Localization-based Visual Tracking with Convolutional Neural Networks

Abolfazl Moridi
School of Electrical and Computer Engineering
Shiraz University, Iran
Email: a.moridi@hotmail.com

Zohreh Azimifar
School of Electrical and Computer Engineering
Shiraz University, Iran
Email: azimifar@cse.shirazu.ac.ir

Abstract—**This paper presents a novel framework for the visual tracking problem. This framework predicts the exact location of the object using a regression. In this work, we first select an approximate region based on object location in the previous frame and then predict the exact location of the object in the current frame by a deep convolutional network that its last layer replaced with a regression. The entire network gets updated due to the occurrence of various challenges during the video. We evaluate our work using 8 challenging benchmark video sequences and shows a significant improvement over state-of-the-art approaches.**

Keywords-visual tracking; localization; convolutional network; deep learning;

I. INTRODUCTION

Computer vision history is now more than half a century old. However, general, robust, and complete satisfactory solutions in real conditions are still impossible. One of these problems is visual tracking that play an important role in computer vision applications such as surveillance and human-computer interaction. That can be formulated for various applications. One of the settings that have been studied for decades is single-object model-free tracking [1]. That only the object location and its extent are determined in the first frame of video and no more knowledge about the object.

This task is considered as a very hard problem due to its challenges in the real world as the illumination, reflection, and motion are irregular, the pose, scale and view frequently change, and the background or foreground may interfere (Figure1). And all these aspects have to be learned from the first frame or when they happen without having a true label. Since learning a visual model from a single example is a major challenge in this task due to lack of sufficient information from object or environment.

In the last few years, deep neural networks have led to breakthrough results on a variety of pattern recognition problems, such as computer vision. These networks are composed of several levels of non-linear transformations that aimed to learn useful and very abstract features [2]. Architectures with several levels naturally provide such sharing and re-use of components: the lower layers features (like edge detectors) and middle layers features (like object parts) that are useful to detect a variety of objects and other visual tasks [2]. Since, the hierarchical architectures are appropriate concepts to obtain generality and adaptability for visual tracking problems. Learning new objects is well done due to the availability of generic features in lower layers and relatively well features in upper layers. In this work, we use convolutional neural networks those are very similar to ordinary neural networks. That make the explicit assumption that the inputs are images, they benefit the fact in order to induct properties into architecture that vastly reduces the amount of parameters in the network and also increase generalization ability [3].

Figure 1. Some challenges of the dataset. It is seen that sequences are in real conditions.

In most previous works tried to solve object localization problem using classification (Figure 2). Classifying candidates of the frame that indicate the selected region of that object. The candidate that covers the object scope is considered as positive sample; and scope that does not include parts of object or includes parts of backgrounds in addition to object is considered negative sample. This reduces final accuracy in the problem. Also selecting desirable candidates that show object properly is an important challenge, and selecting inappropriate candidates may lead to drifting problem, because of the imperfect classification of the target and background.

This paper proposes a novel localization-based visual tracker (LVT) with an emphasis on robustness. First, we train model of multi-layer architecture on data related to the problem due to lack of sufficient information to predict object location. Then, we transfer the learned model to the problem and tune it on our object of interest. Second, unlike previous works, we select only one approximated region where the object is probably

978-1-4673-8789-7/16/$31.00 ©2016 IEEE

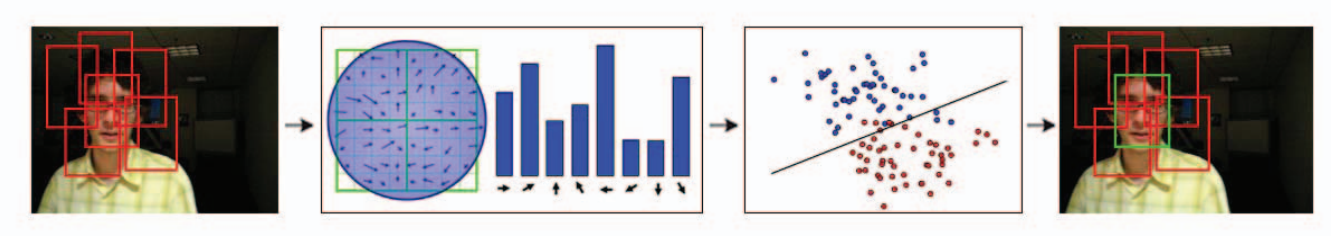

Figure 2. The pipeline of the previous works for visual tracking problem.

located and obtain the exact location of the object based on features obtained by model and a regression that indicate bounding box of the object. Third, since certain dynamic can't be considered for objects and also the camera is moving, we use optical flow [4] instead of usual trackers [6-7] and we consider it appropriate for sequences that don't have fully occlusion problem. Fourth, we compare our proposed framework with DLT [8] and two other works and show a significant improvement over 8 challenging benchmark video sequences.

The rest of the paper is organized as follows. Section 2 refers recent works to a single framework and study disadvantage this framework. Section 3 review convolutional neural networks and introduces the LVT framework. Section 4 describes the experiments of LVT on 8 challenging video sequences.

II. Related Work

In this section, we review the related approaches for tracking problem and will have a brief look at the recent success of deep neural networks.

A simple and common method of visual tracking is template matching that has allocated itself much and different works [9-11]. In this method, the object is described by a template (like a histogram of colors) and a region is selected among available candidates that are closer to the target template. Templates have limited on the modeling object, they can represent only a sample of the object[11]. Since generative trackers were proposed in order to represent more samples. In generative trackers, a generative model is used to describe the object from previous frames. Some of these methods are based on principal component analysis [12], sparse coding [13], and dictionary learning [14]. The generative trackers model only object appearance and may often fail in cluttered background [11]. Since some works [16-18] were preferred using discriminative models instead to generative models. These models learn to explicitly separate the object from the background by a binary classifier. Discriminative trackers need to update itself and collect correct samples at during video and since they have problems such as drifting. Recently Wang et al. [8] presented a model using deep auto-encoders for visual tracking problem. They combining generative and discriminative models philosophy through updates both feature extractor and classifiers simultaneously. They show that their deep learning tracker (DLT) is more accurate and robust.

Deep networks have recently been achieving outstanding performance on challenging visual tasks such as Image classification. Krizhevsky et al. [18] trained deep convolutional network for image classification and reduced drastically error rate on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [19]. Their success lead to applied deep networks, and especially convolutional networks to other tasks in the field of computer vision [21-23].

III. The LVT Tracker

Most previous works tried to solve object localization problem through classifying candidates of the frame that indicate the selected region of that object. In this work, we inspired by [20] trying to predict directly bounding box of the object using convolutional neural networks that its last layer is replaced with a regression.

Convolutional neural networks are particular types of neural network for data processing that have grid-like topology, and unlike a regular neural network, the layers of a convolutional network neurons arrangement includes three dimensions of width, height and depth. Also they have less parameters and higher generalization due to apply convolution operator instead of multiplication matrix:

$$s[i,j] = \sum_m \sum_n I[i-m, j-n] K[m,n] \quad (1)$$

For every i,j-th pixel of the image I, take the sum of products. Each product in (1) is the color value of the current pixel or a neighbor of it, with the corresponding value of the kernel matrix K. The center of the kernel matrix has to be multiplied with the current pixel of the image I, the other elements of the kernel matrix with corresponding neighbor pixels. This accomplish by the kernel smaller than the input lead to sparse interactions and parameter sharing. This means that we need to store fewer parameters, which both reduces the memory requirements of the model and improves its statistical efficiency [23]. In these networks learned kernel matrix parameters by training algorithm in neural networks such as backpropagation.

A convolutional layer addition to convolution operator followed by a pooling function that replaces the output of the net at a certain location with a summary statistic of the nearby outputs. For example, the max pooling operation reports the maximum output within a rectangular neighborhood. In all cases, pooling helps to make the representation become invariant to small translations of the input. This means that if we translate the input by a small amount, the values of most of the pooled outputs do not change. Invariance to local translation can be a very useful property if we care more about whether some feature is present than exactly where it is [23].

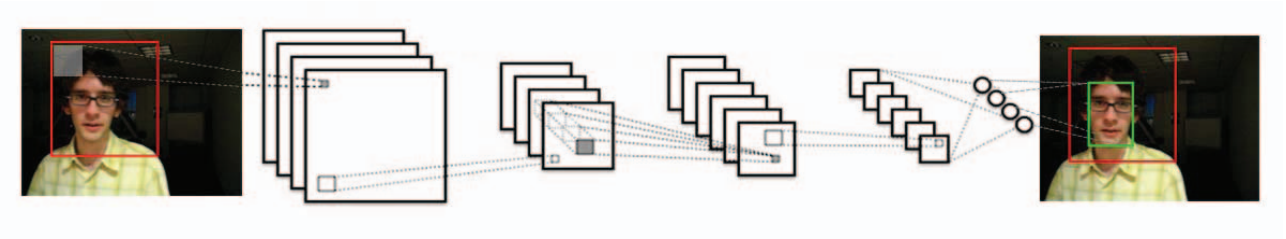

Figure 3. Proposed framework based on localization using convolutional network.

Deep convolutional networks need more data for train due to their multi-layered structure and since are incompatible with visual tracking problem that has only a sample of object. Since, we adopt Krizhevsky model [18] (AlexNet), as pre-trained model which has 5 shareable convolutional layers and 3 fully connected layers, and believe that their lower layers have generic features and can be appropriate for visual tracking problem. We replace its last layer with a regression and add dropout [24] to its third layer in order to improve model generalization capability. Since, features of the model must be so that it covers our problem challenges, we again train entire model on 20 sequences with different objects and environments, and then transfer the obtained parameters to our problem.

Visual tracking is done in terms of the bounding box that defined in the start frame. First, we select regions of object randomly and tune object of interest on our model. Second, we get approximate region in next frame based on object location in current frame using resulting flow between two consecutive frames. Third, we offer approximate region to convolutional network and predict exact location of object by regression (Figure3). Forth, since object appearance changes constantly in successive frames, we need to update our model during run. We update entire network such as DLT [8] including regressor and feature extractor.

IV. EXPERIMENTS

In this section, we compare the performance of LVT with 3 other algorithms: DLT [7], MTT [25], and IVT [12]. We use two criteria of success rate with 50% overlap and central pixel error (Euclidean distance of ground truth with predicted bounding box) on 8 challenging video sequences in order to evaluate our work. The results in Table 1 and 2 show that our work is the most significant among other works.

Table 1. Comparison of trackers using criteria of success rate (in percentage), while the number in parentheses denotes the standard deviation on 10 runs

	Ours	DLT	MTT	IVT
bird2	95.87(2.40)	65.9	9.2	10.2
car4	100(0)	100	100	100
cardark	86.4(6.3)	100	100	100
david	95.51(0.72)	66.1	68.6	92.0
shaking	98.68(0.64)	88.4	12.3	1.1
surfer	92.56(1.29)	86.5	83.8	90.5
trellis	90.29(2.95)	93.6	66.3	44.3
woman	94.36(1.36)	67.1	19.8	21.5
average	**94.2(1.95)**	**83.4**	**57.5**	**57.4**

As Table 1 and 2 show our work due to the use of regression and updating the model during video along with the amount variance. This amount difference in each run of the algorithm due to it is that the regression is very sensitive to its input. However our results show that LVT is still better than other methods.

Table 2. Comparison of trackers using central pixel error (in percentage), while the number in parentheses denotes the standard deviation on 10 runs

	Ours	DLT	MTT	IVT
bird2	10.31(0.66)	16.8	92.8	104.1
car4	4.71(0.42)	6.0	3.4	4.2
cardark	4.7(4.7)	1.2	1.3	3.2
david	6.20(0.26)	7.1	7.8	3.9
shaking	1.2(0.33)	11.5	28.1	138.4
surfer	6.02(0.13)	4.6	6.9	5.9
trellis	10.17(0.42)	3.3	33.7	44.7
woman	7.61(0.53)	9.4	257.8	111.2
average	**6.3(0.93)**	**7.4**	**53.9**	**51.9**

For evaluating our proposed method, 8 videos that is containing various challenges (Figure 1) is selected, this challenges lead to that trackers can't work correctly.

In the bird2 sequence, partially occlusion and deformation are two serious problem. Since MTT and IVT trackers after a few frames will be drifted easily. DLT due to consider tracking as a classification problem, loses part of the object in some frames (Figure 4). However, our tracker can detect the bird correctly in entire sequence.

In the car4, illumination and scale change in during video. However, all four algorithms could track object in entire sequence. For cardark sequence, our tracker can't predict bounding box of the object correctly because object is very dark and small.

In both david and trellis sequences, the target is track a face in indoor and outdoor environments. These sequences are challenging due to frequent change factors such as illumination, pose and view. Since DLT and IVT can't detect object correctly in david and trellis sequences, respectively. However, LVT can detect and track the target accurately in both sequences.

In surfer sequence, images are low resolution, however the performance of all trackers is acceptable and LVT is better than other trackers.

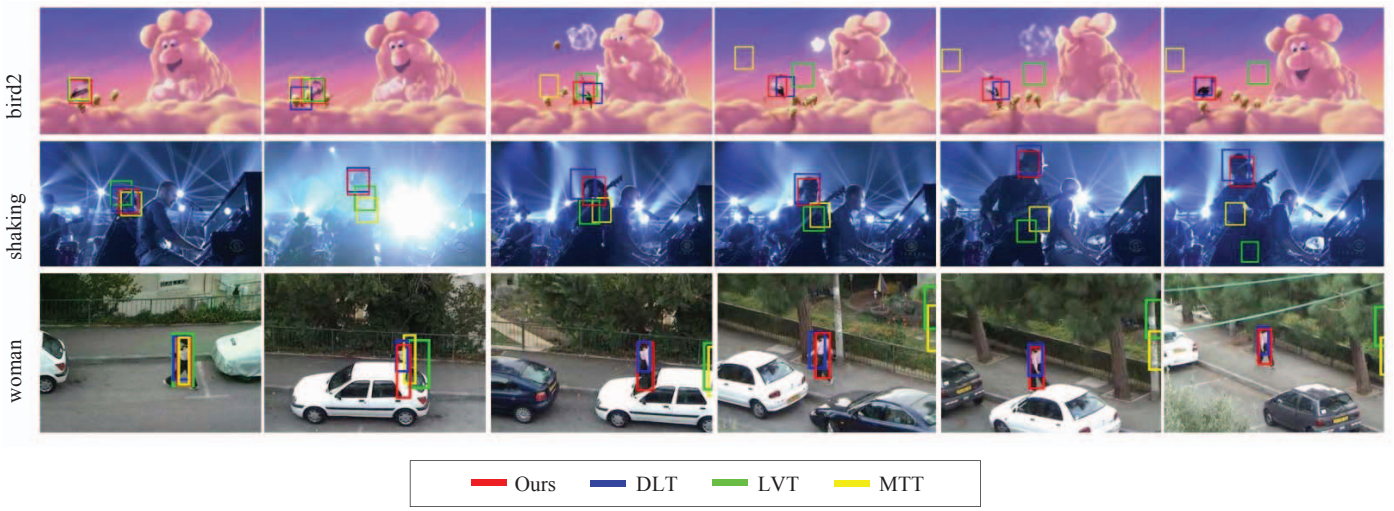

Figure 4. Results of some test sequences in terms of the bounding box.

In shaking sequence, background clutter, partially occlusion, and illumination variation are serious problem. Since MTT and IVT trackers be will drifted easily after few frames. DLT due to consider tracking as a classification problem, loses part of the object in some frames (Figure 4). LVT can detect the shaking sequence accurately and track it in entire sequence.

In woman sequence, due to the occurrence of occlusion in the early frames, MTT and IVT trackers miss the woman and be will drifted easily. DLT can follow the target but only detect part of the object as target. LVT can detect the woman accurately and track it robustly.

V. CONCLUTIONS

This paper proposes a novel framework for object tracking with localization view. The main advantage of this work over previous works is that this framework predict directly bounding box of the object using a regression. In addition to key advantage, we use a deep network with 8 layers because we believe that these networks could quickly adapt to a new object. In this work, we show a significant improvement over state-of-the-art approaches.

REFERENCES

[1] M. Kristan and et al. "The visual object tracking VOT2014 challenge results," In European Conference on Computer Vision Workshop, 2014.

[2] Y. Bengio. "Learning Deep Architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[3] Vision.stanford.edu, "Stanford University CS231n: Convolutional Neural Networks for Visual Recognition," 2015. [Online]. Available: http://vision.stanford.edu/teaching/cs231n/. [Accessed: 01- Dec- 2015].

[4] G. Farneback. "Two-frame motion estimation based on polynomial expansion," In Scandinavian Conference on Image Analysis, 2003.

[5] A. Doucet, D. N. Freitas, and N. Gordon. "Sequential Monte Carlo Methods In Practice," Springer, New York, 2001.

[6] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive Kalman filter," J. Vis. Commun. Image Represent., vol. 17, no. 6, pp. 1190–1208, Dec. 2006.

[7] N. Wang and D.-Y. Yeung. "Learning a deep compact image representation for visual tracking," In NIPS, pages 809–817, 2013.

[8] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," International Joint Conference on Artificial Intelligence, vol. 81, pp. 674–679, 1981.

[9] J. Shi and C. Tomasi, "Good features to track," Conference on Computer Vision and Pattern Recognition, 1994.

[10] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, pp. 564–577, 2003.

[11] Z. Kalal, K. Mikolajczyk, and J. Matas. "Tracking-learning-detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(7):1409–1422, 2012.

[12] D. Ross, J. Lim, R. Lin, and M. Yang. "Incremental learning for robust visual tracking," International Journal of Computer Vision, 77(1):125–141, 2008.

[13] X. Mei and H. Ling. "Robust visual tracking using l1 minimization," In ICCV, pages 1436–1443, 2009.

[14] N. Wang, J. Wang, and D.-Y. Yeung. "Online robust nonnegative dictionary learning for visual tracking," In ICCV, pages 657–664, 2013.

[15] H. Grabner, M. Grabner, and H. Bischof. "Real-time tracking via on-line boosting," In BMVC, pages 47–56, 2006.

[16] B. Babenko, M. Yang, and S. Belongie. "Robust object tracking with online multiple instance learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(8):1619–1632, 2011.

[17] S. Hare, A. Saffari, and P. H. Torr. "Struck: Structured output tracking with kernels," In ICCV, pages 263–270, 2011.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks," In NIPS, pages 1106–1114, 2012.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database," In CVPR, 2009.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation," arXiv preprint arXiv:1311.2524, 2013.

[21] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. "Simultaneous detection and segmentation," In European Conference on Computer Vision (ECCV), 2014.

[22] A. Toshev and C. Szegedy. "Deeppose: Human pose estimation via deep neural networks," In CVPR, 2014.

[23] Y. Bengio, I. Goodfellow and A. Courville, "Deep Learning," 2015. [Online]. Available: http://www.iro.umontreal.ca/~bengioy/dlbook. [Accessed: 01- Dec- 2015].

[24] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. "Dropout:A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, 15(1):1929–1958, 2014.

[25] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. "Robust visual tracking via multi-task sparse learning," In CVPR, pages 2042–2049, 2012.