Motion and Feature Attention Model Based Tracking for Moving Robots

Hong Liu, Huijun He

Key Laboratory of Machine Perception and Moving
Peking University, Shenzhen Graduate School, P.R.China
Email:{liuhong, hehj}@cis.pku.edu.cn

Abstract

Primates' visual system can implement various vision tasks effectively and efficiently by focusing the limited attention resources to relevant information. Motivated by this mechanism, we propose a motion and feature attention model based algorithm in this paper for tracking from moving robot. At first regions of motion attention are obtained by ego-motion compensation and frame difference image projection, and then the most attention-attracting color feature is adaptively selected for each frame as simulation of feature-based attention. Moreover, an attenuation function motivated by foveated vision and a heuristic template update strategy are introduced to increase robustness. Experiments show that our tracking algorithm woks robustly from moving robot under illumination variation conditions in various environments and can reach the speed of 20~25 fps running on 320*240 videos

Key words: visual tracking, moving robot, motion and feature attention, foveated vision

1 Introduction

Generally speaking, two major kinds of methods can be concluded, i.e. deterministic and probabilistic approaches. As the deterministic approach, Comaniciu et al.[1] proposed the mean-shift method which estimates the non-parametric density gradient based on color histogram. As the probabilistic approach, Isard et al. [2] proposed the CONDENSATION algorithm, also called particle filtering or bootstrap filtering.

At present, tracking from moving robot is becoming more and more important because of its many applications in robotic vision, moving vehicle, automated surveillance and so on. There are several high difficulties in such situation for example dramatic illumination variation, apparent motion of background, changing object appearance. Up to now, researchers have proposed several methods to address the problems. JieShao et al.[3] incorporate temporal differencing and shape detection for appearance-based tracking algorithm. They do motion compensation and then operate edge-based shape detection. In [4] the method is based on deformable shape model and a variant of condensation. But they have limitations: in [3], the temporal difference method has the well-know problem of foreground aperture. This will do much harm to the result of detection of candidate motion regions. [4] uses shape model which needs manual labeling and training is very time-consuming. Moreover it doesn't make use of the conspicuous and important color feature.

According to the studies in psychology[5], the mechanism of attention plays a crucial role during visual information processing. It can focus rapidly attention to local region of interest on which the higher level processing will operate. Several visual attention models such as object-based, feature-based, spatial-based have been proposed to explain and simulate primates' visual attention. It should be beneficial to take attention mechanism into account for finding the most discriminating features and attention-attracting regions during tracking.

In this paper, a novel visual attention based tracking algorithm is proposed. It is able to find the regions of motion attention and then adaptively select the most attention-attracting color channel as feature attention. In mode seeking procedure foveated vision is applied so that the attention distribution is highlighted like a spotlight in the target position and the farther it is from the center, the less attention there is. With time going,

object-based attention evolves in the response of the model neurons as a result of heuristic target template update strategy. The main contributions of our algorithm are summarized as follows.

1) Motivated by voice endpoint detection technology we propose a dual thresholds motional regions detection method based on the horizontal and vertical projections of the frame difference image, and by this method the well-known foreground aperture problem [7] in frame difference is overcome to some extent. 2) Effective and efficient feature attention based color channel selection method is proposed to get the most attention-attracting feature. 3) An exponential attenuation function motivated by foveated vision and a heuristic and adaptive template update strategy are introduced to increase the robustness.

The remainder of this paper is organized as follows. Section 2 brings out motion attention detection and color feature selection methods. In section 3 foveated vision and model template update are described. Section 4 shows the experimental results of our proposed algorithm. At last conclusions are made in section 5.

2 Motion and Feature Attention Model

2.1 Motion Attention Detection

Psychological studies show that human eyes are very sensitive to motion. Based on this point, we think that attention aroused by motion contrast is very important during tracking. However when the camera is moving, there exist two kinds of independent motions: the motions of foreground and ego-motion of camera. To get the motional region, it is necessary to estimate the camera motion parameters and then compensate the ego-motion and do frame difference. But the frame difference method has a well-known drawback shown as fig.1: foreground aperture. It means the result is always the boundary but not the whole motion region. To overcome this problem, a dual thresholds motion regions detection method based on the horizontal and vertical projection is proposed motivated by the voice endpoint detection technology.

The motion between two consecutive frames can be assumed to be affine transformation, which means the motion can be decomposed to scale, rotation and translation. Take frame I_{t-1} and I_t for example, let's *(x, y)* and *(x', y')* denote the pixel coordinate in I_{t-1} and I_t, respectively, *s* denotes scaling factor, α is the rotation angle and *(dx, dy)* is the translation displacement. The relationship of the coordinates in two frames is formulated as follows:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = s \bullet \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \bullet \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} dx \\ dy \end{pmatrix} \tag{1}$$

KLT detection method [8] is applied in our algorithm. This method can detect the points with rich texture surrounding regions and can be implemented efficiently using multi-solution pyramid scheme. After this step, *N* pairs of interest points can be obtained. Let $p_i = (x_i, y_i)$ be the *i*-th interest point in frame I_{t-1} and $p'_i = (x'_i, y'_i)$ be the correspondence in frame I_t. Equation (1) can be written as:

$$\begin{bmatrix} x & -y & 1 & 0 \\ y & x & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ dx \\ dy \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} \tag{2}$$

where $s = \sqrt{a^2 + b^2}$, $\alpha = \tan^{-1}(b/a)$. Theoretically, only two corresponding point pairs are enough to determine the linear equations. But because of noise and the points from foreground which are so-called outliers, the solution will be bad. RANSAC (random sample consensus) algorithm is applied to estimate the motion parameter. RANSAC is a robust model parameter estimation algorithm which can even work well when the portion of outliers is close to 50%.

When the fame difference is computed, the result of one frame is shown as an example in fig.1. Obviously, the result is not the whole motion region but the boundary which is the well know foreground aperture problem. To overcome this problem and get the rough motional regions, at first we project the

difference image to horizontal direction and vertical direction. The projected values in motional region are much bigger than the ones in non-motion region. And then two thresholds for projected values are set of which the higher one T_h means the start point and the lower one T_l means potential start point, and another length threshold T_d for the lower threshold duration is also set. If two moving objects are close enough, they will be regarded as one entire region. Thus by doing this the motional regions can be robustly determined in this stage. In the next stages, the attention resource is directed to these detected motion regions which can reduce irrelevant computation remarkably.

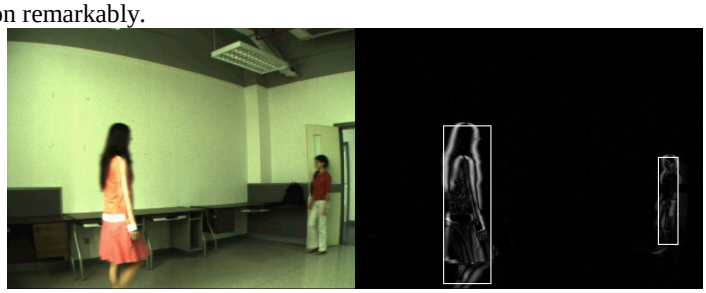

Fig.1: the foreground aperture problem and the motion attentional regions detection result of our method

Additionally, during tracking, if no motional regions are detected at the object position, it means that the object stops moving (because now we have not yet taken occlusion into consideration) and in this situation only color feature works alone. However it is reasonable that in most of the time the object is moving and the time is short when object is immobile. In fact when object isn't moving, only color feature is not insufficient. Consequently, the situation when the object stops moving during tracking doesn't much harm.

2.2 Feature Attention

Once vision task is given, for tracking it means the object to be tracked is specified, attention will be biased to relevant object with irrelevant information suppressed. In this situation it is necessary to find suitable representation of object and attention-attracting feature according to which locations with a high probability of containing the target are obtained. Considering color is an efficient and effective indicative feature for human's visual system, we use color as the feature and color histogram to represent the object.

For tracking from a moving robot, the most difficult problem is the changing illumination condition which leads to instability of color feature and change of object appearance. Obviously if one certain fixed color space is used without changing during the whole process it will fail easily. But if more kinds of color space or more dimension of feature are computed for present the object it will increase the computation burden dramatically which dissatisfies the crucial real-time requirement. To overcome this dilemma, in our algorithm one color channel is selected according to an ad hoc criterion for each frame from the feature pool which is composed of nine channels.

Because the color channel is one-dimension and the size of feature pool is nine, the computation complexity is much less, moreover, the feature is the best one from the pool and is combined with motion attention, our algorithm is not only very computationally efficient but also rather robust and effective. The most similar work is [9], they select the best one from five kinds of color spaces for each frame. But the disadvantage compared with ours is that they still have to compute three-dimension histogram which is not time saving. In addition, they don't make use of motion information and it is not robust enough in some environments. Our algorithm is described in detail as follows.

The feature set is nine color channels $\{R, G, B, H, S, V, r, g, b\}$, which is the decomposition of the three kinds of color spaces: *RGB, HSV, rgb*. As illustrated by fig.2, the discriminating ability of different channels varies obviously. As to how to measure the quality of the feature, the measurement in [9] is adopted. We draw two rectangles as showed by fig.3. The internal one denotes target region and external one denotes background region. The measurement is written as follows.

$$r = \frac{\sum_{(i,j)\in W_{in}} p(i,j)^2}{\sum_{(i,j)\in W_{out}} p(i,j)} \bullet \frac{1}{|W_i|} \tag{3}$$

During tracking, for each frame the color channel with maximum will be selected as the feature. Actually, the main computation amount is at this step. To further speed up the algorithm, take into the continuous changing of environment illumination account, we do not compute all the nine channels for each frame but just every five frames compute all the nine and then discard the four worst channels and keep the rest five best ones. The amount of computation will be cut by nearly 1/3 by conducting this trick.

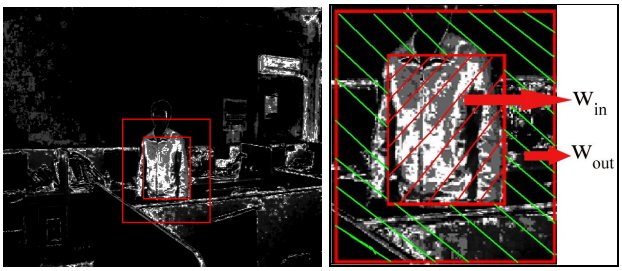

Fig.2: object and background regions denoted by two rectangles

3 Foveated Vision and Model Update

Foveated vision[10] means that the retina has an un-uniform structure with the so-called fovea in the central area which corresponds to the fixation point in vision task. The resolution is the highest in the central area and drops as it is in the peripheral area. The spotlight metaphor is a good explanation for this. Moreover, the smaller the attended region is, the more centralized the attention distribution is. Motivated by this idea, during tracking, the object position is the fixation point which has the highest resolution and attracts most attention. In the algorithm an exponential attenuation function is adopted which decreases with respect to the distance to the object position to simulate the foveated vision. And the scale of the function is controlled by the object region size which is denoted by a rectangle. The analytic form of the function is written as follows.

$$f(p) = \exp(-\frac{\| p - p_0 \|^2}{c \bullet w \bullet h}) \tag{4}$$

where p is one pixel point in the filed of view, p_0 is the fixation point which is the object position of last frame, w and h means the width and high of the rectangle respectively, and c is a constant which can be adapted.

When tracking from a moving robot, it is vital to build a robust template which adapts to the changing of object appearance. In our algorithm a heuristic target template update strategy is employed. For each one newly coming frame, one portion of the template pixels are updated by some pixels selected in tracked region of current frame. Assume that the histogram of the surrounding background region is $hist_{bg}(.)$, it is reasonable that the selected pixels in tracked region for updating should not have too many occurrences in background. Thus a threshold th is set and the pixels in tracked region with $hist_{bg}(.)<th$ are selected to substitute the pixels randomly chosen in template.

4 Experiment Results

Our algorithm is tested on 7 videos which are captured with a hand-held video camera in several kinds of environments. The frame rate is 30 fps and frame size is 320*240 pixels. The target location is initialized

manually by a rectangle. The results show that the speed of our algorithm can satisfy real-time requirement and is robust to illumination variations. Three samples are explained thoroughly as follows.

S1 is the sequence with the target having a black t-shirt and a distracter passing by in extremely similar color. In sequence S2, the target wears a yellow coat of which the color is similar with that of the wall, the post and the door in the background in terms of human eyes' visual perception. There is another person walking aside and the target starts from outside, passes by the hall and at last walks into the hallway. During the process illumination condition undergoes many changes.

Sequence S1 shows the robustness brought by the attenuation function motivated by the foveated vision. With this function it looks like that there is a spotlight in the probability map and the tracking doesn't fail despite of the extreme distracter.

Fig.3. a) Top row: the result of our algorithm with attenuation function. b) Middle row: the attention distribution map. c) Bottom row: the result of the algorithm without attenuation function.

The combination of motion attention and feature attention is validated in sequence S2 shown by fig.4. The target walks through several kinds of illumination conditions. The light yellow color of the coat is similar with some objects in the background. The red ellipse denotes the tracked object region and the green rectangle means the detected motion regions. It can be shown that when the target walks into the hall from frame 456, the light becomes darker.

Fig.4: owing to the combination of motion and feature attention our algorithm can work robustly in several kinds of environmental variations.

Owing to the adaptive template update strategy and feature attention, the algorithm can adapts to the environmental variations and tracks robustly. In frame 537, when the target passes by the post with nearly the same color, because of the combination of motion attention, the attention is only directed to the motion regions and consequently not distracted by the post. Fig.5 shows the quality measurement of each color channel according to which the best channel is selected to compute the feature attention during tracking process.

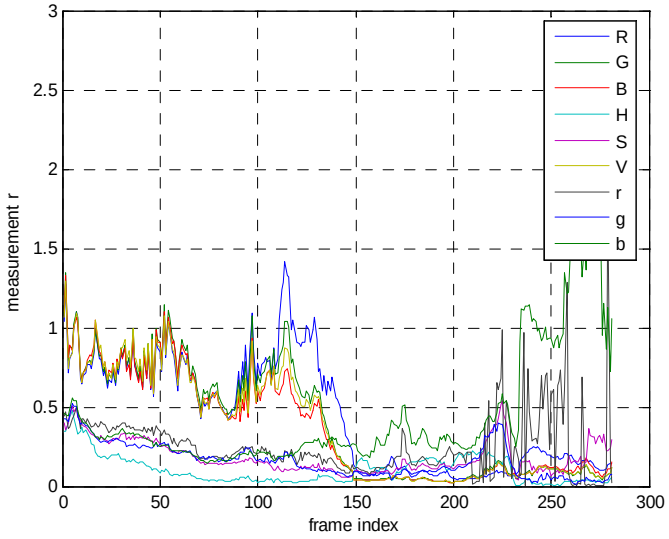

Fig.5. the quality measurement of each color channel during tracking process

5 Conclusions

In this paper, we propose a motion and feature attention model based tracking algorithm from moving robot. The motion attention is obtained by camera's ego-motion compensation and frame difference image projection, and feature attention is computed by adaptive selection of color channels. Moreover an attenuation function motivated by foveated vision and the object template update are introduced. Because the interest point detection is very efficient and only simple one-dimension color histogram is computed for color feature, furthermore motion attention and feature attention work collaboratively; our algorithm is computationally fast and robust to illumination variation from moving robot.

References

[1] D. Comaniciu, P. Meer: Real-Time Tracking of Non-Rigid Objects Using Mean Shift. IEEE Conf. CVPR, Vol.2, pp142-149, 2000.
[2] M. Isard, A. Blake: CONDENSATION – Conditional Density Propagation for Visual Tracking, IJCV1 (29), pp5-28, 1998.
[3] J. Shao, S.K. Zhou, and Q.F. Zheng. "Robust appearance-based tracking of moving object from moving robot," Proc. IEEE *Intl. Conf.* on Pattern Recognition, pp.215-218, vol.4, 2004.
[4] L. Davis, V. Philomin, R. Duraisaimi, "Tracking humans from a moving robot," Proc. IEEE *Intl. Conf.* on Pattern Recognition, pp.171-178, vol. 4, 2004.
[5] L. Itti and C. Koch, "Computational modelling of visual attention," nature reviews neuroscience, pp.194-203, vol. 2, No. 3, 2001.
[6] D.A. Migliore, M. Matteucci, M. Naccari, "A revaluation of frame difference in fast and robust motion detection," Proc. ACM *Intl.* workshop on video surveillance and sensor networks, pp. 215-218, 2006.
[7] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University Technical

Report,April 1991.
[8] H. Stern and B. Efros, “Adaptive color space switching for face tracking in multi-colored lighting environments,” Proc. IEEE *Intl. Conf.* on Automatic Face and Gesture Recognition, pp. 249-254, 2002.
[9] N. Oshiro and A. Nishikawa, N. Maru, F. Miyazaki, “Foveated vision for scene exploration,” Proc. Asian *Conf.* on Computer Vision, pp. 256-263, vol. 1351, 1998.