Spatio-Temporal Stereo Disparity Integration

Sandino Morales and Reinhard Klette

The *.enpeda..* Project, The University of Auckland
Tamaki Innovation Campus, Auckland, New Zealand
pmor085@aucklanduni.ac.nz

Abstract. Using image sequences as input for vision-based algorithms allows the possibility of merging information from previous images into the analysis of the current image. In the context of video-based driver assistance systems, such temporal analysis can lead to the improvement of depth estimation of visible objects. This paper presents a Kalman filter-based approach that focuses on the reduction of uncertainty in disparity maps of image sequences. For each pixel in the current disparity map, we incorporate disparity data from neighbourhoods of corresponding pixels in the immediate previous and the current image frame. Similar approaches have been considered before that also use disparity information from previous images, but without complementing the analysis with data from neighbouring pixels.

Keywords: disparity map, Kalman filter, temporal propagation, stereo analysis, driver assistance

1 Introduction

Disparity (depth) estimation obtained form stereo algorithms is commonly used to provide basic spatial data for complex vision-based applications [Zhihui 2008], such as in vision-based driver assistance [Franke et al. 2005]. Many applications require a very high accuracy of disparity values, often depending on the application context (e.g. accurate disparity discontinuities in driver assistance, and accurate disparities within object regions in 3D modelling).

The following three approaches have been followed to improve disparity estimation. First, the design of new or the improvement of existing strategies for stereo matchers (e.g. local, global, semi-global, or hierarchical). Second, improvements of cost functions (e.g., sum of absolute differences, census transform, or gradient-based cost functions) or smoothness constraints (e.g. Potts model, or truncated linear penalty) used within such a matching strategy. Third, some manipulation of input images (e.g. residuals with respect to some smoothing, or edge detection; see [Vaudrey and Klette 2009]) or some post-processing of obtained disparity maps (e.g. consistency checks or mean filters; see [Atzpadin et al. 2004]).

In this work we discuss post-processing of disparity maps using available temporal information in the context of vision-based *driver assistance systems*

(DAS). The input of such a system is a continuous stream of images, thus it is possible to use the information contained in the *temporal domain* by incorporating disparity values from previous frames into measured disparity value in the current frame. The intention is to reduce the influence of miscalculated disparities in the overall disparity map.

Information contained in the temporal domain has been used before. For example, in [Morales et al. 2009], an 'alpha-blending' of disparity values calculated at current and previous image frames lead to an improved performance for the currently measured values. This model did not yet consider the *ego-motion* induced by the *ego-vehicle* (i.e. the vehicle where the system is operating in; see Fig. 1, left), nor the independent motion of other objects in the scene. [Badino et al. 2007] merged previous and current information using an iconic (pixel-wise) Kalman-based approach [Matthies et al. 1989] and ego-motion information (yaw and speed rate). This approach was designed to improve disparity measurements in regions of input images where visible motion was only induced by ego-motion, and not by motion of other objects. The latter method was extended in [Vaudrey et al. 2008a] by adding a *disparity rate* term to the Kalman filter. The goal was to improve disparity measurements for objects that move relatively to the ego-vehicle in longitudinal direction.

Post processing of disparity maps can also aim at optimizing disparity values by considering disparity values within a neighbourhood of the current pixel (i.e. using data from the *spatial domain*). [Morales et al. 2009] used the same alpha-blending approach for modifying the disparity value at a pixel by using disparity values of the 'north' and 'south' neighbours. Again, no ego- or independent motion was taken into account. In this paper we improve disparity maps by reducing the uncertainty in calculated values using information from both, the spatial and the temporal domain. The Kalman filter used in [Badino et al. 2007,Vaudrey et al. 2008a] is now modified such that data from the spatial domain can also be used. Spatial information is used from the previous and the current disparity maps, and can be taken from an arbitrarily defined neighbourhood. (Actually, those should be not too large.) We perform experi-

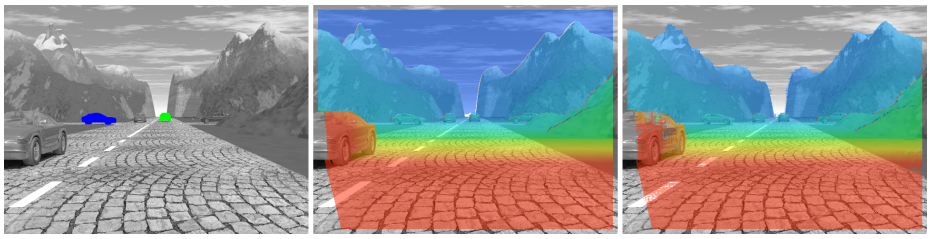

Fig. 1. Frame 47 from a rendered data set. *Left*: Segmented objects used for different approaches. The blue car is static w.r.t the ground while the green car moves away from the ego-vehicle. – Colour encoded (red close, blue far) ground truth (*Middle*) and a disparity map obtained with a semi-global matching algorithm (*Right*).

ments with a computer-generated sequence (i.e. with available ground truth) to compare results using both data domains separately for filtering, or both in combination. The rest of this paper is structured as follows. We start in Section 2 with recalling briefly the structure of Kalman filters. In Section 3 we describe the proposed approach. Section 4 reports and discusses results obtained in our experiments. Section 5 concludes.

2 Kalman Filter

We briefly recall the linear Kalman filter [Kalman 1960] as commonly used for a discrete dynamic system [Kuo and Golnaraghi 2002]. Given the system

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A} \cdot \mathbf{x}_{t-1} + \mathbf{B} \cdot \mathbf{u}_t + \mathbf{v} \\ \mathbf{y}_t &= \mathbf{B} \cdot \mathbf{x}_{t-1} + \mathbf{w} \end{aligned} \tag{1}$$

The Kalman filter is defined in two steps. First, the information from the previous step is incorporated into the filter by generating a *predicted* state

$$\begin{aligned} \mathbf{x}_{t|t-1} &= \mathbf{A} \cdot \mathbf{x}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{A} \cdot \mathbf{P}_{t-1|t-1} \cdot \mathbf{A}^T + \mathbf{Q} \end{aligned} \tag{2}$$

We use the notation $t|t-1$ to denote that we are in an intermediate step between t and $t-1$, while $t-1|t-1$ denotes the state obtained with the Kalman filter at time $t-1$. Matrix $\mathbf{Q}$ represents the process noise variance (obtained from vector $\mathbf{v}$) and $\mathbf{P}_{t|t-1}$, denotes the covariance matrix of the error of $\mathbf{x}_{t|t-1}$ compared to the true value $\mathbf{x}_{t|t}$.

In the second step, the predicted state is corrected using data from the current state (via the *measurement* vector $\mathbf{y}_t$) and the predicted matrix $\mathbf{P}_{t|t-1}$:

$$\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + \mathbf{K}_t \left(\mathbf{y}_t - \mathbf{H} \cdot \mathbf{x}_{t|t-1} \right) \tag{3}$$

where

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \cdot \mathbf{H}^T \left(\mathbf{H}_t \cdot \mathbf{P}_{t|t-1} \cdot \mathbf{H}^T + \mathbf{R} \right) \tag{4}$$

is the *Kalman gain* as derived in [Kalman 1960]. It follows that

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \cdot \mathbf{H}^T) \mathbf{P}_{t|t-1} \tag{5}$$

This is an iterative process. The initial state $x_{0|0}$ and the initial covariance matrix $P_{0|0}$ need to be selected.

3 Approach

The basic idea of our approach is to incorporate disparity information contained in the spatial domain, as reported in [Morales et al. 2009], into the Kalman filter approaches presented in [Badino et al. 2007,Vaudrey et al. 2008a]. These previous Kalman filter approaches were designed to handle different kinds of

moving objects. In [Badino et al. 2007] the interest was on the improvement of disparity values for static objects in the scene, while in [Vaudrey et al. 2008a] the model was about longitudinal movements with respect to ego-motion. We aim at enhancing both methods by adding data from the disparity spatial domain.

By means of a Kalman filter, we merge temporal and spatial information. For simplicity we use an *iconic* Kalman filter [Matthies et al. 1989] (i.e. we use an individual Kalman filter for each pixel under consideration). In each iteration we obtain a new disparity value for each pixel using both spatial and temporal information. For doing so it is necessary to consider the motion of 3D world points that define pixels in the disparity map. Each pixel in an initial disparity map will be followed as the image sequence advances in time (as long as still visible in the current frame). This is done by considering ego-motion and disparity rate.

Let p be a pixel of a disparity map calculated at time t, and $p_1, \ldots, p_n$ are adjacent pixels of it with $n \geq 1$. Our presentation of formulas is for $n = 2$ only; it generalizes for arbitrary neighbourhoods. The pixel p represents a 3D world point P projected into the image plane at time t. Let p also denote the projection of P at time $t+1$ besides knowing that its position on the image plane is actually different due to ego-motion or possible independent motion of P. The disparity value assigned to p is also expected to change through the sequence.

Consider the dynamic system defined at time t by the state vector $\mathbf{x}_t$ and the transition matrix $\mathbf{A}$, given by

$$\mathbf{x}_t = \begin{pmatrix} d_p \\ d_1 \\ d_2 \\ \dot{d} \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} \alpha & \beta & \gamma & \Delta t \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

where d_* denotes disparity values corresponding to p and two adjacent pixels p_1 and p_2; value $\dot{d}$ is the *disparity rate* as introduced in [Vaudrey et al. 2008a]. Values α and β control the interaction of disparity values of pixels p, p_1, and p_2. (In our experiments, we use $\alpha = 0{\cdot}8$ and $\beta = \gamma = 0{\cdot}1$). The parameter Δt represents the time elapsed between two consecutive frames. We assume that the noise vector $\mathbf{v}$ associated to the system (see Section 2) is Gaussian with zero mean and standard deviation σ_d for all disparity values and $\sigma_{\dot{d}}$ for disparity rate.

Measurement data are not available for the disparity rate. Measurement contains all the involved disparity values, thus we can also consider the spatial information in a frame at time t. Therefore, the dimension of the measurement vector $\mathbf{y}_t$ is $n + 1$, and matrix $\mathbf{H}$ equals

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{7}$$

The noise vector $\mathbf{w}$ associated to the measurements taken from the system is assumed to be the same for all of its coordinates: Gaussian with zero mean and with a standard deviation σ_v.

To start the filtering process we need to define the initial state and the initial covariance matrix. The initial state is defined using the disparity values of p and its neighbours calculated with a given stereo algorithm at time $t = 0$. The disparity rate is set to be zero. The covariance matrix is defined by

$$\mathbf{P}_{0|0} = \begin{pmatrix} \sigma_d^2 & \sigma_{dd_1} & \sigma_{dd_2} & \sigma_{d\dot{d}} \\ \sigma_{dd_1} & \sigma_{d_1^2} & \sigma_{d_1 d_2} & \sigma_{d_1 \dot{d}} \\ \sigma_{dd_2} & \sigma_{d_1 d_2} & \sigma_{d_2^2} & \sigma_{d_2 \dot{d}} \\ \sigma_{d\dot{d}} & \sigma_{d_1 \dot{d}} & \sigma_{d_2 \dot{d}} & \sigma_{\dot{d}^2} \end{pmatrix} \tag{8}$$

where, for example, $\sigma_{d\dot{d}} = \sigma_d \cdot \sigma_{\dot{d}}$. Recall that we assumed that all the calculated disparity values have the same variance.

Once the filter has been initialized, we can start the iteration process. Assume that we have already calculated $t-1$ steps and that $\mathbf{x}_{t-1|t-1}$ is available. After the prediction step at time t, the first coordinate of $\mathbf{x}_{t|t-1}$ contains information about a neighbourhood (a 2-neighbourhood in this case) of the disparity map calculated at time $t-1$.

Before applying the Kalman update process it is necessary to update the position of pixel p (i.e. the visible move of the 3D world point P between $t-1$ and t). This is done by calculating the relative motion of P with respect to the ego-vehicle, and this is done in tow steps.

First, the motion induced by the disparity rate is incorporated into P. This motion is away from the ego-vehicle and in the same direction. Thus, only the Z coordinate will be modified. Second, the positional change induced by ego-motion is considered. We use only speed and yaw rate (i.e. the bicycle model [Franke et al. 2005]) and assume that they are noise free. Roll and pitch are not included into our model, but both do have a minor influence for vision-augmented vehicles.

Once the position of p in frame t is known, we calculate the measurement vector $\mathbf{y}_t$ from the disparity map at time t. This vector is then used to generate the corresponding Kalman gain. Note that from the design of our system [see Equation (2)], disparity information from neighbouring pixels of p at time t is included in the measurement vector.

This measurement vector can now be used to calculate the Kalman gain in order to obtain the updated state, so the next iteration at time t can be performed. To avoid noisy states in both steps of the Kalman process, we follow the validation rules suggested in [Vaudrey et al. 2008a].

For adding temporal information to the approach of [Badino et al. 2007] it is necessary to remove the disparity rate term from the state and modify accordingly the rest of the dynamic system presented in Equation (6).

4 Experiments and Results

We perform experiments using the Sequence 1 [Vaudrey et al. 2008b] from the Set 2 from [EISATS 2011]. It is a computer generated sequence representing a driving scenario with available stereo ground truth; see Figure 1, middle. The

ego-vehicle drives straight through the whole sequence. Thus we assume a constant yaw rate of 0 degrees and a speed of 6·99 m/s, calculated from the ground truth and an assumed frame rate of 25 frames per second.

As stated in Section 3, the method proposed in [Vaudrey et al. 2008a] (*dynamic* method from now on) was designed to improve the calculated disparity values from objects moving away but in the same driving direction of the ego-vehicle. We segmented from the test sequence (100 frames) a vehicle whose movement fulfils such requirements. For the experiments using the approach introduced in [Badino et al. 2007] (now called the *static* method) we segmented a static vehicle (53 frames) with respect to the ground. In Figure 1 left, the green vehicle is moving away from the ego vehicle, while the blue one is totally static.

For generating the disparity data we use a semi-global matching (SGM) stereo algorithm [Hirschmüller 2005], with a four-path configuration and the census transform as cost function (see [Herman et al. 2011]). See Figure 1, right, for a sample disparity map. As expected with noise-free stereo pairs, the SGM results were 'very close' to the ground truth, letting almost no room for improvement. All the filtered results were slightly worse than the raw stereo results, as already reported in [Morales et al. 2009]. However, this previous study reported that using either spatial or temporal post-processing leads to improvements when using noisy sequences which resemble real-world data.

Experiments were made for neighbourhoods with $n = 1, 2, 4, 8$ adjacent pixels, both for the static and the dynamic method. For each frame, we calculate the mean of all the disparity values within the corresponding region of interest (i.e. the moving vehicle for the dynamic method and the static vehicle for the static method). We then compare with available ground truth.

The parameters to initialize the Kalman filter for the dynamic method are as follows (for $n = 2$ and analogously for the other cases): The initial state was

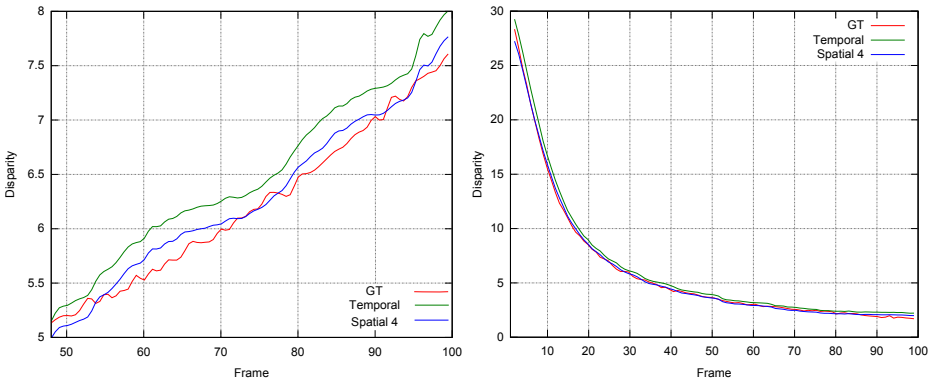

Fig. 2. Average disparity in the region of interest for the static (left) and the dynamic (right) method. Both plots show ground truth (GT), the results using only data from the temporal domain (Temporal), and when using the spatio-temporal approach (Spatial 4) with $n = 4$.

filled up with data from the disparity map calculated for the first available stereo pair; except for the disparity rate term, which was set to zero.

For the matrix $\mathbf{A}$, let $\alpha = 0{\cdot}8$, $\beta = \gamma = 0{\cdot}1$ and $\Delta t = 0{\cdot}04$ (i.e. an assumed frame rate of 25 frames per second).

The entries of the covariance matrix $\mathbf{P}_{0|0}$ were initialized with the following values: $\sigma^2_{d*} = 0{\cdot}3$, assuming an imperfect disparity map. $\sigma_{d*d*} = 0{\cdot}5$, a relative large value to represent large correlation between the pixel under analysis and its neighbours. $\sigma_{d*\dot{d}} = 0{\cdot}0001$, to show a low correlation between the disparity values and the disparity rate. $d*$ represent either d, d_1, or d_2. Finally, $\sigma_{\dot{d}^2} = 1$, a relative large value to express high uncertainty in the initial disparity rate. - The parameter initialization for the static method is analogous. It is only necessary to remove the terms where the disparity rate is involved and modify the matrices accordingly.

The results for the dynamic method show an improvement when using data from the spatio-temporal domain (for $n = 1$, 2, or 4) compared to when just using the temporal domain. See Figure 2. The results improve according to the size of the neighbourhood, being the best for $n = 4$. For $n = 8$, the results were not satisfactory. As expected, a large neighbourhood degrades the final disparity value. The results are summarized in Table 1, presenting the average deviation from the ground truth for the whole sequence and for all the considered settings.

For the static method we obtained similar results. The spatio-temporal approach (for $n = 1$, 2, or 4) shows a better performance than the temporal method. See Figure 2. The best performance was archived when using $n = 4$, and the worst was measured for $n = 8$. However, in this case, for $n = 8$ the results were still better than when just using the temporal approach. Interestingly, the average deviation for the whole sequence for SGM and $n = 4$ is almost the same. See Table 1. For some frames, the average disparity value was closer to the ground truth when using the spatio-temporal approach than with the original SGM algorithm.

5 Conclusions

In this paper we present a method for post-processing disparity maps using a spatio-temporal approach. We use an iconic Kalman filter approach for merging data from both domains.

	SGM	Temporal	$n = 1$	$n = 2$	$n = 4$	$n = 8$
Mobile	0·11	0·41	0·18	0·16	0·15	3·14
Static	0·10	0·28	0·12	0·12	0·10	0·22

Table 1. Average deviation from the ground truth for each one of the methods. SGM stands for the raw stereo results. Temporal is for the exclusively temporal approach. The rest are for the spatio-temporal methods, where n indicates the size of the neighbourhood

Obtained experimental results (this short paper only discusses one sequence given with ground truth) showed improvements compared to original or either only temporal or only spatial post-processing. We suggest to use the combined approach for filtering out noisy values in disparity maps generated for stereo sequences recorded in the real-world.

Future work will quantify improvements on real-world sequences using the prediction-error approach as discussed, for example, in [Morales et al. 2009].

References

[Atzpadin et al. 2004] Atzpadin, N., Kauff, P., and Schreer, O.: Stereo analysis by hybrid recursive matching for real-time immersive video stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *IEEE Trans. Circuits Systems Video Techn.*, **14**:321–334 (2004)

[Badino et al. 2007] Badino, H., Franke, U., and Mester, R.: Free space computation using stochastic occupancy grids and dynamic programming. In Proc. *Dynamic Vision Workshop for ICCV* (2007)

[EISATS 2011] *.enpeda..* Image Sequence Analysis Test Site (EISATS). [Online]. Available: `http://www.mi.auckland.ac.nz/EISATS/` (2011)

[Franke et al. 2005] Franke, U., Rabe, C., Badino, H., and Gehrig, S.: 6D-vision: Fusion of stereo and motion for robust environment perception. In Proc. *DAGM*, LNCS 3663, pages 216–223 (2005)

[Herman et al. 2011] Hermann, S., Morales, S., and Klette, R.: Half-resolution semi-global stereo matching. In Proc. *IEEE Intelligent Vehicles Symp.* (2011)

[Hirschmüller 2005] Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In Proc. *Computer Vision Pattern Recognition*, volume 2, pages 807–814 (2005).

[Kalman 1960] Kalman, R.E.: A new approach to linear filtering and prediction problems. *J. Basic Engineering*, **82**:35–45 (1960)

[Kuo and Golnaraghi 2002] Kuo, B., and Golnaraghi, F.: *Automatic Control Systems.* John Wiley and Sons Inc., New York (2002)

[Matthies et al. 1989] Matthies, L., Kanade, T., and Szeliski, R.: Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. Computer Vision*, **3**:209–238 (1989)

[Morales et al. 2009] Morales, S., Vaudrey, T., and Klette, R.: Robustness evaluation of stereo algorithms on long stereo sequences. In Proc. *IEEE Intelligent Vehicles Symposium*, pages 347–352 (2009)

[Vaudrey et al. 2008a] Vaudrey, T., Badino, H., and Gehrig, S.: Integrating disparity images by incorporating disparity rate. In Proc. *Robot Vision*, LNCS 4931, pages 29–42 (2008)

[Vaudrey et al. 2008b] Vaudrey, T., Rabe, C., Klette R., and Milburn J.: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In Proc. *IVCNZ*, pages 1–6 (2008)

[Vaudrey and Klette 2009] Vaudrey, T. and Klette, R.: Residual images remove illumination artefacts for correspondence Algorithms! In Proc. *DAGM*, LNCS 5748, pages 472–481 (2009)

[Zhihui 2008] Zhihui, X.: *Computer Vision.* InTech, online books (2008)