# Reliable Patch Trackers: Robust Visual Tracking by Exploiting Reliable Patches

Yang Li[1], Jianke Zhu[1], Steven C.H. Hoi[2]
[1]College of Computer Science, Zhejiang University, China. [2]School of Information System, Singapore Management University.

Structural representation has recently been studied actively in tracking community, which has been shown as an effective approach to enhancing the robustness. A grid-like structure [2, 4, 6], typically a bounding box, is employed to represent the target object for tracking. Since most of the tracked target is not strictly a rectangular shape, the bounding box representation often inevitably incorporates background information into the model. This could degrade the overall performance of the tracker. Therefore, the grid-like structure is not the optimal way to represent real objects that are of non-rectangular shape.

In this paper, we propose a novel Reliable Patch Trackers (RPT), which aims to identify and exploit the essential structure across video for tracking. Instead of resorting to hand-crafted structures, a trackable confidence function is proposed to compute and select the reliable patches, which is capable of capturing the underlying object geometry, as shown in Figure 1. To locate those patches, both trackable confidence and motion information are incorporated into the particle filter framework, which are further employed to vote for the target location and estimate the object scale.

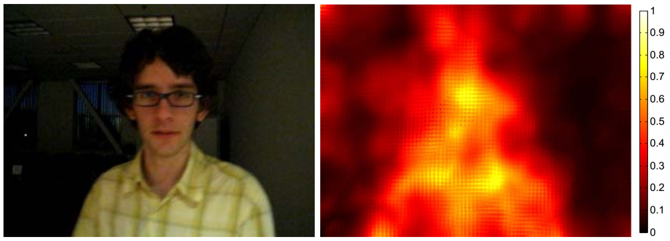

Figure 1: The tracking confidence map of the next frame when the tracker was initialized at different positions with a fixed bounding box.

In general, an image patch is sampled from a bounding box $\mathbf{x} = [x, y, w, h] \in \mathcal{R}^4$. Given the observations in previous frames $z_{1:t-1} = \{z_1, z_2, \ldots, z_{t-1}\}$, the probability density function that determines whether patch $\mathbf{x}_t$ in the current frame is reliable can be formulated as:

$$p(\mathbf{x}_t|z_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|z_{1:t-1})d\mathbf{x}_{t-1} \qquad (1)$$

where $p(\mathbf{x}_{t-1}|z_{1:t-1})$ is the state density function. According to the Bayes rule, it can be recursively calculated as: $p(\mathbf{x}_t|z_{1:t}) = \frac{p(z_t|\mathbf{x}_t)p(\mathbf{x}_t|z_{1:t-1})}{p(z_t|z_{1:t-1})}$ where $p(z_t|\mathbf{x}_t)$ is the observation likelihood, and $p(\mathbf{x}_t|\mathbf{x})$ is the transition density function. Let $\mathcal{N}$ denote Gaussian distribution, $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is defined as:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1}, \Psi(\mathbf{x}_{t-1})). \qquad (2)$$

where $\Psi_1(\mathbf{x}) = [\mathbf{0} \quad \mathbf{E}]\mathbf{x}$ is a function for selecting the image coordinates. $E$ represents a $2 \times 2$ identity matrix. Note that this assumption will allow the reliable parts to move around the object in order to account for the deformations. This will make the tracker more sensitive to the local structures.

Formally, we define a *reliable patch* that has two properties: (1) being trackable; and (2) sticking on the target object. By assuming these two properties are i.i.d., the observation likelihood $p(z_t|\mathbf{x}_t)$ can be formulated as: $p(z_t|\mathbf{x}_t) = p_t(z_t|\mathbf{x}_t)p_o(z_t|\mathbf{x}_t)$ where $p_t(z_t|\mathbf{x}_t)$ denotes the confidence of a patch to be tracked effectively, and $p_o(z_t|\mathbf{x}_t)$ indicates the likelihood that the patch is on the tracked object.

To estimate how likely a patch can be tracked effectively, we adopt the Peak-to-Sidelobe Ratio (PSR) as a confidence metric, which is widely used in signal processing to measure the signal peak strength in response map. Inspired by [1], we generalize the PSR to the template-based tracker as a patch trackable confidence function: $s(\mathbf{X}) = \frac{\max(\mathbf{R}(\mathbf{X})) - \mu_\Phi(\mathbf{R}(\mathbf{X}))}{\sigma_\Phi(\mathbf{R}(\mathbf{X}))}$ where $\mathbf{R}(\mathbf{X})$ is usually a response map. $\Phi$ is the sidelobe area around the peak. $\mu_\Phi$ and $\sigma_\Phi$ are the mean value and standard deviation of $\mathbf{R}$ except area $\Phi$, respectively. It can be easily observed that the function $s(\mathbf{X})$ becomes large when the response peak value is strong. Therefore, $s(\mathbf{X})$ can be treated as the confidence for a patch to measure whether it is tracked properly. Due to the high efficiency and impressive performance, we choose KCF [3] as our base tracker to compute $\mathbf{R}(\mathbf{X})$.

To compute the probability of a patch lying on the tracked object, we exploit the motion information to achieve this goal. Specifically, we track both foreground and background patch particles, and record the relative trajectory for each patch: $\mathbf{V}_t = [\mathbf{v}_{t-k+1}^T, \ldots, \mathbf{v}_t^T]^T \in \mathcal{R}^{2k}$, where $\mathbf{v}_t = \Psi_2(\mathbf{x}_t - \mathbf{x}_{t-1})$ is the relative movement vector and $\Psi_2 = [\mathbf{E}_{2\times2}, \mathbf{0}] \in \mathcal{R}^{2\times4}$ is selective matrix to choose the position vector in the original state. Therefore, we measure the similarity from a patch to its labelled group by formulating a similarity score function as:

$$l(\mathbf{X}) = y_t \left( \frac{1}{N^-} \sum_{j\in\Omega^-} ||\mathbf{V} - \mathbf{V}^{(j)}||_2 - \frac{1}{N^+} \sum_{i\in\Omega^+} ||\mathbf{V} - \mathbf{V}^{(i)}||_2 \right) \qquad (3)$$

where $y_i \in \{+1, -1\}$ is the label to indicate whether $\mathbf{x}_i$ is labelled as object. $\Omega_t^+$ is a set contains the indexes of the positive patch particles and $\Omega_t^-$ for negative ones. $N^+$ and $N^-$ are size numbers responding to the sets, respectively. The function $l(\mathbf{X})$ has the high score when the samplers share the homo-motion in same group and large motion difference in the other group meanwhile the negative score for those wrongly labelled samplers. Thus, we can softly label each sampler again in the sampler set, and focus the patches' emphasis on the "objects". Combining the two clues, the final observation likelihood can be formulated as follows:

$$p(z_t|\mathbf{x}_t) = p_t(z_t|\mathbf{x}_t)p_o(z_t|\mathbf{x}_t) \propto s(\mathbf{X}_t)^\lambda e^{\mu l(\mathbf{X}_t)}. \qquad (4)$$

where $\lambda$ and $\mu$ are coefficients to balance the contributions of the two likelihood.

Figure 2 shows the encouraging experimental result of our proposed tracker. Our conclusion is that by tracking with those more meaningful and reliable patches, the proposed tracker can handle more diverse and challenging situations in visual tracking.

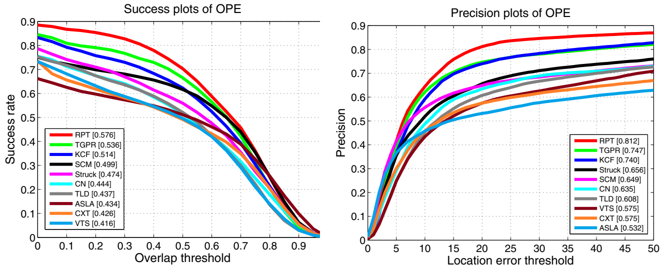

Figure 2: The overall result plots of RPT in the benchmark [5]

[1] D.S.Bolme, J.R.Beveridge, B.A.Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010.

[2] Shengfeng He, Qing-Xiong Yang, Rynson Lau, J. Wang, and M. H. Yang. Visual tracking via locality sensitive histograms. In *CVPR*, 2013.

[3] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015.

[4] Xu Jia, Huchuan Lu, and Ming-Hsuan Yang. Visual tracking via adaptive structural local sparse appearance model. In *CVPR*, 2012.

[5] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

[6] Wei Zhong, Huchuan Lu, and Ming-Hsuan Yang. Robust object tracking via sparsity-based collaborative model. In *CVPR*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.