Robust Pan-Tilt-Zoom Tracking via Optimization
Combining Motion Features and Appearance Correlations

ByeongJu Lee, Kimin Yun, Jongwon Choi, Jin Young Choi
ASRI, Dep. of ECE, Seoul National University
adolys@snu.ac.kr, ykmwww@snu.ac.kr, jwchoi.pil@gmail.com, jychoi@snu.ac.kr

Abstract

This paper proposes a new pan-tilt-zoom (PTZ) tracking method to improve the robustness against occlusions and appearance changes by using motion likelihood map and scale change estimation as well as appearance correlation filter. For this purpose, we introduce a motion likelihood map constructed from motion detection result in addition to the correlation filter. The motion likelihood map is generated by blurring the motion detection result, which shows high probability in the center of target. To combine the correlation filter and the motion likelihood map, we formulate an optimization problem. In addition, to handle the scale change of target, we repeat the combining process for various scale of bounding box. The experiments show that the proposed method outperforms the state-of-the-art methods.

1. Introduction

Visual object tracking with Pan-Tilt-Zoom (PTZ) camera is one of the most important functions in visual surveillance systems. The PTZ object tracking is a branch of visual tracking and has lots of issues such as drastic appearance change, background movement, and background clutter, etc. In the PTZ tracking, the appearance of target may vary drastically due to the motion of target and the change of PTZ view angle. Even for the tracking in a fixed camera, the color of target is changed by unexpected factors such as auto white balance setting and shadow, etc. For the PTZ tracking, there are additional issues which do not occur in a fixed camera, such as real-time camera control issue which is an essential function for PTZ tracking. The computation load of PTZ tracking algorithm must be low due to real time requirements and the PTZ camera must be controlled accurately to keep the target within the field of view.

Various studies have been conducted on visual tracking with several ways [6, 7, 8, 9, 11]. Most of tracking algorithms sacrifice the performance to make the methods operate in real-time. The PTZ tracking should compromise the trade-off between performance and speed to meet the several requirements for real time applications as well as satisfactory performance in actual environments. As for a real-time tracker with estimation of size changes encountered frequently in PTZ settings, Kalal et al. [4] have proposed an algorithm called Tracking-Learning-Detection (TLD) algorithm, which learns the positive and negative models from the object detection results, and estimates the location and size of target using these models. However, TLD sometimes confuses the target and another object whose appear-

![Figure 1: The combination of appearance model and motion map. CFR, MLM, and OPT denote the correlation filter response, the motion likelihood map, and the optimal tracking map, respectively, and the red dot indicates the peak point. The tracking result of each method is shown in bottom row. (a) The correlation filter predicts the position of target as that of tree, but combined result estimates the position of target correctly. (b) The motion detection result cannot decide the position of target without appearance of target, but combined result predicts the position of target.](image-url)
ance is similar to the target, even the object is far from the target. To handle color changes, Danelljan et al. [3] have proposed an algorithm called the Adaptive Color Tracker (ACT), which tracks a target with color information in a tracking-by-detection framework. The ACT is robust to illumination change and drastic appearance change but has weakness to the occlusion, which occurs frequently in PTZ environment.

As for a light computational tracker, Bolme et al. [1] have proposed an algorithm called Minimum Output Sum of Squared Errors (MOSSE) tracker, which trains a correlation filter minimizing the sum of squared error between input image and 2D Gaussian probability map whose peak position indicates the position of target. As an extended approach of MOSSE tracker, Danelljan et al. [2] have proposed an algorithm called Discriminative Scale Space Tracker (DSST), which estimates scale changes using an additional correlation filter with multi-channel inputs. However, the correlation filter algorithms, i.e. MOSSE and DSST, do not fully consider the problems on PTZ tracking, and have been designed to track a target whose shape does not change significantly. To adapt the correlation filter to the PTZ environment, Lee and Yi have proposed an strategy cooperating correlation filter and additional information [5]. Lee and Yi have improved the performance of correlation filter using PTZ sensory data, by estimating the scale change of target, which method referred as MOSSE with size estimation (MOSSE-S). In the MOSSE-S, however, scale estimation is incorrect in case of inaccurate PTZ sensory data in delayed network environments and approximation errors.

In this paper, we propose a PTZ tracking method to improve the performance of the correlation filter based tracker, against background clutter, size variation, and appearance change by using motion likelihood map and scale estimation in addition to the correlation filter. With the motion detection algorithm compensating the movement of PTZ camera [10, 12], we generate the motion likelihood map from the motion in the video, which resolves the problems of correlation filter on PTZ tracking, such as background clutter, and drastic appearance change. To combine the filter response of correlation filter and the motion likelihood map, we formulate an optimization problem. The key idea for this formulation is to adopt an adaptive weighting for the motion matching term. The weighting function is designed to focus on the motion of target motion and neglect the motions of non-target objects. In addition, we suggest a scale estimation method without PTZ sensory data and provide an updating condition of correlation filter to prevent undesired update, i.e. the tracking is mainly affected by the motion features. As shown in Figure 1, the optimization problem reduces the tracking failure cases which frequently happens in tracking when using either appearance or motion.

![Image](image_url)

**Figure 2:** Overview of proposed method.

2. Proposed Method

As shown in Figure 2, we track the target using both appearance and motion of target. We utilize the correlation filter as appearance tracking (Sec. 2.1), and generate the motion likelihood map from the motion in video (Sec. 2.2). To combine the appearance and motion information, we formulate an optimization problem to find optimal tracking position for a given box size (Sec. 2.3). Then, we repeat the tracking procedure for multiple candidate sizes and determine the final position of target based on the score defined in this paper, and define the filter update condition (Sec. 2.4).

2.1. Correlation Filter

Appearance matching in our approach uses correlation filter from MOSSE tracker [1]. The MOSSE tracker learns a correlation filter used to localize the target in a new frame. To learn the filter, the tracker uses a number of grayscale image patches $f^{(1)}, \ldots, f^{(N)}$ of the target appearance as training samples. These are labelled with the desired correlation output $g^{(1)}, \ldots, g^{(N)}$ from the filter, which is Gaussian function and indicates the location of the target in Gaussian probability perspective. Let $F^{(n)}, G^{(n)},$ and $H$ denote the discrete Fourier transforms (DFTs) of $f^{(n)}, g^{(n)},$ and $h$ (correlation filter), respectively. Then the optimal filter of $h$ is obtained by minimizing the sum of element-wise squared errors:

$$H^*_j = \arg \min_{H_j} \sum_{n=1}^{N} |F_j^{(n)} H_j^* - G_j^{(n)}|^2,$$  \hspace{1cm} (1)$$

where the superscript $*$ symbol indicates complex conjugation and the subscript $j$ denotes the $j$-th element. From the Eq. (1), the filter response can be adjusted by the correlation outputs $g^{(n)}$. Eq. (1) is solved with closed form expression
as
\[ H_j^i = \frac{\sum_{n=1}^{N} G_j^{(n)} f_j^{(n)\ast}}{\sum_{n=1}^{N} P_j^{(n)\ast} f_j^{(n)}}. \] (2)

The correlation filter response \( \hat{g} \) for the input patch \( f \) is computed as
\[ \hat{g} = F^{-1}\{F \circ H^i\}, \] (3)

where \( \circ \) symbol denotes the element-wise product and \( F^{-1}\{\cdot\} \) denotes the inverse Fourier transform.

When a target appearance changes, the filter needs to adapt itself in order to follow the changes. The MOSSE filter updated from the frame \( i \) is computed as:
\[ H_j^{(i)} = \frac{A_j^{(i)}}{B_j^{(i)}}, \] (4)
\[ A_j^{(i)} = \eta G_j^{(i)} f_j^{(i)\ast} + (1 - \eta) A_j^{(i-1)}, \] (5)
\[ B_j^{(i)} = \eta F_j^{(i)} f_j^{(i)\ast} + (1 - \eta) B_j^{(i-1)}, \] (6)

where \( \eta \) is a learning rate. This puts more weight on recent frames and lessen the effect of past frames exponentially over time. The filter updating occurs based on a Peak to Sidelobe Ratio (PSR), which is defined as
\[ \text{PSR} = \frac{\hat{g}_{\text{max}} - \mu_{sl}}{\sigma_{sl}}, \] (7)

where \( \hat{g}_{\text{max}} \) is the peak value and \( \mu_{sl} \) and \( \sigma_{sl} \) are the mean and standard deviation of the sidelobe (the pixel out of the 11 \( \times \) 11 window around the peak). PSR is the normalized value, and when PSR is larger than the predefined threshold (set to 3 in our scheme), the MOSSE filter is updated. Since a visual surveillance system is usually installed in outdoor environment, occlusion occurs frequently and the learning rate should be small as \( \eta = 0.05 \) to reduce the influence of the occlusions. Hence, the MOSSE filter learns appearance change slowly, and the filter response shows low peak value when the posture of target changes significantly. To compensate this problem, we use motion features of target, and details are described in Sec. 2.2.

2.2. Motion Likelihood Map

In order to use the motion features of target, we have to segment the target motions from the camera motion in PTZ cameras. To capture the motions of target, we use dual model background subtraction method (Moving Camera Detection in 5.8ms) [10], which is an efficient method that detects a moving target in a moving scene accounting the imperfect estimation of camera movements. The motion detection result is binary, i.e. the foreground pixels are set to 1 and the background pixels to 0.

From the motion detection result, we crop a patch in the same location as the bounding box in appearance based tracker, i.e. MOSSE filter. If the visual surveillance system tracks the entire body via PTZ tracking, the entire body of target exists inside of the bounding box. The Figure 3 shows the ideal example of the cropped motion detection result and the motion likelihood map representing object existence. To support the correlation filter, the motion likelihood map is designed to have high values around the center of object. The likelihood map \( P \) is defined as
\[ P = D \star \mathcal{G}, \] (8)

where \( D \) is a patch of motion detection result cropped inside of bounding box, \( \mathcal{G} \) is Gaussian kernel, and \( \star \) symbol indicates the convolution. The standard deviation of Gaussian kernel \( \mathcal{G} \) is designed as \( \sigma = 15 \) and filter size is set to the half size of the cropped patch. As shown in Figure 3(c), by blurring the motion detection patch, the motion likelihood map shows high values around the torso of target, as we expect.

2.3. Optimization for Tracking

We formulate an optimization problem to determine the tracking position considering the motion likelihood map as well as the appearance correlation. Before formulating the optimization problem, it is required to match the scale of the correlation filter response and that of motion likelihood map. In the correlation filter algorithms, two factors of the filter response is used in tracking: the location of maximum point and the PSR. The location of maximum point indicates the estimated position of target in the current frame, and the PSR means appearance similarity between the shape in the bounding box which the filter predicts and the target appearance which the filter has learned. The location of maximum point and PSR are not affected by the amplitude of Gaussian functions \( g^{(n)} \), which is used to learn correlation filter in Sec. 2.1. Hence, we match the scale of the filter response and motion likelihood map by adjusting the amplitude of correlation outputs, i.e. \( g^{(n)} \in [0,1] \).

To combine the scaled correlation filter response and motion likelihood map, we formulate an optimization problem which finds the optimal tracking map \( R^* \) in pixel-wise as
\[ R^*_j = \arg \min_{R_j} (R_j - \hat{g}_j)^2 + \lambda(D_{sl})(R_j - P_j)^2, \] (9)
where $\lambda(\cdot)$ is the weight function for motion, which varies depending on $D_{\text{all}}$ denoting the motion detection result of full image. The Eq. (9) is solved with a closed form expression as

$$R_j^* = \frac{1}{1 + \lambda(D_{\text{all}})}(b_j + \lambda(D_{\text{all}})p_j),$$

(10)

The motion detection result $D_{\text{all}}$ includes non-desired motions, i.e. the motion of non-target object and errors arising from noise in input video. We design the weight function to compensate the non-desired motion as

$$\lambda(D) = a \exp(-b \text{FR}(D)),$$

(11)

where $\text{FR}(D)$ means the foreground ratio of $D$, and $a$ and $b$ are the constant values indicating dependency on motion. To focus on the target object and neglect the non-target object, the weight on motion should be reduced for the frame with a high foreground ratio. Likewise the preceding, for noisy motion detection result giving a high foreground ratio, the weight of motion would be reduced. To give less weight on the motion than appearance, we set $a = 0.25$ and $b = 10$, which means that the major factor (more than 80%) is appearance and motion becomes minor factor (less than 20%).

### 2.4. Scale Estimation & Filter Update

Since the correlation filter does not support the scale change, additional method is needed to handle the scale change. We handle the scale change by repeating the processes in Sec. 2.1, 2.2, and 2.3 for the different scales of input patches. We change the scale of bounding box from 0.75 to 1.25 times by 0.05 interval from the bounding box in the previous frame. We give a penalty to the excessive scale change and find the optimal scale $s^*$ as

$$s^* = \arg\max_s (1 - |1 - s|)(\max \ R^{(s)}),$$

(12)

where $s$ is the scale, and $(\max \ R^{(s)})$ means the maximum element in the optimal tracking map $R^{(s)}$ at the scale $s$. The $1 - |1 - s|$ term is the penalty term for excessive scale change.

After finding the optimal scale $s^*$, we find the location of maximum point of the optimal tracking map $R^{(s^*)}$ as

$$j^* = \arg\max_j R^{(s^*)}_j,$$

(13)

where $j^*$ denotes the location of maximum point in $R^{(s^*)}$. As shown in Figure 1, the combination result shows that the motion and the correlation filter support each other to enhance the performance of PTZ tracking.

In addition to PSR based condition for update of correlation filter, described in Sec. 2.1, we establish one more condition to prevent the false influence of motion likelihood map to the update. If the final tracking position is determined mainly by the motion features, the correlation filter is not updated. That is, the correlation filter is updated only when the final tracking position is located near the position estimated by the correlation filter as well as the PSR condition is satisfied. The location estimated by the correlation filter $k^*$ is defined as

$$k^* = \arg\max_k \hat{g}^{(s^*)}_k,$$

(14)

Then, the additional condition is explicitly given by

$$|j^* - k^*| < 5,$$

(15)

which means that the final decision is done mainly by the appearance features and the tracked appearance is deserved to be used for the update of the correlation filter.

### 3. Experiments

In our experiments, the pan and tilt control for the camera was performed with the proportional-integral-derivative (PID) control scheme to maintain the current position of target around the center of image. In addition, the zoom control was done to keep the target size in a desired size. For the experiments, we recorded the four video clips. The videos were captured during tracking a target using the proposed method with PTZ control. All the methods were except our method experimented in the recorded videos.

Figure 4 shows illustrative scenes from the four video clips (video 1, 2, 3, 4 in order from the top). In video 1, one person walks in front of trees. He is not occluded by the trees, but the color of his cloth is similar to that of the trees in some frames and the edges of him almost disappears in those frames. He crosses over the guardrail in 387-415 and 548-583 frames, and so his pose changes significantly. In video 2, two persons appears, and the paths of them intersect. They have similar appearances, i.e. they wear the black cloth, and the heights of them are similar to each other. The only discriminative parts between them are the faces of them. One person is fully occluded by the other in 650-680 frames. In video 3 and 4, one person runs freely and he is partially occluded in some frames. The target runs left and right, and his posture changes drastically with slight occlusion. In video 4, the target dashes upward and runs behind a goalpost. The quality of video is poor, and some edges disappear when the camera moves left.

To evaluate the performance of the proposed method, we compared our method with ACT [3], DSST [2], TLD [4], and MOSSE-S [5] for the four video clips. ACT, DSST, and TLD were not designed for the PTZ environment, but each of them has advantages for a certain situation. ACT shows good performance for illumination changes, and DSST and TLD predict the scale of target accurately. MOSSE-S includes scale estimation using PTZ sensory data. Since we did not have the PTZ sensory data in video 1 and 2, the results for video 1 and 2 was experimented by MOSSE tracker.
We compared the performance of trackers with two measurements. First, we compared the average pixel distance between the center point of ground truth bounding box and that of tracked bounding box. As the average center distance is smaller, the performance of tracker is better. Another measurement is the average ratio of overlap region between the ground truth bounding box and the tracked bounding box. The ratio is calculated by

\[
\Phi(\Lambda_{\text{GT}}, \Lambda_{\text{TR}}) = \frac{1}{T} \sum_{t=1}^{T} \frac{A_{\text{GT}}^{t} \cap A_{\text{TR}}^{t}}{A_{\text{GT}}^{t} \cup A_{\text{TR}}^{t}},
\]

where \( \Lambda \) indicates the set of tracking results \( i.e. \Lambda = \{A\}_{t=1}^{T} \), and subscript GT and TR indicate ground truth and tracked result, respectively. The \( A_{\text{GT}}^{t} \) and \( A_{\text{TR}}^{t} \) denote the bounding box of ground truth and tracked bounding box at \( t \)-th frame.

The qualitative comparison results are shown in Figure 4. In video 1, the proposed method succeeds on tracking, \( i.e. \) the tracking algorithm finds the target correctly, while the other methods miss the target. The comparison methods fail on tracking when the target cross over the guardrail where the appearance of target changes significantly, or walk in front of trees where the color of target is similar to that of trees. In video 2, the correlation filter based algorithms, \( i.e. \) DSST and the proposed method, succeed in tracking. However, ACT and MOSSE fail in tracking after the target is fully occluded by another person, and TLD confuses the target and another person. In video 3, all the trackers succeed in tracking, but MOSSE-S fails in scale estimation due to the approximation error. In video 4, the proposed method and MOSSE-S succeed in tracking. Since the input image is blurred by camera movement, ACT and DSST fail in tracking. The quantitative comparison is shown in Figure 5. The center distance measurement of each tracking algorithm varies greatly depending on the accuracy of tracking. When tracking failure occurs, the bounding box moves to the corner of scene and the center distance becomes big. The MOSSE is the baseline method for the proposed method, but the MOSSE fails in tracking for video 1 and 2, and our method shows ten times improvement to the MOSSE. Especially, as shown in Figure 5(c), our method outperforms other methods in the average perfor-
Figure 5: **Quantitative comparison.** (a) The average center distance in log scale. (b) The average overlap ratio. (c) The overlap-center distance plot. The smaller center distance in (a) and the bigger overlap ratio in (b) indicate the better performance. The tracker is better if it resides closer to the top-right corner of (c). The filled mark indicates the average result of each method. (best shown in color)

4. Conclusions

We have proposed a PTZ tracking scheme adaptively combining appearance correlation filter response and motion likelihood map to improve the robustness against occlusions and appearance changes. The proposed adaptive weight in optimization formulation can handle the trade-off between appearance and motion to compensate each other, as a result, it improves the performance significantly. Furthermore, to handle the scale change problem, our method repeats the combining processes at the various scales and determines the optimal scale. The experiments were performed in four videos recorded with PTZ camera control, which includes the cases of pose change, occlusion, background clutter, and scale change. We validated our method in both quantitative and qualitative evaluations. The experiment results demonstrates that our method outperforms the state-of-the-art tracking methods.

Acknowledgements

This work was partly supported by the IT R&D program of MOTIE/KEIT[10041610, The development of automatic user information(identification, behavior, location) extraction and recognition technology based on perception sensor network(PSN) under real environment for intelligent robot], S-1 corporation, and the Brain Korea 21 Plus Project.

References


