On the Benefit of State Separation for Tracking in Image Space with an Interacting Multiple Model Filter

Stefan Becker(✉), Hilke Kieritz, Wolfgang Hübner, and Michael Arens

Fraunhofer IOSB, Gutleuthausstr. 1, 76275 Ettlingen, Germany
stefan.becker@iosb.fraunhofer.de
http://www.iosb.fraunhofer.de

Abstract. When tracking an object, it is reasonable to assume that the dynamic model can change over time. In practical applications, Interacting Multiple Model (IMM) filter are a popular choice for considering such varying system characteristics. The motion of the object is often modeled using position, velocity, and acceleration. It seems obvious that different image space dimensions can be considered in one overall system state vector. In this paper, the fallacy of simply extending the state vector in case of tracking an object solely in image space is demonstrated. Thereby, we show how under such conditions the effectiveness of an IMM filter can be improved by separating particular states. The proposed approach is evaluated on the VOT 2014 dataset.

Keywords: Interacting Multiple Models · Visual tracking

1 Introduction

An important component of tracking is the filtering problem in which estimates of object's state are computed while observations are progressively received. The estimation process is in general modeled using a Bayesian formulation [2]. For many filters, i.e. the well-known Kalman filter [1] or nonparametric methods such as particle filters [3], the posterior probability can be recursively updated by applying a perception model and a motion model. Under real world conditions, the object motion can change over time and it is impossible to define a unique motion model which captures all different motions the object can execute. An elegant way of dealing with motion uncertainties and capturing the complex dynamics of objects is the Interacting Multiple Model (IMM) filter [4]. It has been successfully employed in several applications [5,6]. The IMM approach can be used to fuse several models in one context by weighting each model from a set of models as possible candidates. Each model contributes to the final distribution depending on its current weight. In most cases, the motion is modeled by a bank of standard Kalman filters per object and the dynamics are described in 3D space. However, there exist several scenarios where objects are solely tracked on directly observed image space information. For example person tracking without

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 3–11, 2016.
DOI: 10.1007/978-3-319-33618-3_1

available external calibration. The goal of this paper is to reveal some fallacy when applying an IMM filter restricted to such information. We show how a separation of the state space vector improves the overall system accuracy. After some basic concepts of an IMM filter are described in Sect. 2, we show how a basic IMM setup with three standard motion models should be modified for a better image space object tracking. The results achieved on the public available VOT 2014 dataset are presented in Sects. 3 and 4 contains a conclusion.

2 Interacting Multiple Model Filter

In this section the basic concepts of the IMM filter and a reference IMM configuration for the evaluation are described. For a more detailed description see for example Hartikainen and Särkkä [7] or Bar-Shalom et al. [8]. As mentioned, it is reasonable to assume that the dynamic model of an object can change from time to time. As a solution, a system is considered to be composed of multiple independent models, where the currently active model is one from a discrete set of n candidate models ($M = \{M^1, \ldots, M^n\}$). The IMM filter is a popular choice for practical applications. Some prior probability μ_0^j for each model M^j and the state transition probability between time index $k-1$ and k from model i to model j (denoted by $p_{ij} = P(M_k^j | M_{k-1}^i)$) are assumed to be known. The transition probability matrix p_{ij} can be interpreted as a first order Markov chain characterizing the mode transitions. Hence systems of this type are commonly referred to Markovian switching systems (Bar-Shalom et al. [8]). Thus, the model or mode transition can be characterized by a first order Markov chain and described as transition probability matrix p_{ij}. The closed form solution for the state estimation problem of a discrete-time IMM filter can be written as follows:

$$x_k = F_k^j x_{k-1} + w_k^j \tag{1}$$

$$y_k = H_k^j x_k + r_k^j \tag{2}$$

Here, x_k is the state of the object and the effective model in time step $k-1$ is denoted by j. F_k is the state transition matrix which applies the effect of each system state parameter at time $k-1$ on the system state at time k. H_k is the measurement model matrix that maps the state parameters into the measurement domain. $w_k \sim N(0, Q_k)$ is the process noise and $r_k \sim N(0, R_k)$ is the measurement noise. For our goal to only rely on the directly observed information y_k, we use the image space coordinates and the scale of the object as measurement. This information can be obtained from every object detector following the sliding window paradigm. Although the detectors differ in many aspects, the output of such a sliding window based detector is a rectangular bounding box centered at the object location. Here (x, y) is the center position in the image space and s the scale. For describing the overall state of our reference IMM configuration the corresponding velocities $(\dot{x}, \dot{y}, \dot{s})$ and acceleration are used $\ddot{x}, \ddot{y}, \ddot{s}$. The most common linear motion models are the constant position model (CP),

the constant velocity model (CV), and the constant acceleration model (CA). In our experiments, we choose an IMM filter configuration, which consist of these three basic models. When the object remains at the same position the velocity and acceleration are reduced to zero since the object is not moving. Thus, the transition matrix for a state vector including the 9 mentioned states ($x_k = (x, y, s, \dot{x}, \dot{y}, \dot{s}, \ddot{x}, \ddot{y}, \ddot{s})$) for the constant position motion model is defined as

$$F_k^{CP} = \begin{bmatrix} I_{3\times3} & 0_{3\times6} \\ 0_{3\times6} & 0_{6\times6} \end{bmatrix}. \tag{3}$$

The constant velocity model is used in most tracking approaches and can be then be defined as

$$F_k^{CV} = \begin{bmatrix} I_{3\times3} & I_{3\times3}T & 0_{3\times3} \\ 0_{3\times3} & I_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \end{bmatrix}. \tag{4}$$

Here, T is the number of discrete time steps. In literature, several assumptions on how to model the acceleration process of an object are proposed (see Li and Jilkov [9]). Here, a CA model is considered as

$$F_k^{CA} = \begin{bmatrix} I_{3\times3} & I_{3\times3}T & \frac{1}{2}I_{3\times3}T^2 \\ 0_{3\times3} & I_{3\times3} & I_{3\times3}T \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \end{bmatrix}. \tag{5}$$

The IMM filter basically consists of three major steps: interaction (mixing), filtering and combination. In the interaction stage and under the assumption that a particular model is the right model at the current time step, the initial conditions for this model are obtained by mixing the state estimates produced by all filters. In detail, the mixing probabilities $\mu_k^{i|j}$ for each model M^i and M^j are calculated as $\mu_k^{i|j} = \frac{1}{\bar{c}_j} p_{ij}\mu_{k-1}^i$ with $\bar{c}_j = \sum_{i=1}^n p_{ij}\mu_{k-1}^i$. Thereby, μ_{k-1}^i is the probability of model M^i in the time step $k-1$ and $\bar{c}_j$ a normalization factor. For each filter the mixed mean and covariance is computed as follows:

$$m_{k-1}^{0j} = \sum_{i=1}^{n} \frac{1}{\bar{c}_j} \mu_k^{i|j} m_{k-1}^i \tag{6}$$

$$P_{k-1}^{0j} = \sum_{i=1}^{n} \mu_k^{i|j} \left(P_{k-1}^i + (m_{k-1}^i - m_{k-1}^{0j})(m_{k-1}^i - m_{k-1}^{0j})^T \right) \tag{7}$$

Here, m_{k-1}^i and P_{k-1}^i are the updated mean and covariance for model i at time step $k-1$.

Then in the filtering stage, for each individual model conditioned on the current active mode, a standard Kalman filtering (KF) is done. Correspondingly a prediction $\left[m_k^{-,i}, P_k^{-,i}\right] = KF_p(m_{k-1}^{0j}, P_{k-1}^{0j}, F_k^i, Q_k)$ and update step $\left[m_k^i, P_k^i\right] = KF_u(m_k^{-,i}, P_k^{-,i}, H_k^i, R_k^i)$ is applied. Initialization is done with m_{k-1}^i and P_{k-1}^i. Then the model probabilities $\mu_k^i = \frac{1}{c}\Lambda_k^i \bar{c}_i$ are adapted according to

the likelihood of the measurement for each filter Λ_k^i. Where $c = \sum_{i=1}^{n} \Lambda_k^i \bar{c}_i$ is a normalizing factor.

The final step of the IMM filter is combination. There, the combined estimate for the state mean and covariance is computed as follows:

$$m_k = \sum_{i=1}^{n} \mu_k^i m_k^i \tag{8}$$

$$P_k = \sum_{i=1}^{n} \mu_k^i \left(P_k^i + (m_k^i - m_k)(m_k^i - m_k)^T \right) \tag{9}$$

3 IMM Configuration and Evaluation

In this section, we evaluate the effectiveness of different IMM filter configuration in terms of state separation for the case of tracking the object only with directly observed image space information. The desired states for the IMM filter for tracking were determined in Sect. 2. Besides the center position in the image space (x, y) and the scale (s) of the object, the IMM filter uses the corresponding velocities $(\dot{x}, \dot{y}, \dot{s})$ and acceleration $(\ddot{x}, \ddot{y}, \ddot{s})$. The discrete set of motion models consists of three basic models, in particular CP, CV, and CA. Intuitively, one would simply set the state vector set to

$$x_{k,IMM\ 1} = (x, y, s, \dot{x}, \dot{y}, \dot{s}, \ddot{x}, \ddot{y}, \ddot{s}). \tag{10}$$

Thus, only one IMM filter is required for monitoring all desired states. For a standard Kalman filter, a separation of the states and additionally required filter is redundant. Due to the characteristics of an IMM filter, not only a poor choice of single motion model, but in addition a careless extension of the states can lead to a non optimal performance. An optimal filtering behavior using a multiple model system requires an optimal filter for every possible model sequence. Hence, some kind of approximations are needed in practical applications. For an IMM filter, this is done by conditioning all filters on the currently active model and the final state estimate is obtained by merging the results of all elemental filters. Hence, a poor estimate of active model affects the weighting of the mixed inputs. A combining of the image coordinates and the scale in one state vector can thereby result in errors for the calculation of the model probabilities, especially when combining the scale with the image position. For example, the scale change of an object can be constant while the object is moving. Thus the best fitting model for describing the scale is CP, whereas this model is a poor fit for the image position. Therefore, we propose to use an extra IMM-filter instead of one. Hence, the scale and the corresponding velocity and acceleration are estimated independent from the position states and their derivatives. This leads to the following IMM configuration:

$$x_{k,IMM\ 2} = (x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}); (s, \dot{s}, \ddot{s}). \tag{11}$$

A separation of the scale with an additional filter seems obvious, but when tracking with directly observed image space data, a split into independent image coordinates may appear to be at first not required. In order to show the benefit of such an IMM set up, we recommend an IMM configuration as follows:

$$x_{k,IMM\ 3} = (x, \dot{x}, \ddot{x}); (y, \dot{y}, \ddot{y}); (s, \dot{s}, \ddot{s}). \tag{12}$$

Here, three IMM filter are used to describe the x position, y position, scale, and the corresponding derivatives. Hence, every motion along the image axes is captured with a separate filter.

Evaluation is done on the VOT 2014 dataset [10]. This dataset is a selection of 25 widely-used object tracking sequences. Although the dataset is originally designed to compare different appearance or visual tracker, it includes a variety of different object motions. Figure 1 shows the first frame of exemplary sequences where the unified bounding box of the object is highlighted in green.

Fig. 1. Example tracking sequences for evaluation from the VOT 2014. The first frame with the unified bounding box of the object is shown for the sequences "bicycle", "jogging", "surfing", "woman". (Color figure online)

The main feature of the IMM filter is the ability to estimate the state of a dynamic system with several behavior modes which can switch from one to another. Besides that, the IMM filter is a good compromise between performance and complexity [11]. The overall performance depends on a number of design parameters. The most critical design parameters are the model set structure, process and measurement noises, initial state, and the jump structure with transition probabilities. Nonetheless, the above described basic IMM setup with 3 standard motion model is suboptimal for some scenarios from the VOT 2014 dataset [10], we keep the combination of one constant position, one constant velocity, and one constant acceleration model fixed. In practice, the transition probability matrix is often assumed known and is chosen a priori. As stated in Bar-Shalom [8], an ad-hoc approach is to fill the diagonals with values close to one. We set the diagonals to 0.99 and the other transition values to 0.005. Because the CV model is the mostly used in tracking approaches, we set the initial model probability μ_0^i in favor of this model to 0.98 and to 0.01 for the other models. The measurement and process noise is modeled as additive white noise. In the experiments the standard deviation of both noises was varied between 1, 2, 5, and 10. Here, only the diagonals of process noise covariance matrix Q and measurement noise covariance R include non-zero values.

For every image sequence, the first 10 frames are excluded and used for initializing of the filters. The update interval t_{update} for getting a new measurement for the filter was varied between every frame, every third frame and fifth frame. Since, the standard output of object detectors are a rectangular bounding box centered at the object location, we use the ground truth bounding boxes from the VOT 2014 dataset to simulate the output of an object detector and for evaluating the prediction accuracy. Performance measures aim at summarizing the extent to which the trackers prediction agrees with the ground truth annotation. In Cehovin et al. [12], a general definition of an object state description in a sequence with length N is established based on the center of the object and the region of the object at time k. In case of tracking an object in image space the region is usually described by a bounding box. From the IMM filter, we use the predicted states center location x, y and scale s to calculate an unified bounding box A_k^O. With this predicted objects region form the tracker and the ground-truth region an overlap can be calculated as $\frac{A_k^O \cap A_k^{GT}}{A_k^O \cup A_k^{GT}}$. For the ground truth area A_k^{GT} also an unified bounding box is considered. In general, the width of the enclosing bounding box is more strongly influenced by the body pose of the objects. Hence, a unified bounding box with a width of $\frac{1}{3}$ scale is used. A property of region overlap measures is that they account for both position and size of the predicted and ground-truth bounding boxes simultaneously, and there is no normalization problem. The overlap measure is summarized over an entire sequence by an average overlap. In addition to the average overlap, the number of frames in which the overlap is below a threshold of 0.5 is recorded and used as a second comparative score.

The overall results for the three different IMM configurations are exemplary summarized for $\sigma_w^2 = 2$, $\sigma_r^2 = 5$, $t_{update} = 3$ in Table 1. Other parameter settings may differ slightly, but are equal at their core. This means that the achieved overlap varies and that for some specific sequences the ranking of the IMM configuration changes, but overall it can clearly be noticed that the IMM configuration, that uses separated image space coordinates and scale, outperforms the other configurations. Due to the fact that the motion of objects in some particular sequence is highly non-linear, the chosen combination of motion model is not optimal. Moreover, this can also result in a changed ranking, but the trend towards the third configuration for achieving superior results is clearly visible for all evaluated parameter settings.

When tracking an object without a mapping between measurement domain and the states, the motion in a particular direction is independent from the other direction. Because the elemental filters are conditioned on the best fitting model the final estimate is negatively influenced by a naive extension of the state vector. For combining the scale and its derived changes with the actual motion states this seems obvious. But the presented results show how crucial this is also for mixing between image coordinates. For the majority of the evaluated sequences the average overlap achieved with the separated IMM states is larger than with the other configuration. An improvement can also be perceived by avoiding a combination between dynamics and scale. Thus the second

Table 1. Performance summary for the different IMM filter configurations.

Settings: $\sigma_w^2 = 2$, $\sigma_r^2 = 5$, $t_{update} = 3$						
Sequence	IMM 1		IMM 2		IMM 3	
	Failure rate	Overlap ratio	Failure rate	Overlap ratio	Failure rate	Overlap ratio
Ball	0.270	0.634	0.191	0.679	0.164	**0.695**
Basketball	0.003	0.863	0.003	0.884	0.004	**0.891**
Bicycle	0.304	0.602	0.233	0.641	0.173	**0.692**
Bolt	0.080	0.774	0.044	0.810	0.027	**0.842**
Car	0.340	0.610	0.261	0.642	0.108	**0.710**
David	0.167	0.697	0.152	0.715	0.141	**0.720**
Diving	0.082	**0.793**	0.135	0.749	0.135	0.736
Drunk	0.000	0.931	0.000	0.929	0.000	**0.931**
Fernando	0.018	0.857	0.021	0.852	0.018	**0.859**
Fish1	0.242	0.663	0.144	**0.726**	0.111	0.725
Fish2	0.107	0.745	0.084	0.769	0.064	**0.775**
Gymnastics	0.107	**0.798**	0.138	0.787	0.138	0.786
Hand1	0.262	0.656	0.232	0.664	0.227	**0.665**
Hand2	0.410	0.542	0.379	0.559	0.359	**0.576**
Jogging	0.047	0.769	0.054	0.776	0.047	**0.777**
Motocross	0.092	0.752	0.188	0.745	0.188	**0.755**
Polarbear	0.011	0.848	0.008	0.849	0.011	**0.852**
Skating	0.000	0.866	0.000	0.881	0.000	**0.898**
Sphere	0.295	0.618	0.300	**0.629**	0.316	0.605
Sunshade	0.559	0.426	0.571	0.442	0.484	**0.488**
Surfing	0.111	0.699	0.081	0.746	0.048	**0.773**
Torus	0.150	0.706	0.146	**0.723**	0.142	0.712
Trellis	0.358	0.579	0.276	0.633	0.238	**0.661**
Tunnel	0.129	0.695	0.078	**0.734**	0.101	0.729
Woman	0.051	0.773	0.053	0.794	0.051	**0.805**

configuration (IMM 2) outperforms the naive state extension from configuration one (IMM 1). This state splitting is also recommend when the actual motion is described in 3D. In summary, when relying on direct observed measurement, which is common for 2D Tracking, a naive extension of the state vector in case of tracking with an IMM filter instead of using single Bayes filter should be avoided. However, the fact that independent states are affected by mixing the inputs from the elemental filters, which is a result of the required approximation for an optimal filtering without keeping every possible model sequence, is easily

forgotten when applying IMM filter for direct tracking in 2D. With this simple reminder a better IMM filtering can be achieved. While, the overall performance can be further improved by selecting alternative motion models which better fit to the dynamics of the object in the scene, the awareness of not naively extend the state is also crucial. All states of an IMM state vector should depend on each other and hence each additional independent state and its derivatives should be considered in an additional IMM filter. Hence the conditioning on the current best fitting model can not negatively affects the overall performance. The motion of an object in image space is a very good example where the dynamics along the image axes should be considered independently with an IMM filter.

4 Conclusion

In this paper, we showed how a naive extension of the state space can negatively affect the performance of an IMM filter. The required approximation by merging the output of the elemental filter based on the current best fitting filter affects states which are independent of each other. Especially when tracking an object only based on direct image space measurement, a combination in the state vector of both image coordinates should be avoided. This simple reminder of a common fallacy helps to improve the effectiveness of an IMM filter for considering varying system characteristics. The benefit of this favored design scheme for an IMM filter configuration is shown on the VOT 2014 dataset.

References

1. Kalman, R.E.: A new approach to linear filtering and prediction problems. ASME J. Basic Eng. **82**(1), 35–45 (1960)
2. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics. Intelligent Robotics and Autonomous Agents. The MIT Press, Cambridge (2005)
3. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
4. Blom, H.A.P., Bar-Shalom, Y.: The interacting multiple model algorithm for systems with Markovian switching coefficients. Trans. Autom. Control **33**, 780–783 (1988)
5. Cooper, D.C.: Multiple target tracking with radar applications. Electron. Power **33**(6), 407 (1987)
6. Blackman, S.S., Popoli, R.: Design and analysis of modern tracking system. Artech House radar library (1999)
7. Hartikainen, J., Särkkä, S.: Optimal filtering with Kalman filters and smoothers a Manual for Matlab toolbox EKF/UKF (2008). http://www.lce.hut.fi/research/mm/ekfukf/
8. Bar-Shalom, Y., Kirubarajan, T., Li, X.-R.: Estimation with Applications to Tracking and Navigation. Wiley, New York (2002)
9. Li, X.R., Jilkov, V.P.: Survey of maneuvering target tracking. Part V. Multiple-model methods. Trans. Aerosp. Electron. Syst. **41**(4), 1255–1321 (2005)

10. VOT: Visual Object Tracking Challenge Dataset. In: European Conference on Computer Vision Workshops (2014). http://www.votchallenge.net/vot2014/
11. Gomes, J.B.B.: An overview on target tracking using multiple model methods, PhD thesis, Technical University of Lisbon (2008)
12. Cehovin, L., Kristan, M., Leonardis, A.: Is my new tracker really better than yours? In: Winter Conference on Applications of Computer Vision (WACV), pp. 540–547 (2014)