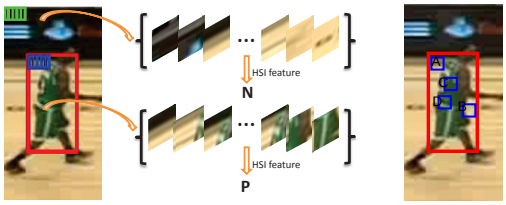
Online Visual Tracking via Coupled Object-Context Dictionary

Mingquan Ye[12]
mingquan.ye@vipl.ict.ac.cn
Hong Chang[1]
hong.chang@vipl.ict.ac.cn
Xilin Chen[1]
xilin.chen@vipl.ict.ac.cn

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China



(a) (b)

Figure 1: Illustration of constructing the coupled dictionaries. The red rectangles in (a) and (b) represent the target bounding boxes. The green squares in (a), which are generated by sliding windows outside the target bounding box, correspond to basis patches involved in the background dictionary $\mathbf{N}$. The blue squares in (a) are generated inside the target bounding box in a similar way, which constitute the noisy target dictionary $\mathbf{P}$.

Motivation: In conventional sparse coding based tracking algorithms, the dictionary is composed of holistic target templates and background templates [2]. The target templates always contain some background parts due to the non-rigidness of the tracked object. The background templates are produced around the labeled target position with big perturbations, but some target contents may still be included. Such circumstances will certainly decrease the discriminative ability of trackers. Found on the above situation, we tend to construct pure background and target dictionaries.

Constructing Pure Coupled Dictionaries: The outside of the target bounding box is pure background, while the inside includes both target and background regions [1]. Hence, we first construct the pure background dictionary with the patches in the outside context region. As shown in Figure 1, the green squares, which are generated by sliding windows in the context region, correspond to basis patches involved in the background dictionary $\mathbf{N} \in \mathbb{R}^{m\times n}$. The blue squares, as shown in Figure 1 (a), constitute a noisy target dictionary $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_l] \in \mathbb{R}^{m\times l}$. Then, we pick out real target patches from $\mathbf{P}$ based on their coding quality with respect to dictionary $\mathbf{N}$.

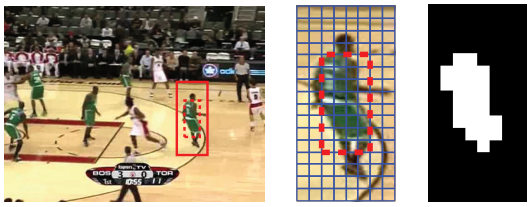
$$\widehat{\mathbf{C}} = \arg\min_{\mathbf{C}} \|\mathbf{P} - \mathbf{N}\mathbf{C}\|_2^2 + \lambda \sum_{i=1}^{l} \|\mathbf{c}_i\|_1 , \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{n\times l}$ is the coefficient matrix with the i-th column $\mathbf{c}_i$ being the sparse coding of $\mathbf{p}_i$. Then the set of reconstruction errors is expressed as $\mathcal{R} = \{\|\mathbf{p}_i - \mathbf{N}\widehat{\mathbf{c}}_i\|_2^2\}_{i=1}^{l}$. As illustrated by Figure 1 (b), the patches A and B belonging to background have small reconstruction errors. In contrast, C and D are parts of the target object, thus have big reconstruction errors. Therefore, we choose l' ($l' = \lfloor \beta \times l \rfloor$ and $\beta \in (0,1)$) basis vectors with top l' highest reconstruction errors to construct the pure target dictionary $\mathbf{P}'$, which is expressed as $\mathbf{P}' = [\mathbf{p}_{i_1}, \ldots, \mathbf{p}_{i_{l'}}]$, where $i_1, \ldots, i_{l'}$ are the indexes of the top l' values in $\mathcal{R}$.

Computing Confidence Map: The target patches should have big and small reconstruction errors when encoded by $\mathbf{N}$ and $\mathbf{P}'$, respectively, while the background patches have the reversed situations. At online tracking in t-th frame, we construct the confidence map of the context region Ω which is centered at the target position of frame $t-1$. Specifically, we normalize and divide Ω into q nonoverlapping patches of same size as those in the dictionaries (as shown in Figure 2 (b)), denoted as $\mathbf{O} \in \mathbb{R}^{m\times q}$. Then the q patches are encoded by $\mathbf{N}$ and $\mathbf{P}'$ as follows:

$$\widehat{\mathbf{C}}_1 = \arg\min_{\mathbf{C}_1} \|\mathbf{O} - \mathbf{N}\mathbf{C}_1\|_2^2 + \lambda \sum_{i=1}^{q} \|\mathbf{c}_{1i}\|_1 , \tag{2}$$

$$\widehat{\mathbf{C}}_2 = \arg\min_{\mathbf{C}_2} \|\mathbf{O} - \mathbf{P}'\mathbf{C}_2\|_2^2 + \lambda \sum_{i=1}^{q} \|\mathbf{c}_{2i}\|_1 , \tag{3}$$



(a) (b) (c)

Figure 2: Illustration of computing the confidence map. (a) The t-th frame image. (b) The context window Ω is divided into nonoverlapping patches, and the red dashed line represents the target bounding box at frame $t-1$. (c) The confidence map of Ω at frame t.

where $\mathbf{C}_1 \in \mathbb{R}^{n\times q}$, $\mathbf{C}_2 \in \mathbb{R}^{l'\times q}$ are the coefficient matrices, $\mathbf{c}_{1i}$ and $\mathbf{c}_{2i}$ are the i-th column vectors of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively.

We define the score (confidence value) of the i-th patch as

$$s_i = \|\mathbf{o}_i - \mathbf{N}\widehat{\mathbf{c}}_{1i}\|_2^2 - \|\mathbf{o}_i - \mathbf{P}'\widehat{\mathbf{c}}_{2i}\|_2^2, i = 1, \ldots, q. \tag{4}$$

In order to alleviate the negative effects caused by outliers, we adjust the patch score as below,

$$s_i' = \begin{cases} 0, \ s_i \le 0 \text{ or } \sum_{j\in\mathcal{N}(i)} \mathbf{1}(s_j \le 0) \ge \tau \\ 1, \ \text{otherwise} \end{cases}, i = 1, \ldots, q, \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\mathcal{N}(i)$ denotes the set of 8 neighbors of the i-th patch. We assign the score of each patch to all the pixels inside it, then the confidence map of the context region Ω can be obtained, as shown in Figure 2 (c).

Bayesian Tracking Framework: Our algorithm is under the Bayesian sequential estimation framework, which performs tracking by solving the maximum a posterior (MAP) problem. The observation model $p(y_t|x_t)$ is modeled by

$$p(y_t|x_t^i) \propto \sum_{(j,k)\in B_i} s_{j,k}, \tag{6}$$

where x_t^i represents the i-th sample (particle) of state x_t, B_i is its corresponding bounding box, and $s_{j,k}$ is the pixel score at location (j,k).

Empirical Results: We provide the precision and success plots over all the 12 tested sequences, as shown in Figure 3.

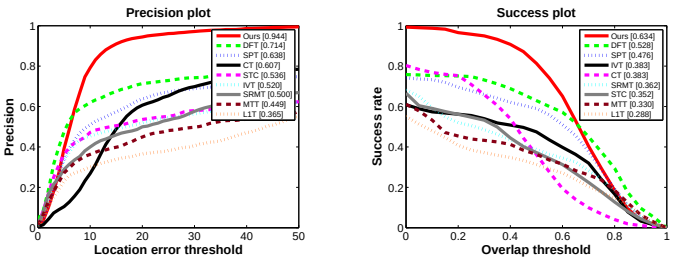


Figure 3: Precision and success plots over all the 12 tested sequences. The precision score of each tracker at 20 pixels is listed in the legend of the left plot. The right plot presents the area-under-the-curve (AUC) score for each method.

[1] Junseok Kwon, Junha Roh, Kyoung Mu Lee, and Luc Van Gool. Robust visual tracking with double bounding box model. In *ECCV*, pages 377–392. Springer, 2014.

[2] Naiyan Wang, Jingdong Wang, and Dit-Yan Yeung. Online robust non-negative dictionary learning for visual tracking. In *ICCV*, pages 657–664. IEEE, 2013.