Fragment-based Variational Visual Tracking

(Invited Paper)

Yi Zhou*†, Hichem Snoussi*, Shibao Zheng†, Cédric Richard*, Jing Teng *
* ICD/LM2S, University of Technology of Troyes,
12, rue Marie Curie, 10000, France
Email: zhouyi21st@gmail.com hichem.snoussi@utt.fr jing.teng@utt.fr
†Institute of Image Communication and Information Processing,
Shanghai Jiao Tong University,
800, Dong Chuan Road, 200240, China
Email: sbzh@sjtu.edu.cn

***Abstract*—We propose a Bayesian tracking algorithm based on adaptive fragmentation and variational approximation. By using the cue of gradient, we fragment the target into disconnected rectangles and reduce the confusion from the background. To handle the uncertainties in real tracking case, we choose the Bayesian framework with a variational implementation. The parameters of the variational inference are updated according to the observation and to the weights of the voting candidates. Experimental results show that our tracker outperforms directive searching and particle filtering. Furthermore, due to the simplicity of calculation, the proposed method can be applied to real-time surveillance systems.**

I. Introduction

Visual tracking in complex environments is one of the key problems in many computer vision applications such as visual surveillance, intelligent robot and wireless camera networks. A visual tracker can be divided into two parts: target representation and localization.

Representation consists of decreasing the dimension of the target in the first frame of the video stream. Due to the simplicity and invariance to scaling and rotation, the histogram is a popular representation method. However, there is an inherent defect of the histogram based representation. Since the spatial information is omitted, the tracker in the later localization step cannot distinguish the objects with almost the same feature histogram. A robust fragments-based method was proposed in [1]. In this method, the window containing the target in the first frame is divided into overlapped sub-regions. By keeping the distance to the center of the main window, each sub-region participates in a voting process in order to estimate the target position in the subsequent frames. However, the fixed sub-region division and unchanged directive searching range in the localization step decreases the tracker's flexibility of handling the abrupt variation of the target trajectory.

In the localization procedure, the tracker generates some candidate regions and then assigns weights to these candidates according to their feature difference from the reference. The candidate with the least difference has the highest similarity to the reference. Two major categories of algorithms are used for localization. The first category of algorithms is based on deterministic searching in a limited region close to the last position. Mean shift method [2] or exhaustive searching [1] are within this category. The second category is based on Bayesian filtering which estimates the posterior distribution of the target position given all the previous frames. The main drawback of the Bayesian approach is the intractable computation of the marginal posterior distribution, mainly due to the nonlinear expression of the likelihood function. The particle filtering is a sequential Monte Carlo approximation allowing the implementation of the Bayesian filtering methods even in highly nonlinear and non-Gaussian models. However, its performance highly depends on the proposal distribution, which is rarely available in real visual tracking applications [3]. As an alternative solution, in this paper, we apply the variational approximation to Bayesian visual tracking based on the fragmentation representation. The variational method is able to approximate the complex Bayesian posterior distribution of the hidden states by minimizing the Kullback-Leibler (KL) divergence to the true distribution at each time step [3], [4].

The main contribution of this paper is to introduce a new framework for efficient visual tracking. In the representation step, we modify the fragmentation in [1] by using the gradient cue, in order to obtain the partitions in the most informative region. By online updating a free form approximation of the filter distribution, the variational approach is proved to outperform the directive searching and the particle filtering. Furthermore, the filtering based method is able to predict the region where the target is likely to appear. Therefore we only compute the integral histogram in the region of interest (ROI), speeding up the tracker.

The remainder of the paper is organized as follows. Section II introduces the target model based on the modified fragment representation. Section III describes the variational approximation in a Bayesian inference framework. Section IV gives the experimental results. Section V concludes the paper.

II. Fragments-based Target Representation

Thanks to its simplicity and independence from scaling and rotation, the color histogram has been widely used for target representation. However, the histogram feature loses the spatial information of the target. In [1], the rectangular window, containing the target, is divided into smaller fixed

978-1-4244-5180-7/09/$26.00 ©2009 IEEE

vertical and horizontal sub-rectangles, indicated in the second column of Figure 1. Each sub-rectangle, called patch, has its

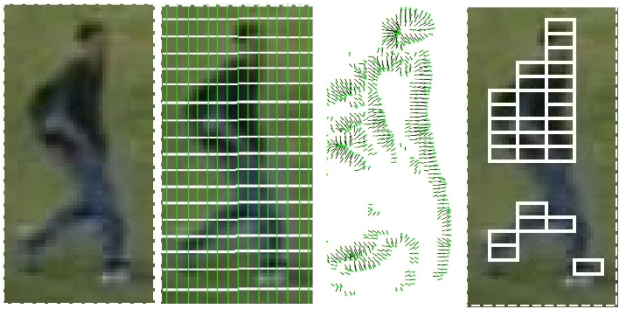

Fig. 1: Fragmentation on a stepping forward pedestrian. The first column is the target in a bounding box; the second column is the fragmentation based on fixed horizontal and vertical rectangles proposed in [1]; the third column shows the gradient information of the target; the last column displays the patches selected by our fragmentation.

own coordinates inside the window, and its own height and width, represented as P_t=(dx, dy, h, w). The computation of histogram is performed in each patch. Therefore, the target in the tracker window was now represented as a series of patches and their respective histograms. Spatial information is efficiently saved through this approach.

However, the main drawback is that the division of the sub-regions is fixed in each tracking step. In visual tracking case, the human target often steps forward. Consequently, the leg's location is variable in the main window. If the fixed patches are used to locate the leg of the target, they would be often viewed as outliers in later comparison. In this paper, in order to locate the patches in more informative places, we use the oriented gradient information of the target [5]. First, the pixel's orientation and magnitude are computed in the main window. Then, we define the patches with suitable width, height and we set a threshold T as the ratio between the cumulative magnitudes cum_M in the main window and the numbers of patches num_P (T=cum_M/num_P). The patch with the higher cumulative magnitudes than the average T is kept as a descriptor. Figure 1 indicates that this method produces patches in the regions containing texture variations.

III. Bayesian Filtering by Variational Approximation

A. State-space Model and Likelihood Model

In this paper, the target state $\boldsymbol{x}_t$ at time instant t is the location of the target in the current frame. In order to model the prior information about the target trajectory, we adopt the model introduced by [3]:

$$\begin{cases} \boldsymbol{\mu}_t & \sim & \mathcal{N}(\boldsymbol{\mu}_t \mid \boldsymbol{\mu}_{t-1}, \bar{\boldsymbol{\lambda}}) \\ \boldsymbol{\lambda}_t & \sim & \mathcal{W}_{\bar{n}}(\boldsymbol{\lambda}_t \mid \bar{\boldsymbol{S}}) \\ \boldsymbol{x}_t & \sim & \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t) \end{cases} \tag{1}$$

where the state of interest $\boldsymbol{x}_t$ has a Gaussian distribution with random mean $\boldsymbol{\mu}_t$ and random precision matrix $\boldsymbol{\lambda}_t$. The mean follows a Gaussian walk reflecting the time correlation of the system trajectory. The precision matrix follows a Wishart distribution. $\bar{\boldsymbol{\lambda}}$, $\bar{n}$ and $\bar{\boldsymbol{S}}$ are hyper-parameters, denoting respectively the Gaussian walk precision matrix, freedom degrees and the precision matrix of the Wishart distribution. In order to capture various motions, we adaptively tune the parameter $\bar{\boldsymbol{\lambda}}$ instead of constructing a fixed motion model. Contrary to [3], where the $\bar{\boldsymbol{\lambda}}$ is fixed, this modification is more robust to motion changing. Details are given in the next section.

To evaluate how likely a candidate region represents the target, we define the likelihood model based on the color histogram measurement. As mentioned before, the model reference is a series of patches with their corresponding color histograms. We compute the Chi-square distance first between the nth patch's histogram of the ith candidate region x^i and the corresponding patch's histogram h_{ref} in the reference as follow:

$$D_n = \Sigma_{j=1}^{N_b} \frac{(h_{n,x^i}(j) - h_{n,ref}(j))^2}{h_{n,x^i}(j) + h_{n,ref}(j)}, \tag{2}$$

where N_b denotes the number of bins in histogram. If there are N patches selected at time t, the whole distance set is $\{D_1, \ldots, D_N\}$ for region x^i. Therefore, the combined distance between the ith candidate and the reference is,

$$D_{x^i} = \sum_{n=1}^{N} D_n. \tag{3}$$

Given M candidates at time t, we define the likelihood of one candidate state x_t^i as,

$$p(y_t \mid x_t^i) \propto exp(-D_{x^i} / \sum_{i=1}^{M} D_{x^i}). \tag{4}$$

B. Bayesian Filtering by Sequential Monte Carlo Approximation

Given the model in section III-A, the Bayesian filtering consists of estimating the posterior marginal probability of the continuous hidden state, i.e. $p(\boldsymbol{\alpha}_t \mid \boldsymbol{y}_{1:t})$, where $\boldsymbol{\alpha}_t = (\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}_t)$ is the extended hidden state, and $\boldsymbol{y}_{1:t}$ denotes all the past frames up to current frame. As the likelihood function is nonlinear and the dimensionality of the observation is highly decreased, the online update of the posterior distribution is intractable. The particle filtering is a popular approximation method based on sequential Monte Carlo procedure. In this methodology the posterior distribution is approximated by a point-mass distribution of a set of weighted samples (called particles) $\{\boldsymbol{\alpha}_t^{(i)}, w_t^{(i)}\}_{i=1}^N$:

$$p_N(\boldsymbol{\alpha}_t \mid \boldsymbol{y}_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta_{\boldsymbol{\alpha}_t^{(i)}}(d\,\boldsymbol{\alpha}_t),$$

where $\delta_{\boldsymbol{\alpha}_t^{(i)}}(d\,\boldsymbol{\alpha}_t)$ denotes the Dirac function. To maintain a sequential schedule, the sampling (proposal) distributions must

respect the following form:

$$\pi(\boldsymbol{\alpha}_{0:t} \mid \boldsymbol{y}_{1:t}) = \pi(\boldsymbol{\alpha}_{0:t-1} \mid \boldsymbol{y}_{1:t-1})\pi(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{0:t-1}, \boldsymbol{y}_{1:t}).$$

Finally, the weights can be computed in a recursive way as:

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\boldsymbol{y}_t \mid \boldsymbol{\alpha}_t^{(i)}) p(\boldsymbol{\alpha}_t^{(i)} \mid \boldsymbol{\alpha}_{0:t-1}^{(i)})}{\pi(\boldsymbol{\alpha}_t^{(i)} \mid \boldsymbol{\alpha}_{0:t-1}^{(i)}, \boldsymbol{y}_{1:t})}. \quad (5)$$

The performance of the particle filter depends on the proposal distribution. Usually, the dynamical transition model is used as the proposal. Therefore, the samples are proportional to the likelihood function. In this case, the samples are selected in the highest value area of the likelihood. If the likelihood model has limited overlapped area with the posterior, the performance would be poor. To improve the performance, one has to use a quasi-optimal proposal distribution. However, this leads to more complex and time consuming algorithm, which is not suitable for real-time visual tracking applications.

Recently, the variational approximation, has been proposed as an alternative to sequential Monte Carlo method. It is based on a deterministic approximation of the posterior distribution of the hidden variables, while keeping tractable and closed analytical forms [6].

C. Variational Approximation

Contrary to the point-mass approximation of the filtering distribution in particle filtering, the variational approach approximates the posterior $p(\boldsymbol{\alpha}_t \mid \boldsymbol{y}_{1:t})$ by a more tractable distribution $q(\boldsymbol{\alpha}_t)$. It is based on minimizing the Kullback-Leibler divergence between the filtering distribution and the approximate distribution:

$$D_{\mathrm{KL}}(q||p) = \int q(\boldsymbol{\alpha}_t) \log \frac{q(\boldsymbol{\alpha}_t)}{p(\boldsymbol{\alpha}_t \mid \boldsymbol{y}_{1:t})} d\boldsymbol{\alpha}_t. \quad (6)$$

with the assumption of a separable distribution $q(\boldsymbol{\alpha}_t) = q(\boldsymbol{x}_t)q(\boldsymbol{\mu}_t)q(\boldsymbol{\lambda}_t)$. Variational calculus leads to the following optimal distribution:

$$\begin{cases} q(\boldsymbol{x}_t) & \propto & \exp \langle \log p(\boldsymbol{\alpha}_t | \boldsymbol{y}_{1:t}) \rangle_{q(\boldsymbol{\mu}_t)q(\boldsymbol{\lambda}_t)} \\ q(\boldsymbol{\mu}_t) & \propto & \exp \langle \log p(\boldsymbol{\alpha}_t | \boldsymbol{y}_{1:t}) \rangle_{q(\boldsymbol{x}_t)q(\boldsymbol{\lambda}_t)} \\ q(\boldsymbol{\lambda}_t) & \propto & \exp \langle \log p(\boldsymbol{\alpha}_t | \boldsymbol{y}_{1:t}) \rangle_{q(\boldsymbol{x}_t)q(\boldsymbol{\mu}_t)} \end{cases} \quad (7)$$

Taking into account the state model (1) and variational approximation $q(\boldsymbol{\alpha}_{t-1})$ at time $t-1$, the filtering distribution is written as,

$$\begin{aligned} p(\boldsymbol{\alpha}_t | \boldsymbol{y}_{1:t}) &\propto p(\boldsymbol{y}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t) p(\boldsymbol{x}_t, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t \mid \boldsymbol{y}_{1:t-1}) \\ &\propto p(\boldsymbol{y}_t | \boldsymbol{x}_t) p(\boldsymbol{x}_t, \boldsymbol{\lambda}_t | \boldsymbol{\mu}_t) \textstyle\int p(\boldsymbol{\mu}_t | \boldsymbol{\mu}_{t-1}) q(\boldsymbol{\mu}_{t-1}) d\boldsymbol{\mu}_{t-1}. \end{aligned} \quad (8)$$

Substituting equation (8) in (7) and considering the space model (1), the separable components of the approximate distribution are obtained in the following form,

$$\begin{cases} q(\boldsymbol{x}_t) & \propto & p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) \mathcal{N}(\boldsymbol{x}_t \mid \langle \boldsymbol{\mu}_t \rangle, \langle \boldsymbol{\lambda}_t \rangle) \\ q(\boldsymbol{\mu}_t) & \propto & \mathcal{N}(\boldsymbol{\mu}_t \mid \boldsymbol{\mu}_t^*, \boldsymbol{\lambda}_t^*) \\ q(\boldsymbol{\lambda}_t) & \propto & \mathcal{W}_{n^*}(\boldsymbol{\lambda}_t \mid \boldsymbol{S}_t^*) \end{cases} \quad (9)$$

where the parameters are updated according to the scheme of equation (10) in [7]. However, due to the non-Gaussian likelihood mode in model (4), the mean and the covariance of the distribution $q(x_t)$ have no closed form. An Importance sampling scheme is thus employed in order to approximate these statistics as follows. The samples are drawn according to the Gaussian $\mathcal{N}(\boldsymbol{x}_t \mid \langle \boldsymbol{\mu}_t \rangle, \langle \boldsymbol{\lambda}_t \rangle)$ and weighted according to their likelihoods:

$$\boldsymbol{x}_t^{(i)} \sim \mathcal{N}(\boldsymbol{x}_t \mid \langle \boldsymbol{\mu}_t \rangle, \langle \boldsymbol{\lambda}_t \rangle),\ w_t^{(i)} \propto p(\boldsymbol{y}_t \mid \boldsymbol{x}_t^{(i)}). \quad (10)$$

The mean and the covariance are then obtained by the following approximations:

$$\langle \boldsymbol{x}_t \rangle = \sum_{i=1}^{N} w_t^{(i)} \boldsymbol{x}_t^{(i)},\ \langle \boldsymbol{x}_t \boldsymbol{x}_t^T \rangle = \sum_{i=1}^{N} w_t^{(i)} \boldsymbol{x}_t^{(i)} \boldsymbol{x}_t^{(i)T}. \quad (11)$$

It is worth noting that the parameters of one factorized distribution are jointly updated with respect to the other remaining components' expectations. In fact, the additional hidden states $\boldsymbol{\mu}_t$ and $\boldsymbol{\lambda}_t$ give more flexibility to the dynamical state. Based on the extended state, candidates can be adaptively hypothesized. In order to handle the uncertainties in an adaptive way, we also propose to tune the hyper-parameters of the prior distributions of $\boldsymbol{\mu}_t$ and $\boldsymbol{\lambda}_t$, based on the distribution $q(\boldsymbol{x}_t)$ approximated by the set $\{w_t^{(i)}, \boldsymbol{x}_t^{(i)}\}$. In this paper, we use the mean $E(w_t)$ and the variance $Var(w_t)$ of the samples weights $w_t^{(i)}$, in order to monitor the hyper-parameters variation. The decreasing of the mean $E(w_t)$ (i.e.lower than a fixed threshold M_w) and/or the increasing of the variance $Var(w_t)$ (i.e. higher than a fixed threshold V_w) indicate that an abrupt event has occurred (for example, the target is speeding up or an occlusion has occurred). In these cases, decreasing the precision matrix $\bar{\boldsymbol{\lambda}}$ in model (1) will increase the range of the search region of the variational filter. Therefore, the tracker is able to localize the target when an irregular event occurs. When the target recovers a regular trajectory, the search region will then be automatically reduced.

IV. Experimental Results

In this section, we present the experimental results on two widely used datasets from PETS2001 and PETS2006 (available at http://ftp.pets.rdg.ac.uk). The histogram was calculated in RGB space with 10x10x10 bins. The parameters in the dynamical model were set to $\boldsymbol{\mu}_0^*$ = $[x_0, y_0]^T$, λ_0^* = diag (0.008 , 0.008), $\bar{\lambda}$ = diag (0.008, 0.008), $\bar{n} = 4$, $\bar{S}$ = diag (5, 5). The threshold for weight mean $E(w_t)$ and weight variance $Var(w_t)$ are respectively set to $M_w = 0.7$ and $V_w = 82$. The range for $\bar{\lambda}$ is defined as $(0.0003 \sim 0.008)\boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix. 60 candidates were produced at importance sampling step in the variational filter. For comparison purpose, we also run the algorithm in [1] and the general particle filter with 200 samples.

Figure 2 shows the limitations of the two concurrent tracking algorithms. As the bounding box of the target in the initial frame contains some background information, the fixed fragmentation tracker lost the target in frame 972. The particle filter, even with more samples, also failed in frame 1086, when the pedestrian changed the direction and the background has a color histogram similar to the top part of the target. Based

Fig. 2: Comparison of tracking performances between the algorithm proposed in [1] (first row), the particle filter (second row) and the proposed tracker (last row). From left to right, frames (840, 972, 1086) were taken from the sequence in PETS2001.

on the gradient cue of the target, our approach avoids the selection of the patches in the background. Furthermore, when the target accelerated, the mean of the weights dropped below the threshold. In this case, our tracker adaptively tunes down the $\bar{\lambda}$ and keeps tracking the real target. With the provided

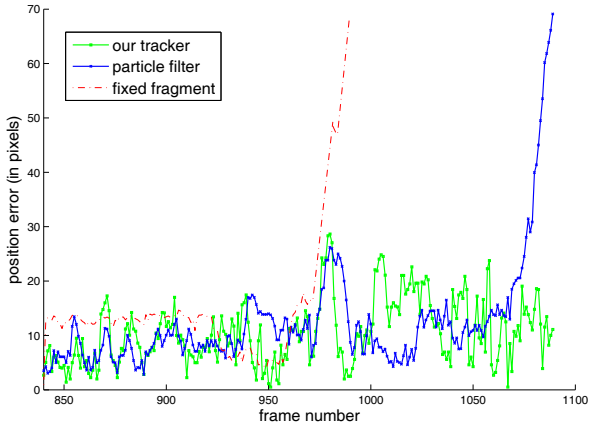

Fig. 3: Position errors w.r.t. ground truth

ground truth, we also compare the position errors of the three trackers in Figure 3. Using much less samples, our tracker has a comparable performance as the particle filter and succeeds when an abrupt event happened.

In Figure 4, there is a severe occlusion when the tracked person encountered an other person with similar appearance. The other challenging aspect of this data set is the fact that when the pedestrian moved closer to the camera, his relative speed increased. The particle filter lost the target when the occlusion happened. However, our proposed algorithm succeeds in tracking the target.

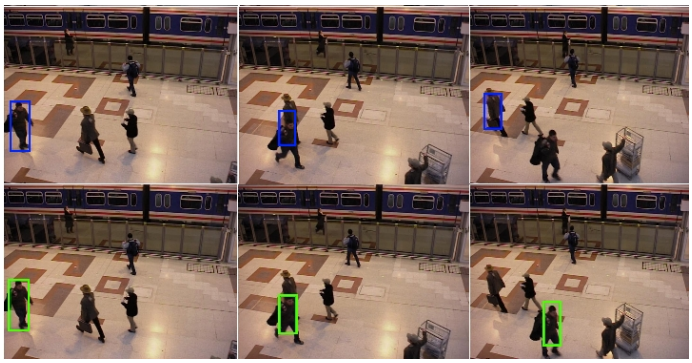

Fig. 4: Tracking results in frames 1040, 1060 (occlusion happened between two persons with similar appearance), 1077 (after occlusion) of PETS2006. The tracker in first row used the particle filter. The second row shows the results of our proposed tracker.

V. Conclusion

Based on gradient information, we use more informative fragments to represent a target. Without the time-consuming background modeling and foreground segmentation, this model works well in real visual surveillance. The variational approach in localization step is adaptively tuned to handle the abrupt trajectory change in real time with less samples, compared to the particle filter. Thanks to these performances, the proposed tracker can be potentially applied to the embedded smart camera systems. One shortcoming of the proposed tracker is that its reference model and scale are not updated during the tracking. This is due to the fact that the tracker can not update the reference in a proper time. In fact, there is a feature dissimilarity even for the same target in different frames. Automatically updating the reference model is one of the perspectives of this work.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2006, pp. 798–805.

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 5, pp. 564–577, May 2003.

[3] J. Vermaak, N. D. Lawrence, and P. Pérez, "Variational inference for visual tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, Jun. 2003.

[4] D. Tzikas, A. Likas, and N. Galatsanos, "The variational approximation for bayesian inference," *Signal Processing Magazine, IEEE*, vol. 25, no. 6, pp. 131–146, November 2008.

[5] D. Marimon and T. Ebrahimi, "Orientation histogram-based matching for region tracking," in *Image Analysis for Multimedia Interactive Services, 2007. WIAMIS '07. Eighth International Workshop on*, June 2007, pp. 8–8.

[6] V. Smidl and A. Quinn, "Variational bayesian filtering," *Signal Processing, IEEE Transactions on*, vol. 56, no. 10, pp. 5020–5030, Oct. 2008.

[7] H. Snoussi and C. Richard, "Ensemble learning online filtering in wireless sensor networks," in *Communication systems, 2006. ICCS 2006. 10th IEEE Singapore International Conference on*, Oct. 2006, pp. 1–5.