Residual LSTM Attention Network for Object Tracking

Hong-In Kim and Rae-Hong Park, *Senior Member, IEEE*

***Abstract*—In this letter, we propose an attention network for object tracking. To construct the proposed attention network for sequential data, we combine long–short term memory (LSTM) and a residual framework into a residual LSTM (RLSTM). The LSTM, which learns temporal correlation, is used for a temporal learning of object tracking. In the proposed RLSTM method, the residual framework, which achieves the highest accuracy in ImageNet large scale visual recognition competition (ILSVRC) 2016, learns the variations of spatial inputs and thus achieves the spatio-temporal attention of the target object. Also, a rule-based RLSTM learning is used for robust attention. Experimental results on large tracking benchmark datasets object tracking benchmark (OTB)-2013, OTB-100, and OTB-50 show that the proposed RLSTM tracker achieves the highest performance among existing trackers including the Siamese trackers, attention trackers, and correlation trackers, and also has comparable performance with the state-of-the-art deep trackers.**

***Index Terms*—Attention network, attention tracker, deep tracker, object tracking, residual long–short term memory (RLSTM), Siamese network, spatio-temporal attention, visual tracking.**

I. Introduction

OBJECT tracking has been considered as a fundamental challenge in a computer vision field, of which the applications include human-computer interaction and surveillance. A traditional object tracking uses spatial and temporal feature-based approaches, such as methods based on edges and colors, and optical flow. Recently, the learning-based approaches have been studied, and their applications to online tracking have been actively investigated for their robustness to training data [1]–[6]. Most recent work includes various online learning methods, which can be categorized into two classes: generative and discriminative models [2]. In generative methods, the appearance model is learned by minimizing the reconstruction errors [3], [4]. In discriminative methods, most online learning focuses on the separation of the foreground and background [5], [6].

A human perception is similar to an attention mechanism [7]. To apply attention to a deep neural network, tentative efforts have been made. The deep Boltzmann machine [8] constructs attention using top–down and reconstruction processes. Using a recurrent neural network (RNN) and a long–short term memory (LSTM) [9], the attention mechanism has been widely applied to sequential tasks [10], [11]. The LSTM is used for as well as video analysis including action recognition [12], [13], video captioning [14], [15], and video hashing [16].

Inspired by the attention mechanism [17], we propose a residual LSTM (RLSTM) attention network for object tracking. This letter has the following contributions: 1) We construct an attention model for object tracking. The Siamese tracker with the residual attention model achieves consistent performance improvement; 2) The RLSTM network is proposed to construct the spatio-temporal attention; 3) Using a rule-based RLSTM learning, the attention is learned more robustly.

The proposed method uses an end-to-end LSTM learning for a visual object tracking, by which the spatio-temporal attention of tracked objects is incrementally propagated. Also, a residual learning framework is combined with the LSTM for the attention between previous features and current features. In experiments, the proposed method is compared with the state-of-the-art trackers (the Siamese trackers, attention trackers, correlation trackers, and deep trackers).

II. Related Work

A. Object Tracking

Recently, a deep learning based on a convolutional neural network (CNN) has been actively studied and successfully applied to feature extraction, image representation, object recognition, and scene labeling [2], [10], [11]. Also, a CNN can abstract a large range of features. The low-level convolutional features contain partial detailed information whereas the high-level ones contain semantic information. An early tracking method with a CNN focused on human tracking [18]. Wang *et al*. [2] combined deep CNN features and correlation filters, constructing a coarse-to-fine search tracker. They used the last layers to handle large appearance changes whereas the first and second layers for precise localization. Tao *et al*. [19] used the Siamese deep neural network to learn a general matching function for their tracker. Recently, Bertinetto *et al*. [20] and Valmadre *et al*. [21] designed the novel fully-convolutional Siamese networks, which were trained end-to-end on the ImageNet large scale visual recognition competition (ILSVRC) 2015 dataset for object detection in a video. However, these two Siamese networks were trained offline without any online update of parameters or templates. Although they were trained on numerous videos, the

Manuscript received March 13, 2018; revised April 30, 2018; accepted May 1, 2018. Date of publication May 11, 2018; date of current version June 8, 2018. This work was supported by the Brain Korea 21 Program for Leading Universities and Students Project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li. *(Corresponding author: Rae-Hong Park.)*

The authors are with the Department of Electronic Engineering, School of Engineering, Sogang University, Seoul 04107, South Korea (e-mail: blacklleye@sogang.ac.kr; rhpark@sogang.ac.kr).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2835768

lack of online update may cause tracking failure. Currently, The RNN is used to learn temporally correlated features. The autoencoder can extract the target features that are robust to severely blurred input. The state-of-the-art method [6] uses the RNN with an LSTM, where spatio-temporal features can be used.

B. Attention Model

Attention [22]–[24] can be trained end-to-end in a convolutional network. An spatial transformer module [22] trained end-to-end by using a differentiable network layer gets the state-of-the-art results on various recognition tasks. Using attention as a scale selection mechanism, an attention to scale [23] achieves the state-of-the-art results in an image segmentation task. The proposed attention structure is inspired by the recent development of localization tasks, e.g., recognition [24], segmentation [10], and human pose estimation [25], [26]. These soft attention methods explore structures with fine-grained feature maps. Using an spatio-temporal attention, the skeleton-based action recognition methods [27], [28] were proposed to better represent the structure of input sequences. Recently, in machine translation, the response at a position by attending all positions is computed by self-attention [29].

Recently, the object tracking methods that use attention models were proposed [30], [31]. Choi *et al*. proposed a structuralist cognitive model (SCM) that considers only spatial attention [30]. Also, they combined an SCM and an attention network that determines an adaptive subset of tracking modules [31]. However, this attention model is constructed by temporally independent frames.

We propose an RLSTM attention network for object tracking. Using an RLSTM memory cell, a temporally-dependent model can be constructed. The proposed RLSTM attention model is compared with existing attention trackers.

III. RLSTM Attention Network

The proposed RLSTM is applied to an attention network for object tracking. Using the Siamese network or any feature extraction network [18], [19], the proposed method obtains feature vectors, $f(x_t)$ and $f(z_t)$, of exemplar image x_t and search image z_t, respectively, and outputs predicted object location and size vector $\mathbf{p}_t$. The proposed memory cell m_t of the RLSTM attention network is learned using $f(x_1)$ and $f(x_t)$ based on a similarity score s_t and similarity score variation Δs_t. The detector compares $f(z_t)$ and h_t, which is an attention model of $f(x_t)$.

A. Residual LSTM (RLSTM)

To construct the soft attention for object tracking, we propose an RLSTM, in which an LSTM is combined with a residual learning framework.

The proposed method uses an RNN for the attention of sequential data, where the RNN has recurrent neurons that have internal memory. It has been successfully applied to the learning of sequential data such as handwriting recognition, speech recognition, and the prediction of location [6]. The RNN predicts location and updates temporal context information, and is applied to online learning. The traditional RNN uses the recurrence equation, which is described as $h_t = g_r(f(x_t), h_{t-1})$,

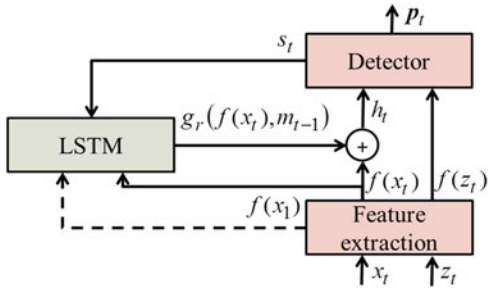

Fig. 1. Block diagram of the proposed RLSTM attention tracker.

where $f(x_t)$ is a current input feature, h_t denotes a current output attention, h_{t-1} signifies the recurrent attention that is a previous output of the hidden state, and g_r represents a recurrent transformation function. Because the traditional RNN [9] uses the $(t-1)$th recurrent input only, it can have the vanishing problem caused by a long-time gradient accumulation of the recurrent and current inputs. Thus, to solve the vanishing problem, the LSTM was proposed [9]. Each memory unit m_t of the LSTM has a forget gate and an input modulation gate to selectively forget its recurrent memory unit m_{t-1} and to learn its current input feature $f(x_t)$, respectively. Using input, input modulation, forget gate, and output gate, the LSTM allows to easily memorizing the context information for the long periods of time without the vanishing problem.

Recently, studies on a CNN have been presented, in which spatial features are extracted using the existing deep convolutional frameworks [24]. The training of deep convolutional frameworks is difficult, and thus a residual learning framework is proposed [24]. This residual learning framework learns the difference between the input and output, which is easy to optimize the deep neural network. This framework uses the residual equation, which is expressed as $h_t = \Psi(f(x_t)) + f(x_t)$, where Ψ represents a residual mapping that learns the difference between the input and output, $h_t - f(x_t)$. This residual learning framework achieves the highest accuracy in many visual recognition tasks.

However, the deep convolutional networks extract only spatial features, and then are complemented by an additional network by incremental learning. Thus, in this letter, the RLSTM is proposed for an spatio-temporal learning that is written as

$$h_t = g_r(f(x_t), m_{t-1}) + f(x_t) \tag{1}$$

where the LSTM g_r [9] learns an spatio-temporal attention, $h_t - f(x_t)$, which is easier to optimize an LSTM g_r than an spatial feature vector $f(x_t)$. Fig. 1 shows the block diagram of the proposed RLSTM attention tracker.

In this letter, we explore the possibility of a leveraging residual network to improve the tracking performance of no attention model (siamese fully convolutional tracker (SiamFC)-3s [20] and an LSTM). The experimental results show that the direct adaptation of a residual network performs well in sequence classification. Fig. 2 shows the area under the curve (AUC) of the success rate of intersection over union (IoU), for no attention model (SiamFC-3s [20] and an LSTM) and the proposed RLSTM attention mode, with 11 challenging attributes of object tracking benchmark (OTB) 2013, which are illumination

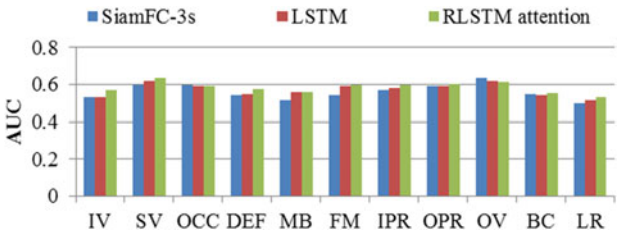

Fig. 2. AUC comparison in terms of different attributes of OTB 2013.

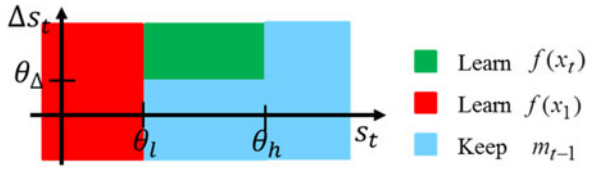

Fig. 3. Rule-based RLSTM learning with the similarity score and similarity score variation.

variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutter (BC), and low resolution (LR). In most attributes, the proposed RLSTM attention model achieves higher accuracy than no attention models. Especially, in the view of DEF attribute, the proposed RLSTM constructs a more robust attention model based on online learning. Also, it considers various motions, including IV, SV, MB, FM, IPR, and OPR attributes.

B. Rule-Based RLSTM Learning

In the proposed RLSTM tracking method, the RLSTM learns target features $f(x_t)$ to obtain an spatio-temporal attention. However, if our attention model learns incorrect target features, the proposed tracking method may lose a target and the RLSTM learns incorrect features. To prevent this problem, the proposed attention method uses an initial target feature vector $f(x_1)$. The proposed RLSTM tracking method uses a rule-based RLSTM learning, which is expressed as

$$m_t = \begin{cases} g_m(f(x_t), m_{t-1}), & \theta_l < s_t < \theta_h,\ \theta_\Delta < \Delta s_t \\ g_m(f(x_1), m_{t-1}), & s_t < \theta_l \\ m_{t-1}, & \text{otherwise.} \end{cases} \tag{2}$$

As shown in Fig. 3, the proposed RLSTM attention tracker decides whether to learn or keep. If the proposed tracker does not need any more learning, it keeps m_{t-1}. If the current feature is reliable, the proposed tracker learns $f(x_t)$, otherwise learns $f(x_1)$. Using these rules, the proposed attention network can have robust learning procedures. We extract correct features, which will be trained by the proposed attention model based on s_t and Δs_t, and construct an online learned memory m_t.

Fig. 4 shows examples with occlusion and the rotation of the rule-based RLSTM learning. Because the second image is very similar to the first image, s_t is higher than θ_h. In this case, m_{t-1} is kept. When the target in the third image is occluded, i.e., s_t is lower than θ_l, the proposed tracker learns $f(x_1)$. The fourth image shows that the target is out of occlusion, and then the proposed tracker learns $f(x_t)$. After the target is horizontally rotated, the proposed tracker sequentially learns $f(x_1)$, $f(x_t)$, and then keeps m_{t-1}. Using this procedure, we can obtain higher accuracy than existing methods such as the Siamese trackers and attention trackers.

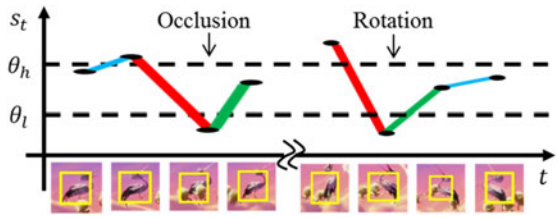

Fig. 4. Examples with occlusion and rotation of the rule-based RLSTM learning.

IV. Experimental Results and Discussions

We use the same optimal parameter set as used in SiamFC-3s [20], of which the initial values of the parameters follow the Gaussian distributions, scaled according to the improved Xavier method [32]. The RLSTM layer is constructed by a single layer with $6 \times 6 \times 192$ memory cell size. The initial learning rate is 10^{-2}, which is annealed geometrically to 10^{-5}. Using a straightforward stochastic gradient descent method, the proposed method is optimized after 70 epochs. Also, 50 consecutive frames that are constructed using 50 000 pairs of an ImageNet video for tracking [33] are trained.

Four experiments are performed to evaluate the proposed RLSTM tracker. The first experiment is conducted to show the superiority of the proposed method to the Siamese trackers and the attention trackers. In the second experiment, we compare the proposed RLSTM tracker with the Siamese trackers and the correlation trackers. The third experiment compares the proposed method with the Siamese trackers and correlation trackers using a temporal robustness evaluation (TRE). The last experiment compares the proposed RLSTM tracker with the state-of-the-art deep trackers.

We experiment a one-pass evaluation (OPE) of the OTB-2013 [34], OTB-100, and OTB-50 [35] benchmarks in terms of the success rate and precision to compare the proposed RLSTM tracker with the Siamese trackers (SiamFC-3s [20] and correlation filter network (CFNet) [21]), attention trackers (structuralist cognitive tracker (SCT) [30] and attention correlation filter network (ACFN) [31]), correlation trackers (Staple [3] and long-term correlation tracker (LCT) [5]), and the state-of-the-art deep trackers Siamese fully convolutional tracker (SiamFC) [20], fully convolutional network tracker (FCNT) [2], sequentially training convolutional tracker (STCT) [36], Siamese instance network tracker (SINT) [19], background-aware correlation filters (BACF) [37], and multi-domain convolutional neural networks (MDNet) [38]). As in the OTB benchmark [34], [35], we measure the performance of the tracker on a sequence in terms of the IoU of the predicted and ground truth rectangles in all frames. The success rate of a tracker at a given threshold corresponds to the IoU computed for a uniform range of 100 thresholds between 0 and 1, effectively constructing the cumulative distribution function. Trackers are compared using

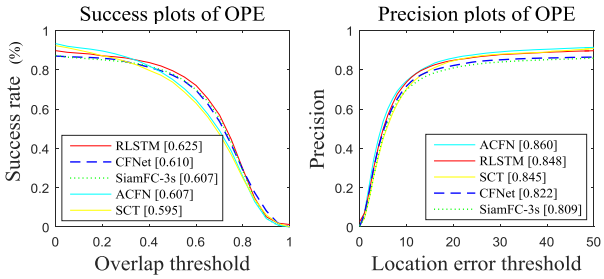

Fig. 5. Success and precision plots of an OPE of the proposed tracker, the Siamese trackers, and the attention trackers (OTB-2013).

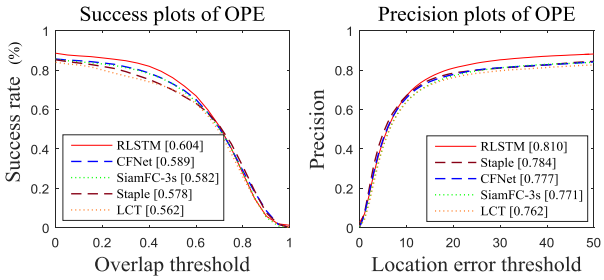

Fig. 6. Success and precision plots of an OPE of the proposed tracker, the Siamese trackers, and the correlation trackers (OTB-100).

TABLE I
TRE IN TERMS OF THE IOU AND PRECISION (OTB-2013 AND OTB-100 DATASETS)

Method	OTB-2013		OTB-100	
	IoU (%)	Precision	IoU (%)	Precision
RLSTM	**63.0**	**85.1**	**61.4**	**81.4**
CFNet	62.6	75.7	60.8	72.7
SiamFC	61.8	75.0	60.5	72.8
Staple	61.7	74.2	60.4	72.8
LCT	59.4	74.2	56.9	68.2

the AUC of the success rate. Mimicking the TRE of an OTB, we choose three equal-spaced points per sequence and run the tracker from each start point to the end point. Differently from the OPE, when the target is lost (i.e., the overlap with the ground truth becomes zero), the tracker is terminated and an overlap of zero is reported for all remaining frames.

Figs. 5 and 6 show the success and precision plots of experimental results, respectively, where the success and precision plots are based on the bounding box overlap and the distance of pixel locations [34], [35]. Also, in legend, their average values are sorted in descending order. Fig. 5 shows that the proposed method has more accurate results for the OTB-2013 dataset than the Siamese trackers and the attention trackers. The proposed RLSTM outperforms the Siamese trackers that have architectures similar to the proposed RLSTM. In the precision plots, an ACFN has a precision similar to the proposed RLSTM. Using the OTB-100 dataset, Fig. 6 shows the comparison of the proposed RLSTM tracker with the Siamese trackers and correlation trackers. These results show that the proposed RLSTM is effective for learning the spatio-temporal variations. Also, it has the highest success rate among the trackers compared.

Table I shows the experimental results of TRE for the OTB-2013 and OTB-100 datasets (bold values represent the best IoU and precision values). Most TRE results have higher IoU and precision than OPE results; however, the TRE results of the proposed method are similar to the OPE results, which are higher than those of existing trackers. This point shows the temporal robustness of the poposed attention model that is spatio-temporally learned.

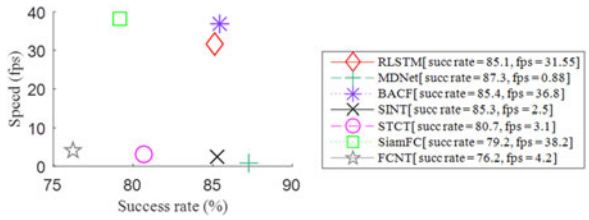

Fig. 7. Comparison of the proposed tracker with the state-of-the-art deep trackers in terms of the tracking speed (fps) and the success rate with IoU > 0.50 (OTB 50).

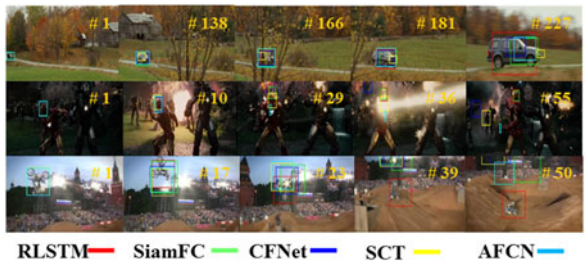

Fig. 8. Tracking results on three challenging sequences (from top to bottom: *CarScale*, *Ironman*, and *MotorRolling*).

Fig. 7 shows the success rate and the tracking speed frame per second (fps) of the proposed method and the state-of-the-art deep trackers. The proposed method has a higher success rate and lower fps than SiamFC. However, it has a higher success rate and fps than the most of existing trackers. The BACF method learns target and background features, whereas the proposed method learns only target features. Then, the BACF has the accuracy comparable to the proposed method.

Fig. 8 shows the some qualitative results of three challenging sequences of the proposed RLSTM method and existing trackers including the Siamese trackers and the attention trackers. The Siamese trackers show a poor performance in an object rotation and deformation (*MotorRolling*). Using the proposed spatio-temporal attention model, the proposed method can successfully track the target. Also, under a fast motion and a background clutter (*CarScale* and *Ironman*), the proposed method tracks more robustly than the existing trackers.

V. CONCLUSION

In this letter, by combining an LSTM and a residual network, an RLSTM method is proposed for object tracking. Unlike typical LSTM, the proposed RLSTM learns spatio-temporal attention based on the residual framework. The spatio-temporal attention is constructed more robustly using a rule-based RLSTM learning with the similarity score and similarity score variation. The experimental results show that the proposed RLSTM tracker has the highest success rate among existing trackers including the Siamese trackers, attention trackers, and correlation trackers, and also has the success rate and tracking speed comparable to the state-of-the-art deep trackers. Future work will focus on the robust RLSTM tracker using a reinforcement learning.

REFERENCES

[1] H. Song, Y. Zheng, and K. Zhang, "Robust visual tracking via self-similarity learning," *Electron. Lett.*, vol. 53, no. 1, pp. 20–22, Dec. 2016.

[2] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 3119–3127.

[3] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 5669–5678.

[4] K. Zhang, Q. Liu, Y. Wu, and M.-H. Yang, "Robust visual tracking via convolutional networks without training," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1779–1792, Dec. 2016.

[5] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Boston, MA, USA, Jun. 2015, pp. 5388–5396.

[6] Q. Li, X. Zhao, and K. Huang, "Learning temporally correlated representations using LSTMs for visual tracking," in *Proc. Int. Conf. IEEE Image Process.*, Phoenix, AZ, USA, Sep. 2016, pp. 1614–1618.

[7] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Int. Conf. IEEE Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 2204–2212.

[8] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. Int. Conf. IEEE Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 1243–1251.

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

[10] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 1520–1528.

[11] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Int. Conf. IEEE Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 2377–2385.

[12] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3D convNet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 634–644, Mar. 2018.

[13] X. Wang, L. Gao, J. Song, X. Zhen, N. Sebe, and H. T. Shen, "Deep appearance and motion learning for egocentric activity recognition," *Nerocomputing*, vol. 275, pp. 438–447, Jan. 2018.

[14] Y. Guo, J. Zhang, and L. Gao, "Exploiting long-term temporal dynamics for video captioning," *World Wide Web*, [Online]. Available: https://link.springer.com/article/10.1007/s11280-018-0530-0

[15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[16] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3210–3221, Jul. 2018.

[17] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Honolulu, HI, USA, Jun. 2017, pp. 3156–3164.

[18] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1610–1623, Oct. 2010.

[19] R. Tao, E. Gravves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 1420–1429.

[20] L. Bertinetto, J. Valmadre, F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 850–865.

[21] J. Valmadre, J. Bertinetto, F. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Honolulu, HI, USA, Jun. 2017, pp. 5000–5008.

[22] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Int. Conf. IEEE Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 2017–2025.

[23] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 3640–3649.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[25] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 483–499.

[26] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Honolulu, HI, USA, Jun. 2017, pp. 5669–5678.

[27] J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, [Online]. Available: https://ieeexplore.ieee.org/document/8101019/

[28] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

[30] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and J. Y. Choi, "Visual tracking using attention-modulated disintegration and integration," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 4321–4330.

[31] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Honolulu, HI, USA, Jun. 2017, pp. 4828–4837.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. Int. Conf. Comput. Vis.*, Washington, DC, USA, Dec. 2015, pp. 1026–1034.

[33] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[34] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Portland, OR, USA, Jun. 2013, pp. 2411–2418.

[35] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 2411–2418, Jan. 2015.

[36] L. Wang, W. Ouyang, X. Wang, and H. Lu, "STCT: Sequentially training convolutional networks for visual tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 1420–1429.

[37] H. K. Galoogahi and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. Int. Conf. Comput. Vis.*, Venice, Italy, Dec. 2017, pp. 1144–1152.

[38] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. Int. Conf. IEEE Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, Jun. 2016, pp. 4293–4302.