Exemplar Loss for Siamese Network in Visual Tracking
Chang Shuo*, Zhang YiFan*, Sai Huang*, Yao Yuanyuan#, Feng Zhiyong*
* Key Laboratory of Universal Wireless Communications, Beijing University of Posts and Telecommunications, Beijing, 100876, P.R. China.
#School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100081, P.R. China.

Abstract: Visual tracking plays an important role in perception system, which is a crucial part of intelligent transportation. Recently, Siamese network is a hot topic for visual tracking to estimate moving targets' trajectory, due to its superior accuracy and simple framework. In general, Siamese tracking algorithms, supervised by logistic loss and triplet loss, increase the value of inner product between exemplar template and positive sample while reduce the value of inner product with background sample. However, the distractors from different exemplars are not considered by mentioned loss functions, which limit the feature models' discrimination. In this paper, a new exemplar loss integrated with logistic loss is proposed to enhance the feature model's discrimination by reducing inner products among exemplars. Without the bells and whistles, the proposed algorithm outperforms the methods supervised by logistic loss or triplet loss. Numerical results suggest that the newly developed algorithm achieves comparable performance in public benchmarks.

Key words: Deep learning; Visual Object Tracking; Siamese Network, Exemplar loss; Logistic loss.

I. INTRODUCTION

Visual tracking plays an important role in perception system to estimate the moving target trajectory. With the bounding box provided in the first frame at a specific video sequence, the visual tracking algorithm should estimate trajectory and scale variation of target in coming video frames. Basically, the major challenges in visual tracking are out-of-plane rotation, fast motion, background clutter, deformation, illumination variation, and scale variation [1]. Therefore, a discriminative feature model is crucial to visual tracking algorithms.

In general, the tracking methods [2][3][4][5][6][7] can be classified into two categories: hand-craft feature based and learning based. The feature model of hand-craft is summarized by people, e.g., histograms of oriented gradients (HOG) [8], scale-invariant feature transform (SIFT) [9]. However, they are not specially designed for visual tracking. Recently, the Siamese learning model with convolutional neural network (CNN) has made a great success in visual tracking [5][6][7]. The domain knowledge of visual tracking can be summarized and the learned Siamese feature model are more robust to diverse challenges in visual tracking. Besides, the Siamese-like algorithms can run in real-time with the help of GPU acceleration.

As shown in Figure 1, there are four different types of image patches to train neural network in Siamese algorithms. From the left to right, they are exemplar $z_a$, positive sample $x_a^+$, background sample $x_a^-$, and exemplar $z_b$. The $z_a$, $x_a^+$, $x_a^-$ patches are all cropped from the same video sequence. Compared to the $z_a$, the exemplar $z_b$ in the right-side of Figure 1 is not the same category as $z_a$, which can be considered as a distractor to $z_b$. However, recent existing algorithms, such as SiamFC [5] and SiamFC-tri [6], mainly focus on the relationship among the “a” patches (i.e., the subscript of patches in Figure 1 is a), while the distractor from exemplar $z_b$ is not considered. Therefore, their feature models may not distinguish the exemplar $z_a$ from distractor $z_b$. Additionally, the feature models' discrimination in these two algorithms is limited.

In this paper, an exemplar loss supervised Siamese network (ESiam) is proposed to improve tracking performance. As supervised by the new loss function, the Siamese feature model can not only keep the properties of SiamFC algorithm, but also separate a specific exemplar from other exemplars. Besides, a new subnetwork is designed to calculate the cross inner products among different exemplars feature level.

The main contributions of this paper are summarized as follows:

- To reduce inner products among different exemplars, a novel exemplar loss integrated with logistic loss is proposed, which can enhance the feature model’s discrimination;
- To calculate the inner products among different exemplars, a cross-inner-product layer is added into SiamFC framework;
- The obtained numerical results suggest that the
newly developed tracking algorithm has made a comparable performance with SiamFC and SiamFC-tri.

The rest of the paper is organized as follows. First, Section II reviews the related works. The main idea of proposed algorithm is described in Section III. In Section IV, the implementation details in model training and test are discussed. The experimental results are depicted in Section V. Finally, we conclude the whole article in Section VI.

II. RELATED WORK

2.1 Hand-Crafted Features Based Algorithms

In the kernelized correlation filter (KCF), Henriques et al. [2] proposes an analytic model to distinguish between the target and the surrounding environment. The feature model used in KCF is raw image pixel value. Different from KCF, a multi-expert using entropy minimization (MEEM) restoration scheme to address the model drift problem in online tracking is made by [4]. Similar to KCF, the CIE Lab color space is used as feature extractor in MEEM. At the same time, a new transfer learning based Gaussian processes regression (TGPR) [3] is introduced. However, these three methods are all based handcrafted features, which are not robust to object deformation.

2.2 Siamese Network-Based Algorithms

For the SiamFC [5] algorithm, it first proposes to use fully convolutional Siamese network for visual object tracking. The framework of SiamFC has two branches: exemplar branch and search branch. The output of exemplar branch is a matching template, where the target is located in center, and the output of search branch has larger size than the exemplar branch. The output of search branch contains the tracking target, but the target’s position is unknown. When the output of exemplar branch correlates with the search branch, the index of maximum value in response map indicates the target position in search region. Based on that, the trajectory of target can be predicted. In practice, the feature model in SiamFC is supervised by logistic loss.

To generate more dataset for Siamese network training, a discriminative correlation filter layer is added into SiamFC framework (SiamDCF) by [7]. However, in Siamese model training, the optimization direction about exemplar template only concentrates on increasing its inner product with positive samples $\langle z, x^+ \rangle$ and reducing the inner product with background samples $\langle z, x^- \rangle$, as shown in Figure 1. But the relationship among different exemplar templates are ignored. In this paper, a novel exemplar loss is introduced to minimize the inner products among different exemplar templates. Because of the novel exemplar loss, the Siamese feature model’s discrimination is enhanced.

In SiamFC-tri [6], a new triplet loss is proposed to enhance the feature model’s discrimination based on the original SiamFC framework. According to the triplet loss definition, more data elements are introduced for model training. Basically, the SiamFC-tri concentrates on enlarging the difference of $\langle z, x^+ \rangle - \langle z, x^- \rangle$. Although the SiamFC-tri has outperformed SiamFC in several benchmarks, the distractors from different exemplar are still ignored. Moreover, the accuracy improvement compared to SiamFC is not significant.

III. PROPOSED METHOD

In this section, we first make a brief introduction
about the overall tracking framework based on the SiamFC. Next, the new added cross-inner-product layer is depicted in details. Finally, the proposed exemplar loss is presented. The overall framework of proposed ESiam is depicted in Figure 2.

3.1 SiamFC Tracking Algorithm

The SiamFC [5] algorithm holds a fully convolutional Siamese neural network to predict the target’s position offset. As depicted in Figure 2, there are two branches in the Siamese network. These two branches share the same weight and structures, while their input image data is different.

Based on the training configurations in SiamFC, the input data size of exemplar branch is $127 \times 127 \times 3$, while the search branch is $239 \times 239 \times 3$ [5]. In tracking inference phase, the input size of search branch is $255 \times 255 \times 3$ [5]. Besides, the training data batch-size is 8. In practice, the 8 exemplar image patches are cropped from different videos, so does the search branch. In our proposed ESiam, we didn’t make any changes in the original settings except the batch-size.

The response map of one image pair (i.e., an image pair is made up of an exemplar image patch and a search image patch) is the correlation outputs between exemplar (the size of it is $6 \times 6 \times 256 \times 1$) and search output (the size of it is $20 \times 20 \times 256 \times 1$). The element $i$ in response map $\upsilon$ are considered to be positive (i.e. $y(i)=1$ ) if the center distance $s(i)$ between candidate sample and ground truth is within $R$:

$$
y(i) = \begin{cases} 
1, & s(i) \leq R \\
-1, & \text{other} 
\end{cases} \quad (1)
$$

As shown in Figure 1, the candidate corresponding to $y(i)=1$ is considered as a positive sample $\{x^+_i\}$, while the other candidates (i.e., $y(i)=-1$ ) are considered as background samples $x^-_i$. According to the logistic loss, the Siamese network in SiamFC is supervised by:

$$
L_1 = \frac{1}{BM} \sum_{b=1}^{B} \sum_{y_b(i)=1}^{M} \log(1+\exp(-y_b(i)\upsilon_b(i))) + \\
\frac{1}{BN} \sum_{b=1}^{B} \sum_{y_b(j)=1}^{N} \log(1+\exp(-y_b(j)\upsilon_b(i))) \quad (2)
$$

where

$$
\upsilon_b(k) = z_b \otimes x_b(k) \quad (3)
$$

$\upsilon_b(k)$ is the score of a single exemplar-candidate pair, where candidate $x_b(k)$ is a sub-region corresponding to search image patch. Symbol $\otimes$ stands for correlation operation, M and N stand for the number of positive pairs and background pairs, respectively. $b$ is the batch index, and $B$ is the batch-size (i.e., the value of B is 8 in model training.). However, the relationship of $\upsilon = z_i \otimes z_j$ is ignored by the logistic loss. To deal with this problem, a new exemplar loss is proposed in this paper.

3.2 Cross-Inner-Product Layer

The cross inner product layer is used to calculate inner products among the outputs from exemplar branch.

![Fig. 3. Two different plots about the properties of basic item in exemplar loss.](image)

For single exemplar template $z_b$, its dimension is $6 \times 6 \times 256$, which can be reshaped as a column vector $9216 \times 1$: $z^b$. In model training, there are $B = K + 1$ exemplar templates of a batch, the inner products of $\hat{z}_{k+1}$ with the other $K$ exemplars are:

$$
r_{K+1} = [\hat{z}_{k+1}^T \cdot \hat{z}_{k+1}^T \cdot \hat{z}_{k+1}^T \cdot \hat{z}_{k+1}^T \cdot \hat{z}_{k+1}^T] \quad (4)
$$

If the gradients of loss function $L(x,z)$ with respect to $r_{K+1}$ is $\frac{\partial L(x,z)}{\partial r_{K+1}}$, the gradients $\frac{\partial L(x,z)}{\partial z_{K+1}}$ is:

$$
\frac{\partial L(x,z)}{\partial z_{K+1}} = [\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_K] \frac{\partial L(x,z)}{\partial r_{K+1}} \quad (5)
$$

For the rest of $K$ exemplar templates, forward and backward propagation are similar to the variable $\hat{z}_{k+1}$. Based on equation (4) and (5), a new type of neural network layer cross-inner-product could be constructed. With the help of proposed cross-inner-product layer, the inner products among different exemplar templates can be calculated. The inner products are the input data of the exemplar loss layer, and the optimization direction of exemplar loss is to minimize the value of input data.

3.3 Exemplar Loss

As depicted in Figure 1, the algorithms of SiamFC and SiamFC-tri mainly concentrate on the relationship among the “a” patches, which are cropped from the same video sequence. The optimization direction is that the value of inner product between exemplar template and positive sample is increasing while reducing the value of inner product between exemplar template and background sample. However, the existing distractors from different exemplar templates are not considered by them, which are challenging for
the models’ discrimination. In this paper, to deal with the problem, a new exemplar loss is proposed:

\[ L_2 = \sum_{k=1}^{B} \max(0, r_k - m) \]  

(6)

where the \( r_k \) is the inner product of the specific exemplar \( z_s \) with the rest of \( K \) exemplars. And the \( m \) is a constant value in model training. The properties of basic item \( y = \max(0, x - m) \) in exemplar loss is shown in Figure (3). The gradient of

\[ \frac{dy}{dx} \]

is 0 in the range of \( (-\infty, m] \), which can relief the overfitting in model training.

By integrating the exemplar loss into logistic loss, a new target function to supervise the siamese network in ESiam is as follow:

\[ L = L_1 + L_2 \]  

(7)

IV. IMPLEMENTATION DETAILS

The proposed algorithm ESiam is implemented in matconvnet on a Dell R7300 server with a NVidia 1080Ti GPU installed. And the cuda and cudnn version are v9.0 and v7.2, respectively. The input data for model training are cropped from ILSVRC-VID [12], which is the same as SiamFC [5]. All video sequences in ILSVRC-VID are cropped to exemplar patches and search patches, which are used as input data for model training. The learning rate for model training decays from 0.01 to 0.00005 in 50 epochs. Besides, there are 10% video sequences are used as validation set. For speed consideration, only three scales are searched to estimate the target’s scale variation.

V. EXPERIMENTAL

5.1 Evaluation on OTB-2015

For the online tracking benchmark 2015 (OTB-2015) [1], it has 100 test video sequences. To compare trackers’ performance, there are two metrics proposed in [1]. And they are the distance precision and area under curve with one-pass evaluation (OPE), respectively. As presented in Figure 4, with a specific location error threshold, it is considered to be a success prediction if the center distance between the predicted bounding box and the ground truth is less than the threshold. In practice, the numbers in the legend of the distance precision plot are the precision value when the location error threshold is 20 pixels. As for the success rate plot, the numbers in the legend stand for the area under curve (AUC), which are also called AUC scores. With a specific overlap threshold \( t \), the value of success rate is the mean intersection-over-union (IOU) of all satisfied tracking results, where the IOU between a tracking result and ground truth is greater or equal to \( t \). The IOU calculation is:

\[ \text{IoU} = \frac{B_{\text{pre}} \cap B_{\text{gt}}}{B_{\text{pre}} \cup B_{\text{gt}}} \]  

(8)

where the \( B_{\text{pre}} \) is a tracking result, and the \( B_{\text{gt}} \) is the ground truth.

The performance plots are displayed in Figure 4. For the baseline tracker SiamFC, its precision score and AUC score are 77.1% and 58.2%, respectively. Compared with SiamFC, the SiamFC-tri with triplet loss has made gains about 1% in distance precision and 0.8% in AUC. However, our proposed algorithm ESiam has made large gains about 2.3% in distance precision and 1.6% in AUC referred to SiamFC, which are twice as much as SiamFC-tri. It’s no doubt that our proposed algorithm makes a larger improvement compared to SiamFC-tri. In general, the Siamese model with the novel exemplar loss has better discrimination. Besides, the Siamese network based algorithms outperform the hand-crafted features based algorithms.

5.2 Evaluation on OTB-50

The OTB-50 [1] is a more challenging evaluation dataset compared to OTB-2015. It has 50 video
sequences. The evaluation results on OTB-50 is depicted in Figure 5. The proposed ISaim outperforms the SiamFC by 3.7% in precision score and 2.6% in AUC score. As for the algorithm SiamFC-tri, the proposed ESaim has made gains about 1.6% in precision score and 1.1% in AUC score, respectively.

The accuracy improvement about proposed ESaim is significant in challenging video sequences. It indicates that the Siamese feature model supervised by the new introduced exemplar loss has better discrimination, which introduce more distractors for model training.

Fig. 5. The distance precision plots and AUC success plots of SiamFC, SiamFC-tri, and ESaim in OTB-50.
Fig. 6. The AUC scores of seven algorithms about six attributes in OTB-50: out-of-plane rotation, fast motion, background clutter, deformation, illumination variation, and scale variation. Best viewed in color.
5.3 Attribute Based Evaluation

According to the OTB-50 [1] office, a video sequence could be classified by different attributes. Due to space limit, only six of them are listed for comparison, including out-of-plane rotation, fast motion, background clutter, deformation, illumination variation, scale variation [1]. They are major challenges for robust visual tracking.

For video sequences with deformation, the proposed ESiam displays a significant improvement compared with SiamFC-tri and SiamFC. As shown in Figure 6, the ESiam has outperformed SiamFC with 4.3%. And the AUC score of SiamFC-tri is not as good as baseline tracker SiamFC. In the cases of out-of-plane rotation and fast motion, the proposed algorithm has also achieved comparable performance. Besides, when the appearance of tracking target undergoes severe illumination variation, the proposed tracker ESiam is still superior to SiamFC-tri, SiamDCF, and SiamFC. Considering the experiment results in different attributes, the excellent performance of ESiam is attributed to the fact that the discrimination of Siamese feature model can be enhanced by the new exemplar loss. With this new exemplar loss, the inner products among different exemplar templates are minimized. Besides, compared with SiamFC-tri and SiamFC, the ESiam has a better performance in scale variation and background clutter.

5.4 Qualitative Evaluation

To analyze the proposed algorithm, three groups of test results are displayed, which are selected from OTB-2015. Due to the limited space, only five image frames in each of video sequences are presented. The qualitative evaluation results in three video sequence (i.e. basketball, human3, girl2) are depicted in Figure 7.

For the basketball sequence, tracking target are surrounded by some similar basketball players. Although all trackers have correctly predicted the target position in frame 632, the SiamFC-tri and SiamFC are confused by the player at the bottom, which can be observed in frame 650, 678, 682. As for the video sequence huamn3, the SiamFC-tri and SiamFC cannot distinguish the children from the target (i.e. male adult.) Not surprisingly, similar phenomena also occur in the video sequence girl2. However, our proposed algorithm is not confused by the other exemplars or distractors. It is because that the feature model in ESiam are supervised with the new exemplar loss, where the distractors from different exemplar templates are considered.

VI. CONCLUSIONS

In this paper, a novel exemplar loss integrated with logistic loss is proposed. This new loss function not only keep the properties of logistic loss, but also consider the distractors from different exemplar templates, which are ignored by SiamFC and SiamFC-tri. Besides, a new subnetwork working in with the proposed exemplar loss is added into the SiamFC framework. Without the bells and whistles, the proposed algorithm outperforms the methods only
supervised by logistic loss or triplet loss. Furthermore, we think that all Siamese-like algorithms can benefit from the new introduced exemplar loss. In addition, with the help of communication technologies [13] [14] [15] in V2X, the visual tracking algorithm can track object through different cameras.

Acknowledgements
This work was supported in part by the National Natural Science Foundation of China under Grant (61801052, 61227801), in part by the National Key Research and Development Program of China under Grant (2018YFF0301202), and in part by the Natural Science Foundation of Beijing Municipality under Grants (4202046, 19L2022).

References
[1] WU Yi, LIM Jongwoo, YANG Ming-Hsuan. Object Tracking Benchmark[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1834-1848.
[2] HENRIQUES João F, CASEIRO Rui, MARTINS Pedro, BATISTA Jorge. High-speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(3): 583-596.
[3] GAO, Jin, Ling, Hai-bin and Hu, Weiming and Xing, Junliang. Transfer Learning Based Visual Tracking with Gaussian Processes Regression[C]/ Proceedings of European Conference on Computer Vision, Munich: Springer, 2014: 188-203.
[4] ZHANG Jianming, MA, Shugao, SCLAROFF Stan. MEEM: Robust Tracking via Multiple Experts using Entropy Minimization[C]/ Proceedings of European Conference on Computer Vision. Zurich: Springer, 2014: 188-203.
[5] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[C]/ Proceedings of European Conference on Computer Vision. Amsterdam: Springer, 2016: 850-865.
[6] DONG Xingping, SHEN Jianbing. Triplet Loss in Siamese Network for Object Tracking[C]/ Proceedings of European Conference on Computer Vision. Amsterdam: Springer, 2018: 459-474.
[7] WANG Qiang, GAO Jin, XING Junliang, ZHANG Mengdan, HU Weiming. DCFNet: Discriminant Correlation Filters Network for Visual Tracking[J]. arXiv, 2017: arXiv:1704.04057.
[8] DALAL N, TRIGGOS B. Histograms of Oriented Gradients for Human Detection[C]/ Proceedings of Computer Vision and Pattern Recognition. Boston: IEEE press, 2005: 886-893.
[9] LOWE D G, et al. Object Recognition from Local Scale-Invariant Features[C]/ Proceedings of International Conference on Computer Vision. Kerkyra: IEEE press, 1999: 1150-1157.
[10] LIN, Tsung-Yi, et al. Feature Pyramid Networks for Object Detection[C]/ Proceedings of Computer Vision and Pattern Recognition: IEEE press, 2017: 2117-2125.
[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet Classification with Deep Convolutional Neural Networks[C]/ Proceedings of International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2012: 1097–1105.
[12] DENG Jia, DONG Wei, SOCHER R, et al. Imagenet: A Large-scale Hierarchical Image Database[C]/ Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE Press, 2009: 248-255.
[13] ZHAO Zhongyuan, XU Mingfeng, LI Yong, PENG Mugen. A Non-Orthogonal Multiple Access-Based Multicast Scheme in Wireless Content Caching Networks [J]. IEEE Journal on Selected Areas in Communications, 2017, 35(12): 2723-2735.
[14] ZHAO Zhongyuan, PENG Mugen, DING Zhiguo, WANG Wenbo, POOR H V. Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(5): 1207-1221.
[15] HUANG Sai, JIANG Yizhou, GAO Yue, FENG Zhiyong. ZHANG, Ping. Automatic Modulation Classification using Contrastive Fully Convolutional NNetwork[J]. IEEE Wireless Communications Letters, 2019, 8(4): 1044-1047.