Visual Attention Based Motion Object Detection and Trajectory Tracking

Wen Guo[1,2], Changsheng Xu[1], Songde Ma[1], and Min Xu[3]

[1] National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] Shandong Institutes of Business and Technology,
Electronic Engineering Department, Yantai, China
{wguo,csxu,sdma}@nlpr.ia.ac.cn
[3] Faculty of Engineering and Information Technology,
University of Technology, Sydney, Australia
min.xu@uts.edu.au

Abstract. A motion trajectory tracking method using a novel visual attention model and kernel density estimation is proposed in this paper. As a crucial step, moving objects detection is based on visual attention. The visual attention model is built by combination of the static and motion feature attention map and a Karhunen-Loeve transform (KLT) distribution map. Since the visual attention analysis is conducted on object level instead of pixel level, the proposed method can detect any kinds of motion objects provided saliency without the affection of objects appearance and surrounding circumstance. After locating the region of moving object, the kernel density is estimated for trajectory tracking. The experimental results show that the proposed method is promising for moving objects detection and trajectory tracking.

Keywords: Visual attention, object detection, trajectory tracking.

1 Introduction

Object detection and tracking is a crowded research area in computer vision community. One of the challenges in object tracking is robust tracking, particularly under various complicated circumstances such as fast motion confrontation, illumination variation, shape deformation and object occlusion. These issues make it difficult to estimate and predict the next state of objects in video. To deal with these issues, various approaches such as mean shift [1], camshaft [2], sequence Karhunen Loeve particle filter [3] are proposed for object tracking. Moreover, the combination of different approaches, such as mean shift and Kalman filter, mean shift embedded particle filter [4], have provided reliable solutions for objects tracking.

Moving object detection is the first step of object tracking. Detection of moving regions provides a focus of attention for tracking. The typical method for moving object detection first models the complex background, and then subtracts the background from the input image to obtain foreground objects. Although the existing

G. Qiu et al. (Eds.): PCM 2010, Part II, LNCS 6298, pp. 462–470, 2010.

methods such as EM- GMM[5], PCA[6], adaptive kernel density estimate[7], have achieved good detection results, most of them cost high computational complexity and lack spatial relationships between neighboring regions. In the past few years, visual attention [8-9,12,13] is used to help object detection, segmentation, recognition and tracking. The basic idea is to use the bottom-up attention to extract useful information about the location, size and shape of objects from images to assist objects detection. However, these approaches are not always suitable for the situations with dynamic background.

In this paper, we propose a novel visual attention based moving object detection approach. We also use kernel density estimation with weight of visual attention to track the object trajectory. The contributions of our work are summarized as follows:

- A continuous symmetry difference of sequence is designed as motion saliency feature to reduce the discrete shadow around objects.
- A Karhunen-Loeve transformation (KLT) distribution map is calculated to obtain the stable static saliency feature to reduce affection by the noise such as illumination, flicker.
- The weight of target presentation based on motion attention detection is calculated for kernel density estimation tracker.

The rest of the paper is organized as follows. The proposed moving object detection and object motion trajectory tracking approaches are described in section 2 and section 3 respectively. The experimental results are reported in section 4. We conclude the paper with future work in section 5.

2 Moving Object Detection

2.1 Motion Attention Detection

Motion attention detection in image sequences aims at finding the regions of moving objects. Temporal analysis is designed for the objects having salient motion. In our work, a continuous symmetry difference of the images in a sequence is proposed to detect the motion attention. For an image $I_i(x,y)$ in a sequence S, we can construct an image set $\{\ldots, I_{(i-n)}(x,y), I_i(x,y), I_{(i+n)}(x,y),\ldots\}$, where two adjacent images have the identical interval. Then we can calculate the difference of adjacent images in the image set using the following formula:

$$dif_{(i,j)}(x,y) = | w * I_i(x,y) - w * I_j(x,y) | \tag{1}$$

where $j = i \pm n$, n is empirically set to 2 or 3 and w is a Gaussian filter function. In order to reduce output noise, values of difference less than a threshold ε are set to zero. In order to remove noise and avoid missing moving object, ε is set to 3 – 10.

$$Sal_{(i,j)}(x,y) = \begin{cases} 1, otherwise \\ 0, if\ dif_{(i,j)}(x,y) < \varepsilon \end{cases} \tag{2}$$

By implementing logical "AND" operation on $Sal_{(i,j)}(x,y)$, the motion saliency $Sal_m(x,y)$ is obtained as follows.

$$Sal_m = \begin{cases} 1, Sal_{(i,i-n)}(x,y) \,\&\&\, Sal_{(i,i+n)}(x,y) = 1 \\ 0, otherwise \end{cases} \tag{3}$$

Here we avoid the conventional average of the two maps in order to reduce the discrete shadow caused by motion. An example of motion attention detection is illustrated in Figure 1.

(a) (b) (c)

Fig. 1. Motion attention detection. (a) Original frame; (b) Simple difference map; (c) Motion saliency map

2.2 Static Attention Detection

Static attention detection aims at finding the conspicuous location of objects. Compared with motion attention detection, static attention should get more weights on the region. In a sense, it sounds like segmenting the object from the background. Similar to the spatial attention analysis method in [10], we use the contrast and information density to generate saliency map. Contrast is an object standing out in the local region. Information reflects objects importance in global background. A 2D difference of Gaussian (DoG) functions for every receptive region [11] is used to model the spatial summation properties of a center-surround cell [11].

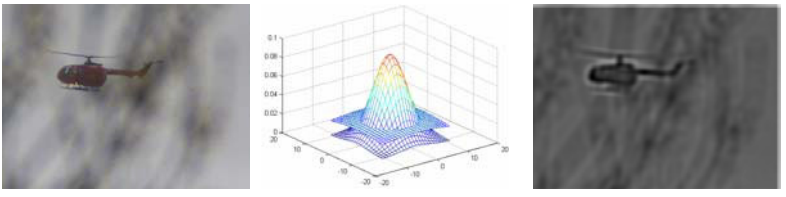
The input image is segmented into different regions using K-means, and the image patches can be obtained as a perceptive unit. If the images are simple, we can also use binarization for segmentation. Then, the neighboring pixels of same color are regarded as a receptive region. The contrast and information density can be calculated as follows:

$$\begin{cases} Con(k) = \sum_{i=1}^{k} d(f(k), f(i)) \times DoG_k \\ ID(x,y) = \log p(f(k)) \end{cases} \tag{5}$$

where k is the total number of the regions in the input image, the radius of $DoG_k(i,k)$ is same with the region, $f(k)$ is the feature of the region k, the function d and p represent the distance between the two features and probability of the features respectively. The static visual saliency is calculated as follows.

$$Sal_s = Con(k) \times ID(x,y) \tag{6}$$

Motion attention detection in an image of a helicopter flying through fog is illustrated in Figure 2.

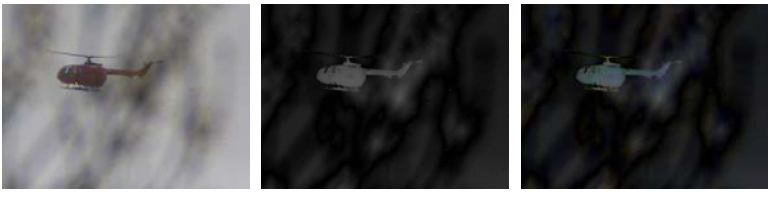


(a) (b) (c)

Fig. 2. Static attention detection. (a) Original frame; (b) 2D DoG map; (c) Static saliency map

2.3 Karhunen-Loeve Transformation (KLT) Distribution Map

Karhunen Loeve Transformation (KLT) is also known as Hotteling, which can reduce the computational cost and extract salient features of the images. For RGB image, every channel is easily affected by illumination variation, and there is high relationship among the channels. KLT can make the new component of the image orthogonally transformed or not related, so that the energy of the image is more stable. A vector set $X = [R_i, G_i, B_i]$ is used to represent an image. The mean vector $\overline{X}$ and covariance matrix Σ of X can be calculated. According to stochastic process theory, we can know the covariance matrix Σ is a 3*3 symmetric matrix. We can gain orthogonal eigenvalue Λ of the matrix and the corresponding eigenvector U, which satisfies the condition of KLT by calculating function $U\Sigma U^T = \Lambda$. X is transformed by $Y = U(X - \overline{X})$ and normalized into [0,1] to generate the KLT distribution map which is represented as $Map_{kl}(x, y)$. Since the eigenvector U is nonsingular matrix, inverse transform can be used to reconstruct image. One image by KLT processing is illustrated in Figure 3.



(a) (b) (c)

Fig. 3. KLT processing. (a) Original frame; (b) KLT distribution map; (c) Reconstruction using KLT

2.4 Visual Attention Region Detection

We have obtained three different feature maps from the previous sub-sections. These results should be combined to generate a final saliency map using an useful method.
If the background of the image sequence is basically still, the motion features play a decisive role. If the background is dynamic, the motion contrast is low in the sequence. The final salient region should pay more attention on spatial attention. We

treat the KLT distribution map equally with the spatial attention, unless the background is distorted by noise, such as smog, illumination, and flicker. Therefore, the final spatiotemporal saliency map of an image is calculated as:

$$Sal(x,y) = \alpha Sal_m + \beta Sal_s + \gamma Map_{kl} \tag{9}$$

The proposed visual attention region detection method can be illustrated in the following framework.

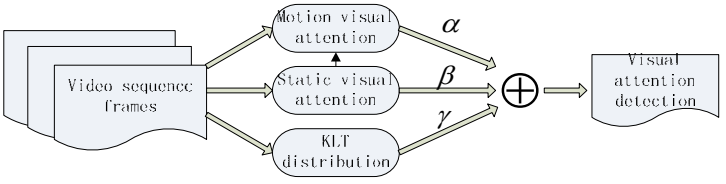


Fig. 4. Framework of visual attention region detection

Considering our goal is to detect the moving objects with background movement as little as possible, we set $\beta + \gamma < 0.5$, and $\alpha = 1 - \exp\{-\varepsilon v\}$, where εv is a constant related to the object moving speed, and all satisfy $\alpha + \beta + \gamma = 1$.

3 Object Motion Trajectory Tracking

Kernel density estimation (KDE) is to find the best kernel density estimations of candidate target near the real object representation by a similarity measure criterion. As a kernel density estimation method, mean shift [1] tracker uses an adaptive step gradient ascent algorithm to estimate a density function. In [1] a Bhattacharyya coefficient is used as a measure criterion to evaluate the similarity between the real and candidate target. Let A be a finite set of n dimensional space $\mathbb{R}$, the mean shift vector of x is achieved as follows.

$$m(x) = \frac{\sum_{i=1}^{n} x_i w_i K(K_i - x)}{\sum_{i=1}^{n} x_i K(K_i - x)} - x \tag{10}$$

where K is a kernel function, w_i is a weight function, and $x_i \in A, x \in \mathbb{R}$.

An efficient object representation is crucial for tracking. The kernel color histogram for target model $q = \{q_u\}_{u=1...m}$ and target candidate $p = \{p_u\}_{u=1...m}$ can be computed as follows:

$$I_u = C\sum_{i=1}^{n} k(\| x_i^* \|^2)\delta\left[b(x_i^*) - u\right] \tag{11}$$

where $I = \{p, q\}$, C is a normalization constant to ensue $\sum_{i=}^{m} q_u = 1$, and δ is the Kronecker delta function.

Considering the pixels near the edge of the target is vulnerable to background interference, in order to reduce the background effect, we use the motion saliency map to pay more attention on the object by large weight w_m, which can take full

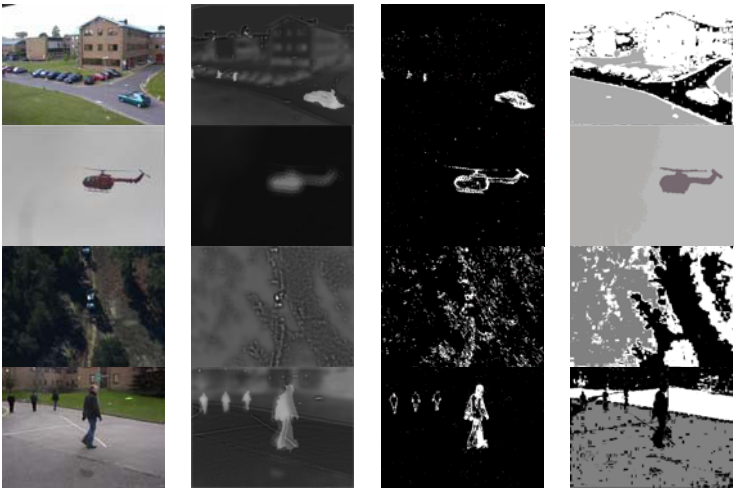
advantage of internal information of object instead of the background. The LoG operator is used to extract the edge of the motion saliency map and further to obtain the edge image $ESal_m(x, y)$. The weight is calculated as:

$$w_m = \begin{cases} 0.9, if\ Sal_m \&\& ESal_m = 1 \\ 0.1, if\ ESal = 1 \\ 0, otherwise \end{cases} \tag{12}$$

And the final target model is $q = w_m q$.

4 Experiments

To illustrate the effectiveness of the proposed method, we evaluate our approach using 200 image sequences from PETS, 10 from egtest and 50 from Internet. In the experiment, moving object is firstly detected using visual attention. Then the KDE method is used to track the object. Here we pick up several representative image sequences for illustration



(a) (b) (c) (d)

Fig. 5. Experimental results (a) original image (b) our results (c) frame difference (d) GMM

In Figure 5, since our approach is to incorporate attention analysis into frame difference, we compare our method with frame difference. We also compare our approach with GMM which can achieve good performance for still background. Figure 5 (a) shows the detection of four different objects, i.e. walking man, moving vehicle, a helicopter flight, and a car with moving background and tree shadow. From Figure 5, we can see that the object motion detection of our method is more credible than the compared results because the results by frame difference have too many discrete shadows and cannot adapt to the background movement, while GMM extracts the objects entangling too much background. It is worth nothing here that the

results are focusing on the visual attention of object detection instead of the object segmentation.

Figure 6 shows the results of two set image sequences of trajectory tracking for a walking man wearing the same color clothes with other people. The first row of Figure 6 is the tracking result in an image set without w_m . When two people overlap, the object lost because of the same color of their clothes. The second row of Figure 6 is the result using our approach. Since our approach considers the weight of the edge using visual attention, we can make use of the object shape to enhance tracking. The motion trajectory is shown in the last column (green dots in the first row and yellow dots in the second row).

We use the center coordinate of the tracking object window as the object trajectory. Ground truth data are generated using mean of several manually selected image regions of the objects. Figure 7 shows the ground truth and our tracking trajectory results (using x and y coordinate) for the video sequences in figure 6.

Fig. 6. The comparison of tracking results of people walking

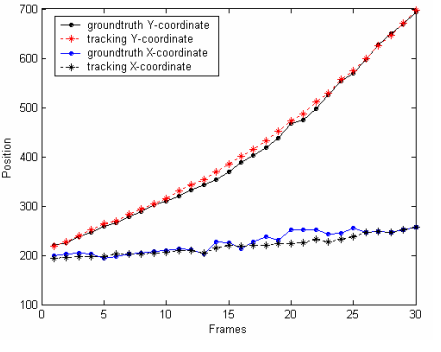


Fig. 7. Comparison results of ground truth trajectory and our tracking trajectory

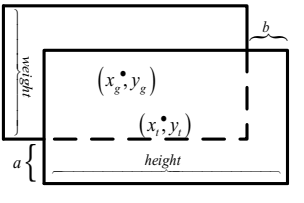


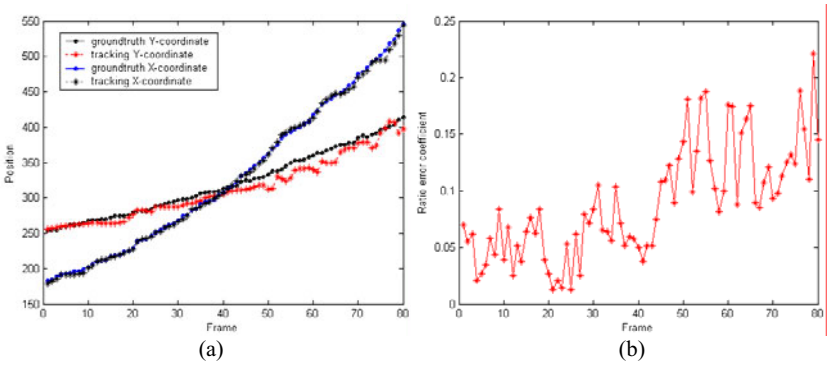
Fig. 8. Illustration of the window

For the quantitative evaluation, the tracking accuracy is evaluated by the ratio of departure area that tracking object window from the ground truth object region to the ground truth object region area. In Figure 8, weight and height are the scale of tracking window, while *a* and *b* are the absolute value of the center coordinate between ground truth and tracking window. The ratio error coefficient function is shown in (13). Figure 9 shows another example of trajectory tracking for a walking

woman and Figure 10 gives the evaluation result where Figure 10(a) is the trajectory tracking results and Figure (b) is the ratio error coefficient.

$$\rho = \frac{a*(height-b)+b*(weight-a)}{weight*height} \tag{13}$$

Fig.9. Tracking result of woman walking



(a) (b)

Fig.10. Tracking result of woman walking. (a) comparison of the trajectory tracking results, (b) ratio error coefficient

5 Conclusion

We have proposed a motion trajectory tracking method based on a novel visual attention model and kernel density estimation. In the visual attention detection, the motion saliency feature, static saliency feature and KLT distribution feature are calculated and combined to detect the object motion. The moving motion can be detected if it attracts more attention than any other parts. The motion saliency is used to improve the moving target model to make the internal information of object more reliable. In the future, we will investigate to incorporate the scale variation in the motion detection and the condition of occlusions in the trajectory tracking.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant No. 60970092, 60970105.

References

1. Comaniciu, D., Meer, P.: Mean-Shift: A Robust Approach toward Feature Space Analysis. IEEE Trans. PAMI, 1–18 (2002)
2. Bradski, G.: Computer vision face tracking as a component of perceptual user interface. In: WACV 1998, Princeton, NJ, pp. 214–219 (1998)
3. Levy, A., Lindenbaum, M.: Sequential Karhunen-Loeve basis extraction and its application to image. IEEE Tran. Image Processing 9, 1371–1374 (2000)
4. Shan, C., Tan, T., Wei, Y.: Real time hand tracking using a mean shift embedded particle filter. Pattern Recognition, 1958–1970 (2007)
5. Friedman, N., Russell, S.: Image segmentation in video sequences: A probabilistic approach. In: 13th Conf. Uncertainty in Artificial Intelligence, pp. 175–181 (1997)
6. Torre, F., Black, M.: Robust principal component analysis for computer vision. In: Proceeding of ICCV 2001, vol. 1, pp. 362–369 (2001)
7. Mittal, A., Paragios, N.: Motion based background subtraction using adaptive kernel density estimation. In: Proceeding of ICCV 2004, pp. 302–309 (2004)
8. Rutishauser, U., et al.: Is bottom-up attention useful for object recognition. In: Proceeding of ICCV, pp. 37–44 (2004)
9. Itti, L., Koch, C., Niebur, E.: A model for saliency based visual attention for rapid scene analysis. IEEE Trans. PAMI 20, 1245–1259 (1998)
10. Liu, H., Jiang, S., Huang, Q., Xu, C.: A Generic Virtual Content Insertion System Based on Visual Attention Analysis. In: Proceeding of the 16th ACMMM, pp. 379–388 (2008)
11. Kruizinga, P., Petkov, N.: Computational model of dot pattern selective cells. Biological Cybernetics 83(4), 313–325 (2000)
12. Zhang, G., Yuan, Z., Zheng, N., et al.: Visual saliency based object tracking. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009. LNCS, vol. 5994, pp. 246–257. Springer, Heidelberg (2010)
13. Michael, D., Martin, U., Martin, H., et al.: Saliency driven total variation segmentation. In: Proceeding of ICCV 2009, pp. 817–824 (2009)