Multi-object stereo filtering in disparity space

Špela Ivekovič[1] and Daniel Clark[2]
1. School of Computing, University of Dundee, Dundee DD1 4HN
2. Joint Research Institute in Signal and Image Processing, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS

Abstract— Stereo tracking refers to the problem of tracking a three-dimensional system from a pair of cameras. The majority of stereo tracking algorithms track the system in 3-D Euclidean space, since the operator is generally interested in knowing the position of the object in the world co-ordinate system. Unfortunately, the projection from the 3-D co-ordinate space onto the image planes is non-linear and the noise is dependent on the position of the system. These facts can seriously impede the reliability of tracking algorithms. We propose a fundamentally different approach by stereo tracking in disparity space for optimal 3-D object state estimation.

Keywords— stereo tracking, 3-D estimation, disparity space, Kalman filter, multi-object tracking, PHD filter

1. Introduction

In stereo tracking, the state of a three-dimensional system is estimated from two-dimensional measurements generated by a pair of cameras. The majority of stereo tracking algorithms track in 3-D Euclidean space, since the operator is generally interested in knowing the state of the object, such as position and velocity, in the world co-ordinate system. Unfortunately, the projection from the 3-D co-ordinate space (the state space) onto the camera image planes (the observation space) is non-linear, violating the Kalman filter assumptions and making it necessary to resort to nonlinear approximations instead. A further difficulty with tracking in 3-D is that the noise in the state estimate is *heteroscedastic*, that is, it changes with the position of the system. The heteroscedasticity is a direct consequence of the nature of stereo reconstruction, as shown in Figure 1. These difficulties can seriously impede the reliability of tracking algorithms. To address these, we propose a fundamentally different approach by stereo tracking in disparity space as an intermediate step to 3-D object state estimation. Stereo-tracking in disparity space has two key advantages over tracking in 3-D Euclidean space: (i) the noise in the state estimate is *homoscedastic*[1], [2], [3], [4], and (ii) the projections onto the observation space (the two image planes) are linear. The immediate consequence of this is that the objects can now be tracked with the usual linear Gaussian assumptions made in the multi-sensor Kalman filter.

We use this result as a basis for stereo multi-object tracking in scenarios with false alarms and missed detections using the iterated-corrector PHD filter [5]. This approach has several advantages over traditional multi-object stereo tracking algorithms: (i) there is no data association for assigning measurements to target tracks, (ii) there is no need to match pairs of measurements corresponding to the same object thereby avoiding the stereo correspondence problem, (iii) the objects do not need to generate observations at each time-step, and (iv) the tracker is robust in scenarios with false alarms.

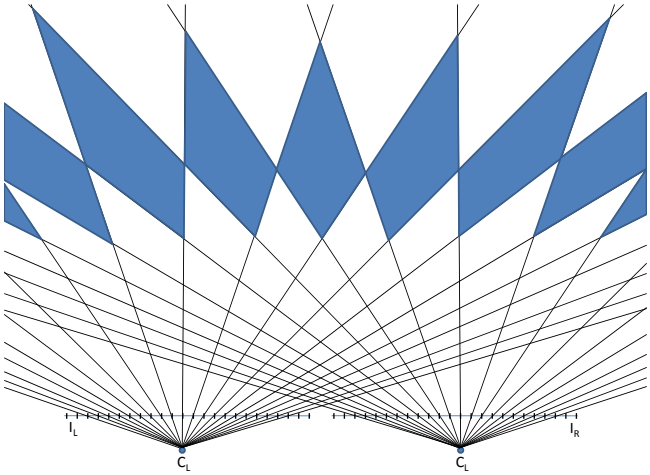

Figure 1. Schematic illustration of the heteroscedastic noise property in stereo-reconstructed 3-D state estimates. The image planes have been split into discrete pixels and the corresponding image rays are shown. As the object moves further away from the sensor, its position uncertainty increases, as illustrated by the quadrilateral formed by the intersecting image rays. The shaded quadrilaterals illustrate the uncertainty in the reconstructed state estimate for a disparity value $d = 3$.

This paper is organized as follows. In Section 2, we describe disparity space and its relation to the two observation planes and 3-D. In Section 3, we present optimal filtering for stereo sensors, describe why it is not possible to track in 3-D directly from stereo camera data and propose optimal linear filtering in disparity space. In Section 4, we describe multi-object filtering and the Gaussian mixture Probability Hypothesis Density (PHD) filter. In Section 5, we present results on simulated data and then conclude in Section 6.

2. Disparity space

We assume that every point expressed in Euclidean coordinates has an equivalent projective space representation in homogeneous coordinates. For reasons of clarity, we emphasise the distinction by using a "hat" to represent points in Euclidean coordinates. To clarify that homogeneous notation is applied to points from $\mathbb{R}^3$, we refer to the corresponding projective space as $\mathbb{P}^3(\mathbb{R}^3)$, while the points from disparity space $\mathbb{D}^3$, when stated in homogeneous coordinates, belong

to $\mathbb{P}^3(\mathbb{D}^3)$.

For example, for 3-D points reconstructed from stereo, we use the following notation:

$$\begin{aligned}\hat{\boldsymbol{p}} &= (x,y,z) \in \mathbb{R}^3 \Rightarrow \boldsymbol{p} = (x,y,z,1) \in \mathbb{P}^3(\mathbb{R}^3),\\ \boldsymbol{p} &= (wx, wy, wz, w) \in \mathbb{P}^3(\mathbb{R}^3) \Rightarrow\\ &\quad \hat{\boldsymbol{p}} = (wx/w, wy/w, wz/w) \in \mathbb{R}^3\end{aligned}$$

Let us now assume that a point $\boldsymbol{p} = (x,y,z,1)$ is viewed by two distinct cameras, left camera with projection matrix P_l and right camera with projection matrix P_r. The images of the point $\boldsymbol{p}$ are defined as $\boldsymbol{p}_l = (u_l, v_l, 1)^T \simeq P_l\boldsymbol{p}$ and $\boldsymbol{p}_r = (u_r, v_r, 1)^T \simeq P_r\boldsymbol{p}$ in the left and right camera's image plane, respectively, where "$\simeq$" denotes equality up to a scale factor. The corresponding points $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$ are related by a *disparity* which, in a general case, is defined as:

$$d(\boldsymbol{p}_l, \boldsymbol{p}_r) = \hat{\boldsymbol{p}}_r - \hat{\boldsymbol{p}}_l = (u_r - u_l, v_r - v_l). \tag{1}$$

In the case of rectified images, the two corresponding points lie on the same scanline, and the disparity simplifies to a displacement along that scanline:

$$d(\boldsymbol{p}_l, \boldsymbol{p}_r) = u_r - u_l. \tag{2}$$

For a rectified stereo pair of images, the disparity space is then defined as a three-dimensional space $\mathbb{D}^3 = \{u, v, d\}$. The so-defined disparity space is a projective space. This can be shown by deriving a projective transformation P_D between $\mathbb{P}^3(\mathbb{R}^3)$ and $\mathbb{P}^3(\mathbb{D}^3)$, as follows.

We assume, without a loss of generality, a specific form of the rectified projection matrices [6]. Let the left and right rectified camera projection matrices, $\tilde{P}_l$ and $\tilde{P}_r$, be written as:

$$\tilde{P}_l = \begin{pmatrix} p^l_{11} & p^l_{12} & p^l_{13} & p^l_{14} \\ p^l_{21} & p^l_{22} & p^l_{23} & p^l_{24} \\ p^l_{31} & p^l_{32} & p^l_{33} & p^l_{34} \end{pmatrix} \tag{3}$$

$$\tilde{P}_r = \begin{pmatrix} p^r_{11} & p^r_{12} & p^r_{13} & p^r_{14} \\ p^l_{21} & p^l_{22} & p^l_{23} & p^l_{24} \\ p^l_{31} & p^l_{32} & p^l_{33} & p^l_{34} \end{pmatrix}. \tag{4}$$

Projecting a point $\boldsymbol{p} = (x,y,z,1)^T$ into left and right view gives the left and right image point, $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$:

$$\boldsymbol{p}_l = \begin{pmatrix} u_l \\ v_l \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{p^l_{11}x + p^l_{12}y + p^l_{13}z + p^l_{14}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ \frac{p^l_{21}x + p^l_{22}y + p^l_{23}z + p^l_{24}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ 1 \end{pmatrix} \simeq \tilde{P}_l\boldsymbol{p} \tag{5}$$

$$\boldsymbol{p}_r = \begin{pmatrix} u_r \\ v_r \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{p^r_{11}x + p^r_{12}y + p^r_{13}z + p^r_{14}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ \frac{p^l_{21}x + p^l_{22}y + p^l_{23}z + p^l_{24}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ 1 \end{pmatrix} \simeq \tilde{P}_r\boldsymbol{p} \tag{6}$$

A point $\boldsymbol{s} \in \mathbb{P}^3(\mathbb{D}^3)$, is defined as:

$$\boldsymbol{s} = \begin{pmatrix} u_l \\ v_l \\ u_r - u_l \\ 1 \end{pmatrix} =$$

$$\begin{pmatrix} \frac{p^l_{11}x + p^l_{12}y + p^l_{13}z + p^l_{14}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ \frac{p^l_{21}x + p^l_{22}y + p^l_{23}z + p^l_{24}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ \frac{p^r_{11}x + p^r_{12}y + p^r_{13}z + p^r_{14}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} - \frac{p^l_{11}x + p^l_{12}y + p^l_{13}z + p^l_{14}}{p^l_{31}x + p^l_{32}y + p^l_{33}z + p^l_{34}} \\ 1 \end{pmatrix}, \tag{7}$$

and the linear transformation P_D, for which

$$\boldsymbol{s} \simeq P_D\boldsymbol{p}, \quad \boldsymbol{s}, \boldsymbol{p} \in \mathbb{P}^3, \hat{\boldsymbol{p}} \in \mathbb{R}^3, \hat{\boldsymbol{s}} \in \mathbb{D}^3 \tag{8}$$

as

$$P_D = \begin{pmatrix} p^l_{11} & p^l_{12} & p^l_{13} & p^l_{14} \\ p^l_{21} & p^l_{22} & p^l_{23} & p^l_{24} \\ p^r_{11} - p^l_{11} & p^r_{12} - p^l_{12} & p^r_{13} - p^l_{13} & p^r_{14} - p^l_{14} \\ p^l_{31} & p^l_{32} & p^l_{33} & p^l_{34} \end{pmatrix} \tag{9}$$

The transformation P_D is the link between $\mathbb{R}^3$ and $\mathbb{D}^3$ which allows us to map the disparity space estimates into 3-D. We take advantage of this by formulating the stereo filtering problem in disparity space. In the next section, we describe the filtering problem for a stereo pair of sensors and present an optimal filter for disparity space.

3. Optimal Filtering for Stereo Sensors

The stereo-sensor Bayes filter

The optimal Bayes filter propagates the posterior density $p_k(x_k|z_{1:k})$ of a system x_k conditioned on the sequence of measurements $z_{1:k} = z_1, \ldots, z_k$ up to the current time step with the following recursion that predicts according to the dynamical model and updates according to the observation model,

$$p_{k|k-1}(x_k|z_{1:k-1}) = \int f_{k|k-1}(x_k|x)p_{k-1}(x|z_{1:k})dx \tag{10}$$

$$p_k(x_k|z_{1:k}) = \frac{g_k(z_k|x_k)p_{k|k-1}(x_k|z_{1:k-1}),}{\int g_k(z_k|x)p_{k|k-1}(x|z_{1:k-1})dx}. \tag{11}$$

The dynamical model is governed by the Markov transition density $f_{k|k-1}(x_k|x_{k-1})$, and the observation model is governed by the conditional likelihood $g_k(z_k|x_k)$ of observing measurement z_k given that the object state is x_k. If we have two independent sensors in stereo, then the Bayes update requires two observation likelihoods, one for each sensor,

$$p_k(x_k|z^{[1]}_{1:k}, z^{[2]}_{1:k}) = \frac{g^{[1]}_k(z^{[1]}_k|x_k)g^{[2]}_k(z^{[2]}_k|x_k)p_{k|k-1}(x_k|z^{[1]}_{1:k-1}, z^{[2]}_{1:k-1})}{\int g^{[1]}_k(z^{[1]}_k|x)g^{[2]}_k(z^{[2]}_k|x)p_{k|k-1}(x|z^{[1]}_{1:k-1}, z^{[2]}_{1:k-1})dx}. \tag{12}$$

The stereo-sensor Bayes update can be calculated either through equation (12) or iteratively, so that the Bayes update in equation (11) is computed for each sensor observation likelihood. The Kalman filter provides an optimal solution to the Bayes filter under linear Gaussian assumptions [7], [8]. In particular, it is assumed that the Markov transition and likelihood are both linear with Gaussian noise, i.e.,

$$f_{k|k-1}(x|\zeta) = \mathcal{N}(x; F_{k-1}\zeta, Q_{k-1}), \tag{13}$$

$$g_k(z|x) = \mathcal{N}(z; H_k x, R_k), \tag{14}$$

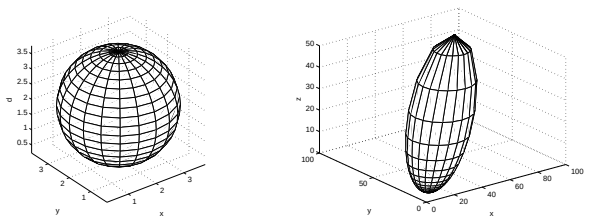

Figure 2. Disparity space covariance (left) and the same covariance mapped into 3-D (right). Heteroscedastic nature in 3-D is evident.

where $\mathcal{N}(\cdot; m, P)$ denotes a Gaussian density with mean m and covariance P, F_{k-1} is the state transition matrix, Q_{k-1} is the process noise covariance, H_k is the observation matrix, and R_k is the observation noise covariance. Now suppose that the posterior at time $k-1$ is the Gaussian $\mathcal{N}(x; m_{k-1}, P_{k-1})$. The predicted mean and covariance of the resulting Gaussian are computed with

$$m_{k|k-1} = F_{k-1} m_{k-1}, \tag{15}$$

$$P_{k|k-1} = Q_{k-1} + F_{k-1} P_{k-1} F_{k-1}^T, . \tag{16}$$

Suppose that the predicted density $p_{k|k-1}(x_k|z_{1:k-1})$ is Gaussian, then the updated mean and covariance are calculated with

$$m_k(z) = m_{k|k-1} + K_k(z - \hat{z}_{k|k-1}), \tag{17}$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1}, \tag{18}$$

and where

$$q_k(z) = \mathcal{N}(z; \hat{z}_{k|k-1}, S_{k|k-1}), \tag{19}$$

$$\hat{z}_{k|k-1} = H_k m_{k|k-1}, \tag{20}$$

$$K_k = P_{k|k-1} H_k^T S_{k|k-1}^{-1}, \tag{21}$$

$$S_{k|k-1} = H_k P_{k|k-1} H_k^T + R_k. \tag{22}$$

The stereo Bayes update for the Kalman filter can be computed by updating with the measurement from each sensor in turn. In the next section, we demonstrate that it is not possible to use the two-sensor Kalman filter for tracking in 3-D.

3-*D Tracking from Stereo*

Let us assume that the unobserved target state x_k at time-step k consists of target's 3-D position and velocity:

$$x_k = (x, \dot{x}, y, \dot{y}, z, \dot{z}, 1, 0)^T \tag{23}$$

Let us suppose that the dynamical model of the target is linear, so that

$$x_k = F_{k|k-1} x_{k-1}. \tag{24}$$

Let the observation space consist of 2 image planes, one for each camera in the stereo pair, and the observation model used to project the predicted state into the observation space be defined as

$$\hat{x}_k = P_k^{[i]} x_k, \tag{25}$$

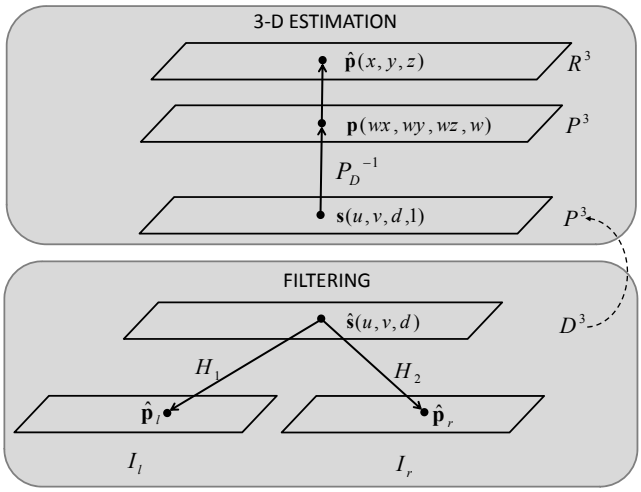

Figure 3. Linear filtering in disparity space. Instead of approximating the non-linear transformation from $\mathbb{R}^3$ to the two image planes, linear filtering can be done in disparity space and the resulting state estimates transformed into $\mathbb{R}^3$ using P_D from Equation (9). This way, the nonlinear relationship is only used in the 3-D estimation step.

where

$$P_k^{[i]} = \begin{pmatrix} p_{11}^i & 0 & p_{12}^i & 0 & p_{13}^i & 0 & p_{14}^i & 0 \\ p_{21}^i & 0 & p_{22}^i & 0 & p_{23}^i & 0 & p_{24}^i & 0 \\ p_{31}^i & 0 & p_{32}^i & 0 & p_{33}^i & 0 & p_{34}^i & 0 \end{pmatrix}, \tag{26}$$

and where the non-zero elements of $P_k^{[i]}$ come from the i^{th} camera projection matrix (see, e.g., Equation (3)).

The restriction that the observation model be linear implies that we can only obtain the projected estimate in homogeneous coordinates ($\mathbb{P}^2$). If we need to know the image coordinates in $\mathbb{R}^2$, we must perform a division by the last coordinate, which contradicts our assumption that the observation model from $\mathbb{R}^3$ to $\mathbb{R}^2$ is linear.

Optimal Linear Filtering in Disparity Space

The nonlinear observation model can be overcome by tracking in disparity space. Disparity space is linked to the 3-D space in which the target really exists via the projective transformation P_D given in Equation (9). If we know the state estimates in disparity space and have a calibrated and rectified camera pair, we are able to compute the state estimates in 3-D space. The trick is in realising that the transformation between disparity space and 3-D does not need to happen during tracking, where the nonlinearity in Euclidean coordinates complicates things. Instead, the tracking itself is done in disparity space only and the resulting estimates transformed into 3-D afterwards, as shown in Figure 3.

When tracking in disparity space, the state now becomes

$$x = (u, \dot{u}, v, \dot{v}, d, \dot{d})^T, \tag{27}$$

where u and v designate image column and row dimensions, respectively, and d denotes the disparity between the point

(u_l, v_l) in the left image and the corresponding point (u_r, v_r) in the right image. We assume a rectified camera setup, which means that the corresponding points are in fact (u_l, v_l) and (u_r, v_l), and the disparity becomes a scalar value, $d = u_r - u_l$.

Contrary to our attempt to track in 3-D space, we now do not have to work with homogeneous coordinates anymore because the observation model becomes a simple orthographic projection, one for each of the cameras:

$$H_k^{left} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \tag{28}$$

$$H_k^{right} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \tag{29}$$

The advantage of tracking in disparity space is not only the linearity of the observation model but also that the noise is homoscedastic. This is not true for 3-D space where the estimates are based on stereo, as the noise in the depth dimension increases with the distance from the camera, as illustrated in Figures 1 and 2.

4. MULTI-OBJECT STEREO FILTERING

Mahler generalised the single-object Bayes filter to multiple objects using the calculus of Finite Set Statistics [9]. Instead of propagating a posterior distribution on a state vector based on observation vectors, the optimal multi-target Bayes filter propagates the multi-target density in state set X_k, conditioned on the sets of observations up to time k, from each independent sensor with the following multi-object generalisation of the stereo Bayes filter recursion

$$p_{k|k-1}(X_k|Z_{1:k-1}^{[1]}, Z_{1:k-1}^{[2]}) = \int f_{k|k-1}(X_k|X) p_{k-1}(X|Z_{1:k}^{[1]}, Z_{1:k}^{[2]})\delta X, \tag{30}$$

$$p_k(X_k|Z_{1:k}^{[1]}, Z_{1:k}^{[2]}) = \frac{g_k^{[1]}(Z_k^{[1]}|X_k) g_k^{[2]}(Z_k^{[2]}|X_k) p_{k|k-1}(X_k|Z_{1:k-1}^{[1]}, Z_k^{[2]})}{\int g_k^{[1]}(Z_k^{[1]}|X) g_k^{[2]}(Z_k^{[2]}|X_k) p_{k|k-1}(X|Z_{1:k-1}^{[1]}, Z_{1:k-1}^{[2]})\delta X}. \tag{31}$$

Note that the integrals above are not standard integrals but *set integrals* from Finite Set Statistics [9] for integrating over set variables. The dynamical model is governed by the multi-object Markov transition density $f_{k|k-1}(X_k|X_{k-1})$ and multi-target likelihoods $g_k^{[i]}(Z_k^{[i]}|X_k)$ for $i = 1, 2$.

The set of objects tracked at time k is modelled by the point process or Random Finite Set (RFS)

$$X_k = \left(\bigcup_{x \in X_{k-1}} S_{k|k-1}(x) \right) \cup \left(\bigcup_{x \in X_{k-1}} B_{k|k-1}(x) \right) \cup \Gamma_k. \tag{32}$$

where $S_{k|k-1}$ is the RFS of targets survived at time t from multi-target state X_{k-1} at time $k-1$, $B_{k|k-1}$ is the RFS of targets spawned from X_{k-1} and Γ_k is the RFS of targets that appear spontaneously at time t. The multi-target measurement from each sensor at time t is modelled by RFS

$$Z_k^{[i]} = K_k^{[i]} \cup \left(\bigcup_{x \in X_k} \Theta_k^{[i]}(x) \right), \tag{33}$$

where $\Theta_k^{[i]}(X_k)$ is the RFS of measurements from multi-target state X_k and $K_k^{[i]}$ is the RFS of measurements due to clutter.

The multi-object Bayes filter is generally intractable due to the inherent complexity of the high-dimensional state space. Mahler proposed the Probability Hypothesis Density (PHD) filter [5] as an approximation to the multi-object Bayes filter. In the PHD filter, instead of propagating the full multi-object posterior, only the first moment is propagated. Under the assumption that the number of targets is distributed according to a Poisson distribution, this reduces the complexity of the approach to linear in the number of targets and the number of measurements.

Let v_k and $v_{k|k-1}$ denote the respective intensities associated with the multi-target posterior density p_k and the multi-target predicted density $p_{k|k-1}$ in the recursion (30)-(31). The prediction equation is given by

$$v_{k|k-1}(x) = \int p_{S,k}(\zeta) f_{k|k-1}(x|\zeta) v_{k-1}(\zeta) d\zeta + \gamma_k(x), \tag{34}$$

where

$$\begin{aligned} f_{k|k-1}(\cdot|\zeta) &= \text{single target transition density at time } k, \\ p_{S,k}(\zeta) &= \text{probability of target existence at time } k, \\ \gamma_k(\cdot) &= \text{intensity of spontaneous births at time } k, \end{aligned}$$

and the update equation is given by

$$\begin{aligned} v_k(x) &= [1 - p_{D,k}(x)] v_{k|k-1}(x) \\ &+ \sum_{z \in Z_k} \frac{p_{D,k}(x) g_k(z|x) v_{k|k-1}(x)}{\kappa_k(z) + \int p_{D,k}(\xi) g_k(z|\xi) v_{k|k-1}(\xi) d\xi} \end{aligned} \tag{35}$$

where

$$\begin{aligned} Z_k &= \text{measurement set at time } k, \\ g_k(\cdot|x) &= \text{single target measurement likelihood at time } k \\ p_{D,k}(x) &= \text{probability of target detection at time } k \\ \kappa_k(\cdot) &= \text{intensity of clutter measurements at time } k, \end{aligned}$$

A specific two-sensor PHD filter has been derived recently [10] directly from the two-sensor Bayes update in equation (31). The advantage of the two-sensor PHD filter over an iterated approach (where we update with each sensor in turn) is that the approximation of the multi-object prediction required to derive the first moment, such as a Poisson

point process, only needs to be made once. In the iterated form we use, the assumption that the predicted multi-object distribution is Poisson is made before each sensor update. However, the disadvantage of the two-sensor PHD filter is that we are required to consider all binary partitions of measurements, i.e., we have to find the corresponding measurements for the targets. Instead, we choose the iterated PHD corrector which has a linear complexity for each measurement set and completely avoids the correspondence problem.

The closed-form solution to the PHD filter under linear-Gaussian assumptions was derived by Vo [11] which enabled efficient practical tracking algorithms to be developed [12]. In the prediction and update stages we assume that the PHD is a Gaussian mixture and the resulting PHD also becomes a Gaussian mixture using the properties of Gaussians for the Kalman filter in Section 3.

5. Simulated Results

Here we demonstrate the approach on simulated data and provide an analysis of the results. The target motions are modelled on a constant velocity model in disparity space. Ideally, one would want to model the target motions in 3-D space, though, as we have shown, the nonlinear transformation between 3-D and disparity space means that Gaussianity and linearity are not preserved. Since depth is inversely proportional to disparity, it is inherent in the nature of the problem that we cannot model the same target dynamics from stereo vision as those normally used with a Kalman filter, though we shall leave consideration of realistic target motion models in disparity space for future work.

The target state variable $x_k = [u, \dot{u}, v, \dot{v}, d, \dot{d}]^T$ is comprised of the positions and velocities of u, v and d, where u and v are defined with respect to the first camera plane and d is the disparity. The state transition model is then

$$x_k = F_{k-1}x_{k-1} + Gw_{k-1}, \tag{36}$$

where

$$F_{k-1} = \begin{pmatrix} 1 & T & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & T & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & T \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, G = \begin{pmatrix} \frac{T^2}{2} & 0 & 0 \\ T & 0 & 0 \\ 0 & \frac{T^2}{2} & 0 \\ 0 & T & 0 \\ 0 & 0 & \frac{T^2}{2} \\ 0 & 0 & T \end{pmatrix}, \tag{37}$$

where $T = 1$, $w_{k-1} \sim \mathcal{N}\left(\cdot; 0, \mathrm{diag}([\sigma_{\mathrm{x}}^2, \sigma_{\mathrm{y}}^2, \sigma_{\mathrm{d}}^2])\right)$, $\sigma_x = 1.0$, $\sigma_y = 1.0$, $\sigma_d = 1.0$.

The sensor observations comprise of the projections of the target positions onto two camera planes. Thus

$$z_k^{[1]} = H_k^{[1]}x_k + \epsilon_k^{[1]}, \tag{38}$$

$$z_k^{[2]} = H_k^{[2]}x_k + \epsilon_k^{[2]}, \tag{39}$$

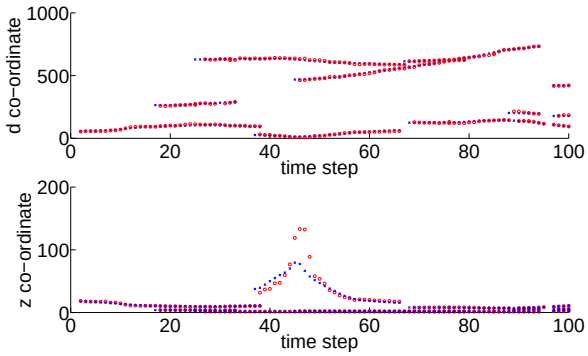

Figure 4. Estimates in disparity (d, upper graph) and depth (z, lower graph) shown side by side (estimates in red, true positions in blue).

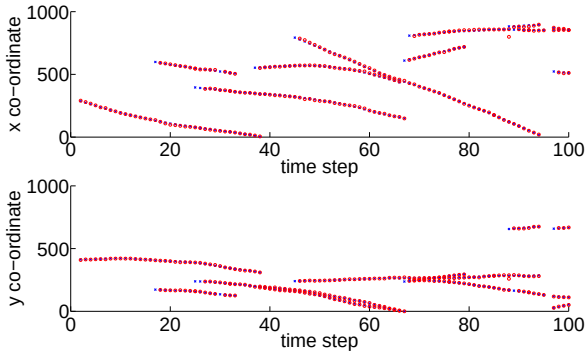

Figure 5. Estimates in u (upper graph) and v (lower graph) dimensions of the disparity space (estimates in red, true positions in blue).

where the projections from the disparity space to the camera planes are

$$H_k^{[1]} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \tag{40}$$

$$H_k^{[2]} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \tag{41}$$

the $P_D : \mathbb{P}^3(\mathbb{R}^3) \rightarrow \mathbb{P}^3(\mathbb{D}^3)$ is:

$$P_D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1000 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{42}$$

the corresponding camera matrices are:

$$P_k^{[1]} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{43}$$

$$P_k^{[2]} = \begin{pmatrix} 1 & 0 & 0 & 1000 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{44}$$

and $\epsilon_k^{[i]} \sim \mathcal{N}(\cdot; 0, R_k^{[i]})$, where $R_k^{[i]} = \mathrm{diag}([\epsilon_{\mathrm{x}}^2, \epsilon_{\mathrm{y}}^2]^{\mathrm{T}})$, $\epsilon_x = \epsilon_y = 5.0$.

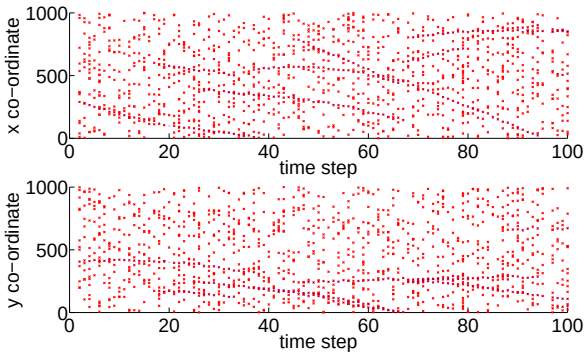

Figure 6. Measurements from sensor 1 over time (measurements in red, true positions in blue).

The birth process has intensity

$$\gamma_k(x) = \sum_{i=1}^{4} w_k^{(i)} \mathcal{N}((x; m_\gamma^{(i)}, P_\gamma)), \tag{45}$$

where $w_k^{(i)} = 0.1$, $m_\gamma^{(1)} = [\ 420, 0, 420, 0,\ 0\ ]^T$, $m_\gamma^{(2)} = [\ 440, 0, 440, 0,\ 2\ ]^T$, $m_\gamma^{(3)} = [\ 460, 0, 460, 0,\ 4\ ]^T$, $m_\gamma^{(4)} = [\ 480, 0, 480, 0,\ 6\ ]^T$, and

$$P_\gamma = \mathrm{diag}([400^2, 5^2, 400^2, 5^2, 400^2, 0.01^2]^{\mathrm{T}}). \tag{46}$$

The probabilities of target survival and detection are $p_{S,k} = 0.99$ and $p_{D,k} = 0.99$, respectively.

Figure 6 shows the measurements in the image planes over 100 time steps and the true positions of the targets projected onto these planes. Figure 5 shows the results of the tracking estimates in x and y, along with the true target positions, and Figure 4 shows the results of the tracking estimates in d and z. As expected, the performance is better in x and y since these are observed simply as a projection onto the image planes. It should be noted that d is not observed, so the fact that we can track in this dimension is a significant advantage of the approach. The results are mapped into z, where the uncertainty is no longer Gaussian distributed. Figure 4 shows that tracking near targets has better performance than far targets, which can be expected since the uncertainty grows with distance from the cameras.

6. Conclusions

This paper presents an optimal solution to the stereo-sensor 3-D tracking problem under linear-Gaussian assumptions by filtering in disparity space. We show that filtering in the 3-dimensional co-ordinate system is not possible with the usual linear-Gaussian assumptions due to the non-linearity of the projection from 3-D to the image planes and the fact that the noise is dependent on the distance of the object from the sensors. However, in disparity space the projection to the image planes is linear and the noise is isotropic and hence we can apply a two-sensor Kalman filter in disparity space and then optimally estimate the target state in 3-D. We apply this result to a multiple-object scenario by using an iterated two-sensor Gaussian mixture Probability Hypothesis Density filter. We demonstrate that it is possible to avoid both the data association problem in multi-object tracking and correspondence problem in computer vision and still be able to accurately determine the correct number of objects with stereo data sets that have false alarms and missed detections. The algorithm is linear in both the number of targets and number of measurements. We illustrate the results of the approach in simulated scenarios. Future work will involve generalising the approach to an arbitrary number of sensors for multi-camera fusion and evaluating the approach on real data. It is anticipated that the use of more sensors from different views will significantly reduce the uncertainty in the estimates.

Acknowledgements

Dr Clark is a Royal Academy of Engineering/ Engineering and Physical Sciences Research Council Fellow.

References

[1] D. Demirdjian and T. Darrell, "Using multiple-hypothesis disparity maps and image velocity for 3-d motion estimation," *International Journal of Computer Vision*, vol. 47, no. 1/2/3, pp. 219–228, 2002.

[2] M. Agrawal and K. Konolige, "Real-time localization in outdoor environments using stereo vision and inexpensive gps," in *International Conference on Pattern Recognition*, 2006, pp. 1063 – 1068.

[3] H. Hattori and N. Takeda, "Dense stereo matching in restricted disparity space," in *Intelligent Vehicles Symposium*, 2005, pp. 118 – 123.

[4] S. Ivekovic and E. Trucco, "Articulated 3-d modelling in a wide-baseline disparity space," in *Proceedings of the 4th European Conference on Visual Media Production*, 2007, pp. 1 – 10.

[5] R. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, No.4, pp. 1152–1178, 2003.

[6] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Machine Vision and Applications*, vol. 12, no. 1, pp. 16–22, 2000.

[7] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[8] Y. C. Ho and R. C. K. Lee, "A Bayesian approach to problems in stochastic estimation and control," *IEEE Trans. AC*, vol. AC-9, pp. 333–339, 1964.

[9] R. P. S. Mahler, "Statistical Multisource Multitarget Information Fusion," *Artech House*, 2007.

[10] R. Mahler, "The multisensor PHD filter: I. General solution via multitarget calculus," *Signal Processing, Sensor Fusion, and Target Recognition XVIII. Proc. SPIE, Volume 7336*, 2009.

[11] B. Vo and W. K. Ma, "The Gaussian Mixture Probability Hypothesis Density Filter," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[12] D. Clark, K. Panta, and B. Vo, "The GM-PHD Filter Multiple Target Tracker," *Proc. International Conference on Information Fusion. Florence.*, July 2006.