LEARNING TEMPORALLY CORRELATED REPRESENTATIONS USING LSTMS FOR VISUAL TRACKING

Qiaozhe Li *Xin Zhao* *Kaiqi Huang*

CRIPAC & NLPR, Institute of Automation Chinese Academy of Sciences,
University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

In this paper, we propose to learn object representations with inference from temporal correlation in videos to achieve effective visual tracking. Unlike traditional methods which perform feature learning either at image level or based on intuitive temporal constraint, we employ the recurrent network with Long Short Term Memory (LSTM) units to directly learn temporally correlated representations of the objects in long sequences. The recurrent network is pre-trained offline with auxiliary data and then online optimized to adapt to the target-specific object. A structured SVM is employed to account for the temporally correlated object appearance as well as distinguish the object from background distraction. Experiment results not only show that the appearance and dynamic patterns of the objects can be characterized via temporally correlated feature learning, but also demonstrate that the proposed tracking algorithm performs favorably against the state-of-the-art methods.

Index Terms— visual tracking, temporally correlated feature learning, Long Short Term Memory, structured SVM

1. INTRODUCTION

Visual tracking is a fundamental problem in computer vision with a wide range of applications. Although significant process has been made over the past decade, it is still a challenge for a tracker to handle the appearance changes caused by pose variation, geometric deformation, illumination change, severe occlusion, and motion blur. To deal with these challenges, it's favorable to develop effective object representations which can capture temporal correlation in videos.

Deep networks have been introduced in visual tracking to learn feature representations. The strong representation power of deep networks relies on the offline training with a large scale of data. Fan et al. [1] present a human tracking algorithm based on an offline trained convolutional neural network. In [2], a stacked denoising autoencoder (SDAE) is trained offline to learn generic natural image features and then used for online tracking. In [3], pre-trained CNN features are used to construct target-specific saliency maps for online tracking. In [4], a feature map selection method is proposed based on the properties of convolutional layers at different levels in a pre-trained VGG net. These trackers mainly focus on the properties of static features learned at image level. However, temporal correlation of the object appearance in videos, which is obviously important for visual tracking, is not incorporated into the network training.

On the other side, some trackers [5, 6, 7] embed temporal cues into convolutional neural network for visual tracking. Based on temporal slowness constraint [8], these trackers assume that the hidden features between consecutive frames should remain largely unchanged. However, the learned features may not account for the complex temporal dynamics of the objects in videos. Besides, long term temporal features may not be characterized by only considering correlation between adjacent frames.

Recently, Recurrent Neural Networks (RNN) using the Long Short Term Memory (LSTM) architecture have been demonstrated success in some sequence learning tasks such as speech recognition [9], video recognition and description [10], and caption generation for images [11]. The LSTM introduces gating units that adaptively control the flow of information across time steps to overcome the vanishing and exploding gradients problem. This makes it possible to learn long-range temporal features in long sequences by constructing "temporal" deep networks.

In this paper, we propose to represent object appearance based on inference with temporal correlation of the objects in videos. The RNN encoder-decoder network with LSTM units is employed for object representations via sequence learning. To propagate spatial-temporal information, the object's sequence is mapped into a fixed length representation with LSTM encoder. This representation is used to describe the object appearance in several past and future frames with LSTM decoders. The LSTM encoder-decoder network is pre-trained offline with auxiliary data and online optimized to adapt to the target-specific object. To account for the learned temporally correlated object appearance as well as exploit contextual information, a structured SVM is proposed to provide discriminative tracking. Experiment results show that the appearance and dynamic patterns of the objects can be characterized via temporal correlated feature learning. They are robust against distraction factors such as illumination change,

motion blur, and severe occlusion. Besides, tracking results on 24 challenging videos demonstrate that the proposed tracker performs favorably against the state-of-the-art trackers.

2. PROPOSED ALGORITHM

2.1. Temporally correlated feature learning

It's important to capture the temporal correlation of object appearance during tracking. We adopt the RNN encoder-decoder network [12] with LSTM units for feature learning, since it has the ability to propagate information through time. We first describe the LSTM unit [13, 14], which is the building block of the RNN encoder-decoder network. The LSTM unit includes input gate i_t, forget gate f_t, output gate o_t, and memory cell c_t. Let $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function. Given the current frame $\mathbf{x}_t$, the previous hidden states $\mathbf{h}_{t-1}$, and previous cell states $\mathbf{c}_{t-1}$, the LSTM units at time t are updated as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_{xi}\mathbf{x}_t + W_{hi}\mathbf{h}_{t-1} + W_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \\
\mathbf{f}_t &= \sigma(W_{xf}\mathbf{x}_t + W_{hf}\mathbf{h}_{t-1} + W_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \\
\mathbf{c}_t &= \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tan(W_{xc}\mathbf{x}_t + W_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{o}_t &= \sigma(W_{xo}\mathbf{x}_t + W_{ho}\mathbf{h}_{t-1} + W_{co}\mathbf{c}_t + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t \tan(\mathbf{c}_t)
\end{aligned}$$

The weight matrices from the cell to gate vectors (e.g. W_{ci}) are diagonal, whereas the rest are dense. With the additional gates and memory units, the LSTMs can discover long term features which can characterize the object appearance even in complex situations.

The RNN encoder-decoder network consists of one encoder and two decoders. The encoder maps the sequence of the tracked object into a fixed length of representation. To get the temporally correlated object appearance, we unroll this representation into a set of sequence with LSTM decoders. In this paper, we directly use grayscale raw pixels as input. Two kinds of decoder LSTM is employed. The first decoder reconstructs the input to project the learned feature representation in the past sequence. The second decoder predicts the future frames to extrapolate the object appearance based on the observed sequence. Given the tracked object sequence $\langle \mathbf{x}_{t-k+1}, ..., \mathbf{x}_t \rangle$ with k frames, the whole encoding-decoding process can be written as:

$$\begin{aligned}
\langle \hat{\mathbf{x}}_{t-k+1}, ..., \hat{\mathbf{x}}_t \rangle &= \Psi^{-1}(\Psi(\langle \mathbf{x}_{t-k+1}, ..., \mathbf{x}_t \rangle)) \\
\langle \hat{\mathbf{x}}_t, ..., \hat{\mathbf{x}}_{t+m} \rangle &= \Gamma(\Psi(\langle \mathbf{x}_{t-k+1}, ..., \mathbf{x}_t \rangle))
\end{aligned} \tag{1}$$

where Ψ denotes encoding transformation, Ψ^{-1} denotes reconstruction transformation, and Γ denotes prediction transformation.

Similar to trackers [15, 2, 4, 7], the RNN encoder-decoder network is also pre-trained offline with auxiliary data. In our work, we use the sequences of objects cropped out from ground-truth bounding box in NLPR dataset [16], which contains hundreds of different pedestrians in multiple views with various appearance. We transfer the offline learned feature to adapt to the target-specific object via online learning. The sequences for online learning are truncated from the previous tracking results with fixed length. Logistic output with the cross entropy loss function is used for network training.

2.2. Online visual tracking

It's favorable to characterize the temporally correlated object appearance as well as exploit contextual information during tracking. In this paper, we integrate the learned object appearance by LSTM network into structured SVM [17] for target localization. The feature extracted from image $\mathbf{I}$ that correspond to the state $\mathbf{y}$ is denoted as $\phi(\mathbf{I}; \mathbf{y})$. The score that measures the similarity between the observed image and the target is defined as:

$$s(\mathbf{y}; \mathbf{I}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{I}; \mathbf{y}) \tag{2}$$

where $\mathbf{w}$ is the parameter representing the appearance. The optimal object state is found by maximizing Eq. (2). Similar to [18, 19, 20], the score $s(\mathbf{y}; \mathbf{I}, \mathbf{w})$ of a true positive sample at state $\mathbf{y}$ should be larger than any other state $\hat{\mathbf{y}}$ by at least a margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$, and Δ is the task loss. The structured SVM loss ℓ is defined as the maximum violation of the task loss by states $\hat{\mathbf{y}}$:

$$\ell(\mathbf{w}; \mathbf{I}, \mathbf{y}) = \max_{\hat{\mathbf{y}}} [s(\mathbf{w}; \mathbf{I}, \hat{\mathbf{y}}) - s(\mathbf{w}; \mathbf{I}, \mathbf{y}) + \Delta(\mathbf{y}, \hat{\mathbf{y}})] \tag{3}$$

The task loss $\Delta(\mathbf{y}, \hat{\mathbf{y}})$ is defined based on the bounding box overlap ratio:

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{Area(\mathbf{y} \cap \hat{\mathbf{y}})}{Area(\mathbf{y} \cup \hat{\mathbf{y}})} \tag{4}$$

Different from [18, 19], all the object appearance obtained from LSTM network together with tracked results are regarded as positive appearance for structured SVM at each frame. The parameter $\mathbf{w}$ is updated with the passive-aggressive algorithm similar to [20]:

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\ell(\mathbf{w}; \mathbf{I}, \mathbf{y})}{\|\nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{I}, \mathbf{y})\|^2 + \frac{1}{2K}} \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{I}, \mathbf{y}) \tag{5}$$

where $K \in (0, +\infty)$ is the hyperparameter that controls the "aggressiveness" of the parameter.

3. EXPERIMENTS

We evaluate the performance of our proposed algorithm on 24 challenge sequences [21] with comparison of several state-of-the-art trackers, including DLT [2], CT [22], KCF [23], STC

(a) coke *(b) david1* *(c) human5* *(d) jumping*

Fig. 1: Sequences of learned object appearance by LSTM encoder-decoder network. **1st** row of each sub-figure shows the tracked object sequence by our tracker (*coke*: frame 255 - 274, *david1*: frame 406 - 425, *human4*: frame 332 - 351, *jumping*: frame 294 - 313). **2nd** row of each sub-figure shows the sequence of learned object appearance by LSTM network at corresponding frame (*coke*: frame 264, *david1*: frame 415, *human4*: frame 341, *jumping*: frame 303).

Table 1: Comparison of center location error (in pixels). The best and second best results are shown in red and blue fonts.

Sequence	DLT	CT	KCF	STC	TGPR	Struck	MST	SSVM	Ours
blurface	158.6	130.8	8.4	125.1	8.6	75.5	6.3	13.8	7.9
blurowl	4.4	99.0	183.4	132.0	19.1	31.4	38.3	4.2	4.6
boy	2.4	26.3	2.9	29.2	2.9	3.5	2.3	2.3	2.1
car4	2.7	81.1	9.9	10.9	5.2	8.6	1.9	2.2	1.7
cardark	1.3	72.6	6.0	2.9	2.3	0.9	2.3	1.1	1.1
coke	14.2	37.5	18.6	74.4	13.9	12.0	15.1	11.9	7.4
crowds	6.0	370.5	3.1	6.4	3.9	23.6	3.1	5.2	4.1
david	27.8	11.6	8.1	12.9	5.9	42.8	3.8	15.8	3.4
faceocc2	9.3	10.6	7.7	10.1	7.9	5.9	5.9	16.1	4.5
human2	19.4	88.8	106.6	348.5	105.9	122.0	107.4	21.1	22.9
human4	52.9	297.3	131.7	329.1	63.2	327.0	8.3	145.5	12.8
human5	15.8	257.8	175.4	154.8	10.5	239.1	21.4	5.8	4.5
human6	165.2	126.6	107.6	152.4	112.5	85.4	64.0	94.7	21.6
human7	2.5	16.7	48.1	34.8	6.8	8.2	4.6	2.6	2.8
jogging	112.9	120.0	88.2	121.4	106.3	62.0	5.7	109.3	5.1
jogging2	159.7	117.8	144.4	159.6	210.1	107.6	4.8	158.6	3.9
jumping	45.1	46.1	26.1	67.3	93.7	6.5	6.9	26.1	4.9
singer1	6.5	23.6	12.8	5.5	98.1	14.5	4.1	3.2	3.2
subway	168.9	12.1	3.0	158.1	5.1	4.5	2.2	5.6	5.3
surfer	25.3	33.4	8.7	182.0	7.7	6.6	4.3	4.1	3.8
sylvester	10.1	13.4	12.9	9.8	5.6	6.3	7.3	10.3	6.2
walking	2.3	5.0	4.0	7.9	4.8	4.6	1.6	1.9	1.7
walking2	2.3	59.0	28.9	14.0	8.2	11.1	1.9	2.7	2.4
woman	49.1	114.7	10.1	26.7	12.5	4.1	9.4	9.9	9.6
Average	38.1	91.6	54.1	93.9	59.8	30.1	20.2	27.3	7.6

Table 2: Comparison of overlap rate (%). The best and second best results are shown in red and blue fonts.

Sequence	DLT	CT	KCF	STC	TGPR	Struck	MST	SSVM	Ours
blurface	26.5	22.8	79.8	32.9	79.1	30.9	85.2	71.7	79.8
blurowl	77.0	11.3	19.5	5.3	60.0	46.9	56.6	83.0	83.6
boy	81.8	39.7	78.5	44.7	74.4	76.3	82.6	81.9	83.1
car4	85.7	23.9	49.7	36.5	52.7	49.6	89.8	88.2	90.2
cardark	78.3	0.3	61.3	77.4	80.2	89.5	79.5	82.9	83.0
coke	57.9	29.1	55.5	10.4	57.1	67.7	52.9	61.2	70.3
crowds	48.3	0.8	79.9	50.3	71.1	33.4	80.1	55.8	67.7
david	35.7	50.5	54.3	52.2	67.1	24.3	74.3	47.8	76.9
faceocc2	73.3	70.9	75.3	69.3	77.7	78.7	80.8	64.6	82.1
human2	69.3	21.6	18.4	11.8	29.6	24.6	27.3	72.3	69.1
human4	23.6	12.1	36.9	12.3	46.4	10.8	59.7	32.7	57.9
human5	29.1	18.3	18.5	20.7	32.5	24.4	50.7	63.0	79.1
human6	20.7	18.1	20.9	16.3	28.2	21.3	44.2	39.7	70.2
human7	80.1	36.1	29.2	25.6	51.5	48.3	75.9	83.4	82.6
jogging	14.3	12.9	19.0	14.5	19.3	17.2	72.2	19.1	81.0
jogging2	19.0	18.3	12.7	14.9	12.7	20.3	75.4	13.3	74.8
jumping	11.9	0.5	27.9	6.8	8.6	61.6	62.0	50.7	71.1
singer1	74.3	34.5	36.1	54.5	27.5	36.5	75.6	85.0	85.1
subway	1.4	55.7	75.7	18.1	59.1	65.8	71.7	61.4	63.1
surfer	37.7	7.5	46.4	10.5	48.6	47.1	67.4	70.3	70.4
sylvester	44.9	62.5	64.9	51.5	74.2	72.7	70.6	64.3	69.7
walking	63.2	28.0	54.4	58.5	60.5	59.2	74.5	69.0	71.2
walking2	80.5	55.6	40.6	52.5	58.4	52.7	80.6	80.2	80.2
woman	49.1	12.7	70.8	35.5	73.1	73.3	69.2	59.6	70.5
Average	51.3	29.9	46.8	33.9	53.6	47.5	66.5	63.7	75.2

[24], TGPR [25], Struck [18], and MST [26] methods. We also construct the SSVM tracker, which performs tracking with online structured SVM and the same parameter updating strategy as our method. The difference between SSVM tracker and our method is that the SSVM tracker does not employ LSTM network to learn temporally correlated object representaions. For a fair comparison, the SSVM tracker does not consider the object's deformable part structure as [20] and updates with 30 positive samples from previous tracking results at each frame. We use tracking results from SSVM tracker in the first 40 frames as the initial training dataset for our LSTM network. Each training sequence contains 20 frames truncated from tracking results with normalized frame size of 32×32 pixels. The recurrent encoder-decoder network has one hidden layer with 1024 LSTM units. The network is online optimized every 20 frames. At each frame, the encoder takes 10 frames of recently tracked object sequence as input. The decoder reconstructs the input and predicts the object appearance in the next 10 frames. The learned object apperance given by decoders together with 10 recent tracking results are regard as positive appearance for online structured SVM.

Fig.1 shows some sequences of learned object appearance by the LSTM network. In the *coke* sequence, the LSTM can "remember" the coke can's appearance though the object suffers severe occlusion in real scene. The *david1* sequence shows the predicted object appearance by LSTM in a illumination varying condition. In the *human4* sequence, the learned object appearance can account for pose variation caused by articulated deformation and is robust against partial occlusion. The *jumping* sequence shows that the deblurred object appearance can be learned with LSTM in a blurred video. These demonstrate that the appearance and dynamic motion of target-specific objects can be characterized via temporally correlated feature learning with LSTM network.

We use two criteria for quantitative evaluation: center location error (CLE) and overlapping rate (OR). CLE measures distances of centers between tracking results and ground truths in pixels. OR is calculated as $\frac{area(B_T \cap B_G)}{area(B_T \cup B_G)}$, which indicates extent of region overlapping between tracking results B_T and ground truths B_G. Table 1 and Table 2 show that our proposed method performs favorably against the state-of-the-art trackers in these two quantitative criteria.

In Fig.2, we compare the tracking results of our proposed algorithm with four trackers (DLT [2], Struck [18], MST [26],

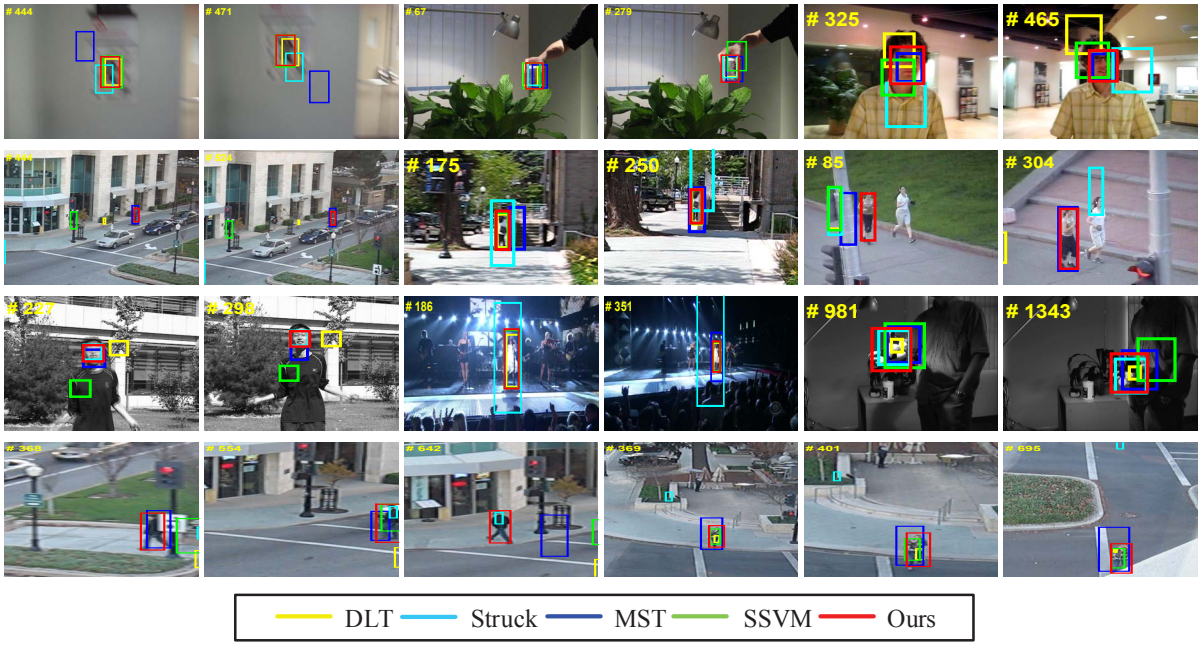

Fig. 2: Comparison of tracking results on 11 challenge sequences (from left to right and top to down are *blurowl*, *coke*, *david*, *human5*, *human7*, *jogging*, *jumping*, *singer1*, *sylvester*, *human4*, and *human6*, respectively) .

and SSVM) related to our method on 11 sequences. Based on generic image feature learning, the DLT tracker fails when objects undergo pose variation (*david1*, *sylvester*), occlusion (*jogging, human5*), and motion blur (*jumping*). The Struck tracker fails in the presence of pose variation (*david1*), severe occlusion (*jogging, human5*), illumination change (*singer1*), and motion blur (*blurowl*) as the hand-crafted features may not characterize the temporal correlation of the object appearance in such complicated situation. Without using LSTM network for temporal feature learning, the SSVM tracker drifts in the presence of pose variation (*david*), motion blur (*jumping*), and severe occlusion (*coke*, *jogging*). The MST tracker is based on the collaboration of long-term and short-term memory with hand-crafted features, and does not perform well in the presence of motion blur (*blurowl*, *jumping*). Besides, all these trackers either fail to follow the target or can not estimate the target scale well in the *human4* and *human6* sequences. These two sequences are challenging because the object appearance varies significantly and continuously caused by articulated deformation. Besides, the contour of human body is irregular so that the cropped 2D object images usually contain distracting background information. On the other hand, our proposed tracker performs well against the state-of-the-art methods on these challenging sequences. This can be attributed to that the learned features with LSTM can capture the temporal correlation of object appearance in videos and is robust against pose variation, illumination change, occlusion, and motion blur.

4. CONCLUSIONS

In this paper, we propose to learn temporally correlated features to represent objects in videos for robust visual tracking. To propagate spatial-temporal information, the LSTM encoder is used to map the object's sequence into a fixed length representation. This representation is used to describe the object appearance in several past and future frames with LSTM decoders. The LSTM encoder-decoder network is pre-trained offline with auxiliary data and online optimized to adapt to the target-specific object. A structured SVM is used to account for the learned temporally correlated object appearance and to discriminate the object from the background. Experiments not only show the effectiveness of the learned features in describing the object appearance in videos but also demonstrated that the proposed algorithm performs favorably against the state-of-the-art methods.

5. ACKNOWLEDGEMENT

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02050000).

6. REFERENCES

[1] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *TNN*, vol. 21, no. 10, pp. 1610–1623, 2010.

[2] N. Wang and D. Yeung, "Learning a deep compact image representation for visual tracking," in *NIPS*. 2013.

[3] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *ICML*, 2015.

[4] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *ICCV*, 2015.

[5] J. Kuen, K. M. Lim, and C. P. Lee, "Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle," *PR*, vol. 48, no. 10, pp. 2964–2982, 2015.

[6] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang, "Video tracking using learned hierarchical features," *TIP*, vol. 24, no. 4, pp. 1424–1435, 2015.

[7] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Learning a temporally invariant representation for visual tracking," in *ICIP*, 2015.

[8] W.Y Zou, A.Y. Ng, S. Zhu, and K. Yu, "Deep learning of invariant features via simulated fixations in video," in *NIPS*, 2012.

[9] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, 2014.

[10] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CVPR*, 2015.

[12] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," *ICML*, 2015.

[13] A. Graves, *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer, 2012.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] Q. Wang, F. Chen, J. Yang, W. Xu, and M.-H. Yang, "Transferring visual prior for online object tracking," *TIP*, vol. 21, no. 7, pp. 3296–3305, 2012.

[16] W. Chen, L. Cao, X. Chen, and K. Huang, "A novel solution for multi-camera object tracking," in *ICIP*, 2014.

[17] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *ICML*, 2004.

[18] S. Hare, A. Saffari, and P.H. Torr, "Struck: Structured output tracking with kernels," in *ICCV*, 2011.

[19] D. Chen, Z. Yuan, G. Hua, Y. Wu, and N. Zheng, "Description-discrimination collaborative tracking," in *ECCV*, 2014.

[20] L. Zhang and L. van der Maaten, "Structure preserving object tracking," in *CVPR*, 2013.

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

[22] K. Zhang, L. Zhang, and M. Yang, "Real-time compressive tracking," in *ECCV*, 2012.

[23] J. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *TPAMI*, vol. 37, no. 3, pp. 583–596, 2015.

[24] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *ECCV*, 2014.

[25] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with gaussian processes regression," in *ECCV*, 2014.

[26] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *CVPR*, 2015.