Search Region Correction via Spectrum Domain for Online Visual Tracking

Katsuya Hotta, Shogo Tokai, Chao Zhang*

University of Fukui, 3-9-1, Bunkyo, Fukui, 910-8507, Japan

*Corresponding Author: zhang@u-fukui.ac.jp

Abstract

Tracking-by-detection methods have drawn extensive attention in recent years. Most algorithms employ a local search region around previously estimated position assuming that the trajectory of moving object is smooth and the displacement of the object between frames is small enough. These, however, may not be valid always. Especially, when the moving objects get out of the local search region due to fast/irregular motion, tracking will fail no matter how well-trained tracker is used. In this paper, we correct the search region by utilizing the energy of the spectrum domain involving motion blur. Also, the proposed approach is light and dependent from core trackers thus the existing tracking-by-detection methods can easily adapt and benefit from the improvement of performance without much change of processing time and methodology. We conduct the experiment on 10 sequences with fast-motion by conventional methods and the proposed method. The proposed method outperforms the conventional methods even the object moves fast and is blurred.

Keywords: visual tracking, fast motion, motion blur, energy of spectrum domain.

1. Introduction

Visual Tracking algorithms with bounding boxes as the output have drawn extensive attention in recent years⁽¹⁻³⁾. Various challenges, benchmarks and measure methods have contributed to easy comparison and standardization of evaluation⁽⁴⁻⁷⁾. Various methods have tried to learn more powerful features^(8,9), build more discriminative classifiers^(10,11), exploiting robust patches for tracking^(3,12), etc. Besides, another important issue needs to be addressed is that “how to determine labeled samples for training and

unlabeled samples for the test”. In order to decrease the computational complexity and exclude irrelevant samples, most of the existing methods collect samples with observation based on recent frames. Specifically, one conventional way is to set a search region based on the previously estimated location, and crop training/test samples within the region. The limitation is also obvious because when the estimation of the previous location is not correct enough, learning models will be updated with noisy and potential hard false-positives, which will lead to the tracking drift problem. Also, objects with fast motion can easily escape from the search region and the algorithms will lose the target completely. Enlarging the search region can sometimes relieve the problem of fast motion⁽¹³⁾, but can also involve other false-positives and increase the risk of tracking drift specifically when the target and background display similar visual cues⁽¹⁴⁾.

As one of the non-negligible influencing conditions, motion blur widely exists in visual tracking tasks especially with the fast-motion sequences taken in slow shutter speed. The illumination changes over time are integrated within the exposure time, which lead to the loss of sharp features and significant appearance change. Such “smoothed” pixels usually increase the difficulty in training models and have a negative impact on target inference. Besides, the motion blur itself can vary from linear to affine, from rotational to space-variant over frames, which is hard to be modeled. Also, blurred images bring uncertainty to visual annotation, not to say an accurate estimation of the target position.

There are several methods about the correction of the search region by utilizing optical flow. One example⁽¹⁵⁾ is to use a branching model based on the judgment of the tendency of the motion blur’s change to estimate the correction vector in the current frame. However, this method has no assurance

for the correct location of the object in the current frame since it does not take the movement of the object in the previous frame into account. As a result, it cannot compare the motion blur in two frames precisely.

It is necessary to overcome the drawback that only global motion blur can be exploited. We propose a novel approach that can utilize local motion blur as well by estimating the object in the current frame. Also, we update the observation model according to the degree of motion blur in order to improve the accuracy of existing methods.

2. Proposed Method

2.1 Framework of Tracking-by-detection

Most tracking-by-detection methods⁽¹⁶⁻¹⁸⁾ set a search region based on the previously estimated position, and crop training/test samples within the region. Samples are used to update the observation model in order to adapt to the change of the object appearance.

In general, the object position is initialized manually as a bounding box B_t at first frame I_t , where $t = 1$. B_t contains four elements of a rectangle surrounding the object, which are top left coordinates (i, j) , width w , and height h of the bounding box within the image. Then, we use an enlarged region of B_t , as the search region windows S_t by assuming the displacement of the object between two frames is small. However, objects with fast motion can easily escape from the search region. In that case, the algorithms will lose the target completely.

2.2 Overview of Proposed Method

To improve the above framework, we add preprocessing that can change the search region and observation model with the information estimated from motion blur. Specifically, we estimate the approximate object's position with fast motion according to corrected optical flow, which is estimated using the spectrum domain features. The energy in the spectrum domain between two frames is analyzed. It is possible to estimate the displacement of the current object and the degree of motion blur between frames, and thus a tracker can track the object with fast motion using corrected search region. On the other hand, the classifier updates the observation model considering the motion blur. The overview of our method is shown in Fig. 1, we will explain the details in the following sections.

2.3 Search Region Correction by Optical Flow

Fig. 1 Overview of the proposed method.

In order to correct the search region misaligned due to the motion blur, a global displacement vector of the object should be estimated in each frame. As a solution, we estimate it by analyzing the optical flow vectors within the bounding box.

Optical flow is assembled of vectors representing the motion of an object. We employ Farneback⁽¹⁹⁾ because it is robust with large displacements of the object by iterative and multi-scale displacement estimation.

Furthermore, we calculate the average of the optical flow vectors within B_t as the representative motion vector, assuming the object has linear motion (i.e. we assume the object moves linearly in the image). The average optical flow $\overline{of}_t$ is calculated as follows

$$\overline{of}_t = \frac{1}{wh} \sum_{y=j}^{j+h} \sum_{x=i}^{i+w} of_t(x, y), \quad (1)$$

where of_t is the representative optical flow vector at B_t . From this, the search region S_t considering the motion of the object is estimated as follows,

$$S_t = shift(S_t, \overline{of}_t), \quad (2)$$

Where shift function $shift(S_t, \overline{of}_t)$ calculates the search region that shifts S_t by $\overline{of}_t$.

However, the optical flow will be influenced by motion blur especially in the case of fast motion. A method which is robust with motion-blur is needed.

2.4 Motion blur

The situations of motion blur can be quite complicated, as they may be space-variant and nonlinear⁽²⁰⁾. Although in the real-world application, most motion blur is non-linear, it is reasonable to assume that the object's motion is linear

Fig. 2 Average success plots over all sequence.

between two successive frames because the exposure time is usually short. The proposed method assumes that motion within B_t is space-invariant. A simulation example of motion blur generation function we use in this paper is shown in Fig. 2. Both sides of the object are transparent by blending into the background when the object moves during the exposure time.

2.5 Analyze Blurred Image by Spectral Domain

Since the motion blur can usually be modeled by the convolution of point spread function (PSF) with the original sharp image, the motion information can be recovered by estimating the parameters of PSF⁽²²⁾. PSF $h(x, y)$ can be modeled as a rectangular pulse whose orientation, width corresponds to the direction and degree of the blur. Regardless of the orientation, the PSF is determined as follows,

$$h(x, y) = \begin{cases} 0 & y \neq 0, -\infty \leq x \leq \infty \\ 1/l & y = 0, -l/2 \leq x \leq l/2 \end{cases}, \quad (3)$$

where l is the length of the blur. The blurred image I_{blur} is generated by convoluting the PSF with the original image I ,

$$I_{blur} = I * h. \quad (4)$$

As the orientation can be determined by the direction of the global optical flow vector calculated in equation (1). Thus, The task left is to estimate l .

We analyze blurred image with different degrees of the blur in the spectrum domain. At first, PSF is translated into the optical transfer function (OTF) represented by H . Each $H(u, v)$ is calculated by taking fast fourier transform (FFT) on the PSF. Secondly, the blurred image after FFT G_{blur} can be generated by doing multiplication with H and G . G is defined by the FFT of I .

$$G_{blur} = G \cdot H. \quad (5)$$

Taking a one-dimensional case as an example, the PSF of motion blur has the following characteristic,

$$\sum_{i=-\infty}^{\infty} h_i = 1. \quad (6)$$

The signal f taken FFT F_k is computed as

$$F_k = \sum_{i=0}^{N-1} f_i e^{-jk\omega i}, \quad (7)$$

where $e^{-jk\omega i}$ is a normal orthogonal base. Thus, the maximum value of h_i taken FFT H_k is 1. From equation (5), (6), and (7), the sum of the power spectrum is

$$\sum_{k=0}^{N-1} |G_{blur k}|^2 \leq \sum_{k=0}^{N-1} |F_k|^2. \quad (8)$$

This means the sum of the power spectrum decreases as the blur increases. The sum of the power spectrum is defined as the average energy of f_i according to Parseval theorem :

$$1/N \sum_{i=0}^{N-1} |f_i|^2 = \sum_{k=0}^{N-1} P_k, \quad (9)$$

where P_k is power spectrum.

In summary, the average energy of the image equals the sum of the power spectrum, which decreases as the blur increases.

2.6 Object Position and Blur's Degree Estimation by Average Energy of Image

As mentioned in Sec. 2.3 and Sec. 2.4, situations of motion blur cannot obtain accurate flow due to motion blur. Thus, we propose to use the estimated average energy (equation (9)) to correct the search region and update the observation model. Specifically, object position in the current frame is estimated by comparing the current candidate positions and previously estimated position with respect to average energy.

Firstly, we create a blurred image set $BT_{t-1,L}$, which contains different degrees of blurred images. From the previously estimated object within bounding box I_{t-1} . Each element of $BT_{t-1,L}$ is bt_l which is calculated as follows,

$$bt_l = I_{t-1} * h_l, \quad (10)$$

where h_l is the motion filter with blur length l . Thus, $BT_{t-1,L}$ has consisted as follows,

$$BT_{t-1,L} = \{bt_1, bt_{1+step_b}, \dots, bt_L\}, \quad (11)$$

where $step_b$ is blur length interval. In fact, bt_l is calculated directly from the spectral domain using equation (5), which is low calculation cost than performing DFT directly on the blurred image.

Secondly, we detect the current candidate positions using equation (1). Corrected optical flow of_n' is calculated as follows,

$$of_n' = \overline{of_t} + 1.5(n-1) \overline{of_t} / step_f \quad n = 1, 2, \dots, count \quad (12)$$

Fig.3 Average success rate overall sequence.

Fig.4 Some examples of tracking scenes by CT method.

Where $\text{count} = \lceil \overline{\mathbf{of}_t} \rceil / \text{step}_f + 1$ which is decided step_f and step is interval divided.

Furthermore, we calculate the difference of average energy between the images in $BT_{t-1,L}$ and the images located at $B_{t,n} = (x + \mathbf{of}_n'(x), y + \mathbf{of}_n'(y), w, h)$ in order to obtain approximate position of the object. Criterion for evaluation is the equality of average image energy, which is evaluated as follows,

$$(h_l, \mathbf{of}_n') = \sum_{i=1}^l \sum_{j=1}^{\text{count}} \min(BT_{t-1,i}, B_{t,j}), \quad (13)$$

where function $\min(BT_{t-1,i}, B_{t,j})$ calculate difference of average of energy between $BT_{t-1,i}$ and $B_{t,j}$, which can therefore estimate h_l and $\mathbf{of}_n'$. Thus the search region is corrected by $\mathbf{of}_n'$ and the observation model is updated by h_l .

3. Experiments

3.1 Experiments conditions

We evaluate our tracking algorithm with 3 tracking-by-detection methods on 10 sequences in benchmark

datasets which belong to fast-motion category⁽⁵⁾. The 3 trackers are real-time compressive tracking (CT)⁽¹⁶⁾, distribution fields for tracking (DFT)⁽¹⁷⁾, and exploiting the circulant structure of tracking-by-detection with kernels (CSK)⁽¹⁸⁾. We check whether the proposed method can improve the accuracy of conventional 3 methods with respect to fast motion. Each method is tested with four different search region sizes. We carry out the experiment on the condition that all the parameters of trackers are fixed. We conduct the experiment with MATLAB written program, which runs on Core (TM) i7-7700 3.60 GHz CPU with 12 GB RAM.

In this paper, we evaluate the accuracy of conventional 3 trackers with/without the proposed method by intersection over union (IoU)⁽²³⁾, which is calculated as follows,

$$\text{IoU} = \frac{\text{area}(BB_{re} \cap BB_{gr})}{\text{area}(BB_{re} \cup BB_{gr})}, \quad (14)$$

where BB_{re} is the previously estimated bounding box and BB_{gr} is the ground truth Bounding box. If IoU is larger than 0.5 in one frame, the tracking result is considered as a success. Each method is initialized with the same position.

3.2 Experimental Results

The average success rates of three methods are shown in Fig. 3. Fig. 4 shows some quantitative results of tracking scenes by CT method and our method. Sequence-dependent success rates of the conventional method and ours are plotted in Fig. 5, 6, 7, which show the relationship between the IoU threshold and the success rate. In Fig. 3, our method improves the performance of all the original methods by correcting the search regions when the object moves fast. Also, the success rates of conventional methods decrease when the search region becomes small. However, with the help of our method, the conventional methods can achieve a

Fig. 5 Examples of success rates in different search regions by CSK.

Fig. 6 Examples of success rates in different search regions by CT.

Fig. 7 Examples of success rates in different search regions by DFT.

high success rate. In conclusion, our approach can improve the accuracy of conventional methods even the search region is small and the blur is strong, as is shown in Fig. 4. There are two reasons for conventional methods losing the target. First, the object escapes from the search region because the displacement of the object is too large. Secondly, the object's appearance in a blurred image which is of low quality is significantly different from the observation model. However, our method can determine the start position of the classifier and match the degree of the motion blur to make the search region fit the object position by analyzing the spectrum domain.

Also, for the sequence which DFT succeeds in tracking originally, there still exists a small increase in success rate by adopting the proposed method. Although enlarging the search region can increase the success rate, the possibility that tracking drift occurs will also increase since the search region will contain more background. Nevertheless, the proposed method does not perform well when the appearance of the object changes largely in two frames because of the spectrum information's limitation.

4. Conclusions

This paper presents a preprocessing method for tracking algorithms to deal with the fast motion with strong motion blur that degrades the image by correcting the search region. Also, we update the observation model according to the degree of the current motion blur to further approach the current object's appearance. We conduct the experiment on the fast-moving objects and succeed in improving the accuracy of tracking by adopting the proposed approach to conventional methods. However, when the object does not have sufficient features, there exists a possibility that our method does not perform well. To cope with this situation, as a future work, considering the scaling of the object can possibly lead to a better solution.

References

- (1) Smeulders Arnold WM, Chu Dung M, Cucchiara Rita, Calderara Simone, Dehghan Afshin, and Shah Mubarak : "Visual tracking: An experimental survey", IEEE transactions on pattern analysis and machine intelligence, Vol. 36, No. 7, pp. 1442-1468, 2014
- (2) Bertinetto Luca, Valmadre Jack, Henriques Joao F, Vedaldi Andrea, and Torr Philip HS : "Fully-convolutional siamese networks for object tracking", European conference on computer vision, pp. 850-865, 2016
- (3) He Zhenyu, Yi Shuangyan, Cheung Yiu-Ming, You Xinge, and Tang Yuan Yan : "Robust object tracking via key patch sparse representation", IEEE transactions on cybernetics, Vol. 47, No. 2, pp. 354-364, 2017
- (4) Li Annan, Lin Min, Wu Yi, Yang Ming-Hsuan, and Yan Shuicheng : "Nus-pro: A new visual tracking challenge", IEEE transactions on pattern analysis and machine intelligence, Vol. 38, No. 2, pp. 335-349, 2016
- (5) Wu Yi, Lim Jongwoo, and Yang Ming-Hsuan : "Object tracking benchmark", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, No. 9, pp. 1834-1848, 2015
- (6) Cehovin Luka, Leonardis Ales, and Kristan Matej : "Visual object tracking performance measures revisited", IEEE Transactions on Image Processing, Vol. 25, No. 3, pp. 1261-1274, 2016
- (7) Liang Pengpeng, Blasch Erik, and Ling Haibin : "Encoding color information for visual tracking: Algorithms and benchmark", IEEE Transactions on Image Processing, Vol. 24, No. 12, pp. 5630-5644, 2015
- (8) Ma Chao, Huang Jia-Bin, Yang Xiaokang, and Yang Ming-Hsuan : "Hierarchical convolutional features for visual tracking", Proceedings of the IEEE International Conference on Computer Vision, pp. 3074-3082, 2015
- (9) Li Hanxi, Li Yi, and Porikli Fatih : "Deeptrack: Learning discriminative feature representations online for robust visual tracking", IEEE Transactions on Image Processing, Vol. 25, No. 4, pp. 1834-1848, 2016
- (10) Nam Hyeonseob, and Han Bohyung : "Learning multi-domain convolutional neural networks for visual tracking", Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, pp. 4293-4302, 2016
- (11) Ning Guanghan, Zhang Zhi, Huang Chen, Ren Xiaobo, Wang Haohong, Cai Canhui, and He Zhihai : "Spatially supervised recurrent convolutional neural networks for visual object tracking", Circuits and Systems (ISCAS), 2017 IEEE International Symposium on, pp.1-4, 2017
- (12) Li Yang, Zhu Jianke, and Hoi Steven CH : "Visual object tracking performance measures revisited", IEEE Transactions on Image Processing, Vol. 25, No. 3, pp. 1261-1274, 2016
- (13) Xu Lingyun, Luo Haibo, Hui Bin, and Chang Zheng : "Real-time robust tracking for motion blur and fast motion via correlation filters", Sensors, Vol. 16, No. 9, pp. 1443, 2016
- (14) Galoogahi H Kiani, Fagg Ashton, and Lucey Simon : "Learning background-aware correlation filters for visual tracking", Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Honolulu HI USA, pp. 21-26, 2017
- (15) Katsuya Hotta, Shogo Tokai, and Chao Zhang : "Search Domain Correction via Frequency-domain based Image Blur Analysis in Online Visual Tracking", Vision Engineering Workshop, 2017
- (16) Zhang Kaihua, Zhang Lei, and Yang Ming-Hsuan : "Real-time compressive tracking", Journal, pp. 864-877, 2012
- (17) Sevilla-Lara Laura, and Learned-Miller Erik : "Distribution fields for tracking", Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1910-1917, 2012
- (18) Henriques Joao F, Caseiro Rui, Martins Pedro, and Batista Jorge : "Exploiting the circulant structure of tracking-by-detection with kernels", European conference on computer vision, pp. 702-715, 2012
- (19) Farneback Gunnar : "Two-frame motion estimation

- based on polynomial expansion”, Scandinavian conference on Image analysis, pp. 363-370, 2003
- (20) Jia Jiaya : “Single image motion deblurring using transparency”, Computer Vision and Pattern Recognition 2007 CVPR'07 IEEE Conference on, pp. 1-8, 2007
- (21) Dai Shengyang, Yang Ming, Wu Ying, and Katsaggelos Aggelos K : “Tracking motion-blurred targets in video”, Journal, Vol. 25, No. 3, pp. 1261-1274, 2016
- (22) Cehovin Luka, Leonardis Ales, and Kristan Matej : “Visual object tracking performance measures revisited”, Image Processing, 2006 IEEE International Conference on, pp. 2389-2392, 2006
- (23) Everingham Mark, Van Gool Luc, Williams Christopher KI, Winn John, and Zisserman, Andrew : “The pascal visual object classes (voc) challenge”, International journal of computer vision, Vol. 88, No. 2, pp. 303-338, 2010