Hierarchical Convolutional Features for Visual Tracking

Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang

Shanghai Jiao Tong University · UC Merced · University of Illinois at Urbana-Champaign

ICCV15 International Conference on Computer Vision

Introduction

Challenges:

- Significant appearance variations caused by deformation, abrupt motion, illumination changes, background clutter, heavy occlusion, out-of-view, etc.

Contributions:

- Use the rich feature hierarchies of CNNs as target representations for visual tracking, where both semantics and fine-grained details are simultaneously exploited to handle large appearance variations and avoid drifting;
- Adaptively learn linear correlation filters on each CNN layer to alleviate the sampling ambiguity, and infer the target location using the multi-level correlation response maps in a coarse-to-fine fashion.

Our Observation

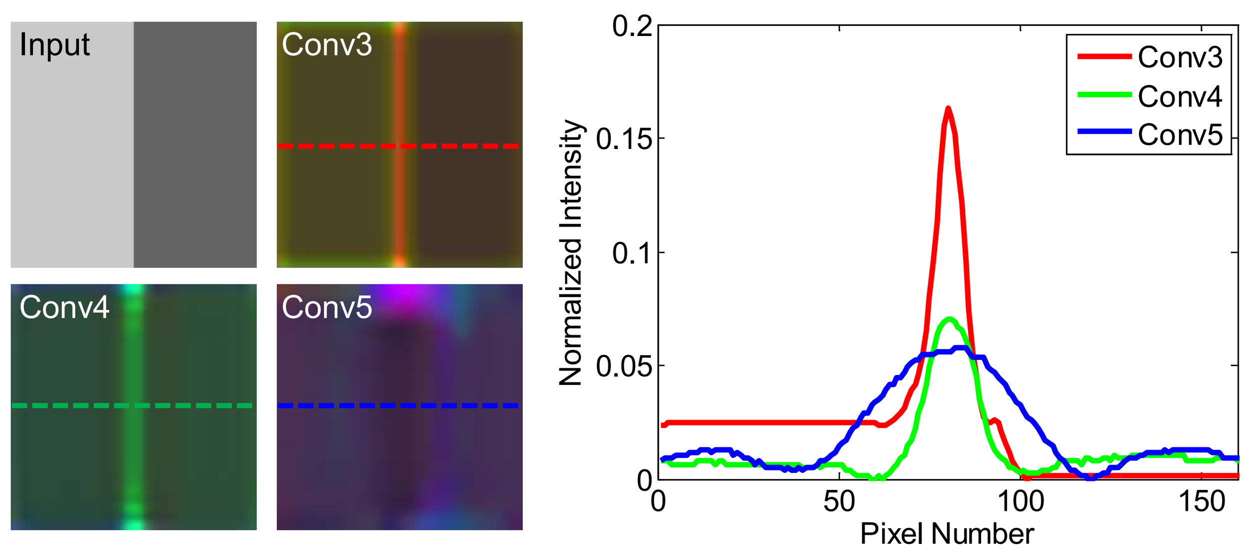


Fig. 1. Visualization of the CNN features using VGG-Net-19 [1].

- The *conv5-4* layer is less effective to locate the step edge due to its low spatial resolution;
- The *conv3-4* layer is more useful for precise localization.

Method Overview

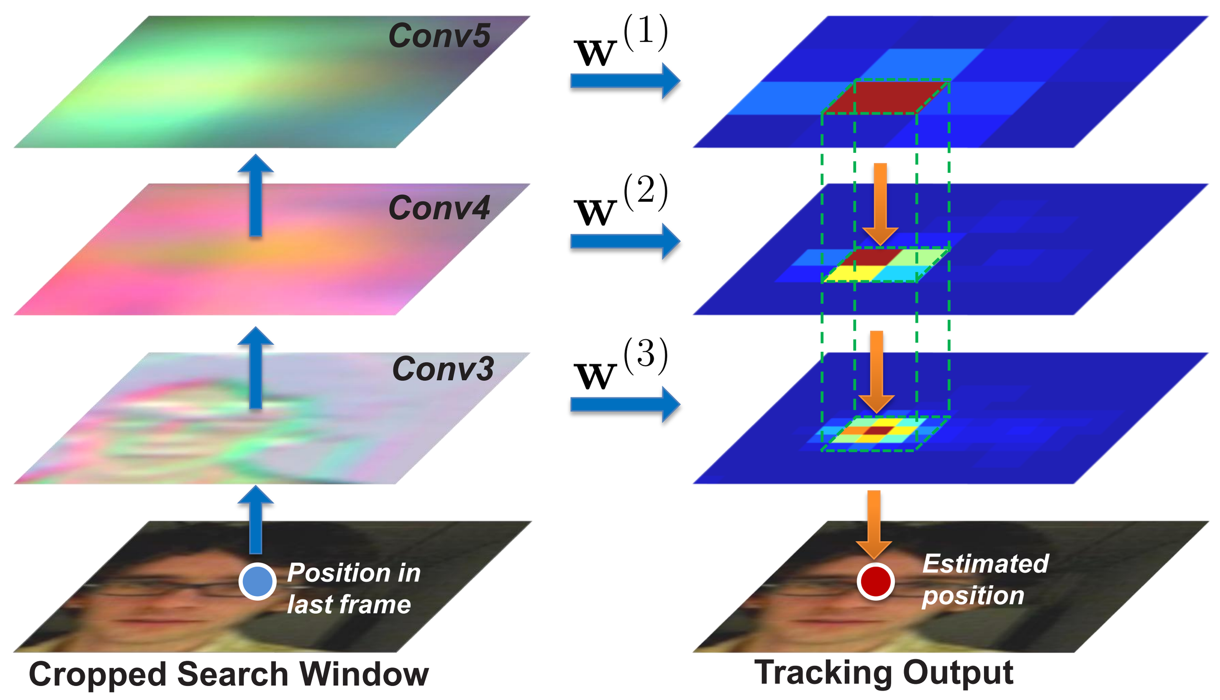
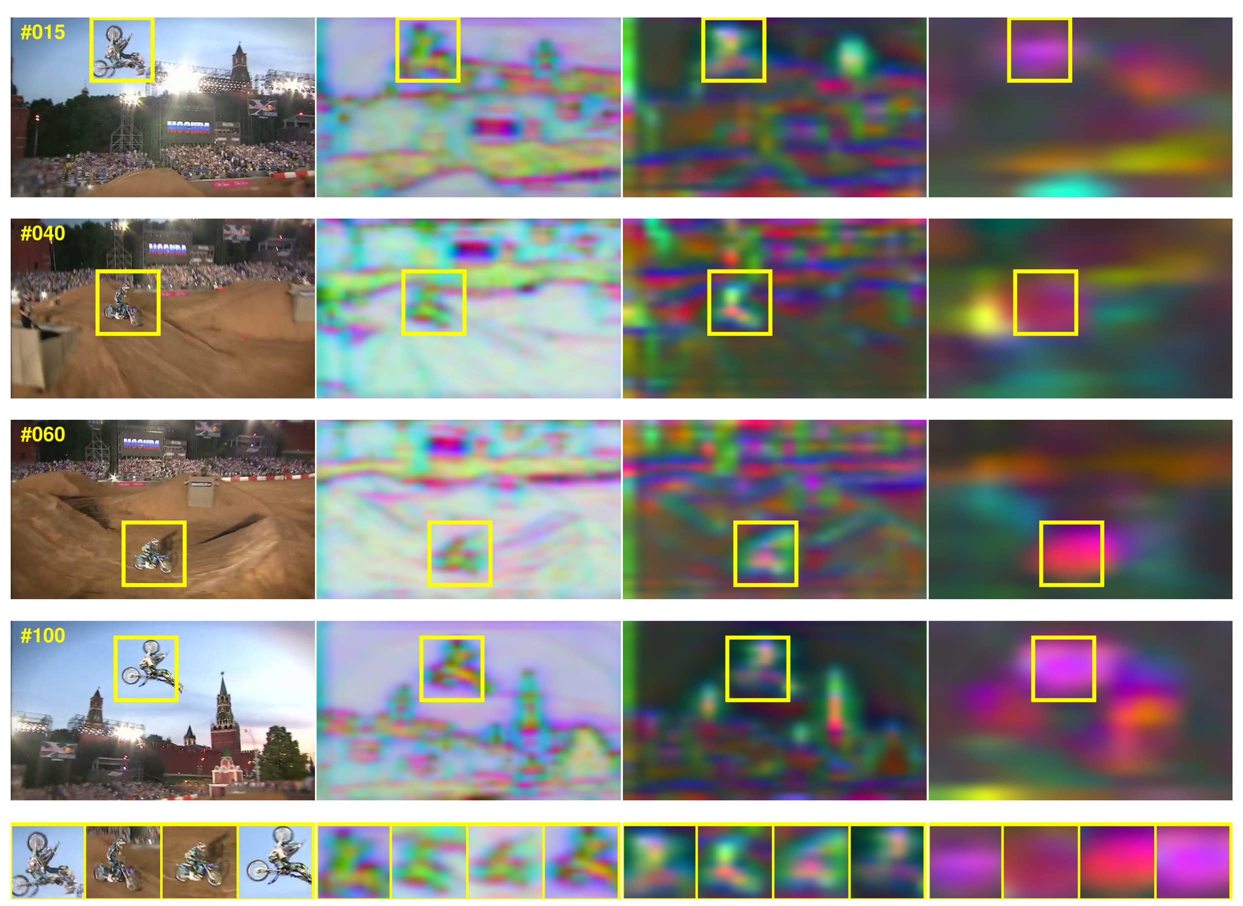


Fig. 2. Main steps of the proposed algorithm.

- Crop the search window;
- Compute the response map for each layer;
- Estimate translation hierarchically.

CNN Features



(a) Input (b) *conv3-4* (c) *conv4-4* (d) *conv5-4*

Fig. 3. Visualization of convolutional layers.

- (d) is with semantic abstraction, and is robust to appearance changes;
- (b) (c) contains more fine-grained spatial details, and is helpful for precise localization;
- It is important to exploit the merits of all layers for robust visual tracking.

Correlation Filters

A correlation filter $\mathbf{w}$ is trained from all the circularly shifted input $\mathbf{x}$ with a Gaussian function label $y_{m,n}$

$$\min_{\mathbf{w}} \sum_{m,n} \|\mathbf{w}\cdot\mathbf{x}_{m,n} - y(m,n)\|^2 + \lambda\|\mathbf{w}\|_2^2, \qquad (1)$$

where $\mathbf{w}\cdot\mathbf{x}_{m,n} = \sum_{d=1}^{D} \mathbf{w}_{m,n,d}^{\top}\mathbf{x}_{m,n,d}$. Using the FFT trick, (1) is minimized in the frequency domain on the d-th ($d \in \{1,\dots,D\}$) channel as

$$\mathbf{W}^d = \frac{\mathbf{Y}\odot\bar{\mathbf{X}}^d}{\sum_{i=1}^{D}\mathbf{X}^i\odot\bar{\mathbf{X}}^i + \lambda}. \qquad (2)$$

Given an new image on the l-th layer with feature $\mathbf{z}$ of size $M\times N\times D$, the response is:

$$f_l = \mathcal{F}^{-1}\Big(\sum_{d=1}^{D}\mathbf{W}^d\odot\bar{\mathbf{Z}}^d\Big). \qquad (3)$$

Overall Performance

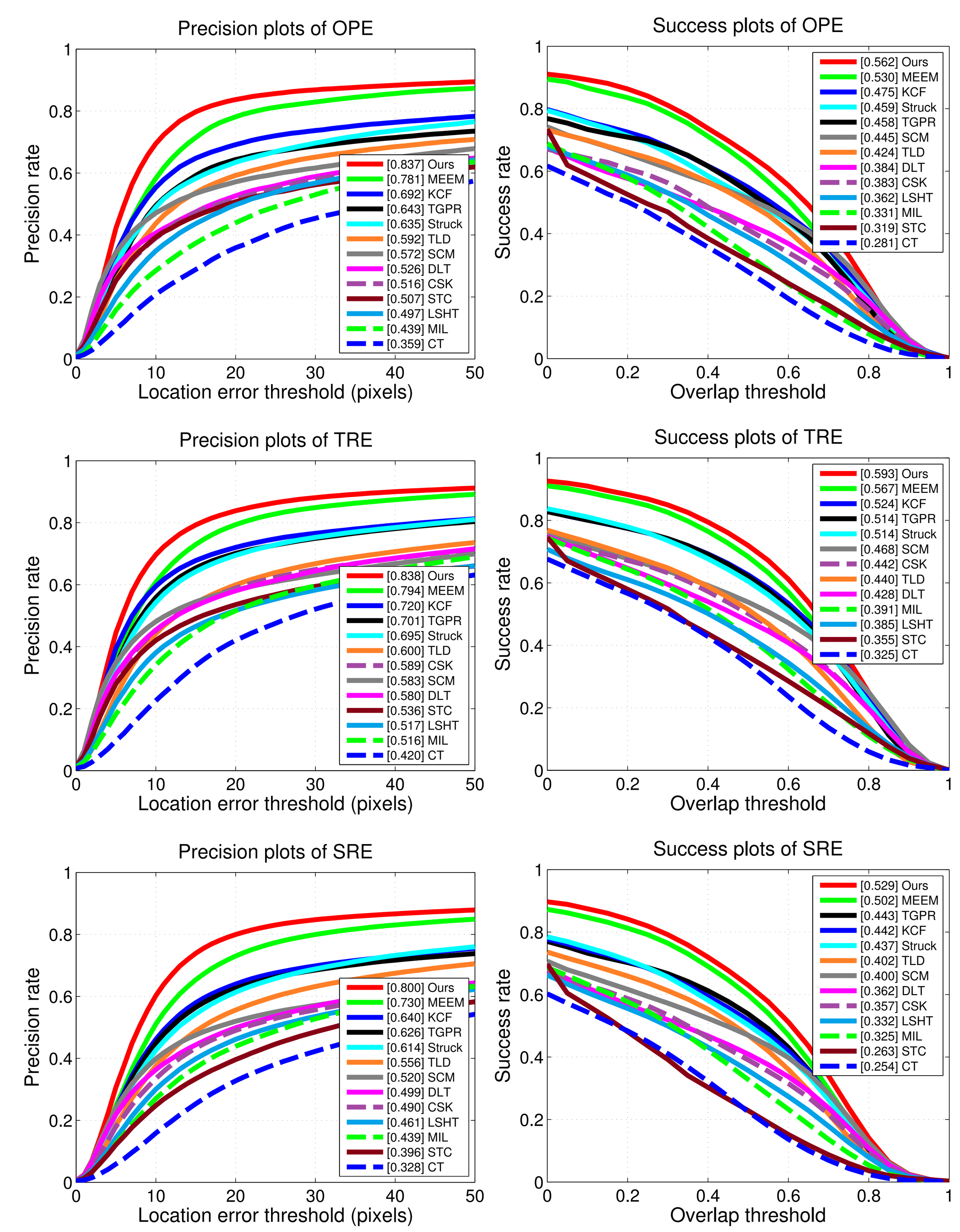


Fig. 4. Distance precision and overlap success plots on OTB-100 [3] using OPE,TRE and SRE.

More Results

Table 1. Results of distance precision (DP) rate at a threshold of 20 pixels, overlap success (OS) rate at a threshold of 0.5 and center location error (CLE) on OTB-50 (I) [2] and OTB-100 (II) [3].

		Ours	DLT	KCF	STC	Struck	SCM	CT	LSHT	CSK	MIL	TLD	MEEM	TGPR
DP rate (%)	I	**89.1**	54.8	74.1	54.7	65.6	64.9	40.6	56.1	54.5	47.5	60.8	83.0	70.5
	II	**83.7**	52.6	69.2	50.7	63.5	57.2	35.9	49.7	51.6	43.9	59.2	78.1	64.3
OS rate (%)	I	**74.0**	47.8	62.2	36.5	55.9	61.6	34.1	45.7	44.3	37.3	52.1	69.6	62.8
	II	**65.5**	43.0	54.8	31.4	51.6	51.2	27.8	38.8	41.3	33.1	49.7	62.2	53.5
CLE (pixel)	I	**15.7**	65.2	35.5	80.5	50.6	54.1	78.9	55.7	88.8	62.3	48.1	20.9	51.3
	II	**22.8**	66.5	45.0	86.2	47.1	61.6	80.1	68.2	305	72.1	60.0	27.7	55.5
Speed (FPS)	I	11.0	8.59	245	**687**	10.0	0.37	38.8	39.6	269	28.1	21.7	20.8	0.66
	II	10.4	8.43	243	**653**	9.84	0.36	44.4	39.9	248	28.0	23.3	20.8	0.64

Ablation Study

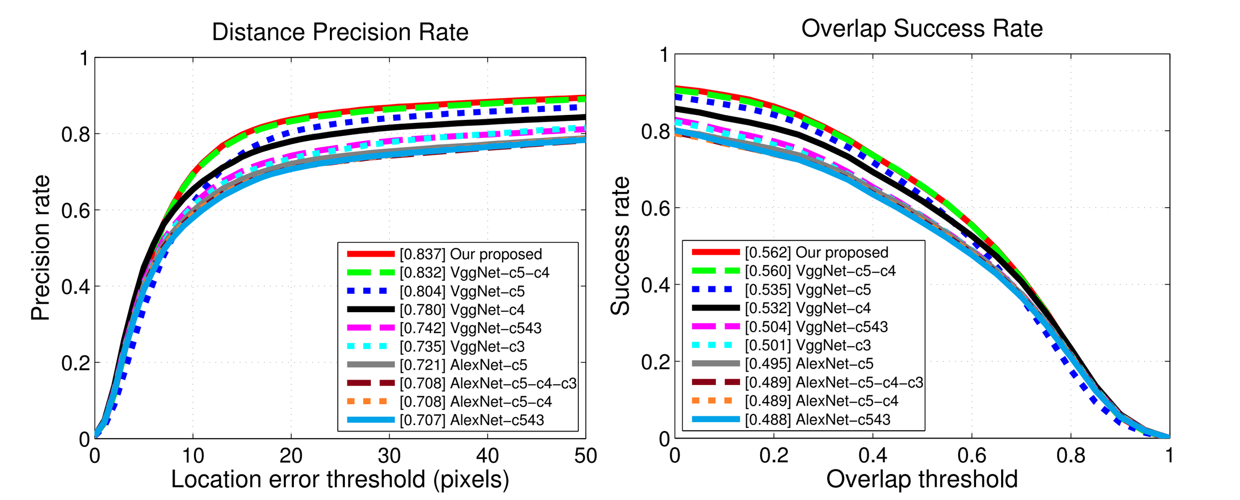


Fig. 5. Results with VGG-Net and AlexNet. c5,c4 and c3: each single convolutional layer; c5-c4: the combination of *conv5* and *conv4*; c543: the concatenation of three convolutional layers.

Discussion

Our method performs favorably against state-of-the-art methods:

- CNN features (e.g., VGG-Net) learned with category-level supervision are effective in discriminating targets from background;
- The deeper layer (*conv5-4*) is insensitive to appearance changes, and is weighted more than earlier layers (*conv3-4* and *conv4-4*).

Reference

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

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

[3] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *TPAMI*, PrePrints.

Code available at `https://github.com/jbhuang0604/CF2`