Pose Evaluation Based on Bayesian Classification Error

Qifeng Liu, Jianguang Lou, Weiming Hu, Tieniu Tan
National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences
{qfliu,jglou,wmhu,tnt}@nlpr.ia.ac.cn

Abstract

Pose evaluation is a fundamental issue in image processing and computer vision. In this paper, we propose a new method called BCE for pose evaluation based on Bayesian classification error. Various image cues are incorporated to depict an object including object shape, side region statistics and temporal information. Then a PEF (Pose Evaluation Function) is constructed based on Bayesian classification error, and an efficient algorithm to calculate it is developed. We test our new method with real outdoor image sequences, and use two criteria to compare it with two other representative ones. It is shown that our new method leads to better performance with respect to localization accuracy and robustness against general clutter and occlusion.

1 Introduction

Tracking can be cast as an attempt to find the optimal estimation of the state in a discrete-time dynamic system [11]. Let X_t be the state of the object and Z_t the observation at time t. From the Bayesian point of view, it is required to construct the *a posteriori* probability of X_t:

$$p(X_t \mid Z_{1:t}) = p(Z_t \mid X_t)p(X_t \mid Z_{1:t-1})/p(Z_t \mid Z_{1:t-1}) \tag{1}$$

where $Z_{1:t}$ is $\{Z_i \mid i = 1,\ldots,t\}$, $p(Z_t \mid X_t)$ is the image likelihood, $p(X_t \mid Z_{1:t-1})$ is the prediction of X_t, and $p(Z_t \mid Z_{1:t-1})$ is typically treated as a constant. So tracking is a maximum *a posteriori* (MAP) problem to search X_t which maximizes $p(X_t \mid Z_{1:t})$.

There are two main classes of approaches computing the optimal Bayesian solution: one is Kalman filter [6] and the other is particle filter [1]. For whatever tracking filters we use, the image likelihood $p(Z_t \mid X_t)$ is absolutely necessary. We always adopt some metrics, instead of the image likelihood, to evaluate the goodness-of-fit between the given state and the image data. This process is also called pose evaluation. Pose evaluation is very important because it can heavily affect the performance of the tracker in terms of accuracy, robustness and efficiency. With respect to accuracy, a simple but powerful argument is that if there is always a great bias between the extremum of the PEF and the true pose, what an elaborate update module or tracking filter can do is to make X_t stably converge at a wrong point.

The reason why we only focus on pose evaluation in this paper is that we think: although many excellent methods have been proposed for object tracking, it still seems that

BMVC 2003 doi:10.5244/C.17.42

pose evaluation has not received enough attention, and need to be studied further (i.e., about localization accuracy). In fact, pose evaluation is closely related with pose refinement and thus difficult to improve. A slight extension of the PEF may be more reasonable, but on the other hand, may sharply increase the difficulty of pose refinement. That is an important reason why most pose evaluation methods would rather only use simple metric such as distance error than exploit more information.

In this paper, we propose a method for pose evaluation based on Bayesian classification error (BCE). We depict an object using shape model, side region and temporal information. The concept of classification in pattern recognition is introduced here and then the PEF is constructed based on BCE. Using two criteria we test and compare the method with two other pose evaluation methods under various complex conditions, and it is shown that our new method leads to better performance with respect to localization accuracy and robustness against general clutter and occlusion.

2 Related Work

Here we mainly review the approaches for pose evaluation based on 3D models. There are two main categories: one is with geometric primitive extraction and the other without geometric primitive extraction. The former is classical and widely used [10, 7, 5, 15, 3]. The distance error can be in the form of point-to-point [5], point-to-line [15, 3], or line-to-line [10, 7]. However, the methods based on the distance error seem simple and intuitive and have inherent shortcomings: the process of primitive extraction and grouping is computationally expensive; the extracted primitives are easily unstable due to occlusion, clutter and noise, and so on.

To overcome the above shortcomings, a few excellent methods without primitive exaction have been proposed. In Iconic algorithm [14], directional derivatives of the image intensities besides the projected modal lines are combined into the pose evaluation function (PEF) in a statistical framework. This method is very fast but the curve of the PEF is not smooth and the PEF is hard to be refined by gradient-based optimization method. Alternatively, Kollnig et al. [8] estimate pose parameters by directly matching polyhedral models with the norms of synthetic gradients to image gradients. Haag et al. [4] improve their previous work by combining edge element and optical flow estimates. However, sometime optical flow constraint is violated because there are often strong light reflexes and shadows on the very smooth surfaces of vehicles. In addition, for investigating the limits of model-based methods, they do not emphasize the computational cost.

On contrast to 3D model based methods, there are much more strategies adopted by the approaches that work purely in the picture domain. We attempt to introduce some of these strategies into our 3D model based methods. There are two main 2D approaches. One is based on region information including texture or motion properties [13]. The other is based on boundary information, which is usually implemented using active contour models [12]. In addition, there are some valuable attempts in integrating the region and boundary information under a common framework [12, 16].

3 Object Representation

In object tracking, the class of methods based on 3D models provides a powerful and natural way to exploit prior knowledge of the shape and appearance of specific objects. We adopt this basic idea in our work [15, 14]. Camera models (i.e., intrinsic and extrinsic camera parameters) and object models (i.e., 3D geometric descriptions of vehicles) are established off-line. In model coordinate system, let $L^{(3)}$ be the set of 3D line segments that outlines the shape of the object. Given pose parameter P, $L^{(3)}$ is projected onto the 2D image plane to obtain a set of 2D line segments of the object $L^{(2)}$. Furthermore, $L^{(2)}$ is sampled at a fixed interval to obtain a set of points as shown in Figure 1: $S = \{S^{(i)} \mid i = 1,\ldots,M\}$. The relationship among $L^{(3)}$, $L^{(2)}$ and S can be summarized as the simplified mapping: $L^{(3)} \xrightarrow{P} L^{(2)} \xrightarrow{sampling} S$.

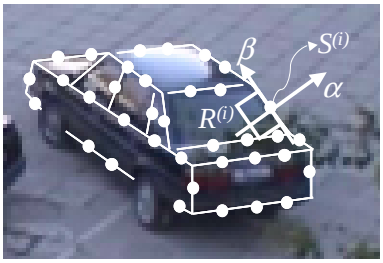


Figure 1: Object representation. White lines are the model projection $L^{(2)}$. For one sample point $S^{(i)}$, there is a rectangular $R^{(i)}$ inside the object. More information about the $\alpha - O - \beta$ coordinate system is indicated in Fig. 2.

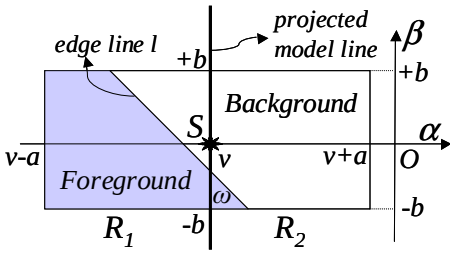


Figure 2: Information around S. S is a sample point shown in Fig. 1. The β-axis axis is parallel to the projected line from $L^{(2)}$, and the line l coincides with the image edge direction of the object near S. ω is the angle between two lines.

The statistic information of the regions located inside the boundary of the object is also considered. As shown in Figure 1, for each sample point $S^{(i)}$, there is a small rectangular region $R^{(i)}$ on the object. Because $R^{(i)}$ only occupies very small area of the object, we can give a reasonable assumption that there is no illumination change in $R^{(i)}$ and the variation of the intensity within $R^{(i)}$ is due to physical white noise. Therefore we can model the intensity within $R^{(i)}$ as a Gaussian distribution with mean $\mu^{(i)}$ and standard deviation $\sigma^{(i)}$. In RGB space, $\mu^{(i)}$ and $\sigma^{(i)}$ are in the form of matrix. So given the pose P, the projection of the object O_P can be represented by the following:

$$O_P = \{(S^{(i)}, \mu^{(i)}, \sigma^{(i)}) \mid i = 1,\ldots,N\} \tag{2}$$

Note that the size of $R^{(i)}$ should be chosen modestly, for too small size will be inadequate for training $(\mu^{(i)}, \sigma^{(i)})$, and too large size will violate the Gaussian assumption. Here we let the size of $R^{(i)}$ be 8*8 experientially.

4 Bayesian Classification Error

Now we focus on a point $S^{(i)}$ of O_P , and for simplicity all the superscripts (i) are omitted in this section. Figure 2 is obtained by enlarging the local area around the point $S^{(i)}$. We define the $\alpha - O - \beta$ coordinate system as shown in Figure 2. The β-axis axis is parallel to the projected model line from $L^{(2)}$, and the line l coincides with the image edge direction of the object near S. The coordinates of S are $(v, 0)$. R_1 and R_2 are two small rectangular regions beside S, and B and F represents the background and the foreground, respectively. Let $f(\alpha, \beta)$ be the intensity at (α, β), then $X_B = \{ f(\alpha, \beta) \mid (\alpha, \beta) \in B\}$, and X_F , X_{R_1}, and X_{R_2} have similar definitions. According to the analysis in Section 3, we assume that X_B and X_F are normally distributed with parameters (μ_B, σ_B) and (μ_F, σ_F), respectively.

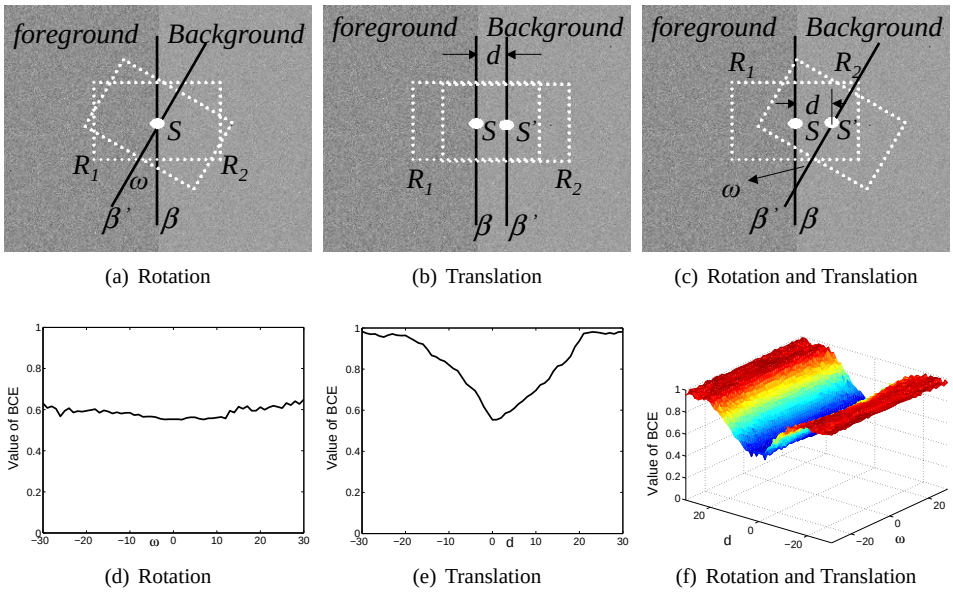
Now the key problem is how to measure the discriminabilities of X_{R_1} and X_{R_2} in a suitable metric. Here we apply hypothesis test and use Bayesian classification error (BCE) as the metric. This idea is similar to that of Basman's work on region segmentation and corer detection [2]. The difference lies in the definition of hypothesis. We define two hypotheses. Let H_0 be *null hypothesis* and H_1 *alternative hypothesis*.

$$\begin{cases} H_0 = \{S \in l \cap \omega = 0\} \quad or \quad H_0 = \{R_1 = F \cap R_2 = B \\ H_1 = \{S \notin l \cup \omega \neq 0\} \end{cases} \tag{3}$$

where ω is the angle between l and the axis β. H_0 has explicit meaning that the projected model line where S lies coincides with the image edge of the object, and therefore R_1 belongs to the foreground and R_2 background. We use BCE to measure the discriminabilities of X_{R_1} and X_{R_2}. It is clear that the discriminabilities reflect the confidence of H_0. Under hypothesis H_0, $X_{R_1} \sim N(\mu_1, \sigma_1)$ and $X_{R_2} \sim N(\mu_2, \sigma_2)$, where (μ_1, σ_1) and (μ_2, σ_2) are approximated by their unbiased estimates learnt from X_{R_1} and X_{R_2}. Thus BCE $E = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ would be calculated [11].

It is a fundamental issue that whether BCE E can be a reasonable metric for pose evaluation or not. In fact, given the same task and classifier, the value of BCE E is mainly determined by two conditions: the selection of the data set for training and the assumed statistical model of the distribution. In our problem, they are essentially related to the position of S, and further to the quality of the given pose P. There is a clear chain that the better the given pose P, the better the two conditions are satisfied, the more (μ_1, σ_1) and (μ_2, σ_2) are similar to (μ_B, σ_B) and (μ_F, σ_F), and finally the smaller BCE $E = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ is. So E reflects the reliability of H_0 and will converge at $E = (\mu_B, \sigma_B, \mu_F, \sigma_F)$ that is the minimum corresponding to the best point S and the best pose P.

The above analysis could explicitly be demonstrated by the experiments shown in Figure 3. First we synthesize an image consisting of two parts: the background and the foreground whose intensities are modelled as Gaussian distributions with parameters $(\mu_B = 145, \sigma_B = 15)$ and $(\mu_F = 160, \sigma_F = 10)$ respectively. Then we rotate the β-axis (marked by ω in Figure 3(a)), and translate the point S (marked by d in Figure 3(b)), and execute both rotation and translation simultaneously (shown in Figure 3(c)). We plot their BCE E shown in Figure 3(d), 3(e), and 3(f). From Figure 3(e), and 3(f), we can see that even when the foreground and background are very similar with lots of noise, the curves are still comparably smooth and have conspicuous peaks corresponding to the true pose exactly.



(a) Rotation (b) Translation (c) Rotation and Translation

(d) Rotation (e) Translation (f) Rotation and Translation

Figure 3: Simulated tests and results of BCE curves

5 PEF Construction

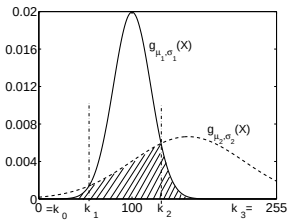


Figure 4: Bayesian classification error in a general case. The shadow area indicates the value of BCE

BCE is only theoretical but not practical to compute except in very special cases. Here we design an efficient scheme to tackle the problem of computing BCE. In general, BCE is formulated as:

$$E = \int_{\Re_1} p(X \mid R_2)P(R_2)dX + \int_{\Re_2} p(X \mid R_1)P(R_1)dX \tag{4}$$

where $\Re_1$ and $\Re_2$ are the decision areas using Bayesian classifier, and $P(R_1)$ and $P(R_2)$ are prior probability. In our problem, we have:

$$P(R_1) = P(R_2) = 0.5 \tag{5}$$

$$P(X \mid R_1) = g_{\mu_1,\sigma_1}(X) \qquad P(X \mid R_2) = g_{\mu_2,\sigma_2}(X) \tag{6}$$

$$g_{\mu,\sigma}(X) = (\sqrt{2\pi}\sigma)^{-1} exp\{-0.5(X-\mu)^2/\sigma^2\} \tag{7}$$

It is difficult to determine $\Re_1$ and $\Re_2$ in an explicit way. Here, we first analyze a special case (shown in Figure 4), and then give general procedures to dynamically compute E (shown in Table 1). If the curves of the two Gaussian functions have two cross points in $[0,255]$, we have $\Re_1 = [0,k_1] \cup [k_2,255]$ and $\Re_2 = [k_1,k_2]$, where k_1 and k_2 are the solutions of the equation $g_{\mu_1,\sigma_1}(X) = g_{\mu_2,\sigma_2}(X)$, which is equivalent to $AX^2 + BX + C = 0$ where

$$\begin{cases} A = \sigma_1^2 - \sigma_2^2 \\ B = 2\mu_1\sigma_2^2 - 2\mu_2\sigma_1^2 \\ C = 2\sigma_1^2\sigma_2^2 \ln(\sigma_2/\sigma_1) + \mu_2^2\sigma_1^2 - \mu_1^2\sigma_2^2 \end{cases} \tag{8}$$

In order to determine $\Re_1$ and $\Re_2$, two things must be done: one is to calculate the solutions k_1 and k_2; and the other is to make the relationship clear among $g_{\mu_1,\sigma_1}(X)$, $g_{\mu_2,\sigma_2}(X)$, 0, 255, k_1 and k_2. From Figure 4, we find a useful cue that the two functions are integrated by turns. According to that, we implement a dynamic method to determine the expression of E as shown in Table 1.

From the computational viewpoint, Equation 9 has at least two advantages:

- The basic component is $e = \int g_{\mu,\sigma}(X)$, which can be obtained directly by looking up a table and make pose evaluation efficient.
- $\partial e/\partial\mu$ and $\partial e/\partial\sigma$ can be worked out, so we can calculate the gradient of E for pose refinement.

Table 1: The procedures of computing E

Input: ($\mu_1, \sigma_1, \mu_2, \sigma_2$).
Output: BCE value E.
Step1: Compute k_1 and k_2 from $g_{\mu_1,\sigma_1}(X) = g_{\mu_2,\sigma_2}(X)$. Let $\eta \in \{1,2\}$ be the number of the solutions. Note that when $k_1 = k_2$, $\eta = 1$.
Step2: Maybe some solutions do not in $[0,255]$. Within $[0,255]$, there are $\xi \in \{0,\ldots,\eta\}$ solutions .Then we obtain: $k_0 = 0 < k_1 < \ldots < k_\xi < k_{\xi+1} = 255$.
Step3: Set $\lambda = k_1/2$, and if $g_{\mu_1,\sigma_1}(\lambda) > g_{\mu_2,\sigma_2}(\lambda)$,exchange μ_1 with μ_2 and σ_1 with σ_1.
Step4: Calculate:

$$E = \sum_{j=0}^{\xi} \int_{k_j}^{k_{j+1}} g_{\mu_1,\sigma_1}^{((j+1)mod2)}(X) * g_{\mu_2,\sigma_2}^{(jmod2)}(X) dX \tag{9}$$

So when j is an even number, the integrand is $g_{\mu_1,\sigma_1}(X)$, and when j is an odd number, the integrand is $g_{\mu_2,\sigma_2}(X)$.

Finally, under the widely used assumption [14, 4] that all the contour points are independent, the Bayesian Errors of these points in the set S are combined in the logarithmic way:

$$E(P) = \sum_{i=1}^{N} \ln(E^{(i)}) \tag{10}$$

Under the visual consistency assumption [12], that is the observed intensities of the same position of the object at time t and time $t-1$ should be the same if there is no illumination change, the temporal information is considered and Equation 10 can be extended into:

$$E(P_i) = \sum_{i=1}^{N} \ln(E_t^{(i)}(\mu_{1,t}^{(i)}, \sigma_{1,t}^{(i)}, \mu_{2,t}^{(i)}, \sigma_{2,t}^{(i)}) \cdot (1 - E_t^{(i)}(\mu_{1,t}^{(i)}, \sigma_{1,t}^{(i)}, \mu_{F,t-1}^{(i)}, \sigma_{F,t-1}^{(i)}))) \qquad (11)$$

where $(\mu_{F,t-1}^{(i)}, \sigma_{F,t-1}^{(i)})$ is obtained from the tracking result at time $t-1$.

6 Experiments

Figure 5: Real world scene

We apply our method to road vehicle localization in the context of visual surveillance. A part of the image sequence for test shown in Figure 5, is with complex conditions, e.g. illumination variation, clutter with foliage and branches, and occlusion by buildings and vehicles (The white curves indicate the object vehicles). The sequence is much more challenging than the standard one provided by PETS'2000. And it is similar to or even more complex than that in the previous work [4, 14, 15]. .

For comparison, we select two other methods: PLS [15] and Iconic [14]. The reason for this selection is that the former is our previous work based on primitive extraction, and the latter is the excellent work from the VIEWS group of Reading University based on no primitive extraction. We implemented the three algorithms (Iconic, PLS and BCE) by Visual C++ 6.0 on one 866MHz CPU.

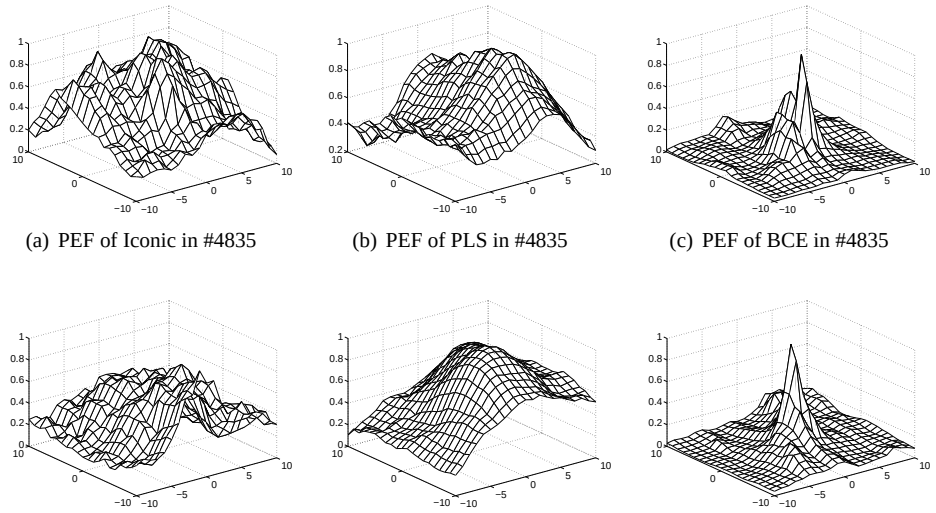
Under GPC (Ground Plane Constraint) [14], the pose P of the vehicle is comprised of transnational parameters x and y and one rotational parameter θ. The continuous PEF is computed at different poses, and then depicted as a discrete set:

$$\{E(P_{i,j,k}) \mid -10 \leq i, j, k \leq +10\} \qquad (12)$$

where $P_{i,j,k} = (x^* + i\lambda, y^* + i\lambda, \theta^* + k)^T$, and $P^* = (x^*, y^*, \theta^*)^T$ is the "true" pose determined by eye [14], and λ is the fixed interval for sampling the transnational parameters. In the 3D parameter space the set is rendered as a surface, which is shown in Figure 6.

We adopt the following two common criteria for comparing the performance of the three methods [9].One is computational cost: measured by the time T for which the PEF is computed once; the other is localization accuracy. Let the pose $P_E = (x_E, y_E, \theta_E)^T$ be the extremum of the PEF. And the localization error consists of two parts: translation error $E_T = \| (x^*, y^*)^T - (x_E, y_E)^T \|$ and orientation error $E_O = \| \theta^* - \theta_E \|$. The comparison results are shown in Figure 7, from which we could see:

- comparably accurate for localization (However, it is inevitable that BCE will fail when occlusions and clutter are serious as well as PLS and Iconic)
- quick for real time tracking;
- with conspicuous peak of its curve showing robustness against general occlusions and clutter.



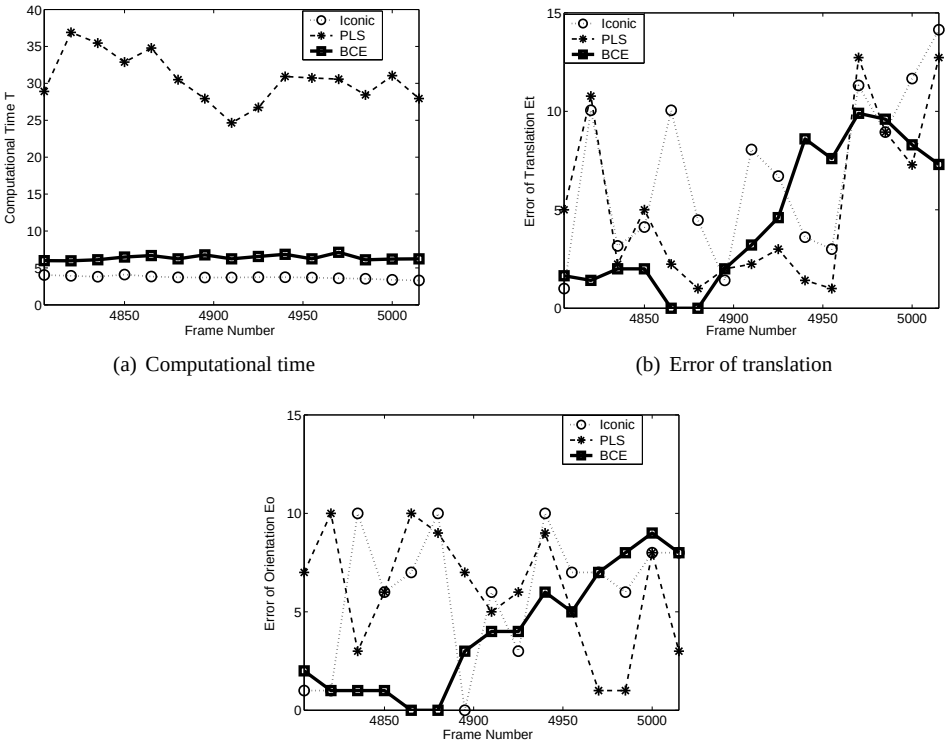
(a) PEF of Iconic in #4835 (b) PEF of PLS in #4835 (c) PEF of BCE in #4835

(d) PEF of Iconic in #4865 (e) PEF of PLS in #4865 (f) PEF of BCE in #4865

Figure 6: PEF surfaces in Frame 4835 and 6865. For each subfigure, the vertical axis is PEF value and the two horizontal axes are i and j shown in Equation 12. And for comparison, the PEF values of Iconic and PLS are normalized from 0 to 1 and then the curves of PLS and BCE are reversed.

7 Conclusions

In this paper, we have proposed a method for pose evaluation based on Bayesian classification error. Various image cues are incorporated to depict an object including object



(a) Computational time

(b) Error of translation

(c) Error of orientation

Figure 7: Comparison results from Frame 4805 to 5015. (a) shows that the computational cost of BCE is close to that of Iconic and much lower than that of PLS, and (b)(c) show that BCE is better than others in terms of localization accuracy with modest clutters.

shape, region statistics and temporal information. Then a novel PEF is defined based on BCE, and an efficient algorithm to calculate it is developed. We have tested our algorithm with outdoor image sequences and have compared it with two other representative methods. Both theoretical and experimental results have demonstrated the validity of the proposed method.

The statistical framework of most pose evaluation methods including ours is based on maximum likelihood estimation under the assumption that contour points are independent. Our future work will attempt to remove this assumption and focus on how to formulate the relations among several adjacent contour points and embed the relations into the PEF for better performance.

8 Acknowledgments

This work was supported by NSFC under Grant 69825105 and 60105002, Natural Science Foundation of Beijing under Grant 4031004, the National 863 High-Tech R/D Program of China under Grant 2002AA117010, and Institute of Automation, Chinese Academy of

Sciences under Grant 1M02J04.

References

[1] Arulampalam, Maskell, Gordon, and Clapp. A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing, Special Issue on Monte Carlo Methods*, 50(2):174–188, February 2002.

[2] A. Basman, J. Lasenby, and R. Cipolla. Applications of the creep-and-merge system: Corner detection. In *British Machine Vision Conference*, 1998.

[3] T. Drummonk and R. Cipolla. Real-time visual tracking of complex structures. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(7):932–946, 2002.

[4] M. Haag and H-H. Nagel. Combination of edge element and optical flow estimates for 3d-model-based vehicle tracking in traffic image sequences. *International Journal of Computer Vision*, 35(3):295–319, 1999.

[5] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[6] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transaction of the ASME-Journal of Basic Engineering*, 82:35–45, March 1960.

[7] D. Koller, K. Daniilidis, and H-H. Nagel. Model- based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 10(3):257–281, 1993.

[8] H. Kollnig and H-H. Nagel. 3d pose estimation by directly matching polyhedral models to gray value gradients. *International Journal of Computer Vision*, 23(3):283–302, 1997.

[9] Q. F. Liu, J. G. Lou, W. M. Hu, and T. N. Tan. Comparison of model-based pose evaluation algorithms in traffic scenes. In *International Conference on Image and Graphics*, Hefei, China, 2002.

[10] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[11] S. D. Ma and Z. Y. Zhang. *Computer Vision*. Science Press, Beijing, 1998.

[12] Nikos Paragios and Rachid Deriche. Geodesic active regions for motion estimation and tracking. In *IEEE International Conference in Computer Vision*, pages 688–694, Greece, 1999.

[13] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *International Conference on Computer Vision*, pages 1154–1160, Bombay, India, 1998.

[14] T. N. Tan, G. D. Sullivan, and K. D. Baker. Model-based localization and recognition of road vehicles. *International Journal of Computer Vision*, 27(1):5–25, 1997.

[15] H. Yang, J. G. Lou, H. Z. Sun, W. M. Hu, and T. N. Tan. Efficient and robust vehicle localization. In *International Conference on Image Processing*, 2001.

[16] Z. Y. Zhang, Y. Wu, and Z. C. Liu. Side statistics and maximum discriminant analysis for real time tracking. In *Asian Conference on Computer Vision*, Melbourne, 2002.