Structural Sparse Tracking

Tianzhu Zhang[1,2], Si Liu[3], Changsheng Xu[2,4], Shuicheng Yan[5], Bernard Ghanem[1,6], Narendra Ahuja[1,7], Ming-Hsuan Yang[8]

[1]Advanced Digital Sciences Center. [2]Institute of Automation, CAS. [3]Institute of Information Engineering, CAS. [4]China-Singapore Institute of Digital Media. [5]National University of Singapore. [6]King Abdullah University of Science and Technology. [7]University of Illinois at Urbana-Champaign. [8]University of California at Merced.

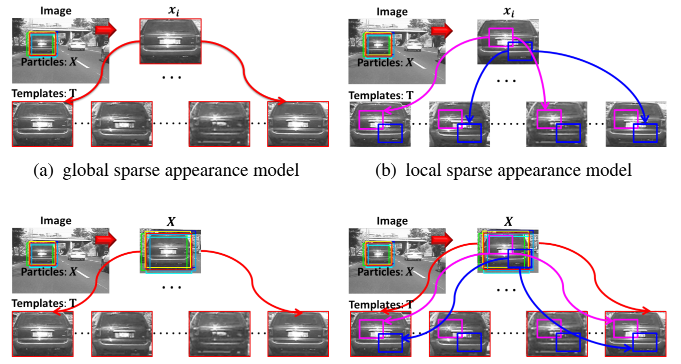
Recently, sparse representation based generative tracking methods have been developed for object tracking [1, 2, 3, 4, 5, 6, 7, 9, 10, 11]. These trackers can be categorized based on the representation schemes into **global**, **local**, and **joint** sparse appearance models as shown in Figure 1. Given an image with the n sampled particles $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_n]$ and the dictionary templates $\mathbf{T}$. (a) *Global sparse appearance model* [3, 4, 6, 7, 9]. These trackers adopt the holistic representation of a target as the appearance model and tracking is carried out by solving ℓ_1 minimization problems. As a result, the target candidate $\mathbf{x}_i$ is represented by a sparse number of elements in $\mathbf{T}$. (b) *Local sparse appearance model* [2, 5]. These trackers represent each local patch inside one possible target candidate $\mathbf{x}_i$ by a sparse linear combination of the local patches in $\mathbf{T}$. Note that, the local patches inside the target candidate $\mathbf{x}_i$ may be sparsely represented by the corresponding local patches inside *different* dictionary templates. (c) *Joint sparse appearance model* [1, 10, 11]. These trackers exploit the intrinsic relationship among particles $\mathbf{X}$ to learn their sparse representations jointly. The joint sparsity constraints encourage all particle representations to be jointly sparse and share the same (few) dictionary templates that reliably represent them. (d) The proposed *structural sparse appearance model* incorporates the above three models together. Our model exploits the intrinsic relationship among particles $\mathbf{X}$ and their local patches to learn their sparse representations jointly. In addition, our method also preserves the spatial layout structure among the local patches inside each target candidate, which is ignored by the above three models [1, 2, 3, 5, 6, 7, 9, 10, 11]. Using our model, all particles $\mathbf{X}$ and their local patches are represented with joint sparsity, i.e., only a few (but the same) dictionary templates are used to represent all the particles and their local patches at each frame. Note that, the local patches inside all particles $\mathbf{X}$ are represented with joint sparsity by the corresponding local patches inside the *same* dictionary templates used to represent $\mathbf{X}$.

In this paper, we use the convex $\ell_{p,q}$ mixed norm, especially, $\ell_{2,1}$ to model the structure information of $\mathbf{Z}^k$ and $\mathbf{Z}_i$ and obtain the structural sparse appearance model for object tracking as

$$\min_{\mathbf{Z}} \frac{1}{2} \sum_{k=1}^{K} \left\| \mathbf{X}^k - \mathbf{D}^k \mathbf{Z}^k \right\|_F^2 + \lambda \|\mathbf{Z}\|_{2,1}, \tag{1}$$

where $\mathbf{Z} = \left[\mathbf{Z}^1, \mathbf{Z}^2, \cdots, \mathbf{Z}^K\right] \in \mathbb{R}^{m \times nK}$, $\|\cdot\|_F$ denotes the Frobenius norm, and λ is a tradeoff parameter between reliable reconstruction and joint sparsity regularization. The definition of the $\ell_{p,q}$ mixed norm is $\|\mathbf{Z}\|_{p,q} = \left(\sum_i \left(\sum_j |[\mathbf{Z}]_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}$, and $[\mathbf{Z}]_{ij}$ denotes the entry at the i-th row and j-th column of $\mathbf{Z}$. Figure 2 illustrates the structure of the learned matrix $\mathbf{Z}$.

The comparison results on benchmark [8] are shown in Figure 3. The results show that our SST tracker achieves favorable performance than other related sparse trackers [2, 6, 7, 9, 10]. Compared with other state-of-the-art methods, our SST achieves the second best overall performance.



(a) global sparse appearance model (b) local sparse appearance model

(c) joint sparse appearance model (d) structural sparse appearance model

Figure 1: Sparse representation based trackers [1, 2, 3, 4, 5, 6, 7, 9, 10, 11]. These methods are grouped based on their sparse appearance models.

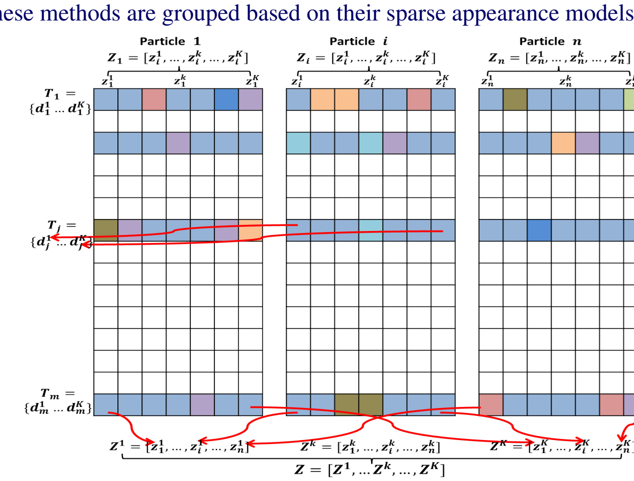
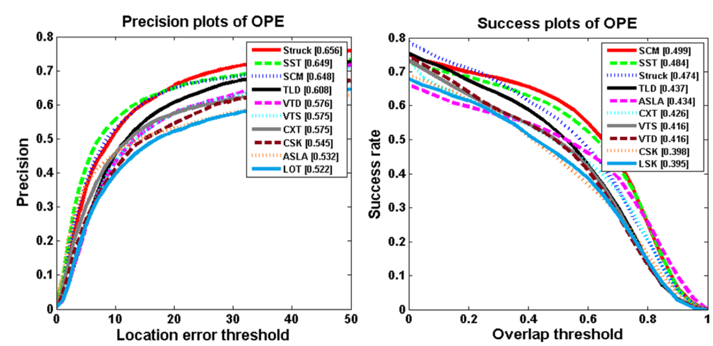


Figure 2: Illustration for the structure of the learned coefficient matrix $\mathbf{Z}$.



(a) global sparse appearance model (b) global sparse appearance model

Figure 3: Precision and success plots of overall performance comparison for the 51 videos in the benchmark [8] (best-viewed on high-resolution display).

[1] Zhibin Hong, Xue Mei, Danil Prokhorov, and Dacheng Tao. Tracking via robust multi-task multi-view joint sparse representation. In *ICCV*, 2013.

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

[3] Hanxi Li, Chunhua Shen, and Qinfeng Shi. Real-time visual tracking with compressed sensing. In *CVPR*, 2011.

[4] Baiyang Liu, Lin Yang, Junzhou Huang, Peter Meer, Leiguang Gong, and Casimir Kulikowski. Robust and fast collaborative tracking with two stage sparse optimization. In *ECCV*, 2010.

[5] Baiyang Liu, Junzhou Huang, Lin Yang, and Casimir Kulikowski. Robust visual tracking with local sparse appearance model and k-selection. In *CVPR*, 2011.

[6] Xue Mei and Haibin Ling. Robust Visual Tracking and Vehicle Classification via Sparse Representation. *TPAMI*, 33(11):2259–2272, 2011.

[7] Xue Mei, Haibin Ling, Yi Wu, Erik Blasch, and Li Bai. Minimum error bounded efficient l1 tracker with occlusion detection. In *CVPR*, 2011.

[8] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

[9] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *ECCV*, 2012.

[10] T Zhang, B Ghanem, S Liu, and N Ahuja. Robust visual tracking via multi-task sparse learning. In *CVPR*, 2012.

[11] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Low-rank sparse learning for robust visual tracking. In *ECCV*, 2012.

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