Coarse-to-Fine Siamese Network for Tooth Object Tracking

Xiaofeng Wang¹, Houjun Ao², Hongbin Shen¹, Hongbin Cai¹*, Guanghui Lu¹

¹University of Electronic Science and Technology of China, Chengdu, China
²China Aerodynamic Research and Development Center, China

*Corresponding author: caihb@uestc.edu.cn

Abstract. At present, object tracking algorithms based on deep learning have been widely used in the computer vision industry and has great application value. In this paper, we track the tooth, which has the problem of poor texture and high similarity, with the object tracking algorithm. In general, previously existing single object tracking algorithms are derived from the siamese network architecture. These algorithms have the problem of losing a tracked object with poor textures and high similarities. In order to address the issue of object loss, we put forward a novel coarse-to-fine siamese network (CFSN) for tooth object tracking. The coarse detection network is based on the features extracted by the Feature Pyramid Networks (FPN) and the Region Proposal Network (RPN). The fine detection network has a similar structure to the coarse detection network, and the main purpose is to fine-tune the results of the previous stage. Based on the coarse-to-fine network, a framework of multi-task is proposed, that is, the mask prediction branch is added to the fine detection network, which lead to significant improvements in tracking efficiency. Through the comprehensive experimental evaluation of the tooth dataset, we find that the CFSN model is more effective in tracking tooth objects than the existing methods.

Keywords: object tracking, siamese network, coarse-to-fine, multi-task, tooth object tracking

1. Introduction

Traditional visual tracking is the sequential positioning of moving objects in video. Object tracking has a pivotal position in the field of computer vision and has gained more and more attention from researchers. In any video application that requires contextual reasoning based on the target, it models the target and its contextual correspondence. In this article, we apply object tracking technology to tooth object tracking. Currently, during dental treatment, patients can only infer their own treatment results based on the dentist's descriptions and pre-existing cases. Therefore, an intuitive solution is urgently needed for a dental diagnosis. Similar to online dressing, we have raised questions that can be applied to items such as tooth replacement and whitening. As shown in Fig. 1, the algorithm can provide necessary help for dentists and patients, to speed up the process of specifying a treatment plan. Object tracking can be viewed as consisting of two tasks: classification and estimation. The first task is to obtain an initial position through classification, while the second task assigns a score to the target.
area. The current method has a big difference in the second task. In summary, We can roughly divide the previous methods into three categories. The first category assumes that the scale/ratio of the target change in a fixed manner between adjacent frames, including the Discriminative Correlation Filter (DCF) and SiamFC [1]. However, the assumption of constant scale/ratio of an object is usually invalid in reality. Therefore, the second method calculates the possible multiple-scale boxes. ATOM proposes to use gradient descent to initialize the bounding boxes multiple times, which has produced an important improvement in accuracy. But the second method has the disadvantage of heavy computation and too many additional parameters. The third category is the SiamRPN tracker family [2], [3], [4], which uses the Region Proposal Network (RPN) [5] to achieve both speed and accuracy. However, the pre-set anchor box leads to similar scores for different teeth, which seriously leads to the robustness of the model.

**Fig. 1.** The diagram of the application of tooth object tracking technology. Tooth object tracking technology can be used as a visualization to provide guidance when patients select replacing tooth, whitening tooth, or other functions.

**Fig. 2.** The diagram of the coarse-to-fine siamese network. Our basic network structure uses the siamese network, and the input has two branches, a template branch and a tracking branch. The coarse detection network can roughly locate the tooth and give a credible score. The fine detection network merge the results of the coarse detection network and output the optimized results.

This paper proposes to apply object tracking technology to tooth tracking. The tooth has the characteristics of blurry texture and high similarity, which will cause the position of the tracked object to shift. To solve the offset problem, a better balance is achieved between the real-time performance and accuracy of the algorithm. This paper is based on SiamFC [1] and SiamRPN [2] to track the tooth. better balance is achieved between the real-time performance and accuracy of the algorithm. This paper is based on SiamFC [1] and SiamRPN [2] to track the tooth. We put forward a novel method in this
article called Coarse-to-Fine Siamese Network (CFSN) for tooth object tracking. As shown in Fig. 2, this paper proposes to use a two-stage network (coarse detection and fine detection) to solve the problem of high similarity scores. This problem is caused by low texture features and high similarities. We combine the Feature Pyramid Networks (FPN) [6] and RPN [5] models to give the approximate locations and scores of the tooth in the coarse detection stage. Then, the proposed method combines results of coarse detection and uses a fine detection network to optimize the final detection results. We design a multi-task module with reference to the Mask-RCNN [7] method, in order to evaluate the mask in the fine detection stage to tune the results of our network. Our coarse-to-fine detection network provides a tuning network to optimize our results. As shown in Fig. 4, our result is superior to the SiamFC [1] and SiamRPN [2] methods. Extensive experiments are based on our proposed tooth dataset and are superior to the sota solutions.

In conclusion, we list the main contributions as follows:

- We design a coarse-to-fine detection network to address the problem of fewer target texture features and high similarity in tooth object tracking.
- We propose to use multi-task learning to enhance the fine detection stage, further tune our output results, and obtain better results.
- We achieved competitive results on the tooth dataset, compared to the previous methods. To prove the effectiveness of our proposed method, we also provide comprehensive experimental results.

2. Related Work

2.1 Detection Framework

Visual tracking usually has many things in common with object detection models, and many visual tracking algorithms also refer to some methods of object detection. For example, SiamRPN [2] achieves a win-win situation of speed and accuracy with the help of RPN structure. This structure is first proposed in Faster-RCNN [5]. Faster-RCNN [5] is the first to apply Convolutional Neural Network (CNN) to complete the prediction of proposals, and many subsequent object detection networks [7], use the RPN structure and the anchor boxes set. However, due to the introduction of anchor frames, the number of model parameters has increased significantly. And the anchors set in advance directly affect result and require heuristic adjustments. Therefore, now the main research direction of researchers has appeared anchor-free detectors, like predicting bounding boxes at points near the center of an object or detecting and grouping a pair of corners of a bounding box. To obtain size information of different sizes, some models also use a feature pyramid structure [6] to use the semantic information of different network layers.

2.2 Single Object Tracking

Visual tracking technology is usually divided into single target tracking and multiple target tracking. In this paper, we propose to solve the problem of tooth object tracking with target tracking technology. At present, the proposed method utilize the siamese network, which has occupied most of the single target tracking. The siamese network treats tracking roughly as a matching problem, which uses the target of the first frame as a template to match other frames. For the first time, SiamFC [1] proposes to treat a tracking problem as a matching problem, which opens the way for the siamese network to be applied to object tracking. SiamRPN [2] combines the siamese network with the RPN [5] structure in the detection field, achieving a win-win situation of speed and accuracy. DasiamRPN [3] improves model's robustness from three perspectives of training data, template update, and search area. DasiamRPN [3] enables the SiamRPN [2] network to adapt to long-term tracking. SiamMask [8] just has one more segmentation branch than SiamRPN [2], that is, one more learning task. SiamRPN++ [4] and SiamDW [9] can be said to push the siamese series network to a new peak, solving the problem of how to deepen the siamese network. SiamFC++ [10] proposes four aspects of content to improve target tracking, including decomposing classification and state estimation, removing classification branch
disambiguation, tracking no longer rely on prior knowledge such as aspect ratio, and increasing the estimated quality score branch. Due to the fuzzy texture features and high similarity of the tooth, the previous methods are difficult to apply to tooth object tracking task. Therefore, our proposed CFSN model is based on the SiamFC [1] and SiamRPN [2] models, proposing to use a coarse-to-fine detection structure to solve the problem of objects with similar scores. Based on SiamRPN [2], we propose to use FPN [6] network to enhance the feature representations of the tooth. The proposed CFSN method combines the results of coarse detection and uses a fine detection network to optimize the final detection results, alleviating the problem of similarity scores for teeth. Based on the Mask-RCNN [7] model, we propose to use the multi-task module in Fig. 3 to enhance the fine detection network to adjust the network results.

3. Proposed method

As shown in Fig. 2, this paper proposes to use a coarse-to-fine detection network to solve the problem of high similarity scores for tooth object tracking. The coarse detection network first performs rough positioning and classification of the tooth, and the fine detection network fine-tunes the classification scores and target boxes. In the fine detection stage, it is proposed to use a multitasking module, which helps the fine detection network to obtain better results.

3.1 Coarse-to-Fine Detection

The proposed CFSN model uses the siamese network structure, using the feature representations obtained by the ResNet network. We combine the FPN structure [6] and the RPN structure [5] to use the features extracted by the feature extraction network. After ROI Align, we input the feature distribution into a classification branch and a positioning branch, respectively. The classification branch uses two-class cross-entropy loss for training, as bellow:

\[
L_{C_{cls}} = - \log \left[p_i p_i^* + \left(1 - p_i\right)\left(1 - p_i^*\right)\right] \tag{1}
\]

where \(p_i\) and \(p_i^*\) denote the probability of anchor prediction as to the object and the ground truth. We train the positioning branch to use the regression function to determine the length and width of the center and the object. As follow:

\[
L_{C_{reg}} = \text{smooth}_{\sigma_i} \left(t_i - t_i^*\right) \tag{2}
\]

where \(t_i = \{t_x, t_y, t_w, t_h\}\) denote the offset of the predicted anchor, \(t_x, t_y, t_w, t_h\) denote the center abscissa of the anchor, the center ordinate of the anchor, the width of the anchor and the height of the anchor, respectively. \(t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}\) is the same as \(t_i\), representing the actual offset of the anchor.

And \(\text{smooth}_{\sigma_i} (x) = \begin{cases} 0.5x^2 \times \frac{1}{\sigma^2} & \text{if } |x| < \frac{1}{\sigma^2}, \text{ where } x \text{ and } \sigma \text{ denote input and hyperparameter,} \\ |x| - 0.5 & \text{otherwise} \end{cases}\)

respectively. The coarse detection network initially outputs a target anchor and its corresponding score, and then the fine detection network uses the extracted features to further fine-tune the output of the coarse detection. The fine detection network is similar to the coarse detection network, and both have a classification branch and a positioning branch. And the training method of the two branches is consistent with the coarse detection stage. We weight and fuse the scores and coordinates of the candidate regions for coarse detection and fine detection to obtain the final results.
3.2 Multi-Task Module

Based on the existing methods [7], [8], we propose to use multi-task learning to enhance the tracking ability of the fine detection network. We add a segmentation branch in the fine detection stage and use the results of the segmentation branch to obtain better positioning and classification effects, as shown in Fig. 3. We train our segmentation branch in the same way as the previous method, and the formula is as follows:

\[
L_{mask} = \sum_n \left( \frac{1 + y_n}{2wh} \sum_y \log \left( 1 + e^{-c_n y^{ij}} \right) \right)
\]  

where \( y_n \in \{\pm 1\} \) denotes a ground-truth binary label of the response of a candidate window (RoP). This RoP refers to the concept of SiamMask [8]. And \( wh \) denotes the size of a pixel-wise ground-truth mask \( c_n \). Let \( c_n^{ij} \in \{\pm 1\} \) denote the label corresponding to pixel \((i, j)\) of the object mask in the \( n \)-th candidate RoW. And \( m_n^{ij} \) denote prediction frame, the weight of classification network and the weight of feature extraction network, respectively. The segmentation branch uses shallow features to be more sensitive to positions and edge gradients, making the target mask more accurate.

![Fig. 3. The architecture of the multi-task module. On the basis of the original fine detection network, the proposed CFSN method adds a segmentation branch.](image)

![Fig. 4. The visualized results of different methods. Both SiamFC and SiamRPN have errors in tracking object lost. The proposed CSFN method also has this problem in the case of coarse detection network. Through the fine detection network, the model successfully tracked the tooth object again.](image)
3.3 Loss Function

We use the previously mentioned losses, namely the losses of rough detection network and fine detection network, as our objective function and formulate them in the following form:

\[ L = \alpha L_{c_{\text{cls}}} + \beta L_{c_{\text{reg}}} + \gamma L_{f_{\text{cls}}} + \delta L_{f_{\text{reg}}} + \eta L_{\text{mask}} \tag{4} \]

where \( \alpha, \beta, \gamma, \delta \) and \( \eta \) denote denote hyper-parameters for balancing the relation of loss. \( L_{c_{\text{cls}}} \) and \( L_{f_{\text{reg}}} \) are similar to \( L_{c_{\text{cls}}} \) and \( L_{c_{\text{reg}}} \), representing the loss functions corresponding to the two branches of the fine detection network. We use the Adam \([11]\) optimization algorithm to train the above loss functions.

### Table 1. The results of tooth object tracking

| Method          | Accuracy | FPS |
|-----------------|----------|-----|
| SiamFC+AlexNet  | 0.584    | 86  |
| SiamFC+VGG16    | 0.601    | 80  |
| SiamFC+ResNet50 | 0.546    | 59  |
| SiamFC+ResNet101| 0.323    | 50  |
| SiamFC+ResNet41 | 0.618    | 65  |
| SiamRNP+AlexNet | 0.646    | 160 |
| SiamRNP+VGG16   | 0.650    | 148 |
| SiamRNP+ResNet50| 0.593    | 109 |
| SiamRNP+ResNet101| 0.351   | 92  |
| SiamRNP+ResNet41| 0.665    | 120 |
| CFSN(ours)      | 0.705    | 83  |

4. Experiments

4.1 Implementation Details

The proposed CFSN method uses ResNet41 as the backbone. In the data preprocessing stage, we use rotation, scaling, noise addition, and brightness changes to expand the number of datasets. We train the proposed CFSN model to use the Adam \([11]\) optimization algorithm to optimize in an end-to-end manner. We set 9 kinds of Anchor box generation rules for tooth, and the specific rules are anchor _rarios = \([0.5, 1, 2], [0.4, 0.5, 1, 1.5], [0.5, 0.618, 1, 1.3, 1.618] \) , which corresponding generation rules are called \( A_3, A_4, A_5 \). We set anchor _scales = \([2, 4, 8]\) , batch_size is 8, and set \( \alpha, \beta, \gamma, \delta \) and \( \eta \) is set to 0.5,1,0.5,1 and 0.8.

4.2 Evaluation and Datasets

We use the customized evaluation index of online Object Tracking Benchmark (OTB) as the standard, that is, the Average Overlap Rate (AOR). In the experiment, we use "accuracy" instead. The overlap rate refers to the overlap ratio between the predicted box and the ground truth, and the formula is as follows:

\[ O = \frac{\text{Area}_{\text{pre}} \cap \text{Area}_{\text{gt}}}{\text{Area}_{\text{pre}} \cup \text{Area}_{\text{gt}}} \tag{5} \]

where \( \text{Area}_{\text{pre}} \) and \( \text{Area}_{\text{gt}} \) denote the area of the prediction box and the area of the ground truth, respectively. And the average overlap rate is calculated from \( O_{\text{ave}} = \frac{1}{N} \sum O \). Since there is no related literature on the tracking of tooth, all the data in this paper comes from anonymous user tooth videos collected by the backstage of Software.
4.3 Experimental Results

Extensive experiments have been done on the tooth dataset to compare with previous methods. In TABLE 1, the SiamFC [1] and SiamRPN [2] methods have different results with different backbones. When the backbone is ResNet101, the accuracy of SiamFC [1] and SiamRPN [2] is reduced to the lowest. With the deepening of the network, Frames Per Second (FPS) also shows a downward trend, where the larger the FPS value, the faster the model processing speed. When using the ResNet41 as the backbone, SiamFC and SiamRPN reach the highest correct rate and have a competitive FPS. The proposed method achieves superior results than previous methods, and the Accuracy reaches 0.705. Moreover, the FPS reaches 83, which satisfies the applicability of the method. Fig. 4 shows that, when using the highest scores as tracking tooth, SiamFC [1] and SiamRPN [2] have emerged tracking tooth object are lost. However, the proposed CFSN method can tune the output of the network with the fine detection network. Compared with the previous method, the proposed CFSN method achieves a competitively tracking effect without losing tooth.

Table 2. The results of using FPN different layer features

|                  | Accuracy | FPS |
|------------------|----------|-----|
| CFSN+ResNet41    | 0.665    | 120 |
| CFSN+ResNet41+FPN_p1 | 0.670    | 116 |
| CFSN+ResNet41+FPN_p2 | 0.668    | 115 |
| CFSN+ResNet41+FPN_p3 | 0.665    | 117 |

4.4 Further Analysis

In TABLE 2, we have done ablation experiments for FPN networks. The FPN module extracts the shallow features of tooth, which can better detect low-level semantics such as tooth texture. The model using the p1 layer has achieved relatively competitively results, with accuracy and FPS of 0.670 and 116 respectively. The accuracy of the network can be improved by adding the FPN module, and the speed is flat.

Table 3. The results of using different anchor box generation rules

|                  | Accuracy | FPS |
|------------------|----------|-----|
| CFSN+ResNet41+FPN_p1 | 0.670    | 116 |
| CFSN+ResNet41+FPN_p1+A_3 | 0.665    | 125 |
| CFSN+ResNet41+FPN_p2+A_3 | 0.661    | 119 |
| CFSN+ResNet41+FPN_p3+A_3 | 0.672    | 115 |

In TABLE 3, the proposed CFSN method has been extensively experimented with under different anchor box generation rules. Each anchor of A_j only generates 9 anchor boxes, so the FPS is the highest. A_5 sets special parameters for the shape characteristics of tooth. Therefore, the setting of A_3 is more in line with the task of tooth tracking and obtains the best accuracy.

Table 4. The results of whether to use multi-module

|                  | Accuracy | FPS |
|------------------|----------|-----|
| CFSN w/o mask    | 0.694    | 100 |
| CFSN w mask      | 0.705    | 83  |

In TABLE 4, we test the proposed multi-task network. It can be seen that adding the Mask branch effectively increases the accuracy by 0.01. Although the Mask branch is only added in the fine detection stage, the speed loss is still obvious. A large amount of calculation of the convolutional layer of the Mask branch causes a certain speed loss.
5. Conclusion
In this paper, we are concerned that the application of object tracking methods for tooth object. We propose a novel coarse-to-fine siamese network (CFSN) structure, including a coarse detection network and a fine detection network. The proposed CFSN method is based on Convolutional Neural Networks and combines the FPN structure and the RPN structure. The fine detection network predicts more detailed positions and more accurate scores of tooth based on the results of the coarse detection network, effectively solving the problem of losing tooth. In addition, we propose to use a multi-task learning framework in the fine detection network, that is, to add a segmentation branch to predict and rotate the object box to improve the tracking accuracy. We conduct an experiment on the dataset we have collected. Comprehensive experimental results, compared with sota methods, indicate the proposed CFSN method shows effectiveness and prominence for tooth object tracking.

References
[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, “Fully-convolutional siamese networks for object tracking,” in European conference on computer vision. Springer, 2016, pp. 850–865.
[2] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with siamese region proposal network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.
[3] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, “Distractor-aware siamese networks for visual object tracking,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 101–117.
[4] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “Siamrpn++: Evolution of siamese visual tracking with very deep networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4282–4291.
[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.
[6] T.-Y. Lin, P. Doll’ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[7] K. He, G. Gkioxari, P. Doll’ar, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[8] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, “Fast online object tracking and segmentation: A unifying approach,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 1328–1338.
[9] Z. Zhang and H. Peng, “Deeper and wider siamese networks for real-time visual tracking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4591–4600.
[10] Y. Xu and Z. Wang, “Li, x.; yuan, y.; and yu, g. 2020. siamfc++: Towards robust and accurate visual tracking with target estimation guidelines.”AAAI.
[11] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.