The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Target Tracking for Moving Robots Using Object-based Visual Attention

Yuanlong Yu, George K. I. Mann and Raymond G. Gosine

Abstract— **Visual tracking is a quite challenging issue for a moving robot due to the appearance changes of both the background and targets, large variation of motion, partial or full occlusion and so on. However, humans are capable to cope with those difficulties to achieve satisfactory tracking performance. Thus this paper presents a biologically-inspired method of visual tracking for moving robots by using object-based visual attention mechanism. This tracking method consists of four modules: pre-attentive segmentation, top-down attentional biasing, post-attentive completion processing and online learning of the target model. Experimental results in natural and cluttered scenes are shown to validate this general and robust tracking method.**

I. INTRODUCTION

There are mainly three types of challenging issues in the visual tracking task. The first challenge is caused by the cluttered and dynamically changing background since 1) background contains a variety of clutters that share some features with the target, and 2) discrimination between foreground and background will change dynamically during tracking. To cope with this issue, two requirements for building the target model should be satisfied: robustness and discriminability. Robustness means that the target model can represent various instances of the target in different viewing conditions. Several probabilistic models [1], [2], [3], [4] and subspace appearance models [5], [6] are proposed to improve robustness. Discriminability means that the target model can be discriminated from the background. Collins et al. [7], [8] proposed a method for online selecting a discriminative feature from a set of color features. However, there are three shortcomings in that feature selection method. Firstly, some important features, such as contour, are not included in the candidate feature set. Secondly, a rectangle or an ellipse is used to approximately outline the target and surrounding regions such that outliers included in both regions would interfere the feature selection. Thirdly, the selected feature is locally discriminative as only a small background region around the target is used for feature selection.

The second challenging issue is the ability of automatical recovery in the case of tracking failure. Besides background clutters, a large variation of motion and full occlusion during several sequential frames are another two major reasons causing tracking failure. Although some methods [9], [10], [11] have been proposed to accommodate abrupt motion, the occurrence of tracking failure cannot be absolutely eliminated based on the fact that the designed tracking systems cannot accommodate all possible reasons that cause failure. Thus an automatical recovery mechanism is required. This paper proposes that two components are necessary for the recovery mechanism: validation and global search. For each frame, the estimated target state needs to be validated. If it is incorrect, a global search then attempts to detect the correct target in the entire image. Thus it is required to online select a target's feature that are discriminative over the entire image.

The third challenging issue is the completion of the tracked target. It is required in robotic applications, since a complete target region can provide important information for robot's following actions, such as grabbing. Unfortunately, most tracking methods only use primitive geometric shapes, e.g., an ellipse [2], [3], to represent the target.

The perception behavior of humans inspires a novel approach to the above issues in a unified framework. Object-based attention theory posits that some pre-attentive processes serve to segment the field into discrete objects, followed by an attention process that selects one object at a time [12]. Integrated competition (IC) hypothesis [13] is further proposed to model object-based attention: By directing attention to a conspicuous cue of an object, it produces a competitive advantage over the whole object.

Therefore, this paper presents a biological-inspired visual tracking method based on IC hypothesis. The tracking process is modeled as an object-based attentional selection procedure. This new tracking method attempts to contribute in the following aspects: 1) Adaptivity and effectiveness: The discriminative feature of the learned target model can be automatically online selected for tracking in order to cope with cluttered and dynamically changing environment; 2) Robustness: It has the ability to automatically recover tracking failure caused by any reasons; and 3) Target completion: By combination of pre-attentive segmentation, attentional selection and post-attentive completion processing, the complete target region is achieved. Another top-down attention based tracking approach has also been presented in [14]. The difference is that our method is object based.

The remainder of this paper is organized as follows. The framework of the proposed target tracking method is given in section II. Pre-attentive segmentation module is presented in section III. Online learning of the target model is presented in section IV. The tracking process is described in section V. Experimental results are finally given in section VI.

II. FRAMEWORK OF PROPOSED TRACKING METHOD

This tracking method is developed by extending our previously proposed object-based attention model [15]. It consists of four modules: pre-attentive segmentation, top-down

This research is funded by Natural Sciences and Engineering Research Council of Canada (NSERC) and Memorial University of Newfoundland.

The authors are with Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, A1B 3X5, Canada. {y.yu,gmann,rgosine}@mun.com

attentional biasing, post-attentive completion processing and online learning of the target model, as shown in Fig. 1.

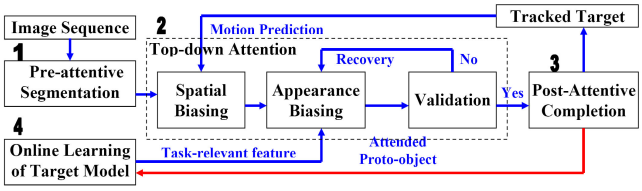

Fig. 1. Proposed target tracking framework using object-based attention.

The pre-attentive segmentation module extracts a set of *pre-attentive features* and then divides the scene into homogenous *proto-objects* in a unsupervised fashion.

Following pre-attentive segmentation, the top-down attentional biasing module, consisting of spatial biasing, appearance biasing and validation, is carried out. Spatial biasing estimates a predicted region based on the target region at previous moments. Using the discriminative *task-relevant feature* deduced from the target model, appearance biasing then evaluates a proto-object based attentional activation map, which represents the likelihood of each proto-object to be the tracked target. The proto-object with the maximal activation is selected for validation. If it is validated to be the target, it is put into the post-attentive completion module to yield a precise and complete target region. Otherwise, it means an occurrence of tracking failure, the recovery mechanism is triggered by carrying out the appearance biasing procedure again over the entire image to globally search for the target.

After the post-attentive completion processing, the complete target region (i.e., a tracked instance of the target) is used for online learning. In the first tracking frame, the target model is initialized by using only one type of supervision information: the trainer specifies which proto-objects belong to the target. In the following tracking frames, the tracked instance of the target is used to update the target model such that it can accommodate changes of the environment.

III. Pre-attentive Segmentation

A. Pre-attentive Feature Extraction

Pre-attentive features include intensity F_{int}, red-green F_{rg}, blue-yellow F_{by}, local orientation energy F_{o_θ} and contour F_{ct}. Given the 8-bit RGB color components r, g, b, multi-scale intensity $F_{int}(s)$, red-green $F_{rg}(s)$ and blue-yellow $F_{by}(s)$ are created using the method proposed in [16], where s denotes the spatial scale.

An oriented Laplacian pyramid [17] (a log-Gabor like filter) is used to extract multi-scale orientation energy $F_{o_\theta}(s)$ in four orientation directions $\theta \in \{0°, 45°, 90°, 135°\}$.

Contour feature $F_{ct}(s)$ is approximated by using the total orientation energy, which is obtained by applying a pixel-wise maximum operator over those four directions:

$$F_{ct}(\mathbf{r}_i, s) = \max_{\theta \in \{0°,45°,90°,135°\}} F_{o_\theta}(\mathbf{r}_i, s) \qquad (1)$$

where $\mathbf{r}_i$ represents a pixel at scale s.

B. Pre-attentive Segmentation

This paper simulates pre-attentive segmentation as a hierarchical accumulation procedure by extending irregular pyramid techniques [18], [19]. Each level of the irregular pyramid is built by accumulating similar nodes at the level below, resulting that the final global segments emerge in this process as they are represented by single nodes at some levels. Details of the pre-attentive segmentation algorithm can be seen in our previous work [20].

IV. Target Model and Learning

A. Structure of Statistical Target Model

The proposed target model $\mathbf{O}$ consists of global coding $\mathbf{O}_{gb}$ using contour feature and local coding $\mathbf{O}_{lc}$ using intensity, red-green, blue-yellow and local orientations. Each coding includes two descriptors: appearance $\mathbf{O}^a$ and salience $\mathbf{O}^s$. The appearance descriptor represents appearance values of each feature. The salience descriptor represents discriminability of a feature dimension of the target in contrast to the background, and therefore it is used to deduce the task-relevant feature.

B. Post-attentive Features

Based on the hypothesis that the statistical structure of the perceived data is recoded for high-level processing [21], a set of statistical quantities, termed as *post-attentive features*, are estimated within the complete target region. They are dependent on the structure of target models, thereby consisting of global post-attentive features $\tilde{\mathbf{F}}_{gb}$ and local post-attentive features $\tilde{\mathbf{F}}_{lc}$. Each one also consists of appearance component $\tilde{\mathbf{F}}^a$ and salience component $\tilde{\mathbf{F}}^s$.

Appearance components are estimated using the statistics of corresponding pre-attentive features.

Salience components are estimated using the conspicuity values of corresponding feature dimensions. The conspicuity values are calculated from the bottom-up attention mechanism [16]. At first, center-surround differences in terms of each feature dimension are calculated:

$$F'_f(c, s) = |F_f(c) \ominus F_f(s)| \qquad (2)$$

where $\ominus$ represents across-scale subtraction, $c = \{0, 1, 2\}$ and $s = c + \delta$ with $\delta = \{2, 3\}$ represent the center scale and surround scale respectively, and $f \in \{int, rg, by, o_\theta, ct\}$.

Those center-surround differences in terms of the same feature dimension are normalized and combined together at scale 2, termed as *working scale*, using across-scale addition to yield a location-based conspicuity map F^s of that feature dimension:

$$F^s_f = \mathcal{N}\left(\frac{1}{6}\bigoplus_{c=0}^{2}\bigoplus_{s=c+2}^{c+3}\mathcal{N}\big(F'_f(c, s)\big)\right) \qquad (3)$$

where $\mathcal{N}$ denotes the normalization operator, $\bigoplus$ is across-scale addition, consisting of reduction of each normalized center-surround difference to the working scale and point-by-point addition, and $f \in \{int, rg, by, o_\theta, ct\}$.

It can be seen that the conspicuity values can be used to measure discriminability between the target and the background in the entire scene. Thus the task-relevant feature deduced from them is globally discriminative.

1) Global Post-attentive Feature: This paper uses B-Spline techniques [22] to represent a contour curve $\mathbf{C}$:

$$\mathbf{C} = f_c(\mathbf{W}\mathbf{X} + \mathbf{Q}_0) \tag{4}$$

where

$$\mathbf{Q}_0 = \begin{pmatrix} \mathbf{Q}_0^x \\ \mathbf{Q}_0^y \end{pmatrix} = \begin{pmatrix} x_1, x_2, ..., x_P \\ y_1, y_2, ..., y_P \end{pmatrix} \tag{5}$$

where (x_1, y_1),...,(x_P, y_P) are coordinates of control points along the contour, and P is the number of control points.

The control point vector $\mathbf{Q}_0$ characterizes the target's basic shape, which is the metric for shape discrimination between the target and distractors. The state vector $\mathbf{X}$ represents the spatial transformation (e.g., translation, rotation and scaling) of a contour instance $\mathbf{C}$ with respect to $\mathbf{Q}_0$. Using $\mathbf{Q}_0$ and $\mathbf{X}$ together, the shape of an object can be described. Since $\mathbf{X}$ is estimated based on the actual observation, this paper builds the global coding $\mathbf{O}_{gb}$ by using control point vector $\mathbf{Q}_0$.

At the beginning of learning, control points along the complete target region are extracted at the working scale by using the method proposed in our previous work [23]. The global post-attentive feature at each extracted control point is finally estimated at working scale, all of which comprise a set denoted as $\{\tilde{\mathbf{F}}_{gb}\}$:

$$\tilde{\mathbf{F}}_{gb} = (\tilde{F}_x^a, \tilde{F}_y^a, \tilde{F}_{ct}^s)^T \tag{6}$$

where appearance components $\tilde{F}_x^a$ and $\tilde{F}_y^a$ are spatial coordinates of a control point, and the salience component $\tilde{F}_{ct}^s$ is the average of conspicuity values in terms of contour feature (i.e., F_{ct}^s) around the control point. Each entry is sent into the learning routine sequentially.

2) Local Post-attentive Feature: It is also a set, denoted as $\{\tilde{\mathbf{F}}_{lc}\}$. Each entry of that set consists of statistical appearance and salience values in terms of intensity, red-green, blue-yellow and local orientations within a proto-object belonging to the complete target region, denoted as:

$$\tilde{\mathbf{F}}_{lc} = (\tilde{F}_{int}^a, \tilde{F}_{rg}^a, \tilde{F}_{by}^a, \tilde{F}_{o_\theta}^a, \tilde{F}_{int}^s, \tilde{F}_{rg}^s, \tilde{F}_{by}^s, \tilde{F}_{o_\theta}^s)^T \tag{7}$$

Each $\tilde{\mathbf{F}}_{lc}$ is also sent into the learning routine sequentially.

Appearance and salience components of intensity, red-green and blue-yellow are averages of pre-attentive feature values F and conspicuity values F^s respectively in terms of the corresponding feature dimension within the proto-object. The appearance component of local orientations is a histogram with 4 bins, each of which represents an orientation direction. Each pixel within the proto-object is accumulated into the corresponding bin according to its orientation direction. The salience component of local orientations is a 4×1 mean vector. For each pixel within the proto-object, its conspicuity value F^s in terms of orientation energy in its orientation direction is accumulated to the corresponding entry of that mean vector.

C. PNNs based Target Models

In order to improve robustness, probabilistic neural networks (PNNs) [24] are used to model the probabilistic distributions of the local coding and global coding respectively. Both PNNs have the identical structure, consisting of three layers. The input layer receives $\tilde{\mathbf{F}}_{lc}$ or $\tilde{\mathbf{F}}_{gb}$. The hidden layer is composed of radial basis functions (RBFs), each of which represents the distribution of a local parts or a global control point of that target. The output layer is the mixture distribution by combination of all local parts or global control points of that target.

The RBFs in the hidden layer are represented by a multi-dimensional Gaussian distribution $\mathcal{G}$:

$$q_n(\tilde{\mathbf{F}}) = \mathcal{G}(\tilde{\mathbf{F}}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \tag{8}$$

where $\boldsymbol{\mu}_n = (\boldsymbol{\mu}_n^a, \boldsymbol{\mu}_n^s)^T$ and $\boldsymbol{\Sigma}_n$ denote the mean vector and covariance matrix of a RBF indexed by n and $\tilde{\mathbf{F}} \in \{\tilde{\mathbf{F}}_{lc}, \tilde{\mathbf{F}}_{gb}\}$.

Assuming that feature distributions are independent, standard deviation (STD) vector $\boldsymbol{\sigma}_n = (\boldsymbol{\sigma}_n^a, \boldsymbol{\sigma}_n^s)^T$ can be obtained from the diagonal entries of covariance matrix $\boldsymbol{\Sigma}_n$. The mean vector and STD vector of a RBF in local PNNs can be also represented as: $\boldsymbol{\mu}_n^{a/s} = (\mu_{n,int}^{a/s}, \mu_{n,rg}^{a/s}, \mu_{n,by}^{a/s}, \mu_{n,o_\theta}^{a/s})^T$, $\boldsymbol{\sigma}_n^{a/s} = (\sigma_{n,int}^{a/s}, \sigma_{n,rg}^{a/s}, \sigma_{n,by}^{a/s}, \sigma_{n,o_\theta}^{a/s})^T$, where a/s means a or s. The mean vector and STD vector of a RBF in global PNNs can be also represented as: $\boldsymbol{\mu}_n^a = (\mu_{n,x}^a, \mu_{n,y}^a)^T$, $\boldsymbol{\mu}_n^s = \mu_{n,ct}^s$, $\boldsymbol{\sigma}_n^a = (\sigma_{n,x}^a, \sigma_{n,y}^a)^T$, $\boldsymbol{\sigma}_n^s = \sigma_{n,ct}^s$.

The output layer of local or global PNNs is:

$$p(\mathbf{O}) = p(\tilde{\mathbf{F}}) = \sum_n \pi_n q_n(\tilde{\mathbf{F}}) \tag{9}$$

where $\mathbf{O} \in \{\mathbf{O}_{lc}, \mathbf{O}_{gb}\}$ and π_n is the weight of a RBF. π_n is estimated based on the occurrence rate of a RBF and thereby can represent the activity of a RBF. The condition $\sum_n \pi_n = 1$ is imposed on all RBFs in either local or global PNNs.

D. Online Learning

Since the number of RBFs might be dynamically changed during tracking, this paper proposes an online incremental learning algorithm by using both maximum likelihood estimation (MLE) and a Bayes' classifier.

The learning algorithm regards a control point or a proto-object as a pattern. A Bayes' classifier is developed to classify the training pattern to an existing RBF in terms of appearance components. In the classifier, prior probabilities are set identical for all existing RBFs. If the training pattern is labeled, both appearance and salience descriptors of the corresponding existing RBF are updated based on MLE; Otherwise, a new RBF is created. Two thresholds τ_1 and τ_2 are introduced to determine the minimum correct classification probability to an existing RBF of local PNNs and global PNNs respectively. Meanwhile, τ^- is also introduced to avoid misclassifications. Thus it is used to determine the STD once a new RBF is created and to adjust the STD of all RBFs after each learning routine. In order to keep track of the most recent target's states, some inactive RBFs are discarded. τ_π is the pre-defined threshold to determine whether a RBF is active or inactive.

Algorithm 1 Online learning routine of PNNs

Given a local or global pattern $\tilde{\mathbf{F}} \in \{\tilde{\mathbf{F}}_{lc}, \tilde{\mathbf{F}}_{gb}\}$:
$\forall n$: Calculate $q = q_n^a(\tilde{\mathbf{F}}_{lc}^a)$ or $q = q_n^a(\tilde{\mathbf{F}}_{gb}^a)$
if $\exists n$: q is maximal and $q \geq \tau_1$ or $q \geq \tau_2$ **then**
  // Update the existing RBF indexed by n
  $(\boldsymbol{\sigma}_n)_{temp} = \left[(\boldsymbol{\sigma}_n).^2 + (\boldsymbol{\mu}_n).^2 + (\tilde{\mathbf{F}}).^2\right]/2$;
  $\boldsymbol{\mu}_n = (\tilde{\mathbf{F}} + \boldsymbol{\mu}_n)/2$;
  $\boldsymbol{\sigma}_n = \left[(\boldsymbol{\sigma}_n)_{temp} - (\boldsymbol{\mu}_n).^2\right].^{-\frac{1}{2}}$;
  $\pi_n = \pi_n \times 2$;
else
  // Create a new RBF
  $N = N + 1$ or $P = P + 1$; Set $n = N$ or $n = P$;
  $\boldsymbol{\mu}_n = \tilde{\mathbf{F}}$; $\pi_n = \min_{1 \leq j \leq n-1} \{\pi_j\}$;
  $\forall f$: $\sigma_{n,f}^{a/s} = \min_{1 \leq j \leq n-1} \left\{ \sqrt{-|\mu_{n,f}^{a/s} - \mu_{j,f}^{a/s}|^2/\tau^-} \right\}$;
end if
// Adjust STD of all RBFs
$\forall n, \forall f$: $\sigma_{n,f}^{a/s} = \min \left\{ \sigma_{n,f}^{a/s}, \sqrt{-|\tilde{F}_f^{a/s} - \mu_{n,f}^{a/s}|^2/\tau^-} \right\}$
$\forall n$: $\pi_n = \pi_n / \sum_j \pi_j$; *// Normalize weights of RBFs*
for n=1:N or n=1:P **do**
  if $\pi_n < \tau_\pi$ **then**
    Discard the n-th RBF; *// Discard inactive RBFs*
  end if
end for
$\forall n$: $\pi_n = \pi_n / \sum_j \pi_j$; *// Re-normalize weights of RBFs*

Algorithm 1 shows the learning routine of either local or global PNNs. In the algorithm, q_n^a is the probability in terms of appearance in the RBF level, N and P respectively represent the number of existing local RBFs and global RBFs, a/s means a or s, $f \in \{int, rg, by, o_\theta\}$ for local PNNs, $f \in \{x, y\}$ or $f \in \{ct\}$ for global PNNs, and $.^2$ denotes element-to-element vector square.

V. Tracking Process

A. Spatial Top-down Biasing

Due to the variation of target's motion, the target dynamics is difficult to estimate. Thus this paper only predicts a large region centered at the target position at the last moment as the predicted region. Appearance top-down biasing is performed in that region.

B. Appearance Top-down Biasing

1) Automatical Feature Selection: The proposed tracking method uses the target's salience descriptor to automatically select a task-relevant feature dimension. This is implemented by finding out a feature dimension that has great salience:

$$(f_{rel}, j_{rel}) = \arg \max_{f \in \{ct, int, rg, by, o_\theta\}} \max_j \frac{\mu_{j,f}^s}{1 + \sigma_{j,f}^s} \quad (10)$$

where f_{rel} is the task-relevant feature dimension and j_{rel} is the index of the task-relevant part. For local features $f \in \{int, rg, by, o_\theta\}$: $j \in \{1, 2, \ldots, N\}$. For the global feature $f \in \{ct\}$: $j \in \{0\}$, $\mu_{0,ct}^s = \sum_{n=1}^P \mu_{n,ct}^s / P$, and $\sigma_{0,ct}^s = \left\{ \sum_{n=1}^P \left[(\sigma_{n,ct}^s)^2 + (\mu_{n,ct}^s)^2\right]/P - (\mu_{0,ct}^s)^2 \right\}^{-\frac{1}{2}}$.

2) Attentional Template: The target's appearance descriptor in terms of the task-relevant feature dimension is used to build an attentional template so as to estimate appearance top-down biases.

If f_{rel} is contour, an attentional template $\mathbf{F}_{ct}^t$ is built as:

$$\mathbf{F}_{ct}^t = \begin{pmatrix} \mu_{1,x}^a, \ldots, \mu_{P,x}^a, \mu_{1,y}^a, \ldots, \mu_{P,y}^a \\ \sigma_{1,x}^a, \ldots, \sigma_{P,x}^a, \sigma_{1,y}^a, \ldots, \sigma_{P,y}^a \end{pmatrix}^T \quad (11)$$

If f_{rel} is intensity, red-green or blue-yellow, an attentional template $\mathbf{F}_{int}^t$, $\mathbf{F}_{rg}^t$ or $\mathbf{F}_{by}^t$ is built as:

$$\mathbf{F}_f^t = \begin{pmatrix} F_f^{t,\mu} & F_f^{t,\sigma} \end{pmatrix} = \begin{pmatrix} \mu_f^{a,j_{rel}} & \sigma_f^{a,j_{rel}} \end{pmatrix} \quad (12)$$

where $f \in \{int, rg, by\}$.

If f_{rel} is the orientation in a direction θ, an attentional template F_o^t is built using that direction:

$$F_o^t = \theta \quad (13)$$

3) Estimation of Top-down Biases: If the task-relevant feature dimension is contour, a probabilistic method inspired from active contour techniques [22] is proposed to estimate the bias. Details of the biasing algorithm in terms of contour can be seen in our previous work [23]. A location-based top-down bias map in terms of contour is finally achieved as:

$$B_{ct}(\mathbf{r}_i) = \begin{cases} 1 & \text{if } \mathbf{r}_i \in \mathbf{R}_{g_{max}} \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where B represents the top-down bias, $\mathbf{r}_i$ is an image pixel at working scale, and $\mathbf{R}_{g_{max}}$ is the proto-object that has the maximal posterior probability.

If the task-relevant feature dimension is intensity, red-green, blue-yellow and orientation, location-based top-down biases are estimated respectively as:

$$B_f(\mathbf{r}_i) = \exp(-\frac{1}{2} \frac{|F_f(\mathbf{r}_i) - F_f^{t,\mu}|^2}{(F_f^{t,\sigma})^2}) \quad (15)$$

where $f \in \{int, rg, by\}$.

$$B_{o_\theta}(\mathbf{r}_i) = \begin{cases} F_{o_\theta}(\mathbf{r}_i)/255 & \text{if } \theta = F_o^t \\ 0 & \text{Otherwise} \end{cases} \quad (16)$$

4) Proto-object based Attentional Activation Map: This map, denoted as S, is obtained by combination of location-based appearance top-down biases within each proto-object:

$$S(\mathbf{R}_g) = \sum_{\mathbf{r}_i \in \mathbf{R}_g} B_{f_{ref}}(\mathbf{r}_i)/N_g \quad (17)$$

where N_g is the number of pixels in the proto-object $\mathbf{R}_g$. The focus of attention (FOA) is directed to the proto-object that has the maximal proto-object based activation.

C. Validation

The validation procedure is based on Bayes' theorem using the appearance component of the local post-attentive feature $\tilde{\mathbf{F}}_{lc}^a$ of the attended proto-object. If the probability $p^a(\tilde{\mathbf{F}}_{lc}^a)$ in terms of appearance is above the predefined threshold τ_v, the attended proto-object is confirmed. Otherwise, the recovery mechanism is triggered to carry out another appearance top-down biasing procedure over the entire image.

D. Post-attentive Completion Processing

This paper proposes that the precise and complete target region can be achieved using local and global descriptors of the target model around the attended proto-object. The detailed routine can be seen in our previous work [20].

VI. Experiments

This proposed tracking method is tested in three tasks. Performance of our method is also compared with CamShift algorithm [25]. Three videos are obtained by a moving robot in three scenes with different settings.

A. Tasks

The first task is to track one moving human (i.e., target) in scene 1, in which the background shares some features with the tracked target. The objective of this task is to show the adaptivity of our method in the sense that it can adaptively track the object by automatically selecting a discriminative feature. The online learned salience descriptors of the target in scene 1 are shown in Fig. 2(a), which indicates that contour is the task-relevant feature. The tracking result of our method is shown in Fig. 3(e) - 3(h): Our method succeeds to track the target when it passes by the red board. Results of Camshift algorithm are shown in Fig. 3(i) - 3(l): It fails to track the target when it is passing by the red board, since the red board share hue values with the target.

The second task is to track one moving human (i.e., target) in scene 2, in which there is full occlusion during several sequential frames. The objective of this task is to show that our method can automatically recover the tracking after the full occlusion. The learned salience descriptors of the target in scene 2 are shown in Fig 2(b), which indicates that contour is the task-relevant feature. The tracking result of our method is shown in Fig. 4(e) - 4(h): Our method succeeds to track the target after it goes through the full occlusion. Results of Camshift algorithm are shown in Fig. 4(i) - 4(l): The tracking region covers almost the whole scene after the target goes through the occlusion, so CamShift algorithm fails to recover the tracking after the full occlusion.

The third task is to track one moving human (i.e., target) in scene 3 in which there is another moving human (i.e., distractor). One objective of this task is to show that our method is robust to variations of lighting on the target. The other objective is to show our method can provide the completion of the tracked target that includes several parts. The learned salience descriptors of the target in scene 3 are shown in Fig. 2(c), which shows that red-green of the part 2 (i.e., the upper body of the target) is the task-relevant feature. The tracking result of our method is shown in Fig. 5(e) - 5(h): Our method succeeds to track the target and achieves target completion. Results of Camshift algorithm are shown in Fig. 5(i) - 5(l): It fails to track the target when the target passes by the blue door (Fig. 5(l)).

B. Performance Evaluation

Tracking precision P_{TPR} is one type of performance evaluation. It is calculated as a true positive rate: $P_{TPR} = nTP/nP$ where nTP is the number of frames in which the target is correctly detected and nP is the total number of frames in a video.

Target completion is another type of performance evaluation. It is calculated by using both true positive rate $C_{TPR} = A_{TP}/A_{real}$ and false positive rate $C_{FPR} = A_{FP}/A_{real}$, where A_{real} is the pixel number of the real target, A_{TP} is the number of pixels that are both in the tracked region and in the real target, and A_{FP} is the number of pixels that are in the tracked region but not in the real target.

Performance evaluation of our method and Camshift algorithm is shown in Table I. It can be seen that the tracking performance and target completion performance in our method are both better than those in CamShift algorithm.

VII. Conclusions

This paper has presented a target tracking method using object-based visual attention mechanism. Compared with other tracking methods, this proposed method is capable to cope with the difficulties including appearance changes of the background and the target, large variation of motion, partial and full occlusion and so on.

TABLE I
Performance Evaluation

Task	Method	Frm #	P_{TPR}	C_{TPR}	C_{FPR}
1	Ours	44	100.00 %	92.71 %	6.70 %
	CamShift	44	11.36 %	39.34 %	2.09 %
2	Ours	42	64.29 %	91.60 %	8.13 %
	CamShift	42	26.19 %	36.38 %	3.32 %
3	Ours	65	96.92 %	97.80 %	2.48 %
	CamShift	65	80.00 %	92.50 %	5.09 %

References

[1] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.

[2] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," in *IEEE Int. Conf. CVPR*, 2001, pp. 415–422.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, 2003.

[4] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *IEEE Int. Conf. CVPR*, 1998, pp. 232–237.

[5] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comp. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.

[6] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 10, pp. 1025–1039, 1998.

[7] R. T. Colins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1631–1643, 2005.

[8] Z. Yin and R. Collins, "Spatial divide and conquer with motion cues for tracking through clutter," in *IEEE Conf. CVPR*, 2006, pp. 570–577.

[9] N. Bouaynaya and D. Schonfeld, "A complete system for head tracking using motion-based particle filter and randomly perturbed active contour," in *SPIE, Image and Video Communications and Processing*, 2005, pp. 864–873.

[10] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low-frame-rate video: A cascade particle filter with discriminative observers of different lifespans," in *IEEE Conf. CVPR*, 2007, pp. 1–8.

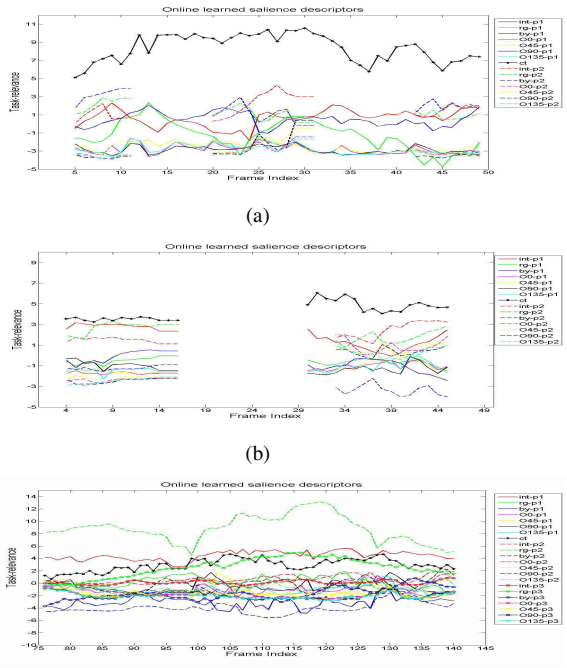

Fig. 2. Online learned salience descriptors of the target representations in the three tasks. (a) In task 1; (b) In task 2; (c) In task 3.

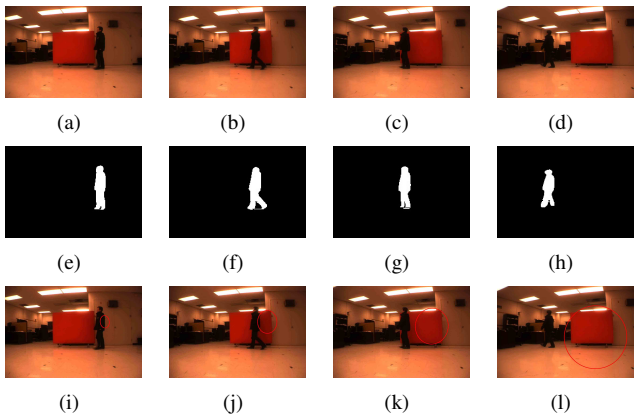

Fig. 3. Tracking results of task 1: Tracking of a moving human by the moving robot in scene 1, in which the background shares some features with the target. (a)-(d) Original images from video 1; (e)-(h) The complete tracking regions obtained from the proposed method; (i)-(l) Tracking results using CamShift algorithm; Red ellipses represent the tracking regions.

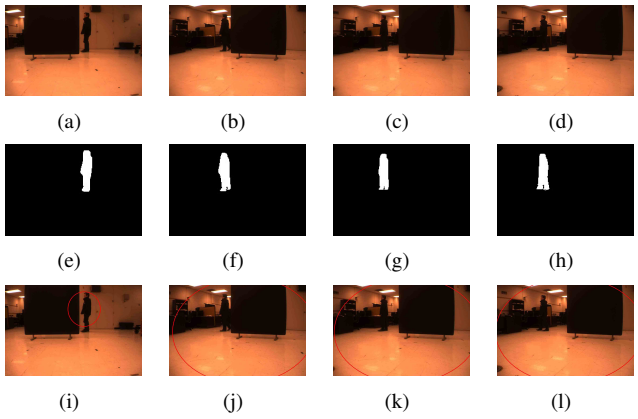

Fig. 4. Tracking results of task 2: Tracking of a moving human by the moving robot in scene 2, in which the full occlusion exists. (a)-(d) Original images from video 2; (e)-(h) The complete tracking regions obtained from the proposed method; (i)-(l) Tracking results using CamShift algorithm; Red ellipses represent the tracking regions.

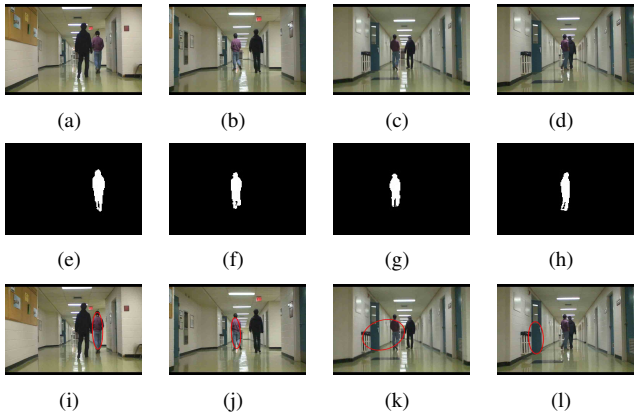

Fig. 5. Tracking results of task 3: Tracking of a moving human by the moving robot in scene 3. (a)-(d) Original images from video 3; (e)-(h) The complete tracking regions obtained from the proposed method; (i)-(l) Tracking results using CamShift algorithm; Red ellipses represent the tracking regions.

[11] Y. Lao, J. Zhu, and Y. F. Zheng, "Sequential particle generation for visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 9, pp. 1365–1378, 2009.

[12] J. Duncan, "Selection attention and the organization of visual information," *J. Exp. Psych.: Gen.*, vol. 113, pp. 501–517, 1984.

[13] J. Duncan, G. Humphreys, and R. Ward, "Competitive brain activity in visual attention," *Curr. Opin. Neurobiology*, vol. 7, pp. 255–261, 1997.

[14] S. Frintrop and M. Kessel, "Most salient region tracking," in *Proc. IEEE ICRA*, 2009, pp. 1869–1874.

[15] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An object-based visual attention model for robots," in *Proc. IEEE ICRA*, 2008, pp. 943–948.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998.

[17] A. G. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson, "Overcomplete steerable pyramid filters and rotation invariance," in *IEEE Int. Conf. CVPR*, 1994, pp. 222–228.

[18] A. Montanvert, P. Meer, and A. Rosenfeld, "Hierarchical image analysis using irregular tessellations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 307–316, 1991.

[19] E. Sharon, A. Brandt, and R. Basri, "Fast multiscale image segmentation," in *IEEE Int. Conf. CVPR*, 2000, pp. 70–77.

[20] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An autonomous visual perception model for robots using object-based attention mechanism," in *IEEE Int. Conf. Robotics and Biomimetics*, 2009, pp. 1474–1479.

[21] A. N. Redlich, "Redundancy reduction as a strategy for unsupervised learning," *Neural Computation*, vol. 5, no. 2, pp. 289–304, 1993.

[22] A. Blake and M. Isard, *Active contour*. Springer, 1998.

[23] Y. Yu, G. K. I. Mann, and R. G. Gosine, "Modeling of top-down influences on object-based visual attention for robots," in *IEEE Int. Conf. Robotics and Biomimetics*, 2009, pp. 1021–1026.

[24] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, pp. 109–118, 1990.

[25] G. R. Bradski, "Compter vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, no. Q2, 1998.