A BAYESIAN METHODOLOGY FOR VISUAL OBJECT TRACKING ON STEREO SEQUENCES

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

A general Bayesian post-processing methodology for performance improvement of object tracking in stereo video sequences is proposed in this paper. We utilize the results of any single channel visual object tracker in a Bayesian framework, in order to refine the tracking accuracy in both stereo video channels. In this framework, a variational Bayesian algorithm is employed, where prior knowledge about the object displacement (movement) is incorporated via a prior distribution. This displacement information is obtained in a preprocessing step, where object displacement is estimated via feature extraction and matching. In parallel, disparity information is extracted and utilized in the same framework. The improvements introduced by the proposed methodology in terms of tracking accuracy are quantified through experimental analysis.

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

1. INTRODUCTION

Efficient visual object tracking is very useful in semantic video analysis, human-computer interaction, surveillance, etc. [1]. In this paper, we use the term "object" to refer to any entity to be tracked, including faces or other human parts. Face tracking in particular, in conjunction with face detection, is required as a preprocessing step for a number of human-centered video analysis tasks such as face clustering and recognition or facial expression recognition. Tracking can be formulated in a stochastic Bayesian framework, see [2]. In this work, a post-processing methodology is introduced, which is formulated on a Bayesian framework, with the aim to accurately localize an object in stereo videos, by combining the tracking results of a single channel tracking algorithm, applied independently on the two channels of a stereo video. Combination of multiple tracking information has also been proposed in [3], where a Monte Carlo stochastic sampling algorithm is employed. A similar approach is presented in [4], where multiple trackers are combined, by exploiting only the probability density function of the new target position of each tracker. The major novelty of our proposed framework is that the tracking information, which the proposed post-processing algorithm combines, comes from the left and right stereo video channels. The framework also utilizes object displacement information over time, obtained through SIFT feature matching [5], as well as disparity information between object appearances in left and right video frames. An overview of the proposed methodology is illustrated in Figure 1.

Object displacement and disparity information is obtained prior to post-processing, by using the initial tracking results. Displacement and disparity are manifested in the proposed methodology as probability distribution parameters. To this end, two Student's-t distributions are used. The reasoning for adopting the Students'-t distribution is that it is flexible enough to model the temporally varying statistical properties of the data [6], [7].

Based on these two models, a variational Bayesian approximate inference algorithm [6] is employed, in order to bypass the intractability problem that appears when exact inference is attempted. The ultimate goal is to obtain more accurate estimates of the ideal object ROI coordinates in both channels.

The rest of the paper is organized as follows. The observation and prior models are described in Subsections 2.1 and 2.2 respectively, whereas the variational Bayesian inference algorithm is derived in Subsection 2.3. Experimental evaluation of the proposed post-processing framework is provided in Section 3. Finally, Section 4 concludes the paper.

2. METHOD DESCRIPTION

2.1. Observation Model

With the proposed methodology we aim to estimate the ideal (unknown) positions of the object ROIs in each video frame and channel. We model as random variables (observation model) only the ideal object ROIs in the left video channel,

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.

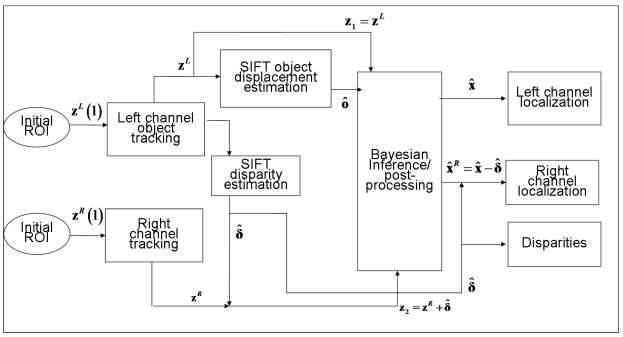

Fig. 1. Overall diagram of the proposed methodology.

which are inferred by the Bayesian post-processing algorithm. On the other hand, the right video channel ROIs are estimated after the Bayesian inference process.

Let N be the total number of frames for each channel (left and right). A ROI in the i-th left channel frame, is assumed to be a rectangle, defined 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 what follows. 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 containing the coordinates of all ideal object ROIs over the entire video. The right channel coordinates $\mathbf{x}^R(i)$ are given by $\mathbf{x}^R(i) = \mathbf{x}(i) + \boldsymbol{\delta}(i)$, where $\mathbf{x}^R(i)$ and the disparity vector:

$$\boldsymbol{\delta}(i) = [\delta_1(i), \delta_2(i), \delta_1(i), \delta_2(i)], \tag{1}$$

are four element vectors. All coordinates are assumed to be real numbers for optimization convenience, as seen in Subsection 2.3. Note that $\boldsymbol{\delta}(i)$ are estimated in a pre-processing step.

An object is tracked by initializing a single-view object tracker on the first frame of the left and right video channels and applying it in both channels, independently. We denote by $\mathbf{z}^L(i), \mathbf{z}^R(i)$ (Figure 1) the extracted ROI coordinates obtained by this procedure in the left and right video channel, respectively. These coordinates are assumed noisy observations of $\mathbf{x}(i)$. However, in the proposed model, we use $\mathbf{z}_1(i) = \mathbf{z}^L(i)$, $\mathbf{z}_2(i) = \mathbf{z}^R(i) - \boldsymbol{\delta}(i)$, which are direct and indirect observations of $\mathbf{x}(i)$, respectively, where

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

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

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

the vector of all extracted ROI coordinates.

The observed ROI generation procedure mentioned above is modelled by assuming that the extracted ROIs are noisy observations of the ideal ROIs $\mathbf{x}$. In more detail, we assume that $p(\mathbf{z}|\mathbf{h})$ is given by:

$$p(\mathbf{z}|\mathbf{h}) \propto \prod_{i,k} \exp\left(-\frac{\lambda_{\mathbf{b}} d_k(i) b_k(i)}{2} \|\mathbf{z}_k(i) - \mathbf{x}(i)\|_2^2\right), \tag{3}$$

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, that will be explained later in this Subsection. Moreover, $\mathbf{u}$ are random variables introduced and explained in Subsection 2.2 and $\mathbf{b}$ denotes the set of all inverse variances $b_k(i)$ of $\mathbf{z}_k(i)$, appearing in (3):

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

We impose a Gamma hyper-prior distribution [6] on $b_k(i)$:

$$p(b_k(i)) \propto b_k(i)^{\nu_{\mathbf{b}}/2-1} \exp(-\nu_{\mathbf{b}} b_k(i)/2), \forall k, i, \tag{4}$$

where $\nu_{\mathbf{b}} > 0$. It should be noted that if we integrate out the $\mathbf{b}$ variables from $p(\mathbf{z}_k(i), b(i)|\mathbf{x}(i), \mathbf{d}(i))p(b_k(i))$, based on (3) and (4), we obtain a Student's-t distribution [6].

The key feature of this model is the flexibility of each $b_k(i)$ to vary with i and k. Thus, this model is used with the purpose to moderate the influence of ROIs coming from highly inaccurate tracking results (such as object localization failures due to occlusions 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.

Binary variables $d_k(i)$ in (3) take values $d_k(i) = 0$ or 1, with $\sum_{k=1}^{2} d_k(i) = 1$. $d_k(i) = 1$ indicates that the ROI $\mathbf{z}_1(i)$, and not $\mathbf{z}_2(i)$, is actually the observation of the ideal ROI $\mathbf{x}(i)$. For $\mathbf{d}$, we assume a multinomial prior, with parameters $\pi_k = \frac{1}{2},\ k = 1, 2$.

2.2. Prior model

For $\mathbf{x}$, we adopt a Student's-t distribution. Mathematically, this means, first, that the conditional prior distribution that generates $\mathbf{x}$ is a Gaussian distribution, given by:

$$p(\mathbf{x}|\mathbf{u}) \propto \prod_{i=2}^{N} \exp(-\frac{\lambda_{\mathbf{x}}}{2} u(i) \|\mathbf{x}(i) - \mathbf{x}(i-1) - \mathbf{o}(i)\|_2^2), \tag{5}$$

where $\mathbf{o}(i)$, $i = 2, \ldots, N$, play the role of the mean (expected value) of the temporal object displacement between two consecutive ROIs. These are computed after the initial single-channel tracking step and before the Bayesian inference, using SIFT feature extraction and matching [5], as presented in Figure 1. Variables $\mathbf{u} = [u(1), \ldots, u(N)]^T$ are generated independently through a Gamma distribution:

$$p(u(i)) = \text{Gamma}(u(i); \nu_{\mathbf{x}}/2, \nu_{\mathbf{x}}/2),\ i = 2, \ldots, N, \tag{6}$$

except $u(1)$, which is assumed to be zero. Note that if we integrate out $\mathbf{u}$ from the joint probability $p(\mathbf{x}, \mathbf{u})$, a multivariate Student's-t distribution [6] is obtained.

In (5), every displacement $\mathbf{o}(i)$ is used in the model as the mean of a Gaussian distribution. Specifically, it is assumed to be the expected value of the difference $\mathbf{x}(i) - \mathbf{x}(i-1)$. We first define $\mathbf{o}(i) = [o_1(i), o_2(i), o_1(i), o_2(i)]^T$ as a 4×1 vector. The values of its elements are (ideally) the differences:

$$\mathbf{o}(i) = \mathbf{x}(i) - \mathbf{x}(i-1),\ i = 2, \ldots, N. \quad (7)$$

$\mathbf{o}(i)$ consists, in essence, of two variables. We could have used four variables, in order to make the model more accurate, i.e. to model alterations of $\mathbf{x}(i-1)$ to $\mathbf{x}(i)$ in every one of its four elements, but we avoided that for the sake of simplicity. However, this type of modeling does not prevent the estimate of $\mathbf{x}(i)$, which is influenced by $\mathbf{z}^L$, $\mathbf{z}^R$, $\boldsymbol{\delta}$ and not only $\mathbf{o}$, to have all four variables varying over time (i.e., the produced ROIs are of varying size and aspect ratio).

2.3. Variational Bayesian Inference

We employ the variational Bayesian methodology, in order to obtain an approximate posterior for the hidden variables $\mathbf{h}$, in a tractable manner [6]. In this way, we avoid the exact inference intractability problem, by using, for example, the expectation maximization (EM) algorithm. Following this methodology, we utilize the always positive Kullback-Leibler (KL) divergence between $q(\mathbf{h})$ and $p(\mathbf{h}|\mathbf{z})$ [6], in order to define the upper bound L of the log-likelihood:

$$L\left(q(\mathbf{h}), \theta\right) = \log p(\mathbf{z}; \theta) - KL\left(q||p\right) \geq \log p(\mathbf{z}; \theta) \quad (8)$$

q, which plays the role of the inferred posterior, and θ are estimated by iteratively minimizing the bound with respect to q and θ. We adopt the *mean-field* approximation, which is a common practice in the variational framework. Specifically, $\mathbf{x}$, $\mathbf{u}$, $\mathbf{d}$ and $\mathbf{b}$ are assumed independent in the inferred posterior.

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

Next, we give the update included in each iteration of the variational Bayesian algorithm. $\mathbf{x}_c,\ c = 1, 2, 3, 4$, are $N \times 1$ vectors that are subsets of $\mathbf{x}$ and contain respectively the coordinates $x_1(i)$, $x_2(i)$, $x_3(i)$ and $x_4(i)$, $\forall i$, as defined in Subsection 2.1. The update of each of their posteriors is given by:

$$q^{(t)}(\mathbf{x}_c) = N\left(\boldsymbol{\mu}_c^{(t)}, \mathbf{C}_{\mathbf{x}}^{(t)}\right),\ c = 1, 2, 3, 4, \quad (9)$$

$$\boldsymbol{\mu}_c^{(t)} = \mathbf{C}_{\mathbf{x}}^{(t)} \mathbf{y}_c, \left(\mathbf{C}_{\mathbf{x}}^{(t)}\right)^{-1} = \mathbf{B}^{(t)} + \lambda_{\mathbf{x}}^{(t-1)} \mathbf{Q}^T \mathbf{U}^{(t)} \mathbf{Q}, \quad (10)$$

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

$$[\mathbf{B}^{(t)}]_{(i,i)} = \lambda_{\mathbf{b}}^{(t-1)} \sum_{k=1}^{2} \hat{d}_k(i) \hat{b}_k(i), [\mathbf{U}^{(t)}]_{(i,i)} = \hat{u}(i), \quad (11)$$

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. Moreover, $\mathbf{y}_c,\ c = 1, 2, 3, 4$ are $N \times 1$ vectors with elements:

$$\mathbf{y}_c(i) = \lambda_{\mathbf{b}} \sum_{k=1}^{2} \hat{d}_k(i) \hat{b}_k(i) z_{k,c}(i) + \lambda_{\mathbf{x}} [\mathbf{Q}^T \mathbf{U}^{(t)} \mathbf{o}_c](i), \quad (12)$$

where, $[v](i)$ denotes the i-th element of a vector v. Also, $\mathbf{o}_c$ is a vector containing all $o_c(i)$, $c = 1, 2, 3, 4$, for $i = 1, \ldots, N$. Finally, $\mathbf{Q}$ is the $N \times N$ first order difference operator. The posteriors for $u(i)$ and $b_k(i)$ are:

$$q^{(t)}(u(i)) = \text{Gamma}\left(u(i); \alpha_{\mathbf{u}}(i), \beta_{\mathbf{u}}(i)\right), \quad (13)$$

$$q^{(t)}\left(b_k(i)\right) = \text{Gamma}\left(b_k(i); \alpha_{\mathbf{b}}(i), \beta_{\mathbf{b}}(i)\right),\ \forall k, i, \quad (14)$$

where $\alpha_{\mathbf{u}} = \nu_{\mathbf{x}}/2 + 1/2$ and $\alpha_{\mathbf{b}}(i) = \nu_{\mathbf{b}}/2 + \hat{d}_k(i)/2$,

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

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

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

Thus, using the mean of a Gamma distribution formula [6]:

$$\hat{u}(i) \equiv \langle u(i) \rangle_{q^{(t)}(u(i))} = \frac{\alpha_{\mathbf{u}}(i)}{\beta_{\mathbf{u}}(i)},\ \forall k, i. \quad (15)$$

The update for $\hat{b}_k(i)$ is similar with the above. We also set:

$$\hat{d}_k(i) = \pi_k = 0.5,\ \forall k, i. \quad (16)$$

The updates for $\lambda_{\mathbf{x}}$ and $\lambda_{\mathbf{b}}$ are found by maximizing L.

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

3. EXPERIMENTAL PERFORMANCE ANALYSIS

The proposed Bayesian post-processing tracking algorithm was evaluated on two stereo sequences, by performing three stereo tracking experiments. The single channel (SC) tracker [8] was used to track objects (one face in each video and a hand) independently in the left and right channel of the stereo sequences. The tracking algorithm was initialized by a user selected ROI in the first video frames of the left/right channels and no automatic object detection was performed. Using the tracking results obtained by the SC tracker, we employ the SIFT feature extraction and matching procedure, so as to estimate the ROI coordinate displacements $\mathbf{o}$ and disparities $\boldsymbol{\delta}$.

The output of the post-processing methodology includes the estimates of the left channel ROI coordinates $\hat{\mathbf{x}}$. The right channel ROI coordinates are obtained by:

$$\hat{\mathbf{x}}^R(i) = \hat{\mathbf{x}} + \boldsymbol{\delta}(i),\ i = 1, \ldots, N. \quad (17)$$

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

In order to measure the tracking accuracy, the Average Tracking Accuracy (ATA) [9] metric, denoted by $\hat{a}$ was used:

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i \bigcap G_i|}{|D_i \bigcup G_i|}, \tag{18}$$

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

The accuracy of tracking results in terms of the ATA metric is presented in Table 1 for the SC and SBP algorithms. The results demonstrate that the SBP algorithm provides higher tracking accuracy than SC.

In addition to the refinement of tracking results, in this work, we extract coarse disparity values $\hat{\boldsymbol{\delta}}(i)$ between the two object ROIs in the left and right video frames, where again a SIFT matching procedure is employed. The proposed approach avoids the time consuming application of dense disparity estimation algorithms over the entire video frames. We estimate the average disparity $\hat{\gamma}(i)$ in an object ROI, by using the median of the the dense disparity values obtained through the application of the method in [10] within an object ROI, and we use them as reference disparities in the comparison with the proposed coarse disparity estimation method. In the Table 1, the mean and standard deviation (std) of the differences $\hat{\delta}(i) - \hat{\gamma}(i)$ are provided. We can see that estimated disparities $\hat{\delta}$ are very close to the "ground truth" disparities $\hat{\gamma}$ in terms of the mean and standard deviation of their differences.

Table 1. Tracking performance (ATA) of SC and SBP.

Video name	N	SC	SBP	$\hat{\delta}(i) - \hat{\gamma}(i)$	
				$mean$	std
Badminton, left	499	0.560	**0.562**	-2.11	1.62
Badminton, right	499	0.522	**0.525**		
Poker (hand), left	499	0.461	**0.513**	0.52	1.01
Poker (hand), right	499	0.481	**0.538**		
Poker (head), left	599	0.714	**0.726**	1.06	0.79
Poker (head), right	599	0.701	**0.747**		

4. CONCLUSIONS

An object tracking Bayesian post-processing methodology for stereoscopic sequences was presented in this paper. The methodology refines the outputs of standard tracking algorithms, by exploiting, the left and right channel tracking results. Moreover, object displacement over time, as well as disparity information, were exploited successfully to this end. The refined tracking results are significantly better than those provided by the initial, single-channel tracking algorithm. In the future, we plan to improve the stochastic model.

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

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

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] C. Riechert, F. Zilly, and P. Kauff, "Real time depth estimation using line recursive matching," in *European Conference on Visual Media Production (CMVP)*, November 2011.