Visual Tracking by Sampling Tree-Structured Graphical Models

Seunghoon Hong and Bohyung Han

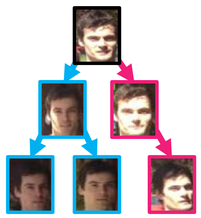
Department of Computer Science and Engineering, POSTECH, Korea

POSTECH POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

ECCV '14 Zürich, September 6-12th, 2014

Problem

- **Objective:** Identifying tree-structured graphical model for tracking based on characteristics of an input video



- Multi-modality is handled by an independent branch
- Failures are isolated at local subtree
- Difficult frames are located near leaf nodes

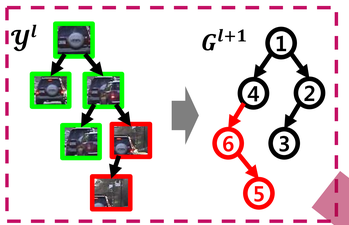
- **Challenge:** Difficult to obtain a good tree without tracking

Our Approach

- **Joint tree learning and tracking by MCMC sampling**

$$\hat{G} = \underset{G^l}{\operatorname{argmin}} -\log p(\mathcal{Y}^l|G^l), \quad l = 1, \dots, M$$

Generate M samples by iterating following procedures

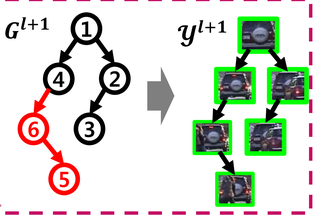


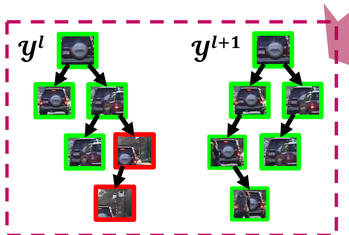
1 Proposal step

Propose a new tree by modifying subset of edges from previous sample

2 Tracking

Perform tracking on the new sample





3 Acceptance step

Probabilistically accept new sample by comparing quality of tracking with previous sample

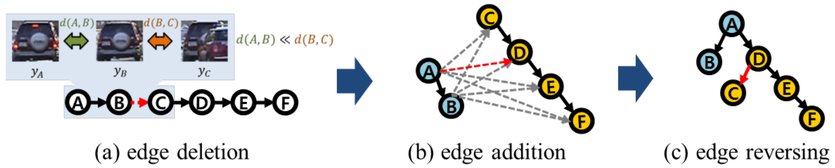
Tree learning by MCMC sampling

- Gradually improve the quality of tree by *proposal* and *acceptance* step

1 Proposal step

- Propose a new tree by deleting and adding an edge from the previous tree

$$Q(G^*; G) = P_{delete} * P_{add}$$



(a) edge deletion (b) edge addition (c) edge reversing

3 Acceptance step

- Accept a new tree sample by acceptance ratio α

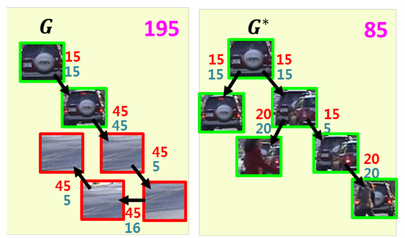
$$\alpha = \min\left[1, \frac{(-\log p(\mathcal{Y}^*|G^*))^{-1} Q(G; G^*)}{(-\log p(\mathcal{Y}|G))^{-1} Q(G^*; G)}\right]$$

$-\log p(\mathcal{Y}|G)$: The quality of tree for tracking

$$-\log p(\mathcal{Y}|G) = \sum_i c_i = \sum_i \max(d(i, p_i), c_{p_i})$$

$d(i, p_i)$: Edge cost based on template distance

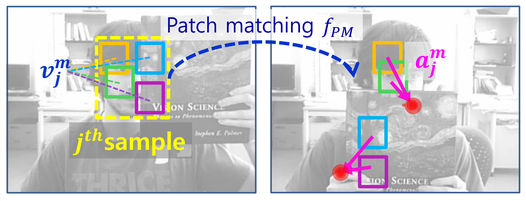
c_{p_i}: Accumulated cost from root to each node



2 Tracking on tree by Bayesian filtering

$$p(x_t|\{z_i\}_{i=1,\dots,t}, G) \propto p(z_t|x_t) \int_{x_{p_t}} p(x_t|x_{p_t}) p(x_{p_t}|\{z_i\}_{i=1,\dots,p_t}, G) dx_{p_t}$$

- Propagate posterior by patch-matching and voting process[1]



Parent frame p_t Child frame t

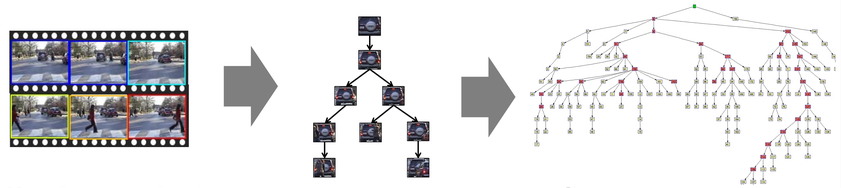
$$\approx \sum_{x_{p_t}^j \in \mathbb{S}_{p_t}} p(Z_t|x_t) p(x_t|x_{p_t}^j)$$

$$\approx \sum_{x_{p_t}^i \in \mathbb{S}_{p_t}} \sum_m \mathcal{N}(x_t|f_{PM}(v_j^m) - a_j^m, \Sigma)$$

[1] Hong, S., Kwak, S., Han, B.: Orderless tracking through model-averaged posterior estimation. In: ICCV. (2013)

Efficient Hierarchical Approach

- **Learn a tree for small number of key frames first, and extend the tree for entire frames later**



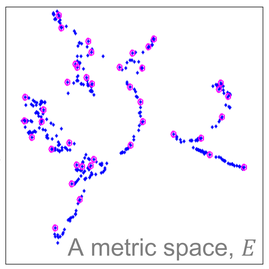
1 Key frame selection Learning tree structure for key frames **2 Extending the tree to entire frames**

1 Key frame selection

- Computing distance between all pairs of frames

$$D(I_1, I_2) = \frac{1}{n_1} \sum_{P \in I_1} \min_{Q \in I_2} d(P, Q) + \frac{1}{n_2} \sum_{Q \in I_2} \min_{P \in I_1} d(P, Q)$$

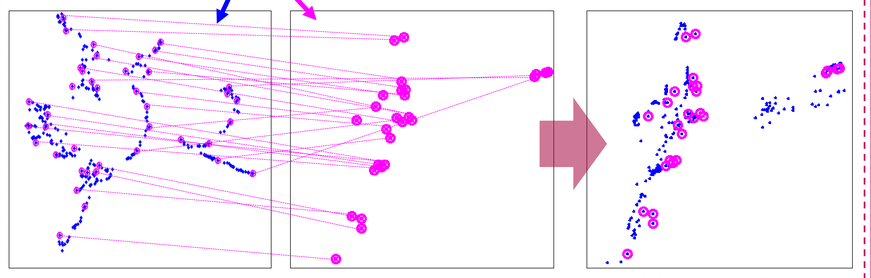
- Embedding frames into a metric space by ISOMAP
- Selecting k most representative frames by centroid of k-means clustering result



2 Tree extension by semi-supervised manifold alignment

Find a joint embedding space considering both scene and target distance

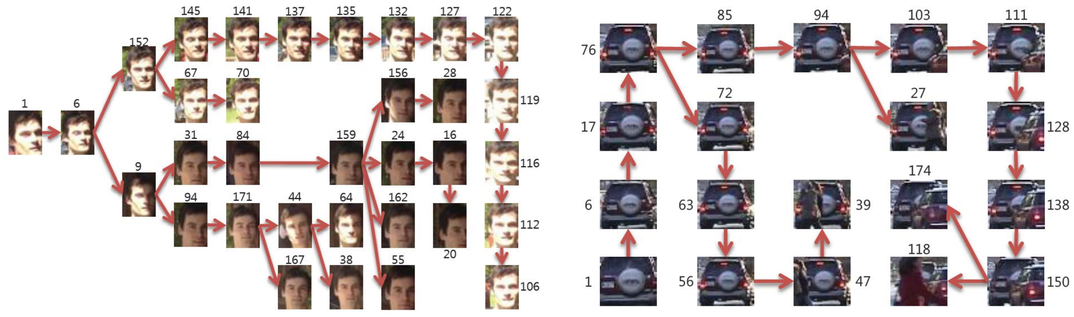
$$\min_{s,t} \Phi(s, t) \equiv \mu \sum_{i \in \mathcal{K}} |s_i - t_i|^2 + s^T L^s s + t^T L^t t$$



Embedding of **all frames** based on scene distance | Embedding of **key frames** based on target distance | Joint embedding of all frames

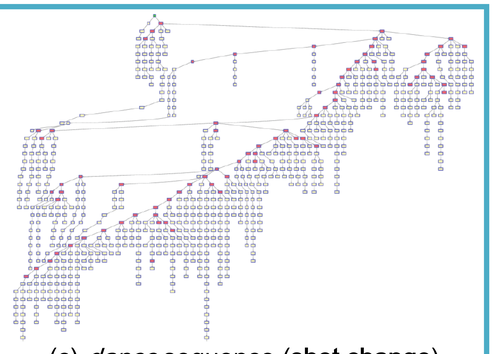
Experiments

- Identified tree for key frames



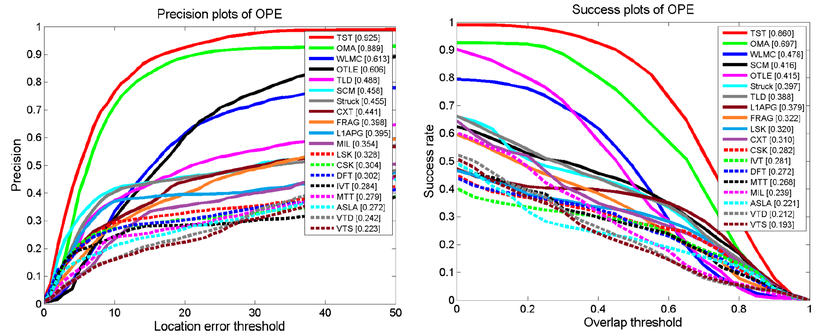
(a) *sunshade* sequence (**bi-modal appearance**) (b) *campus* sequence (**occlusion**)

- Extended tree for entire frames



(c) *dance* sequence (**shot change**)

- **Quantitative results**



(a) Precision (b) Success ratio