement latency become coherent, leading to divergence. Meanwhile, the GP observation model is trained to fit at best the training set within which strong measurement variations in the measurements are observed for similar state values. Intuitively we understand that some information is missing when trying to fit a precise model of measurement generation (note however that in optimal conditions, this is not the case). This is were the benefits of the RVM-based approach come into play: the contextual knowledge compensates for the missing information and allows recognition of situations where the Kalman filter was unable to apply the correct rejection. The estimation output for the RVM-based filter on the validation set is shown in Fig.6. This figure also illustrates the role of the heteroscedastic observation noise in the reliability of the final estimate uncertainty. As can be seen, the noise models provide realistic and consistent evaluation of the observation noise such that the altitude truth remains inside the final uncertainty boundaries. Conversely, as noticed in [5], and as observed in these experiments, filters using homoscedastic observation noise such as the GP-PF and the Kalman filter provide lower and less variable uncertainties, leading to less consistent estimate output distribution.

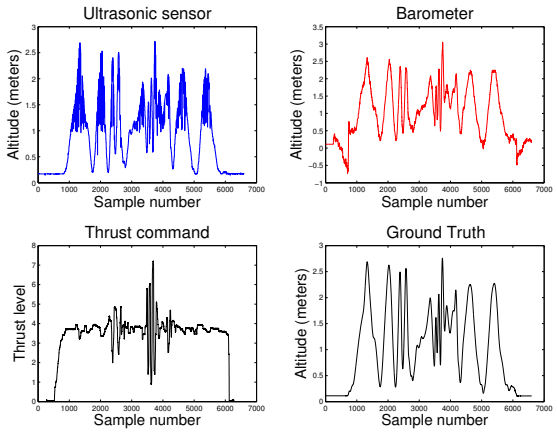

Fig. 4. Validation set

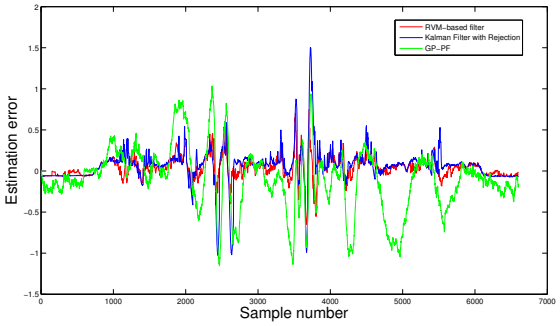

Fig. 5. Estimation error for the three filter instantiations

VI. Conclusion

A. Summary

This article presented a simple approach for introducing context-dependent knowledge within the observation model of a Bayes filter. As such, the approach is generic as it

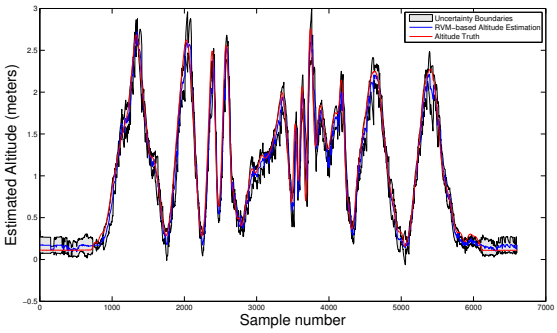

Fig. 6. RVM-based filter estimation output and associated 3 sigma uncertainty boundaries

can be directly applied regardless the version of the Bayes filter. Based on contextual information defined by the current measurements (and any additional information if required), a first classification model is trained. Its role consists in incorporating measurements in the estimation process when they are known to help in improving the final estimate. Conjointly, heteroscedastic (in our case context-dependent) observation noise models are also learned and help the filter in keeping a realistic estimation uncertainty level. The approach is evaluated and compared to two standard and state-of-the-art methods. Evaluation shows that context-dependency allows the approach to outperform other filters in both terms of estimation and uncertainty consistency. This work can be seen as an extension of the approach proposed in [3], but still offers room for improvement, especially through the introduction of a dynamic model which could potentially greatly improve context identification. However, the current implementation has the advantage of relying on classical RVM training and inference methods.

B. Discussion

In this work, RVM models have been chosen instead of GP for computational efficiency reasons. There is a actually a strong relationship between these two models, as RVM can be seen as a special case of GP [14]. Both models provide probabilistic outputs, *i.e.* each prediction is provided with an associated uncertainty. This specificity is a key feature and could be used to protect the system when it enters part of the input space where too few or no data has been used for training. Analysis of the prediction uncertainty can then allow the designer of the system to implement a *failsafe* behavior, either by asking the robot to stop, or by switching to a default standard estimation filter. It should however be noticed that in this situation, the RVM model is known to decrease drastically its uncertainty when entering unexplored parts of the input space [14], while GP behave more logically by increasing the uncertainty. This specific aspect requires further investigation about the applicability of such safety systems.

GP-PF and the RVM-based filter are both based on non-parametric models capable of representing dense and complex information. However, the principal difference between both approaches is that the GP-PF is based on a state-dependent model, where the RVM-based filter only exploits the contextual information. Introduction of the state value as an additional contextual information has been tested and proved to give lower performance for our approach. This is a direct consequence of the fact that relying on perceptual information is a much more robust foundation than using the state value. Clearly, a state-dependent model trained with the ground truth is more likely to produce erroneous predictions, as errors on the system state are unavoidable at runtime (hence differing from values seen in the training set). Conversely, using the measurement inputs ensures the RVM-based filter to be more reliable as its prediction capabilities do not depend on the own filter performance. Especially, one has to notice that the probabilistic models applied here do not take into account the possible existence of noise over the input variables. Investigating the exploitation of such models [15] might be an interesting direction for future work.

REFERENCES

[1] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, 2013, pp. 3923–3929.

[2] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[3] A. Ravet, S. Lacroix, G. Hattenberger, and B. Vandeportaele, "Learning to combine multi-sensor information for context dependent state estimation," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, 2013, pp. 5221–5226.

[4] Y. Bar-Shalom and X.-R. Li, *Estimation with Applications to Tracking and Navigation*. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[5] J. Ko and D. Fox, "Gp-bayesfilters: Bayesian filtering using gaussian process prediction and observation models," *Autonomous Robots*, vol. 27, no. 1, pp. 75–90, 2009.

[6] P. Pfaff, C. Stachniss, C. Plagemann, and W. Burgard, "Efficiently learning high-dimensional observation models for monte-carlo localization using gaussian mixtures," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, 2008, pp. 3539–3544.

[7] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sept. 2001.

[8] R. Turner, M. P. Deisenroth, and C. E. Rasmussen, "State-space inference and learning with Gaussian processes," in *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, Y. W. Teh and M. Titterington, Eds., 2010, pp. 868–875.

[9] M. Lzaro-gredilla and M. K. Titsias, "Variational heteroscedastic gaussian process regression," in *In 28th International Conference on Machine Learning (ICML-11*. ACM, 2011, pp. 841–848.

[10] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic gaussian process regression," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 393–400.

[11] I. Psorakis, T. Damoulas, and M. A. Girolami, "Multiclass relevance vector machines: sparsity and accuracy." *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1588–1598, 2010.

[12] P. Abbeel, A. Coates, M. Montemerlo, A. Y. Ng, and S. Thrun, "Discriminative training of kalman filters," in *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2005.

[13] P. Brisset, A. Drouin, M. Gorraz, P. Huard, and J. Tyler, "The paparazzi solution," in *Micro Aerial Vehicles*, 2006. [Online]. Available: http://paparazzi.enac.fr

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[15] A. McHutchon and C. E. Rasmussen, "Gaussian process training with input noise." in *NIPS*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 1341–1349.