VARIATIONAL BAYESIAN INFERENCE FOR STEREO OBJECT TRACKING

Giannis Chantas, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki, GR 54124, Greece, e-mail:{nikolaid,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper, we deal with object tracking in stereo video sequences. We introduce a Bayesian framework for utilizing the results of any conventional single channel object tracker, in order to accomplish the refinement of the tracking accuracy in the left/right video channel. In this Bayesian framework, a variational Bayesian algorithm is employed to this end, where a priori information about the object displacement (movement) over time is incorporated by means of a prior distribution. This a priori information is obtained in a preprocessing step, in which the object displacement over time is estimated. Experiments demonstrate the efficiency of the proposed post-processing methodology in terms of tracking accuracy.

Index Terms— Stereo Tracking, Variational Inference, Student's-t

1. INTRODUCTION

Object tracking is an important problem in semantic video analysis, surveillance, etc. [1], [2]. In the following, we use the term object to describe any entity to be tracked, including faces or human bodies. In many works, tracking is formulated in a stochastic Bayesian framework [3]. In this work, a Bayesian post-processing methodology is introduced, which provides accurate object localization in stereo videos, by combining the tracking results of a single channel tracking algorithm, when applied independently on both channels of a stereo video.

The combination of multiple tracking information has also been proposed in [4], where a Monte Carlo stochastic sampling algorithm is employed, in order to combine the multiple tracker results into a single object localization. A similar work is presented in [5], where multiple trackers are combined, by exploiting only the probability density function of the new target position of each tracker, in order to yield one object location estimate on each video frame. The major difference between these two works and the proposed framework is that the tracking information, which is provided to the proposed post-processing algorithm, comes from the left and right stereo video channels, i.e., the proposed algorithm is tuned for object tracking in stereo videos. This work is similar in spirit with [6], but with two main differences: the use of object displacement and disparity information, extracted before the Bayesian inference, algorithm, and the use of the Student's-t distribution to model the new information. The displacement and disparity information is extracted based on a SIFT feature matching technique [7]. Thus, in this way, valuable luminance information is incorporated to the post-processing problem.

In more detail, in this work, we aim to exploit, in a Bayesian post-processing framework, the tracking results of a single-channel tracker, applied on both left and right stereo video channels, independently. In addition, in this framework, information about object displacement over time information, as well as disparity between object appearances in left and right video frames, is efficiently exploited. An abstract overview of the proposed post-processing methodology is illustrated in Figure 1. Object displacement and disparity information is obtained prior to post-processing, by using the initial tracking results.

In this framework, two distributions are defined. The first distribution is a stochastic observation model that treats the coordinates pinpointing the region of interest (ROI) of the tracked object as random variables. More specifically, it is assumed that the ROIs resulting from single left/right channel tracking are noisy observations of the ideal ROI coordinates, where the noise follows the Students'-t distribution [8].

The object displacement information is incorporated in the Bayesian framework by a prior distribution, which models the difference between consecutive ideal ROI coordinates. We adopt a Students'-t distribution, which models the accuracy of the object displacement estimated values.

The reasoning for adopting the Students'-t distribution in both of the above mentioned stochastic models is that it is flexible enough to adapt to the temporarily varying statistical properties of the data, in contrast to the Gaussian distribution [8], [9], [10]. The key feature of the Student's-t distribution is that the variance of the modeled variables is assumed to vary

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^o 287674 (3DTVS). The publication reflects only the authors' views. The EU is not liable for any use that may be made of the information contained herein. The authors would also like to thank the Fraunhofer Heinrich Hertz Institute for providing the stereo videos used in the experimental section. The videos belong to the project MUSCADE.

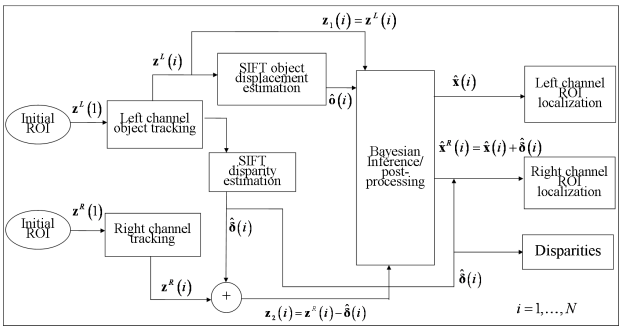

Fig. 1: Abstract overall diagram of the proposed methodology.

over time (temporal variability). Thus, this model can adapt to the temporally varying tracker and displacement estimation accuracy.

Based on these two models, a variational Bayesian approximate inference algorithm [8], [11] (since exact inference is intractable) that provides more accurate estimates of the ideal object ROI coordinates is employed.

The rest of the paper is organized as follows. The observation and prior models are described in Section 2 and 3, respectively. The variational Bayesian inference algorithm is derived in Section 4. In Section 5, experiments are provided that demonstrate the efficiency of the proposed post-processing algorithm. Finally, Section 6 concludes the paper.

2. OBSERVATION MODEL

Since we deal with stereo sequences, the observation model concerns the left and right channels of a stereo video. The goal of the proposed methodology is to estimate the ideal (unknown) positions of the object ROIs in each video frame. To this end, we model as random variables the ideal object ROIs in the left video channel. Hence, their inference is an explicit estimate of them. Regarding the right video channel ROIs, they are estimated using the inferred left-channel ROIs and the disparities. A ROI in the i-th left channel frame, where the total number of frames is N for each channel, is assumed to be a rectangle, defined uniquely by the upper-left and lower-right vertex coordinates $[x_1(i), x_2(i)]^T$ and $[x_3(i), x_4(i)]^T$, respectively. This definition holds for every other type of ROI, mentioned in the rest of the paper. We denote by $\mathbf{x}(i) = [x_1(i), x_2(i), x_3(i), x_4(i)]^T$ the i-th ROI ideal coordinates and by $\mathbf{x} = [\mathbf{x}(1)^T, \mathbf{x}(2)^T, \ldots, \mathbf{x}(N)^T]^T$ the vector that contains the coordinates of all ideal object ROIs to be estimated over the entire video. Moreover, we assume that the right channel coordinates $\mathbf{x}^R(i)$ can be calculated by $\mathbf{x}(i)$, according to $\mathbf{x}^R(i) = \mathbf{x}(i) + \boldsymbol{\delta}(i)$, where $\mathbf{x}^R(i)$ and $\boldsymbol{\delta}(i)$, which is the dispariry vector, are four element vectors. These coordinates are assumed to be real numbers for optimization convenience, as seen in Section 4. Note that $\boldsymbol{\delta}(i)$ are estimated in a pre-processing step.

Let us track an object by initializing a single-view object tracker on the first frame of the left and right video channels, independently, and running it in both channels. We denote by $\mathbf{z}_1(i), \mathbf{z}_2(i)$ the extracted ROI coordinates obtained by this procedure in the left and right video channel, respectively. They provide two observations of $\mathbf{x}(i)$, namely $\mathbf{z}_1(i) = \mathbf{z}^L(i)$, $\mathbf{z}_2(i) = \mathbf{z}^R(i) - \hat{\boldsymbol{\delta}}(i)$, where

$$\mathbf{z}_k(i) = [z_{k,1}(i), z_{k,2}(i), z_{k,3}(i), z_{k,4}(i)]^T, \qquad (1)$$

for $k = 1, 2,\ i = 1, \ldots, N$. We denote by

$$\mathbf{z} = [\mathbf{z}_1(1), \ldots, \mathbf{z}_1(N), \mathbf{z}_2(1), \ldots, \mathbf{z}_2(N)]$$

the vector of all extracted ROI coordinates.

To model the observed ROI generation procedure mentioned above, we assume that the extracted ROIs are noisy measurements of the ideal ROIs described by $\mathbf{x}$. Precisely, we assume that $p(\mathbf{z}|\mathbf{h})$ is given by:

$$p(\mathbf{z}|\mathbf{h}) \propto \prod_{i,k} \exp\left(-\frac{\lambda_{\mathbf{b}} d_k(i) b_k(i)}{2} \|\mathbf{z}_k(i) - \mathbf{x}(i)\|_2^2\right), \quad (2)$$

where $\mathbf{h} = \{\mathbf{x}, \mathbf{b}, \mathbf{d}, \mathbf{u}\}$ are the model hidden variables. $\mathbf{d} = \{d_k(i) : \forall i, k\}$ are binary random variables, explained later in this section. Also, $\mathbf{u}$ are random variables introduced and explained in detail in Section 3. By $\mathbf{b}$ we denote the set of all the inverse variances $b_k(i)$ of $\mathbf{z}_k(i)$, appearing in (2):

$$\mathbf{b} = \{b_k(i) : \forall i, \forall k\}.$$

A Gamma prior distribution [8] is imposed on each $b_k(i)$:

$$p(b_k(i)) \propto b_k(i)^{\nu_{\mathbf{b}}/2-1} \exp(-\nu_{\mathbf{b}} b_k(i)/2), \qquad (3)$$

where $\nu_{\mathbf{b}}$ is a parameter with positive value. We have to note that if we integrate out the $\mathbf{b}$ variables in the joint probability distribution $p(\mathbf{z}, \mathbf{b}|\mathbf{x}, \mathbf{d})$, based on (3) and (2), we take a Student's-t distribution [8].

This model is used with the purpose to moderate the influence of ROIs coming from highly inaccurate tracking results (e.g., object localization failures due to occlusion or fast movement). Indeed, a very small value of $b_k(i)$ moderates the influence of $\mathbf{z}_k(i)$ ROI on the estimate $\hat{\mathbf{x}}(i)$, since its variance is very large.

The binary variables $d_k(i)$ in (2) take values $d_k(i) = 0$ or 1, with $\sum_{k=1}^{2} d_k(i) = 1$. These variables indicate which ROI between $\mathbf{z}_1(i)$ and $\mathbf{z}_2(i)$ is the most or least noisy observation of the ideal ROI $\mathbf{x}(i)$. For $\mathbf{d}$, we assume a multinomial prior, with parameters $\pi_k = \frac{1}{2},\ k = 1, 2$.

3. PRIOR MODEL

As a prior model of $\mathbf{x}$, we adopt a Student's-t distribution, which implies a two-level generation process of the data we

want to model. The conditional prior distribution that generates $\mathbf{x}$ is a Gaussian distribution, given by:

$$p(\mathbf{x}|\mathbf{u}) \propto \prod_{i=2}^{N} \exp(-\frac{\lambda_{\mathbf{x}}}{2} u(i) \|\mathbf{x}(i) - \mathbf{x}(i-1) - \mathbf{o}(i)\|_2^2), \quad (4)$$

where $\mathbf{o}(i), i = 2, \ldots, N$, play the role of the mean (expected value) of the temporal object displacement between two consecutive ROI coordinates. These are computed after the initial single-channel tracking and before the Bayesian inference, using SIFT feature extraction and matching, as shown in Figure 1. The variables $\mathbf{u} = [u(1), \ldots, u(N)]^T$ are generated independently by a Gamma distribution:

$$p(u(i)) = \text{Gamma}(u(i); \nu_{\mathbf{x}}/2, \nu_{\mathbf{x}}/2), \; i = 2, \ldots, N, \quad (5)$$

except $u(1)$, which is assumed always to be zero. Notice that, if we integrate out $\mathbf{u}$ from the joint probability $p(\mathbf{x}, \mathbf{u})$, we take a product of multivariate Student's-t distributions [8].

In (4), the displacements $\mathbf{o}(i)$ play the role of the expected value of the difference $\mathbf{x}(i) - \mathbf{x}(i-1)$. To analyze this, we first note that $\mathbf{o}(i) = [o_1(i), o_2(i), o_1(i), o_2(i)]^T$ is a 4×1 vector. The values of its elements are (ideally) the differences:

$$\mathbf{o}(i) = \mathbf{x}(i) - \mathbf{x}(i-1), \; i = 2, \ldots, N. \quad (6)$$

We have assumed for simplicity that the ROI corresponding to $\mathbf{x}(i)$ in the i-th video frame is of the same size with that in the $(i-1)$-th video frame. It must be noted, that this constraint does not hold for the estimation of $\mathbf{x}$, denoted by $\hat{\mathbf{x}}$, as explained in the next section.

4. VARIATIONAL BAYESIAN INFERENCE

The Bayesian paradigm dictates that we should estimate the model variables $\mathbf{h}$ by taking their expectation with respect to their posterior, which can be obtained by the Bayes rule. However, in our case, as in most models of interest, finding this expectation is intractable.

Thus, the variational Bayesian methodology can be employed, in order to obtain an approximate posterior for the hidden variables $\mathbf{h}$, which provide tractable computations, [8]. Following this methodology, using the always positive Kullback-Leibler (KL) divergence between $q(\mathbf{h})$ and $p(\mathbf{h}|\mathbf{z})$ [8], we define the upper bound of the log-likelihood:

$$L(q(\mathbf{h}), \theta) = \log p(\mathbf{z}; \theta) - KL(q||p) \geq \log p(\mathbf{z}; \theta), \quad (7)$$

Estimation of q, which plays the role of the inferred posterior, and θ is performed by iteratively minimizing the bound with respect to q and θ. For q, minimization of the bound is achieved when $q(\mathbf{h}) = p(\mathbf{h}|\mathbf{z})$. However, the expectation of $\mathbf{h}$ w.r.t $p(\mathbf{h}|\mathbf{z})$ and minimization of L w.r.t θ is intractable. Thus, we adopt the *mean-field* approximation, in order to perform approximate Bayesian inference, which is a common practice in the variational framework [8], [11]. Specifically:

$$q(\mathbf{h}) = \prod_{l=1}^{4} q(\mathbf{h}_l). \quad (8)$$

where $\mathbf{h}_1 = \mathbf{x}$, $\mathbf{h}_2 = \mathbf{u}$, $\mathbf{h}_3 = \mathbf{b}$ and $\mathbf{h}_4 = \mathbf{d}$. In other words, $\mathbf{x}$, $\mathbf{u}$, $\mathbf{d}$ and $\mathbf{b}$ are assumed to be independent in the inferred posterior. Then, iterative inference algorithm consists of two steps at each iteration, given by the following equations:

$$q^{(t)}(\mathbf{h}_l) = \underset{q(\mathbf{h}_l)}{\text{argmin}} \, L\left(q^{(t-1)}(\mathbf{h}_l), q(\mathbf{h}_{\setminus l}), \theta^{(t-1)}\right), \; \forall l, \quad (9)$$

$$\theta^{(t)} = \underset{\theta}{\text{argmin}} \, L\left(q^{(t)}(\mathbf{h}), \theta\right), \quad (10)$$

where t is the iteration step and $\mathbf{h}_{\setminus l}$ denotes the set $\mathbf{h} - \mathbf{h}_l$. Due to the use of the mean-field approximation, equation (9) results in the update equations:

$$q^{(t)}(\mathbf{h}_l) = \frac{\exp\left(\log \langle p(\mathbf{z}, \mathbf{h}) \rangle_{q^{(t-1)}(\mathbf{h}_{\setminus l})}\right)}{\int_{\mathbf{h}_l} \exp\left(\log \langle p(\mathbf{z}, \mathbf{h}) \rangle_{q^{(t-1)}(\mathbf{h}_{\setminus l})}\right) d\mathbf{h}_l}. \quad (11)$$

The notation $\langle . \rangle_{q(.)}$ is used to denote the expectation with respect to the q distribution. Also, in what follows, we denote by $[\mathbf{A}]_{(i,i)}$ the i-th diagonal element of a matrix $\mathbf{A}$

Next, we derive the variational Bayesian algorithm. To achieve the minimization described in (9), we must perform the updates described in (11). The derivation of the update of $q^{(t)}(\mathbf{x})$:

$$q^{(t)}(\mathbf{x}) = q^{(t)}(\mathbf{x}_1) q^{(t)}(\mathbf{x}_2) q^{(t)}(\mathbf{x}_3) q^{(t)}(\mathbf{x}_4). \quad (12)$$

$\mathbf{x}_c, \; c = 1, 2, 3, 4$, are $N \times 1$ vectors and subsets of $\mathbf{x}$ that contain respectively the coordinates $x_1(i)$, $x_2(i)$, $x_3(i)$ and $x_4(i)$, for all i, as defined in Section 2. The update of each of their posteriors is given by:

$$q^{(t)}(\mathbf{x}_c) = N\left(\boldsymbol{\mu}_c^{(t)}, \mathbf{C}_{\mathbf{x}}^{(t)}\right), \; c = 1, 2, 3, 4, \quad (13)$$

where

$$\boldsymbol{\mu}_c^{(t)} = \mathbf{C}_{\mathbf{x}}^{(t)} \mathbf{y}_c, \left(\mathbf{C}_{\mathbf{x}}^{(t)}\right)^{-1} = \mathbf{B}^{(t)} + \lambda_{\mathbf{x}}^{(t-1)} \mathbf{Q}^T \mathbf{U}^{(t)} \mathbf{Q}, \quad (14)$$

and $\mathbf{B}^{(t)}$ and $\mathbf{U}^{(t)}$ are diagonal matrices with elements:

$$[\mathbf{B}^{(t)}]_{(i,i)} = \lambda_{\mathbf{b}}^{(t-1)} \sum_{k=1,2} \hat{d}_k(i) \hat{b}_k(i), [\mathbf{U}^{(t)}]_{(i,i)} = \hat{u}(i). \quad (15)$$

where $\hat{d}_k(i) \equiv \langle d_k(i) \rangle_{q(\mathbf{d})}$. $\hat{u}(i), \hat{b}_k(i)$ will be defined next. Also, $\mathbf{y}_c, \; c = 1, 2, 3, 4$ are $N \times 1$ vectors with elements:

$$\mathbf{y}_c(i) = \lambda_{\mathbf{b}} \sum_{k=1,2} \hat{d}_k(i) \hat{b}_k(i) z_{k,c}(i) + \lambda_{\mathbf{x}} [\mathbf{Q}^T \mathbf{U}^{(t)} \mathbf{o}_c](i), \quad (16)$$

where, $[v](i)$ denotes the i-th element of a vector v. Also, $\mathbf{o}_c$ is a vector containing all $o_c(i)$, for $i = 1, \ldots, N$.

Lastly, $\mathbf{Q}$ is the $N \times N$ first order difference operator.

The posterior for each $u(i)$, according to (11) is:

$$q^{(t)}(u(i)) = \text{Gamma}\left(u(i); \alpha_{\mathbf{u}}(i), \beta_{\mathbf{u}}(i)\right). \quad (17)$$

where $\beta_{\mathbf{u}} = 0.5(\lambda_{\mathbf{x}}^{(t)} \|\boldsymbol{\mu}^{(t)}(i) - \boldsymbol{\mu}^{(t)}(i-1) - \mathbf{o}(i)\|_2^2 + \nu_{\mathbf{x}})$, $\alpha_{\mathbf{u}} = \nu_{\mathbf{x}}/2 + 1/2$. Moreover,

$$\boldsymbol{\mu}^{(t)}(i) = [\mu_1^{(t)}(i), \mu_2^{(t)}(i), \mu_3^{(t)}(i), \mu_4^{(t)}(i)],\ \forall k, i.$$

Thus:

$$\hat{u}(i) \equiv \langle u(i) \rangle_{q^{(t)}(u(i))} = \frac{\alpha_{\mathbf{u}}(i)}{\beta_{\mathbf{u}}(i)},\ \forall k, i. \quad (18)$$

The same holds for every $b_k(i)$:

$$q^{(t)}\left(b_k(i)\right) = \text{Gamma}\left(b_k(i); \alpha_{\mathbf{b}}(i), \beta_{\mathbf{b}}(i)\right),\ \forall k, i, \quad (19)$$

where $\beta_{\mathbf{b}}(i) = 0.5\nu_{\mathbf{b}} + 0.5\hat{d}_k(i)\lambda_{\mathbf{b}}^{(t-1)} \|\mathbf{z}_k(i) - \boldsymbol{\mu}^{(t)}(i)\|_2^2$, $\alpha_{\mathbf{b}}(i) = \nu_{\mathbf{b}}/2 + \hat{d}_k(i)/2$. Thus:

$$\hat{b}_k(i) \equiv \langle b_k(i) \rangle_{q^{(t)}(\mathbf{b})} = \frac{\alpha_{\mathbf{b}}(i)}{\beta_{\mathbf{b}}(i)},\ \forall k, i. \quad (20)$$

$\hat{d}_k(i)$ are not estimated in this framework. Although $d_k(i)$ are binary, their expected value is in the range $[0, 1]$. We simply set:

$$\hat{d}_k(i) = \pi_k = \frac{1}{2},\ \forall k, i. \quad (21)$$

Finally, the parameters $\lambda_{\mathbf{x}}$ and $\lambda_{\mathbf{b}}$ updates are found by maximizing the bound L, see (10):

$$\lambda_{\mathbf{x}}^{(t)} = \frac{N(\hat{u}(i))^{-1}}{\sum_{i=2}^{N}([\mathbf{Q}\mathbf{C}_{\mathbf{x}}^{(t)}\mathbf{Q}^T]_{(i,i)} + \frac{\|\boldsymbol{\mu}^{(t)}(i) - \boldsymbol{\mu}^{(t)}(i-1) - \mathbf{o}(i)\|_2^2}{4})},$$

$$\lambda_{\mathbf{b}}^{(t)} = \frac{N(\hat{d}_k(i)\hat{b}_k(i))^{-1}}{\sum_{i=1}^{N}\sum_{k=1}^{2}(\frac{\|\mathbf{z}_k(i) - \boldsymbol{\mu}^{(t)}(i)\|_2^2}{4} + [\mathbf{C}_{\mathbf{x}}^{(t)}]_{(i,i)})}.$$

After convergence of the above iterative scheme, we obtain the estimates of the ideal ROI coordinates $\hat{\mathbf{x}}(i) = \boldsymbol{\mu}^{(t)}(i), \forall i$, for a large number of iterations t.

5. SIMULATION EXPERIMENTS

We evaluated the proposed Bayesian post-processing tracking algorithm on two stereo sequences. To obtain initial tracking results, the tracker [12], denoted by SC, was used to track objects (faces in both videos and a hand in one) in the left and right channel of the stereo sequences independently. No object/face detection was performed; instead the tracking algorithm was initialized by a user selected ROI in both of the first video frames of the left/right channels. Using SC tracking results, we employ the SIFT feature extraction and matching technique, in order to estimate the ROI coordinate displacements $\mathbf{o}$ and disparities $\boldsymbol{\delta}$. $\hat{\boldsymbol{\delta}}$ denotes this estimate of $\boldsymbol{\delta}$.

The post-processing output contains the estimates of the left channel ROI coordinates $\hat{\mathbf{x}}$. In addition, we take the right channel ROI coordinates by:

$$\hat{\mathbf{x}}^R(i) = \hat{\mathbf{x}} + \hat{\boldsymbol{\delta}}(i),\ i = 1, \ldots, N. \quad (22)$$

In what follows, the proposed post-processing algorithm that provides the estimates $\hat{\mathbf{x}}^R$ and $\hat{\mathbf{x}}$ is called SBP (Stereo Bayesian Post-processing).

The Average Tracking Accuracy (ATA) [13] metric, denoted by $\hat{a}$, was used to measure tracking accuracy:

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i \bigcap G_i|}{|D_i \bigcup G_i|}, \quad (23)$$

where D_i are the estimated ROI region, while G_i is the ideal (ground truth) ROI region obtained by manual video annotation, for $i = 1, \ldots, N$. D_i corresponds to the area determined by the estimated ROI coordinates $\hat{\mathbf{x}}(i)$ or $\hat{\mathbf{x}}^R(i)$ for the left and right channel, respectively. $|D|$ denotes the pixel number of a ROI D.

In order to demonstrate the performance of the proposed algorithm, we show the accuracy of tracking results in terms of the ATA metric in Table 1 for the SC and SBP algorithms. The results demonstrate that the SBP algorithm provides higher tracking accuracy than SC (in most cases).

The parameters $\nu_{\mathbf{x}}$ and $\nu_{\mathbf{b}}$ are fixed to a predetermined value that provides the best tracking accuracy in terms of ATA (found by trial-and-error).

Table 1: Tracking performance (ATA) in stereoscopic sequences. Tracked object is a head, unless stated otherwise.

Video/Channel	N	SC	SBP
Musicians, left	930	0.673	**0.756**
Musicians, right	930	0.632	**0.724**
Poker, (hand), left	499	0.461	**0.513**
Poker, (hand), right	499	**0.552**	0.538
Poker, left	599	0.692	**0.726**
Poker, right	599	0.699	**0.747**

6. CONCLUSIONS AND FUTURE WORK

We presented an object tracking Bayesian post-processing methodology for stereo sequences, which refines the outputs of standard tracking algorithms, by exploiting, the left and right channel tracking results. Also, object displacement over time, as well as disparity information, was exploited successfully to this end. The refined tracking results are significantly better than those provided by the initial tracking algorithm. In future, we plan to extend the algorithm in order to combine the results of multiple independent trackers. Also, alternative prior distributions will be considered.

7. REFERENCES

[1] N. Nikolaidis, M. Krinidis, E. Loutas, G. Stamou, and I. Pitas, *The Essential Guide to Video Processing*, 2nd ed. Al Bovik, Elsevier, 2009.

[2] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, Dec. 2006.

[3] A. Dore, M. Soto, and C. Regazzoni, "Bayesian tracking for video analytics," *Signal Processing Magazine, IEEE*, vol. 27, no. 5, pp. 46 –55, sept. 2010.

[4] J. Kwon and M. Lee, K., "Tracking by sampling trackers," in *International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1195–1202.

[5] I. Leichter, M. Lindenbaum, and E. Rivlin, "A general framework for combining visual trackers — the "black boxes" approach," *International Journal of Computer Vision*, vol. 67, no. 3, pp. 343–363, May 2006.

[6] G. Chantas, N. Nikolaidis, and I. Pitas, "Variational bayesian inference for forward backward visual tracking in stereo sequences," in *IEEE International Conference on Image Processing*, Orlando, Florida, USA, 30 September - 3 October 2012.

[7] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[9] J. Christmas and R. Everson, "Robust Autoregression: Student-t Innovations Using Variational Bayes," *IEEE Transactions on Signal Processing*, vol. 59, pp. 48–57, 2011.

[10] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders, "Variational bayesian image restoration based on a product of t-distributions image prior," *IEEE Transactions on Image Processing*, pp. 1795–1805, 2008.

[11] M. Beal, "Variational algorithms for approximate bayesian inference," *PhD. Thesis, Gatsby Computational Neuroscience Unit*, 2003.

[12] O. Zoidi, A. Tefas, and I. Pitas, "Visual object tracking based on the object's salient features with application in automatic nutrition assistance," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 25-30 March 2012.

[13] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319 –336, feb. 2009.