IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 5, 2020, accepted April 3, 2020, date of publication May 14, 2020, date of current version June 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994440

Online Tracking and Relocation Based on a New Rotation-Invariant Haar-Like Statistical Descriptor in Endoscopic Examination

HAIFAN GONG[1], (Student Member, IEEE), LIMIN CHEN[1], CHANGHAO LI[1], JUN ZENG[2], XICHEN TAO[1], AND YUE WANG[3], (Student Member, IEEE)

[1]School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]Jiangxi Provincial People's Hospital, Nanchang 330031, China
[3]Second Clinical Medical College, Nanchang University, Nanchang 330031, China

Corresponding author: Limin Chen (Chenlimin@ncu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773051 and Grant 6171101044, in part by the National Innovation and Entrepreneurship Program for College Students under Grant 201910403076, and in part by the Jiangxi Provincial Department of Science and Technology under Grant 20171ACB20007, Grant 20151BBE50046, Grant 20142BBE50035, and Grant 20151BAB207052.

ABSTRACT In the gastrointestinal biopsy, online tracking and relocation of the region-of-interest are essential to early diagnosis and surgical intervention of colorectal cancer. However, it is challenging for the examiner to track and retarget the optical biopsy site due to interfering factors, e.g. violent rotation of the lens, illumination variation, shape deformation, and target long-time-lost. Previous works may not effectively handle the mentioned challenges due to the complexity of gastrointestinal environment and the limitation of data. In this work, we construct an online tracking and relocation framework based on the concept of detection and tracking, which is dramatically adapted to the inherent characteristics of the gastrointestinal biopsy image. To effectively distinguish the target area from the gastrointestinal biopsy, we designed a new rotated invariant Haar-like statistical descriptor which is robust for rotating and illumination changes. The descriptor is based on the sector-ring difference under the circular sampling area. A simplified statistical random forest discriminator based on confidence statistics is proposed to complete the preliminary screening of the potential tracking target. In order to further estimate the location of the target, a supervised support vector machine is introduced to rank the candidate target regions. Based on proposals of Siamese network and the random forest, a location refinement fusion has been proposed to determine the location and the confidence of the tracking area. Extensive experiments on various gastrointestinal videos, which consists of open source and self-collected data, demonstrate that the proposed framework is superior to the mainstreams methods in accuracy and robustness.

INDEX TERMS Relocation, online tracking, Haar-like feature, random forest, Siamese network.

I. INTRODUCTION

Screening for colorectal cancer is conducive to early detection, diagnosis, and treatment of colorectal cancer. It is the key to preventing colorectal cancer and reducing the cumulative mortality of colorectal cancer. The main screening methods for colorectal cancer include endoscopic examination, fecal occult blood detection, and CT colonography. With the development of technologies such as optical coherence tomography (OCT) [1] and narrowband imaging (NBI) [2], the non-invasive optics-based visual examination has replaced conventional biopsy. When it comes to internal gastrointestinal surgery, medical endoscopy is the most common assistant instrument. However, because of the lack of obvious anatomical features, sparse features, and many similar areas, the gastrointestinal video in vivo is not friendly to examiners. Not only is the gastrointestinal tissue is prone to deformation but because of the patient peristalsis and doctor's operation, rapid movement, illumination changes, long-term out of FOV of target and motion blurring all commonly occurring, there are several challenges presented even for the experienced examiner. Therefore, there is an urgent need

The associate editor coordinating the review of this manuscript and approving it for publication was Xianye Ben.

for accurate and robust tracking and relocation system which is semi-automated. By combining this system with manual selection of regions of interest (ROI), or other recognition models of potential gastrointestinal lesions, the gastrointestinal examiner is able to reduce the false detection rate, shorten the examination time, and relieve the pain of the examinee. In addition, the automatic tracking and relocation system is also significant for the development of routinization and popularization of gastrointestinal biopsy. At present, the problem of tracking and relocating gastrointestinal biopsy area is generally solved by a universal tracking algorithm. For instance, Nader Mahmoud *et al*. proposed using ORB-SLAM to track the region of interest in gastrointestinal surgery scenes [3]. An extended monocular SLAM method was proposed by Oscar G. Grasa *et al*. to process images from hand-held standard monocular endoscopes to calculate the motion of endoscopes in real time [4]. Bingxiong Lin *et al*. proposed a parallel tracking mapping (PTAM) framework for stereo tracking in minimally invasive surgery (MIS) [5]. In order to guarantee long-term tracking performance in the gastrointestinal video, we need a robust descriptor to relocate the target area. Generalized descriptors such as LBP [6], SIFT [7], SURF [8], ORB [9], mean projection transform, dual-tree complex wavelets, and Haar-like descriptor [10] are common feature extraction methods. However, there exist challenges like sparse features caused by the poor image quality, light changes, tissue deformation, target disappearance, severe rotation, and occlusion in the gastrointestinal environment, which affects the above-mentioned feature extraction methods' performance.

High-dimensional convolution features enable the precise positioning of tracking regions, but their descriptive regions are often too small to exclude similar areas. The low-dimensional artificial features provide more shape context information to distinguish similar regions, but the accuracy of localization is not high. Therefore, we propose an endoscope visual field tracking and relocation framework that combines artificial statistical features with high-dimensional depth features. The framework includes a new robust statistical descriptor, a simplified random forest discriminator based on confidence statistics, a candidate region ranking and filtering component based on ranking support vector machine, and a location refinement component based on Siamese network and probability fusion. As this framework is designed to handle the strong rotation and long-time retargeting in the gastrointestinal biopsy, we name it rotation-invariant relocatable tracker (RIRT). We evaluate the RIRT on several gastrointestinal videos in vivo. The results show that the RIRT is superior to state-of-the-art methods in both accuracy and robustness while guaranteeing real-time computation.

The main contributions of this work are summarized as follows. We design a new robust Haar-like descriptor called RIBHD to handle the severe rotation and illumination changes that often occur in the gastrointestinal biopsy. Based on RIHBD, a simplified random forest discriminator based on confidence statistics is constructed to realize the preliminary screening of tracking areas efficiently. The proposed features contain sufficient shape context information, so a large number of similar regions are distinguished in the preliminary screening.

The tracking issue is treated as the detection issue in Siamese network, which means there is no need to update the template online. The response graph obtained by cross-correlation operation achieves target detection and boundary box regression, which significantly improves the real-time performance of the algorithm. To further improves the robustness of RIRI framework, we propose a probabilistic fusion method, which fuses the probabilities obtained from the discriminant results of random forest, support vector machine, and the classification branch of Siamese networks.

The proposed RIRT absorbs the advantages of the statistical features and high-dimensional self-learning features. It further improves the accuracy and robustness of tracking and retargeting while granting real-time computing. By evaluating the proposed RIBHD on different scenarios such as NBI, white light, heterogeneous, and extensive rotation, the results show that the RIBHD outperforms than Haar-like statistical descriptor based on rectangular region difference. We evaluate the proposed RIRT and other mainstream trackers like DaSiamRPN, ECO, TLD on extensive endoscopic video sequence. It shows that the proposed RIRT outperforms in f-measurement, EAO, average overlap rate, etc. It is worth noting that we also evaluate the RIRT on challenging retargeting dataset. The results show that the proposed RIRT performs better than other retargeting available trackers.

The rest of this paper is organized as follows. Section 2 compares and analyses the descriptors and system framework related to gastrointestinal biopsy area tracking. Section 3 proposes a RIRT tracking and relocation framework, which combines manual statistical features and deep high-dimensional features. A new Haar-like statistical feature descriptor based on circular sampling region and the sector-ring difference is designed. A weak discriminator is also proposed to allow for preliminary selection of the ROI. After that, we propose a location refinement component based on Siamese network and probability fusion. Section 4 evaluation evaluates the performance of the new rotation invariant descriptor and the RIRT through different types of gastrointestinal videos. Section 5 draws a conclusion.

II. RELATED WORK

There are challenging characteristics of gastrointestinal biopsies, such as a lack of sufficient anatomical features, extensive similar areas, sparse features, and low contrast, inherently. Besides, during the optical biopsy, the potential lesion area isn't labeled, which make it possible for it to be misjudged by examiner. Because changes in illumination, occlusion, target disappearance, rapid movement, and violent rotation often occur in the gastrointestinal biopsy, the difficulty of tracking and location further increased. For these reasons, a robust framework for the online gastrointestinal tissue tracking and relocation is proposed.

A. THE EXISTING ROTATED-INVARIANT HAAR-LIKE DESCRIPTOR

Based on the advantages of the area difference operator, such as better anti-illumination performance, convenient and efficient implementation, we designed a new rotation invariant Haar-like operator called RIBHD to resolve the mentioned challenges. The first application of target detection based on Haar-like descriptor is proposed by Paul Viola *et al*. [10]. The features were extracted by Haar-like feature and the cascaded week classifiers trained by AdaBoost, which reduce the missed detection rate and false detection rate while guaranteeing real-time computation. Barczak *et al*. proposed an efficient rotational detector based on Haar-like feature [11]. They acquired the original feature from the classifier by conversion algorithm and calculated two features to find the approximate equivalent value of any angle to realize the detector. In order to handle the rotating target detection, Shaoyi Du *et al*. constructed a discriminator based on Haar-like feature of 26.565 degrees of rotation, which achieved good results on CMU-MIT dataset [12]. Barczak *et al*. proposed a detection framework based on mobile terminals and sensors [13]. This framework proposed a Haar-like descriptor with rotation invariance at both feature level and classifier level, which efficiently realized the detection of rotating targets. Sadiq *et al*. proposed a new method to detect objects by rotating the Viola-Jones detector at different angles [14]. On this basis, the algorithm is extended to a variety of Haar-like features by means of rotation and asymmetry. Tests on Umist and CMU-PIE datasets show that the algorithm is successful at detecting a target in common scenes under different scales, locations, directions, and illumination conditions [15]. These Haar-like descriptor propulsors handle the rotation of target under the 2D plane. However, the above-mentioned Haar-like descriptor realizes rotated target detection by rotating rectangular sampling area and equivalent substitution. They cannot adapt to the round field of vision, lack of obvious anatomical features and severe rotation in the gastrointestinal examination. Therefore, we propose a new Haar-like descriptor based on circular sampling area and sector-ring difference. The descriptor effectively adapts to the circular field of vision and the anatomical structure of intestinal tissue, and can better handle the severe rotation and illumination changes during the gastrointestinal biopsy.

B. AN OVERVIEW OF THE EXISTING STATISTICAL METHOD

Because of poor imaging quality, blurred motion, and high light of local tissue in gastrointestinal biopsy images, it is challenging to extract the target information from traditional descriptors based on the point and angle feature. Therefore, we use the joint weak classifier based on the descriptor of the region feature statistics mentioned above to make a statistical decision to preliminarily select the region of interest. In the aspect of sample discriminator based on a statistical decision, Zhang *et al*. proposed an object detection method based on local binary pattern (LBP) histogram feature [16]. To deal with the target drift and occlusion, Babenko *et al*. propose an online multi-instance learning tracking method (MIL) [17]. This method constructs a training set decision-maker by constructing a bag of positive samples and a bag of negative samples. In order to resolve the segmentation image patches depending too much on template matching, Dinh *et al*. proposed the CXT based on dense sampling [18]. Ye *et al*. proposed a feature statistical analysis method based on region difference Haar-like descriptor to preliminary screen the candidate tracking regions [19]. On this basis, Ye *et al*. proposed a new differential statistical model in 2017 to achieve real-time retargeting in gastrointestinal biopsy examination [20]. However, when constructing the classifier based on statistical features, the descriptor does not have rotation invariance inherently. Therefore, a large number of positive samples are enhanced by the affine transformation, so that the detector achieves the desired results. We designed a new descriptor with a natural rotation invariant property, which reduces unnecessary data enhancement. Based on this new descriptor, we propose a statistical strategy based on confidence weighted positive and negative samples to handle the sample imbalance.

C. AN OVERVIEW OF THE CONVENTIONAL TRACKING METHOD

Since there are soft tissue deformations, illumination changes, long time out of field-of-view(FOV), scale shifts and occlusion during the gastrointestinal biopsy of the digestive tract [20], [21], higher accuracy and robustness for tracking and repositioning ROI have been put forward. Universal tracking and repositioning frameworks are encountering significant challenges in intestinal gastrointestinal scenarios. It is challenging for conventional generative methods to deal with target tracking and relocation under the gastrointestinal environment. For the Kalman filter [22] and Mean-shift [23], the Kalman filter fails when facing the occlusion of the target, while the fast motion and scale change make the Mean-shift useless for tracking the target. For discriminant generation methods like MIL [17] and structured SVM [24], multi-instance learning is sensitive to training samples, while structured SVM can't deal with the fast motion of gastrointestinal tissue. Even methods based on correlation filter, such as the kernelized correlation filter (KCF) [25] and scaled correlation filter (DSST) [26], which attracts many researchers for their proficient real-time services and accuracy, still fail to cope with endoscope fast motion, occlusion, target deformation and target loss because they only search near the tracking area. Even the discriminant correlation filter CSR-DCF [27] with confidence fails to track because of the local highlight in the gastrointestinal video.

For long-time tracking or tracking with shape change of the target, detection and tracking method should be introduced to handle the deformation and partial occlusion of the target. At the same time, the model is updated continuously through online learning mechanism, which makes the tracking effect more robust and reliable. Table 1 shows the mainstream

TABLE 1. Summary of mainstream trackers for comparison.

Tracker	Appearance	Method
MedianFlow[34]	Image local pixels	Forward-backward error
TLD[28]	LBP and image patches	Detection and tracking
KCF[25]	Image patches generated by rotation matrix	Kernel correlation filter
CSR-DCF[27]	Image patches generated by rotation matrix	Discriminant kernel correlation filter
MIL[17]	Haar-like feature	Multi-instance learning
ECO[34]	Convolution feature and LBP	Factorized convolution operator
OTR[19]	Binary Haar-like feature	Simplified random forest + structured SVM
SiamMask[36]	Convolution feature	Correlation operation + semantic segmentation
DaSiamRPN[39]	Convolution feature	Correlation operation + regional proposal network
RIRT	RIBHD and convolution feature	Confidence statistical random forest + correlation operation

trackers for comparison. TLD [28] proposed by Kalal *et al*. and OTR [19] proposed by Ye *et al*. are representative methods. In the gastrointestinal video, the robustness and tracking success rate of TLD is unsatisfactory, which is mainly caused by server tissue deformation. OTR makes up for the shortcomings of TLD in feature exaction, but its descriptor does not have the natural rotation invariance. To overcome the above weakness, we propose a statistical descriptor that deals with the rotation of ROI, illumination changes, and a large number of similar areas in FOV.

D. AN OVERVIEW OF THE DEEP-LEARNING TRACKING METHOD

With the continuous improvement in computing power and the increase of effective data, the accuracy and robustness of feature extraction methods based on depth learning have been dramatically improved. The first application of deep learning in the tracking field is DLT [29], which was proposed by Naiyan *et al*. in 2013. It combines the idea of off-line pre-training and online fine-tuning to overcome the lack of data in the tracking process to achieve object tracking. Objects of different classes are easily distinguished by high-level CNN features, while the similar distractors in the background can be distinguished by low-level CNN features. Based on the above observation, L Wang *et al*. proposed FCNT [30] which builds feature screening network and two complementary heat-map prediction networks to prevent tracker drift. However, FCNT is not robust to occlusion, which is a very common phenomenon in gastrointestinal biopsy. MDNet [31] is a multi-domain learning framework based on CNN, which achieved the championship of VOT 2015 [32] by separating domain-independent information from domain-specific information to improve performance. TCNN [33] synthetically evaluates the target to be tracked in the current frame by preserving the appearance model of the target in several successful tracking frames through CNN tree. It is robust to the change of the appearance of the target, illumination and the disappearance of the target in a short time. Martin *et al*. proposed ECO [34] in 2017, which combines depth features with shallow features to achieve high-quality tracking performance, but its real-time performance still needs to be improved. In order to avoid the incompleteness caused by artificial features, feature learning is integrated into the process of modeling. These frameworks misbehave due to the changes of illumination and soft tissue deformation in the endoscopic environment, and its time costing to adjust the parameters of the network.

Because of the excellent performance in accuracy and real-time, tracking methods based on Siamese networks, such as SiameseFC [35], SiamMask [36], SiamRPN [37] and SiamRPN++ [38], have attracted extensive attention of researchers. To alleviate the probability obtained by a response graph is unreliable when the target disappears, Zhu *et al*. proposed DaSiamRPN [39]. However, during tracking and retargeting of the gastrointestinal ROI, it is possible for the above-mentioned models to fail due to the inadaptability of pre-learning models. The Siamese network-based trackers search the most similar region in the adjacent region of the former frame, which might not be the ground truth because of the rapid motion and displacement occlusion. At the same time, there are plenty of similar areas in the gastrointestinal images. The Siamese network-based trackers find the maximum response area in the response map by generating multiple anchors. The correlation detection in the feature domain ignores the context information of the target region, which make the trackers usually find a similar region rather than the ground truth. To resolve the tracking fails caused by fast motion in the 2D plane, Jianren Wang *et al*. proposed a tracking method which combines motion estimation with the Siamese network [40]. Because of the uncertainty of motion and local highlight in gastrointestinal biopsy, it is hard for motion estimation methods such as optical flow method [41] to achieve convincing results. Therefore, a simplified random forest discriminator for the global screening of candidate regions is proposed. By extracting more shape context information with the RIBHD, some of these similar regions are distinguished. Based on the candidate regions, the target is positioned accurately by cross-correlation operation in high-dimensional features extracted from Siamese networks.

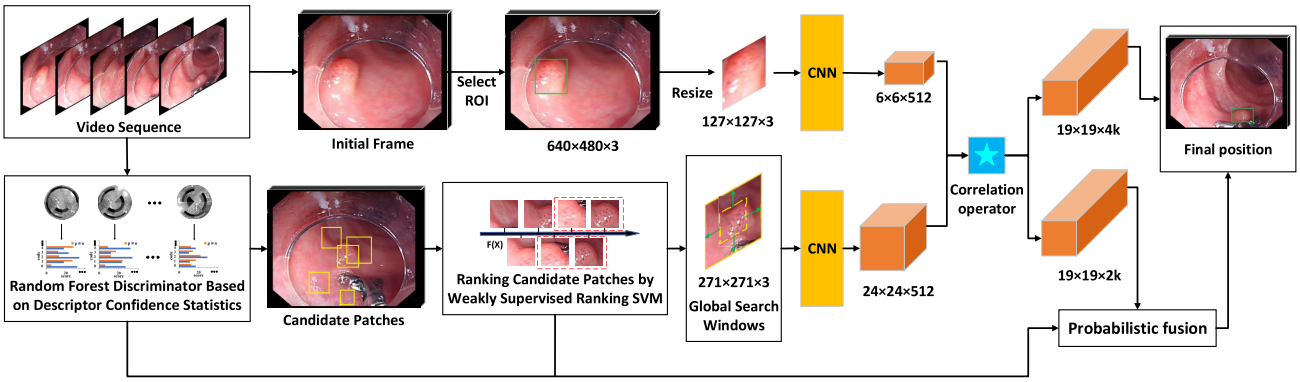


FIGURE 1. Online tracking and retargeting system based on RIBHD statistical discrimination and Siamese network probability fusion.

III. METHODS

In order to track and relocate under the complicated endoscopic environment, we propose the following methods. A new statistical descriptor called RIBHD is first designed for preliminary screening of candidate regions. Then, we construct a confidence-based statistic random forest to screen candidate target regions in the whole image preliminarily. In order to find the image patches close to the target, a ranking support vector machine is introduced. Based on proposals of Siamese network and random forest, a fusion framework is proposed to locate the tracking area and its confidence accurately. The pipeline of the proposed approach is shown in Figure 1.

A. A NEW ROTATE INVARIANT BINARY HAAR-LIKE DESCRIPTOR

Although the conventional Haar-like descriptor guarantees real-time computation, it does not have the natural rotation invariance feature. In our work, a new rotation invariant binary Haar-like descriptor called RIBHD is proposed to extract features in the endoscopic environment. Its advantages are as follows. Haar-like feature extraction is realized by the gray-scale difference of candidate pixel regions and binary coding, which is conducive to ensuring real-time performance. Besides, rather than pixel-level difference, the proposed Haar-like region descriptor is based on regional differences, which is ready to handle the illumination change. Within our new rotation invariant Haar-like descriptor, it is easier to handle the rotation of the vision which often occurs during the endoscopic examination.

The conventional Haar-like feature is extracted by differentiating the gray values of small rectangular windows in horizontal or vertical directions. Although it is robust for changes in illumination, it still has obvious deficiencies in rotation invariance. Thus, the affine transformation is applied to realize the rotation invariance. However, on the one hand, the affine transformation (mainly rotation enhancement) of the original image will increase processing time. On the other hand, expanding the rotating angle of training samples is unreachable to cover the actual angle of endoscope random rotation, which influences the accuracy of the tracker. Although other researchers have proposed rotation invariant Haar-like descriptors [11]–[15], most of these works utilize the equivalent substitution to guarantee rotation invariance, which makes it hard to accurate positioning in the gastrointestinal environment.

Thus, we design a new rotation invariant binary descriptor (RIBHD). The proposed RIBHD not only handle the changes of illumination but also depict the features of image patches that are segmented from the circular view of endoscopy. Based on the idea of gray centroid method and local matching, the proposed RIBHD guarantees the robustness of image translation and rotation. The construction procedure of RIBHD from sample template construction, encoding, rotation invariance implementation, and fast computation is shown below.

1) SAMPLING TEMPLATE CONSTRUCTION AND CODING DESIGN

Conventional rectangular Haar-like descriptors have poor robustness to lens rotation and tissue deformation in endoscopic scenes. Thus, we propose a sampling template based on sector-ring sampling area, as shown in Fig.2(a). We regard a single sector-ring region as an atomic sector-ring and use S to represent it. S is a quaternion vector $[r_0, r_1, \theta_0, \theta_1]$, in which r_0 and r_1 represent the inner and outer diameters of the sector-ring respectively, while θ_0 and θ_1 represent the starting and ending angles of the sector-ring relative to the horizontal right directions, respectively. It is worth noting that the sector is also a special sector-ring, which the inner diameter of the sector-ring is 0. The sampling area consisting of four atomic sector-rings adjacent to each other, which is denoted by A, namely $A = \{S_1, S_2, S_3, S_4 | S_i \textit{ is adjacent to each other}\}$. By randomly selecting several non-overlapping sector-ring sampling areas, the sector-ring sampling template is obtained. We define the sector-ring sampling template as T, where $T = \{A_1, A_2, \cdots, A_k | 1 \leq k\}$, and k is the number of the randomly selected sector-ring sampling area.

In order to obtain a randomly selected and non-overlapping sector-ring sampling area, we need to determine two

FIGURE 2. Construction of sector-ring sampling template. (a) represents the adopted equal area difference method. (b) represents the equal-spacing difference method. (c) represents the atomic sector-ring coding method.

parameters: concentric circle equal fraction a and sector equal fraction b. As the area difference increases, the sampling area of the inner ring is too small while the sampling area of the outer ring is too large, which is not conducive to feature extraction. Thus, we construct a concentric circle with the same area (Fig.2(a)) instead of a concentric circle with the same area difference (Fig.2(b)). The basis of determining parameters a and b is to make the area of the sampling area as small as possible to ensure the performance of feature extraction while taking the computational efficiency into account. After dividing the area, we numbered the pictures from the inside to the outside in anticlockwise direction. As the unique identifier (r, w) of the partitioned patches, it is in the order of increasing r and w, as shown in Fig.2(c). The yellow part is $r = 1$, the orange part is $r = 2$, the blue part is $r = 3$, and the yellow area 1th's identifier is (1, 1). By analogy, we get the unique identifier of each atomic sector ring in the sample candidate region. The sampling template T is constructed by randomly selecting k groups of non-repetitive and adjacent identifiers.

It is worth noting that when constructing templates to extract Haar-like features from target regions, their relative independence should be guaranteed. Therefore, when selecting the set of sampling regions, we control that any identifier of four partitioned regions only appears in a set of sampling identification sets. Because the sampling templates are randomly selected, which is beneficial to reducing the variance of the model, the random forest achieves advantageous generalization results and anti-over-fitting ability without additional sampling template selection.

In order to obtain a binary Haar-like similar feature descriptor, the sum of gray values of the upper and lower parts of the pixels is differentiated, and the sum of gray values of the left and right parts of the pixels is differentiated. The sum of the pixel strengths of the upper left, upper right, lower left, and lower right regions shown in Fig.2(a) are represented by S_1, S_2, S_3, and S_4, respectively. The binary encoding method is as

$$BC_{vertical}(A) = \begin{cases} 1, & S_1 + S_3 - S_2 - S_4 \geq 0 \\ 0, & otherwise, \end{cases}$$

$$BC_{horizontal}(A) = \begin{cases} 1, & S_1 + S_2 - S_3 - S_4 \geq 0 \\ 0, & otherwise, \end{cases}$$

$$BC(A) = 2 \times BC_{vertical}(A) + BC_{horizontal}(A), \quad (1)$$

where A is the select-ring sampling area, namely the area consist of S_1, S_2, S_3, and S_4. If $S_1 + S_3 - S_2 - S_4 \geq 0$ and $S_1 + S_2 - S_3 - S_4 < 0$, we obtain a binary code 10, which is the value of $BC(A)$. By defining the above differential encoding regular, the RIBHD's encoding formula for the current image patch is defined as

$$RIBHD(S) = \sum_{i=1}^{k} BC(A_i) \times 2^{2i-1}, \quad (2)$$

where $BC(A_i)$ represents two Haar-like codes and k is the number of random generated non-overlapping sector-ring sampling areas. It is worth mentioning that all operations here are binary operations. For a sampling template with four randomly selected sampling patches, $k = 4$, if the sequence of codes is 10,00,11,01, the binary coding of the sampling template on the image patches is 10001101. By converting the coding into an integer, we introduce confidence weighting to calculate the distribution of the feature encoding. More details are depicted in Section 3.2.

2) IMPLEMENTATION OF ROTATION INVARIANCE

The circular sampling template does not have inherent rotation invariance. Inspired by the implementation of rotation invariance of ORB [9], which is an efficient algorithm proposed by Ethan Rublee *et al.* to extract and describe the feature point, we transform the original image of three channels into a grayscale image and calculate the grayscale centroid of image patches as

$$\bar{u} = \frac{\sum_{(u,v)\in\Omega} u \cdot f(u, v)}{\sum_{(u,v)\in\Omega} f(u, v)}, \quad (3)$$

$$\bar{v} = \frac{\sum_{(u,v)\in\Omega} v \cdot f(u, v)}{\sum_{(u,v)\in\Omega} f(u, v)}, \quad (4)$$

where $f(u, v)$ is the gray value of the pixel with (u, v) coordinate, Ω is the set of target regions, that is, the set of each pixel contained in the current image patch after downsampling, and $(\bar{u}, \bar{v})$ is the gray centroid coordinate. By connecting the center of the image patch with the gray centroid $(\bar{u}, \bar{v})$ as the spindle, we rotate the image patch to the horizontal direction, as shown in Fig.3. For any image patch used to train the random forest, we need to do this before extracting Haar-like features. Based on the idea of local rotation

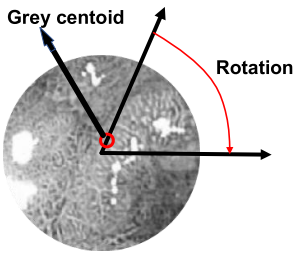


FIGURE 3. Rotating spindle to realize rotation invariance.

angle matching, the rotation invariance of the descriptor is guaranteed.

It is worth mentioning that one of the inherent characteristics of the endoscopic video is that as the motion of objects and the position of external light source changes, there exists more noise in the endoscopic image than in the image of a common scene. More noise means that there exists a deviation between the calculated gray centroid and the real gray centroid, which affects the robust sampling mode proposed. In order to handle and combat this phenomenon, common image filtering methods such as median filtering based on spatial domain, total variational image denoising, etc are used. We then denoise the images by Gauss filter. After that, we downsample the image patches to alleviate the inaccuracy of noised gray centroid calculation in the endoscope environment.

3) FAST COMPUTING IMPLEMENTATION

The conventional Haar-like descriptor calculates the sum of gray values of any rectangular area in the original image in constant time by the integral graph. However, in the proposed RIBHD, it is hard to apply the polar coordinate integration statistics method to calculate the sum of regional gray values, because the coordinate system of the frames differs. Thus, we firstly reduce the size of all the scanned image patches to a quarter of the original size. Accordingly, the size of the pre-determined sampling template is also reduced to a quarter of the original size. Then the gray center of gravity within the image patch is computed by multi-threading method and, and its spindle is rotated to the right horizontal direction. Next, we define the sampling template as a convolution core. At the same time, we stitch all N patches rotated into a 2r*2r*n tensor, where r is the radius of the reduced patches. By calling the bit-wise and reduce operation, the features of each image patches extracted by RIBHD under different sampling templates are obtained simultaneously. We apply the Numba [42] framework during all the previous work, which ensures the real-time performance of the framework.

B. ONLINE SIMPLIFIED RANDOM FOREST CONSTRUCTION BASED ON CONFIDENCE STATISTICS

The target is unpredictable in the endoscopic view due to rapid motion and regions of interest (ROI) disappearance and reappearance. At the same time, there are challenges such as unbalanced positive and negative samples for training and multiple similar regions. Thus, we propose a confidence statistics Haar-like random forest discriminator (CSHRF) to select the locations of ROI preliminarily. We apply the RIBHD to extract the features of image patches under different scales, and then feed the feature vectors to the simplified random forest for weak discrimination. In the learning stage of CSHRF, a confidence weighting strategy is introduced, which cope with the imbalances in positive and negative sample sets. By evaluating the image patches selected from the random forest, the results show that the performance of the CSHRF is better than OTR's binary Haar-like random forest.

In order to obtain patch samples for training the random forest, a sliding fixed-size window scan the endoscopic view field under different scale levels. We mark the scanned image patches as $\{(x, y), R\}$, in which (x, y) denote the center of the image patch, and R is the radius of the circular sampling area. The conventional method of creating the training samples for the random forest is to set a threshold to divide positive and negative samples according to overlap rate. In order to alleviate the impact of small training samples and class imbalance in the training forest, we determine the confidence of positive and negative samples as

$$C_P = \begin{cases} 1, & 0.8 <= olr(pa) <= 1 \\ 0.8, & 0.6 <= olr(pa) < 0.8 \\ 0.6, & 0.4 <= olr(pa) < 0.6, \end{cases}$$

$$C_N = \begin{cases} 0.5, & 0 <= olr(pa) < 0.1 \\ 0.4, & 0.1 <= olr(pa) < 0.2 \\ 0.3, & 0.2 <= olr(pa) < 0.3 \\ 0, & 0.3 <= olr(pa) < 0.4, \end{cases}$$

$$olr(pa) = \frac{S(pa) \cap S(gt)}{S(pa) \cup S(gt)}, \tag{5}$$

where olr represents overlap rate, pa represents the image patch scanned by the sliding window, and gt represents the ground truth of the current frame. C_p and C_n is representing the confidence of a positive sample and the negative sample, respectively. For instance, if the overlap rate between image patch $pa1$ and the ground truth is 75%, it is a positive sample with confidence $C_p = 0.8$. If the overlap rate between image patch $pa2$ and the ground truth is 35%, it is a negative sample with confidence $C_p = 0.2$. In order to alleviate the deterioration of prediction accuracy caused by concept drift, we set the threshold of overlap rate of positive samples to 0.4 and that of negative samples to 0.3. Samples with overlap rates between 0.3 and 0.4 are not included in the statistics. For example, if the overlap rate of an image patch is 0.35, its confidence is 0.

By increasing the weight of positive sample confidence scores which alleviates the imbalance between positive and negative samples, the CSHRF is more accurate and robust. It is worth noting that we constructed a FIFO queue to partially update negative samples in the endoscopic sequence, which further improves the efficiency of the random forest. Meanwhile, the queue assigns the random forest memory effect to handle the relocation.

According to the construction method of sampling template described in Section 3.1, we randomly generate M group sampling template T to extract features from each training sample. It is worth mentioning that each randomly generated sampling template contains k sets of non-overlapping sampling regions. In other words, each sampling template contains $4k$ atomic sector-rings. We choose the atomic sampling number not less than one-fourth of the total number of atomic

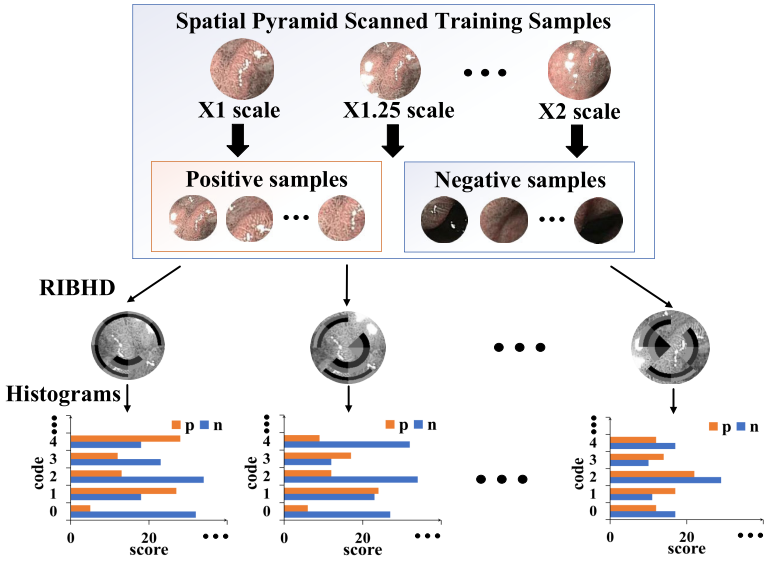


FIGURE 4. Confidence-based random forest based on binary Haar-like statistical descriptor.

sector-rings to ensure the effectiveness of feature extraction. The confidence histograms of positive and negative samples are obtained by the sample template T randomly generated through M groups, and the confidence histograms of $2M$ samples with values ranging from 0 to $2^{2k}-1$ are constructed. These histograms show the distribution of confidence of positive and negative samples sampled by RIBHD. For a positive sample p, the confidence C_p of the positive sample is counted to M confidence histograms. In order to evaluate the probability that the candidate image patches in the next frame belong to the tracking region, we synthesize M confidence histograms to calculate the posterior probability of each candidate image patch. The formulaic probability calculation method is as

$$P(pa_i|d_m) = \frac{\sum C_p}{\sum C_p + \sum C_n},$$
$$P(pa_i|d) = \frac{1}{M}\sum_{m=1}^{M} P(pa_i|d_m), \tag{6}$$

where $\sum C_p$ and $\sum C_n$ are the scores of positive sample confidence histogram and negative sample confidence histogram d_m bits corresponding to sampling template m, respectively. The posterior probability $P(pa_i|d_m)$ of the inquired image patch is retrieved with the binary coding value d_m of the sampling template m. By averaging the probability within M sets of sampling templates, we get the probability of the image patch belonging to the tracking area. It should be noted that when $\sum C_p = \sum C_n = 0$, we assign $P(pa_i|d_m)$ to 0 directly to avoid the denominator being zero.

The construction of a simplified random forest based on confidence statistics is described above. By defining the sampling template shown in Fig.2(a) and combining with rotation processing based on gray centroid method (Fig.3), a random forest classifier (Fig.4) based on RIBHD is established. By using this random forest, the candidate ROI of the next frame image is preliminarily screened to facilitate the next step of location selection and fine discrimination.

C. ONLINE TRACKING AND RELOCATION SYSTEM BASED ON RIBHD

The new descriptor and the simplified random forest enable us to achieve the preliminary selection of global candidate ROI while ensuring the real-time ability. In order to further determine the location of target patches for tracking or relocation, we propose an online tracking and relocation system based on support vector machine candidate region ranking and prediction probability fusion.

1) RANKING CANDIDATE REGIONS OF SUPPORT VECTOR MACHINES BASED ON MEMORY EFFECT

RIBHD-based distribution statistics is a rough estimate of the actual distribution but isn't accurate enough to represent the proximity of each candidate region of the ground truth. Besides, the scene reappearance often occurs in the gastrointestinal biopsy. It is necessary to ensure memory effect in the design of the tracking and relocation system. Therefore, the ranking support vector machine (SVM) with a memory effect is proposed to further rank the patches after the preliminary discriminating. It has further improved the accuracy of the tracking and relocation system by feeding the position of the image patches in front of the sorting into the next component.

Ranking learning has been widely exploited in the application of information retrieval. In recent years, researchers have applied ranking learning on video tracking. Ranking SVM learns the ranking model by using eigenvectors and

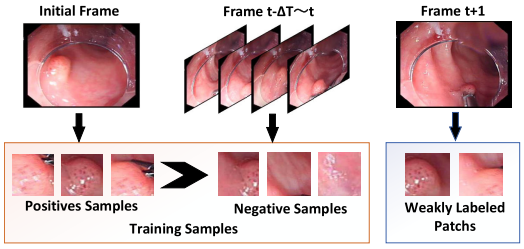


FIGURE 5. Construction of patch ranking samples based on memory effect.

ordered classification as training data. Here, we divide the feature vector set of the training set into a positive sample set $S^P = \{x_i : i = 1, \ldots, N_p\}$ and negative sample set $S^N = \{x_j : j = 1, \ldots, N_n\}$, where N_p and N_n represent the number of positive samples and negative samples for training, respectively. For any $x_i \in S^P, x_j \in S^N$, there are $x_i \succ x_j$, namely x_i ranks higher than x_j. By sorting support vector machines, we hope to learn a sorting function $F(x)$ to satisfy the conditions shown below.

$$F(x) = w^T \Phi(x)$$
$$x_i \succ x_j \Leftrightarrow F(x_i) > F(x_j) \tag{7}$$

Among them, w represents the weight vector and $\Phi(x)$ represents the implicit eigenfunction acting on the eigenvector X. In the training process, we need to minimize the sorting error of the sorting support vector machine and maximize the sorting interval, which is transformed into the following quadratic programming problem-solving form.

$$\min_{w,\eta} \|w\|^2 + C \sum_{ij} \eta_{ij}$$
$$s.t.\ w^T \left(\Phi\left(x_i\right) - \Phi\left(x_j\right)\right) \geq 1 - \eta_{ij}, \eta_{ij} \geq 0$$
$$i = 1, \ldots, N_p, \quad j = 1, \ldots, N_n \tag{8}$$

C is a trade-off parameter to balance interval and training error. The feature of the image patch we selected is the feature of RIBHD after binarization in the initial screening differential stage.

Since there exists challenges like the change of light source, the re-entry of the tracking object in the process of endoscopy tracking and retargeting. If we only use the image of the previous frame as a training sample to sort the candidate image patches selected by the random forest in the next frame, it is difficult to achieve the desired results. Therefore, we set a memory effect time Δt. For the pictures in Δt time interval, different learning rates are set to update the support vector machine online, to achieve target selection under weak supervision.

In the gastrointestinal biopsy, challenges such as rapid visual field movement, occlusion, and tissue deformation often arise. In recent years, the tracking algorithm based on Siamese network has attracted enormous attention because of its high accuracy and fast speed, but it has poor robustness to relocation and motion blurring. Since the endoscopic scene has the above features naturally, it isn't available to achieve convincing performance in our task. Therefore, we combine the global location screening and memory-based sorting method to obtain potential search areas in the global scope.

2) POSITION REFINEMENT COMPONENT BASED ON SIAMESE REGION PROPOSAL NETWORK

The Siamese Region Proposal Network(SiamRPN) evolved from the Fully-convolutional Siamese network(FCSiam). SiamRPN applies off-line training convolution neural network to search the location of the tracking area in the next frame and obtain the location of the detection target boundary box and the confidence of the target. SiamRPN extracts the feature of a template image $z(w \times h \times 3)$ and target area image x through BN-AlexNet. After that, classification and regression branches are obtained, respectively. In order to obtain the response map, the cross-correlation operation between the extracted template area and the similar branches of the target area is carried out as

$$g_\theta(z, x) = f_\theta(z) \star f_\theta(x), \tag{9}$$

where $\star$ stands for correlation operation. Through cross-correlation operation on the depth feature map, the classification branch and regression branch are obtained, and the location and confidence of the tracking area are eventually determined. However, when the tracking target disappears, it is challenging for the Siamese network to locate the target area near the last target search location. The disappearance of the target is very common in endoscopy. Thus, we adopt a Siamese network framework based on DaSiamRPN and SiamRPNBIG. By expanding the search area from 255*255 to 271*271 and the number of channels from 256 to 512, we alleviate the problem that it is challenging to search again when the target disappears. The network structure is shown in Figure 6.

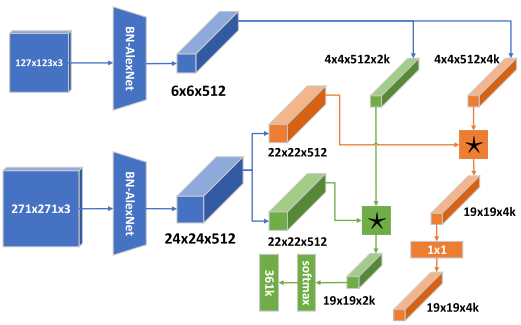


FIGURE 6. Structure of DasiamRPN with SiamRPNBIG. The blue represents the feature extraction based on Siamese network. The green represents the classification branch. The orange represents the bounding box regression branch.

3) PROBABILITY FUSION BASED ON THE MODEL INTEGRATION

The strategy of training DaSiamRPN partly alleviate the problem of unreliable confidence. However, due to the inherent particularity of our gastrointestinal scenarios, their

models trained with common datasets are often unreliable in the process of online tracking and relocation. Therefore, we propose a probabilistic fusion method based on model integration to improve the reliability of confidence in tracking and relocation. The proposed probabilistic fusion method combines the confidence C_{final} of final region searched by Siamese network and the random forest's statistical result C_{hrf} of the final region.

$$C_{final} = \lambda_1 C_{hrf}(pa) + \lambda_2 C_{siam}(pa) \tag{10}$$

Among them, pa represents the final region, C_{hrf} represents the final confidence of pa through RIBHD while C_{siam} represents the final probability obtained by Siamese network screening near pa. λ_1 and λ_2 are super-parameters to control the weight of each component. We set them to 0.4 and 0.6, respectively. The formula for calculating C_{hrf} is as

$$\begin{aligned} C_{hrf}(pa) &= \frac{S_{pa}}{S_{max}}, \\ S_{max} &= \max\{S_1, \ldots, S_j; j \in N+\}, \\ S_j &= \frac{e^{RP_j}}{\sum_{l=1}^{RN} e^{RP_l}}, \end{aligned} \tag{11}$$

where S_{pa} represents the confidence of a specific image patches output by softmax and S_j represents the probability that each image patches filtered by the random forest are a positive sample. RN denotes the candidate region number and RP_l denotes the specific confidence of the current region. Through the above methods, we get more accurate confidence in tracking results and improve the robustness of the whole system.

IV. EXPERIMENTS

We designed three sets of experiments to evaluate the proposed RIBHD statistical descriptors and RIRT in a large number of clinical endoscopy video sequences. In the first set of experiments, we compare the proposed RIBHD statistical descriptor with the statistical descriptor [19] in OTR in different endoscopy scenarios to evaluate the detection performance of the descriptor and its decision maker in endoscopy scenarios. The second and third experiments were designed to evaluate the tracker. In the second group of experiments, we constructed five representative repositioning test sequences to test the repositioning performance of the proposed RIRT framework. In the third group, we collected 10,646 frames of clinical endoscopy data to evaluate the tracking performance of RIRT and other state-of-the-art trackers.

In terms of data sources, part of our data comes from the video sequences we collected and labeled from hospitals, and the other part comes from the data set constructed by Ye *et al*. [19]. The whole data set has 10,646 frames with 640 × 480 resolution, and its rendering mode includes white-light and NBI rendering. It is worth mentioning that in order to describe more precisely the biopsy areas of interest in endoscopy, we re-labeled the potential lesion areas in some video sequences provided by Ye *et al*. The above RIRT framework runs on a self-built Ubuntu 16.04-based workstation, including I7 7800X CPU, 16GB RAM, and Nvidia Titan Xp GPU. In order to ensure the real-time performance of the whole framework, we exploit CPU multithreading and GPU CUDA to accelerate the algorithm. These algorithms are all implemented on the Python 3.6 platform. Our current implementation speed is 26 frames per second on average.

A. EVALUATION METRICS AND PARAMETER CONFIGURATION

In order to evaluate the performance of the RIBHD-based statistical decision-making random forest, we selected four video sequences from the data set under white light, NBI, severe rotation, and light source conversion scenarios. Considering the size difference between the scanned image patch and the ground truth, we define the normalized average overlap rate nao to evaluate the RIBHD-based CSHRF discriminator. The nao is consist of the average overlap rate $\overline{olr}$ and the average size of ground truth $\overline{gts}$. Let TP denotes the number of image patches obtained by CSHRF and $olr(itp)$ (Eq5.) represents the overlap rate between image patches and ground truth, the average overlap rate $\overline{olr}$ is defined below.

$$\overline{olr} = = \frac{\sum_{itp=0}^{TP} olr(itp)}{TP} \tag{12}$$

After that, we let w and h denote the width and height of the real area of the image and idx denote the number of frames in the video sequence, in order to get the definition of $\overline{gts}$ is shown below.

$$\overline{gts} = \frac{\sum_{idx=0}^{n} w_{idx} \times h_{idx}}{n} \tag{13}$$

By defining $\overline{gts}$ and $\overline{olr}$, we get the normalized average overlap rate nao. The pas represents the size of the scanned image patch.

$$nao = \overline{olr} \times \frac{pas}{\overline{gts}} \tag{14}$$

In order to evaluate the tracking performance of the proposed RIRT framework, we select the average overlap rate and location error as the core evaluation criteria. In addition, we use the average overlapping expectation (EAO), precision, recall, and F1-measure to evaluate the performance of each tracker. For a video set tracked by a tracker tk, the precision p_{tk} and recall r_{tk} are defined as follows.

$$p_{tk} = \frac{C(T \cap G)}{C(T)} \tag{15}$$

$$r_{tk} = \frac{C(T \cap G)}{C(T)} \tag{16}$$

T is the frameset calibrated by the tracker in the video set, and G is the frameset containing the target area in the

video set. In order to track the set of frames whose overlap rate is more than 0.5, $C(U)$ represents the number of elements in the set U. F-measure is a weighted harmonic average of precision and recall rate. Here we consider that precision and recall rate is equally important, so we define the form of F1-measure F_{tk} as follows.

$$F_{tk} = \frac{2 \cdot p_{tk} \cdot r_{tk}}{p_{tk} + r_{tk}} \tag{17}$$

It is worth noting that because many models do not directly trade-off the tracking results according to confidence, we uniformly set the confidence threshold to 0.6. If it is less than the threshold, we consider that the current frame does not change the tracking result, and if it is greater than or equal to the threshold, it will be the final tracking result.

The proposed RIRT needs to initialize RIBHD, CSHRF, and Ranking SVM in the first frame. In order to effectively extract the features of the gastrointestinal environment, RIBHD is constructed with ring equal fraction $a = 6$, sector equal fraction $b = 8$, and the area of atomic sector-ring is about 200 square pixels. The numerical experiments show that the number of sampling areas $k = 6$ and the number of decision trees $M = 8$ in constructing CSHRF. In the initialization part of Ranking SVM, we use the first 80 samples to initialize the weight w of the sorted support vector machine according to its real overlap rate. In order to obtain image patches for training the random forest, we use pyramid scanning to obtain circular image patches. Then, we calculate the overlap rate between the surrounding square and the calibrated area to obtain positive and negative samples and their confidence.

B. RIBHD BASED STATISTICAL RANDOM FOREST

In order to evaluate the proposed RIBHD, we constructed a simplified random forest based on RIBHD and compared it with the simplified random forest based on Haar-like used by Ye *et al*. In the process of relocation based on the random forest, we use the t frame to detect the location of the biopsy area in the $t+n$ frame image. For frame t, we use multi-scale circular window scanning to generate about three thousand image patches. After confidence strategy screening, training samples are obtained. The random forest is trained by image patches segmented from images, and the image patches segmented from $t+n$ frame are inputted into the random forest.

Each image patch input into the random forest will get a confidence level. According to the confidence level of the image patch, we rank it in descending order and take the first *tp* image patches for evaluation. Here, the parameters that have a major impact on the screening process include the number of RIBHD sampling areas k and the number of decision trees M (i.e., the number of sampling templates) of CSHRF. The above parameters are selected in a wide range to determine the approximate range of the above parameters. After that, we find out the optimal interval of sampling area number k should be between 3 and 7, and the optimal interval of decision tree number M should be between 4 and 12.

Next, we evaluate the proposed RIBHD in the number intervals of sampling areas, decision trees, and video sequences containing NBI, white light, vigorous rotation, and heterogeneous images. The criterion used here is *nao* (Eq.14). When selecting the number of sampling patches k, we fix the number of decision trees to 6. When choosing the number of decision trees M, we fix the number of sampling patches to 4. The evaluation results are shown in Table 2 and Table 3.

TABLE 2. Average overlapping under different number of sampling patches k. We fix the number of decision trees to 6.

Method-Data Type	3	4	5	6	7
OTR[19]-White Light	0.39	0.44	0.43	0.45	0.37
RIBHD-White Light	**0.43**	**0.45**	**0.48**	**0.50**	**0.49**
OTR[19]-NBI	0.40	0.47	0.47	0.43	0.42
RIBHD-NBI	**0.45**	**0.54**	**0.57**	**0.61**	**0.61**
OTR[19]-Vigorous rotation	0.35	0.34	0.36	0.34	0.23
RIBHD-Vigorous rotation	**0.43**	**0.45**	**0.49**	**0.52**	**0.50**
OTR[19]-Heterogeneous	0.27	0.25	0.26	0.26	0.25
RIBHD-Heterogeneous	0.27	**0.26**	**0.29**	**0.28**	**0.27**
OTR[19]-Average	0.35	0.38	0.38	0.37	0.32
RIBHD-Average	**0.39**	**0.43**	**0.46**	**0.48**	**0.47**

* **Bold** numbers represent the better performance, same as follows.

TABLE 3. Average overlapping under different numbers of trees M. We fix the number of sampling patches to 4.

Method-Data Type	4	6	8	10	12
OTR[19]-White Light	0.42	0.45	0.38	0.39	0.41
RIBHD-White Light	**0.49**	**0.50**	**0.50**	**0.51**	**0.52**
OTR[19]-NBI	0.32	0.33	0.33	0.32	0.33
RIBHD-NBI	**0.57**	**0.61**	**0.61**	**0.62**	**0.63**
OTR[19]-Vigorous rotation	0.29	0.33	0.31	0.27	0.30
RIBHD-Vigorous rotation	**0.47**	**0.51**	**0.53**	**0.54**	**0.55**
OTR[19]-Heterogeneous	0.25	0.25	0.24	0.25	0.27
RIBHD-Heterogeneous	**0.26**	**0.28**	**0.30**	**0.28**	**0.31**
OTR[19]-Average	0.32	0.34	0.32	0.31	0.33
RIBHD-Average	**0.45**	**0.48**	**0.49**	**0.49**	**0.50**

Through the above numerical experiments, the performance of the proposed RIBHD based CSHRF performs much better than that of the Haar-like descriptors proposed by Ye *et al*., which is better than that of random ferns and other descriptors. Eventually, it is determined that CSHRF gets better performance when six sampling areas and ten decision trees are selected. In the absence of data enhancement, the performance of the RIBHD-based random forest exceed that of OTR using data enhancement.

In addition to the above parameters, we also evaluated the effect of the number of pre-screened image patches selected

TABLE 4. Detection rate under different number of pre-screened patches.

Patch's Number	#20	#40	#60	#80	#100	#120	#140
OTR[19]	0.61	0.54	0.53	0.51	0.47	0.45	0.35
RIBHD	**0.67**	**0.65**	**0.64**	**0.62**	**0.62**	**0.59**	**0.56**

on the detection rate. The results are shown in Table 4. It is worth noting that the detection rate here is the average overlap rate of the real area in the first several image patches selected in the initial screening process. As shown in Table 4, the average overlap rate of RIRT frameworks is much higher than that of OTR frameworks. As the number of candidate regions increases, the detection rate decreases gradually. However, the RIRT framework we proposed is always significantly better than the OTR framework. When the number of pre-screened image patches is adjusted from 60 to 80, the detection rate decreases significantly. Therefore, in order to take account of both computational efficiency and detection efficiency, we set the number of pre-screened image patches in RIRT to 60.

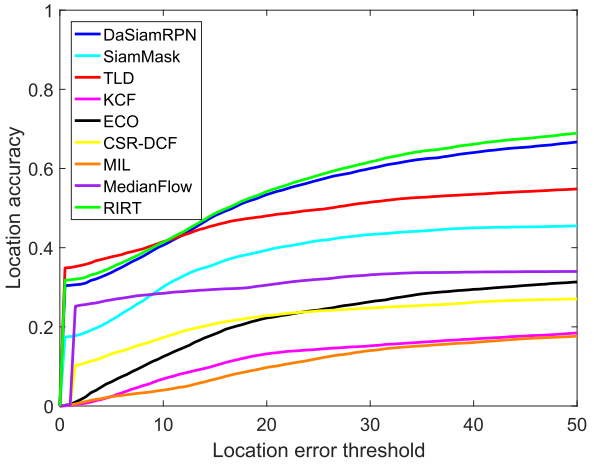
C. ONLINE RETARGETING AND TRACKING

The gastrointestinal tract is prone to tissue deformation; endoscope lens rotation is prone to occur in the operation process; and the inner wall of the digestive tract is smooth, easy to produce specular reflection and high light characteristics. Also, due to the narrow intestinal environment, tracking targets are often not in the field of vision. These characteristics make it a challenge to track and reposition the biopsy tissue area online. At the level of quantitative analysis and qualitative analysis, we compare the proposed RIRT method with other mainstream trackers.

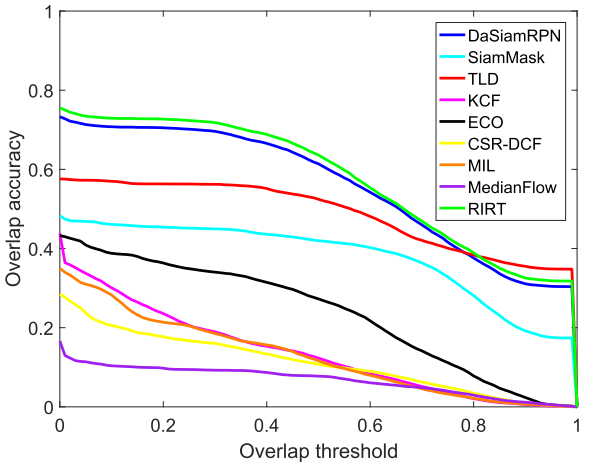
1) QUANTITATIVE ANALYSIS

All the evaluated methods and their evaluation results of location error and overlap are shown in Figure 7, which verifies that the performance of our tracker exceeds that of the mainstream tracker. The location error refers to the linear distance between the center of the tracking area and the center of the real area. The overlap rate refers to the ratio of the intersection and union of the tracking area and the real area. From the figure, the proposed RIRT framework has better tracking performance than DaSiamRPN in the endoscopic scene through its relocation function based on rotation invariant Haar-like descriptor. At the same time, because of the sparse features in endoscope scene, local highlights are easy to be generated. The tracking methods based on correlation filtering (KCF, CSR-DCF), multi-instance learning (MIL), and optical flow (Median Flow) make it difficult to achieve ideal results.

We set the overlap threshold to 0.5 to evaluate the performance of each tracker, and the results are shown in Table 5. The evaluation indexes used here include expected average overlap (EAO), average location error, average overlap,



(a) The accuracy curves of location error.



(b) The accuracy curves of overlap.

FIGURE 7. Plots of accuracy values regarding varying overlap and location errors of trackers.

precision, recall, and F1-measure. EAO is the non-reset overlap expectation of each tracker on an image sequence, and it is one of the most important indexes to evaluate the tracking algorithm accuracy. Because some trackers only give the sample confidence, they do not deal with the target disappearance alone. Therefore, we believe that if the confidence of tracking results is less than 0.2, the current frame does not contain the target. The values of each evaluation index obtained from the whole endoscopy data set are shown in Table 5. It shows that our RIRT framework performs well in EAO, average overlap rate, F1-measure, and other indicators, greatly exceeding TLD, KCF, ECO, and other methods. By extracting shallow features with more context information from RIBHD, the similar areas often encountered in endoscopy are distinguished. In terms of precision index, TLD surpasses other states of the art methods, which may be due to the high threshold set by TLD to make the tracking results as accurate as possible, but at the same time, it dramatically affects its recall rate. Therefore, the F1-measure of the TLD are inferior to those of DaSiamRPN and the proposed RIRT.

TABLE 5. Evaluation of tracking performance.

Tracker	EAO	Location error	Overlap	Precision	Recall	F1-measure
TLD[28]	0.138	183.166	0.487	**0.657**	0.282	0.394
KCF[25]	0.137	258.117	0.134	0.123	0.197	0.152
CSR-DCF[27]	0.113	216.799	0.210	0.134	0.172	0.151
MedianFlow[34]	0.067	271.873	0.314	0.234	0.124	0.162
SiamMask[36]	0.207	193.038	0.287	0.279	0.398	0.328
DaSiamRPN[39]	*0.274*	*131.632*	*0.484*	0.470	*0.527*	*0.497*
ECO[34]	0.235	242.839	0.236	0.273	0.435	0.335
RIRT	**0.276**	**123.767**	**0.492**	*0.573*	**0.678**	**0.621**

* **Bold** numbers represent the best performance, and *italic* numbers represent the second best performance.

TABLE 6. Evaluation of relocation performance.

Tracker	EAO	Location error	Overlap	F1-measure	False detection rate	Relocation rate	Relocation times
TLD[28]	0.055	207.137	0.102	0.203	**0.062**	0.018	1
SiamMask[36]	0.103	221.552	0.188	0.139	0.382	0.058	5
ECO[34]	0.072	219.920	*0.358*	0.216	*0.069*	0.036	6
DaSiamRPN[39]	*0.146*	*191.079*	0.268	*0.295*	0.324	*0.084*	*9*
RIRT	**0.199**	**170.583**	**0.438**	**0.471**	0.302	**0.131**	**12**

* **Bold** numbers represent the best performance, and *italic* numbers represent the second best performance.

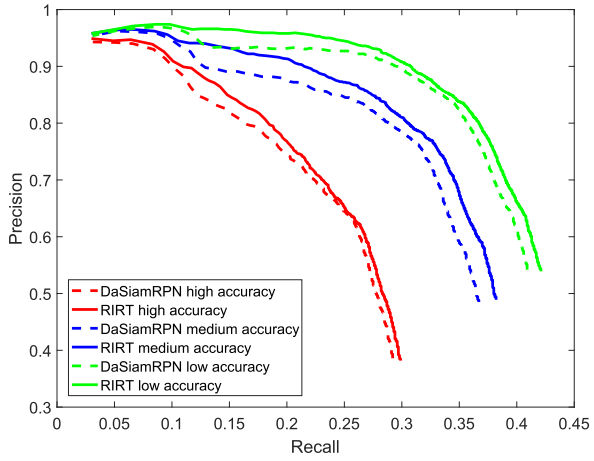


FIGURE 8. The precision-recall curves of DaSiamRPN and RIRT under different overlap rates.

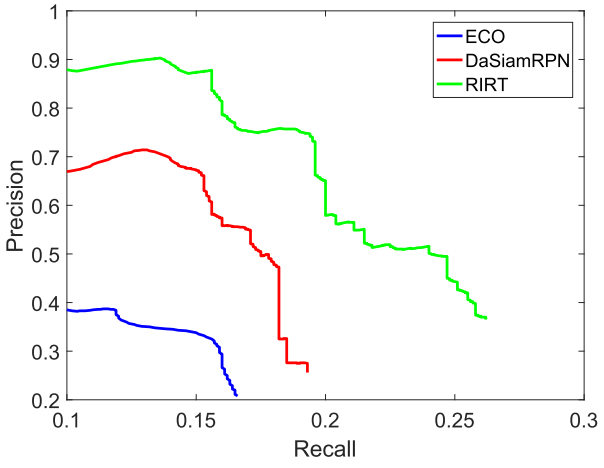


FIGURE 9. The precision-recall curves over the relocation dataset using overlap rates.

Under different overlap rates, we compared the accuracy and recall rates of RIRT and DaSiamRPN. The results are shown in Figure 8. In the PR curve obtained by adjusting the threshold of overlap rate, the overlap rates of positive and negative samples are set to 0.55, 0.45, and 0.35, respectively. From the figure, the P-R curves of the proposed RIRT are better than that of DaSiamRPN. The main reason is that the features from the depth feature map aren't always effective to distinguish the target area from the background, while shallow features extracted by RIBHD could achieve it. We alleviate the unreliable confidence from depth feature calculation by probability fusion strategy.

Relocation is an inevitable problem in the gastrointestinal biopsy. The relocation ability of the tracker is of considerable significance to alleviate patients' pain and shorten the examination time in clinical examination. In order to evaluate the relocation performance of the trackers, we construct five challenging repositioning video sequences to evaluate the repositioning performance of each tracker in endoscopic scenes. Each video sequence contains fifty-five frames, and the first five frames are used to initialize the tracker. In the next fifty frames, the target disappears and replays every five frames, so it needs to be relocated five times. If the overlap rate between the tracking area and the real area of

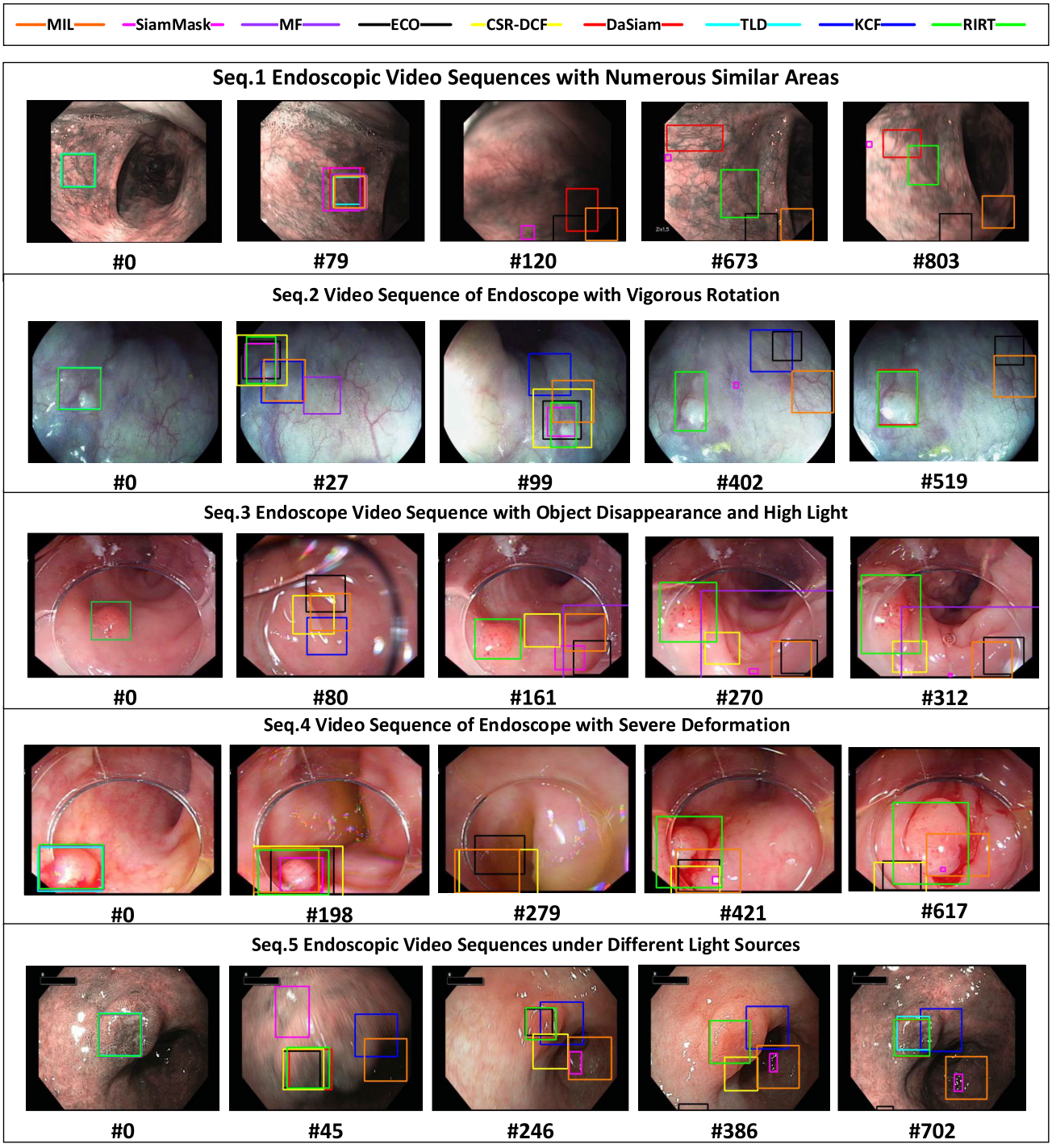


FIGURE 10. Tracking and relocation display of trackers in endoscopic scene.

the relocation frame is higher than 0.35, we consider the relocation successful. By comparing the number of successfully relocated frames with the total number of relocated frames, the relocation rate is obtained. We set the tracking confidence of each tracker to 0.65. If the tracking frame still appears in the target disappearance frame, we believe that the tracker has a false detection. By comparing the number of false detection frames with the total number of frames that the target disappears, the false detection rate is obtained.

Four relocatable trackers are evaluated in terms of relocation rate, false detection rate and above criterion. The results are shown in Table 6. The shallow features extracted by RIBHD are used to discriminate and screen candidate target regions, which makes the repositioning performance

of the proposed RIRT greatly exceed that of the DaSiamRPN tracker. It is worth noting that SiamMask, which has achieved convincing results on various common tracking data sets, has not achieved ideal results in this paper. We think that the main reason is that the ROI of gastrointestinal biopsy often does not have an obvious edge contour. At the same time, the local highlight caused by a smooth intestinal tract leads SiamMask to track the highlight, which further restricts its tracking performance in the gastrointestinal biopsy.

We also draw the P-R curves of the relocatable trackers by regarding the frame with overlap over 0.35 as positive samples, which is shown in Figure 9. Because the TLD and SiamMask fail to relocate the target on the challenging relocation dataset, the results did not appear in the figure. The RIBHD-based confidence-statistical random forest screens candidate tracking regions globally, and resist fast motion ambiguity, which makes our RIRT surpass DaSiamRPN in precision and recall under the challenging relocation dataset.

2) QUALITATIVE ANALYSIS

a: ROBUSTNESS TO ROTATION

In order to deal with the rotation problem in the endoscopic online tracking, we design a new rotation invariant Haar-like descriptor. The random forest based on statistical decision-making constructed by the descriptor effectively rotates the biopsy area. Compared with the traditional way of expanding positive samples by an affine transformation, the proposed descriptor is more sensitively to perceive the rotation angle of the lens. In sequence 2 and 3 of Figure 10, the proposed RIRT is able to handle the significant rotation of the visual field in gastrointestinal biopsy by using Haar-like descriptors with rotation invariance.

b: ROBUSTNESS TO SCALE CHANGES

In terms of scale change, we scan the original video sequence at different scales to ensure that the preliminary screening results of random forest have the ability to cope with scale change. Fine search component based on the Siamese network is used to locate the tracking area more accurately. We use the above methods to deal with the scale change in the tracking process. In sequence 2 and 3 of Figure 10, our tracking performance is better than that of other trackers such as KCF, MIL, CSR-DCF.

c: ROBUSTNESS TO RETARGETING (ROBUSTNESS TO SMALL FOV)

The tissue area of optical biopsy often enters and exits the field of vision during the examination. When the camera moves back to the same biopsy location, it needs to be repositioned, as shown in the figure below. Many trackers used for comparison do not have the ability to judge the disappearance of the tracking target, which leads to the inability to track the correct tracking area after the biopsy area exceeds the field of vision. Even when the target area returns to the field of vision again, the tracker fails to recognize them, as shown in video sequences 1, 3, 4 of Fig.10. Failure trackers include KCF, TLD, MedianFlow, etc.

d: TISSUE DEFORMATION

Compared with robust medical images such as CT, Tissue Deformation is a challenging problem in medical image analysis, because of its natural organ characteristics, the gastrointestinal tract is easily affected by the movement and peristalsis of patients or breathing during the examination process. The robustness of the global deformation method has been proved in vivo experiments by cross-correlation operation of the depth characteristics of Siamese networks, as shown in Fig.10 sequences 3 and 4.

e: ROBUSTNESS TO FALSE POSITIVES

In Fig.10's video sequence 1, the search method based on the Siamese network often gets false examples, such as the location result of the Siamese network in 750 frames and 803 frames, because of the characteristics of intense motion and many similar regions in the endoscope scene. The RIBHD in our RIRT framework contains more contextual information, which makes it difficult to distinguish similar areas through the network effectively.

f: SPECULAR HIGHLIGHTS

Because the endoscope camera is close to the surface of the object in the navigation process, the specular highlight on the image cannot be ignored. For example, sequence 3 and 4 in Fig.10, exemplify how particularities can lead to biopsy site occlusion. In our framework, HRF classification and shape context take part of the biopsy site information (local area comparison and key points) into account, so that it shows excellent performance in the above video sequences.

V. CONCLUSION AND FUTURE WORKS

We propose an online tracking relocation framework, RIRT, which is initially relocated by a new RIBHD descriptor and a simplified random forest based on a statistical decision. A refined component based on Siamese network and probability fusion is used to locate the region of interest accurately. Unlike existing rotational descriptors, the RIBHD, based on circular sector-ring differences, excellently adapts to the characteristics of endoscopic images. Compared with the traditional endoscope tracking framework, the proposed RIRI framework combines the advantages of manual statistical features and high-dimensional self-learning features. More shape context information is obtained by artificial feature extraction to exclude similar regions in endoscope video, and preliminary selection of candidate tracking regions is achieved. Responsibility is obtained by cross-correlation operation of depth features to accurately determine the location of tracking regions. In addition, we fuse the probability of tracking position obtained by each component decision to improve the robustness of the whole system. At the descriptor level, the proposed descriptor and the descriptor in OTR are evaluated under various endoscopic environments and

parameters. The results show that the proposed descriptor exceeds the descriptor in OTR. Besides, the proposed CSHRF might also be helpful in recognition and tracking gait [43]. In terms of overall performance, we evaluate the proposed RIRT with state-of-the-art methods. The results show that our RIRT outperforms the current best methods in terms of EAO, average overlap rate, and average location error. Furthermore, on the self-constructed challenging relocation dataset, we evaluated the relocation-capable trackers. Through the screening of global candidate regions based on RIBHD, our RIRT surpassed other relocation-capable trackers.

The RIRT framework provides technical support for clinical gastrointestinal biopsy doctors to shorten the operation time and improve the accuracy of surgery. Because this algorithm system can automatically locate the biopsy area and reduce the requirement of the surgeon for clinical professional skills, it is also important to popularize gastrointestinal biopsy screening. On this basis, future researchers can combine the cascade system with other components, such as a system that automatically identifies abnormalities through gastrointestinal biopsy, to assist doctors in diagnostic decision-making. In addition, in order to meet the needs of patient review, long-term re-location algorithm is also a very valuable research point.

REFERENCES

[1] C. A. Kousera, S. Nijjer, R. Torii, R. Petraco, S. Sen, N. Foin, A. D. Hughes, D. P. P. Francis, X. Y. Xu, and J. E. Davies, "Patient-specific coronary stenoses can be modeled using a combination of OCT and flow velocities to accurately predict hyperemic pressure gradients," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1902–1913, Jun. 2014, doi: 10.1109/TBME.2014.2310954.

[2] T. Kanemitsu, K. Yao, T. Nagahama, K. Imamura, S. Fujiwara, T. Ueki, K. Chuman, H. Tanabe, O. Atsuko, A. Iwashita, T. Shimokawa, K. Uchita, and T. Kanesaka, "Extending magnifying NBI diagnosis of intestinal metaplasia in the stomach: The white opaque substance marker," *Endoscopy*, vol. 49, no. 6, pp. 529–535, Jun. 2017.

[3] N. Mahmoud, I. Cirauqui, and A. Hostettler, "Orbslam-based endoscope tracking and 3D reconstruction," in *Proc. Int. Workshop Comput.-Assist. Robotic Endoscopy*. Cham, Switzerland: Springer, 2016, pp. 72–83.

[4] O. G. Grasa, J. Civera, and J. M. M. Montiel, "EKF monocular SLAM with relocalization for laparoscopic sequences," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 4816–4821, doi: 10.1109/ICRA.2011.5980059.

[5] B. Lin, A. Johnson, X. Qian, J. Sanchez, and Y. Sun, "Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery," in *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*. Nagoya, Japan: Spinger, Sep. 2013, pp. 35–44.

[6] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006, doi: 10.1109/TPAMI.2006.244.

[7] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. ECCV*, Graz, Austria, May 2006, pp. 404–417.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571, doi: 10.1109/ICCV.2011.6126544.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Dec. 2001, pp. 511–518.

[11] A. L. Barczak, "Toward an efficient implementation of a rotation invariant detector using haar-like features," in *Proc. IVCNZ*, Dunedin, New Zealand, 2005, pp. 31–36.

[12] S. Du, N. Zheng, Q. You, Y. Wu, M. Yuan, and J. Wu, "Rotated haar-like features for face detection with in-plane rotation," in *Proc. Int. Conf. Virtual Syst. Multimedia*, Xi'an, China, Oct. 2006, pp. 128–137.

[13] A. Barczak, C. Messom, and R. Chemudugunta, "Real-time rotationally invariant features for environmental feature detection by mobile robots sensor networks," in *Proc. Int. Workshop Robotic Sensors Environ.*, Oct. 2007, pp. 1–6.

[14] M. Oualla, A. Sadiq, and S. Mbarki, "Rotated Haar-like features at generic angles for objects detection," in *Proc. 3rd IEEE Int. Colloq. Inf. Sci. Technol. (CIST)*, Oct. 2014, pp. 351–355, doi: 10.1109/CIST.2014.7016645.

[15] M. Oualla and A. Sadiq, "Rotated asymetrique Haar features for face detection," in *Proc. 4th IEEE Int. Colloq. Inf. Sci. Technol. (CiSt)*, Oct. 2016, pp. 471–475, doi: 10.1109/CIST.2016.7805094.

[16] H. Zhang, W. Gao, X. Chen, and D. Zhao, "Object detection using spatial histogram features," *Image Vis. Comput.*, vol. 24, no. 4, pp. 327–341, Apr. 2006.

[17] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011, doi: 10.1109/TPAMI.2010.226.

[18] T. B. Dinh, N. Vo, and G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *Proc. CVPR*, Jun. 2011, pp. 1177–1184, doi: 10.1109/CVPR.2011.5995733.

[19] M. Ye, S. Giannarou, A. Meining, and G.-Z. Yang, "Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations," *Med. Image Anal.*, vol. 30, pp. 144–157, May 2016, doi: 10.1016/j.media.2015.10.003.

[20] M. Ye, E. Johns, B. Walter, A. Meining, and G.-Z. Yang, "An image retrieval framework for real-time endoscopic image retargeting," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 8, pp. 1281–1292, Aug. 2017, doi: 10.1007/s11548-017-1620-7.

[21] M. Ye, E. Johns, B. Walter, A. Meining, and G.-Z. Yang, "Robust image descriptors for real-time inter-examination retargeting in gastrointestinal endoscopy," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, Athens, Greece, 2016, pp. 448–456.

[22] B. F. La Scala and R. R. Bitmead, "Design of an extended Kalman filter frequency tracker," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 739–742, Mar. 1996, doi: 10.1109/78.489052.

[23] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002, doi: 10.1109/34.1000236.

[24] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. S. Torr, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016, doi: 10.1109/TPAMI.2015.2509974.

[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015, doi: 10.1109/TPAMI.2014.2345390.

[26] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 1.

[27] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4847–4856, doi: 10.1109/CVPR.2017.515.

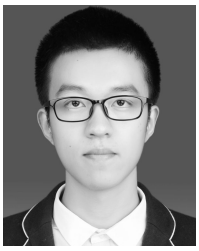
[28] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012, doi: 10.1109/TPAMI.2011.239.

[29] W. Naiyan and D. Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, CA, USA, 2013, pp. 809–817.

[30] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3119–3127, doi: 10.1109/ICCV.2015.357.

[31] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302, doi: 10.1109/CVPR.2016.465.

[32] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, A. Gupta, A. Bibi, A. Lukezic, A. Garcia-Martin, A. Saffari, A. Petrosino, and A. S. Montero, "The visual object tracking VOT2015 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 564–586, doi: 10.1109/ICCVW.2015.79.
[33] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," 2016, *arXiv:1608.07242*. [Online]. Available: http://arxiv.org/abs/1608.07242
[34] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939, doi: 10.1109/CVPR.2017.733.
[35] L. Bertinetto, J. Valmadre, and J. F. Henriques, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 850–865.
[36] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.
[37] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980, doi: 10.1109/CVPR.2018.00935.
[38] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.
[39] Z. Zheng, W. Qiang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 101–117.
[40] J. Wang, Y. He, X. Wang, X. Yu, and X. Chen, "Prediction-tracking-segmentation," 2019, *arXiv:1904.03280*. [Online]. Available: http://arxiv.org/abs/1904.03280
[41] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2756–2759, doi: 10.1109/ICPR.2010.675.
[42] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A LLVM-based Python JIT compiler," in *Proc. 2nd Workshop LLVM Compiler Infrastruct. HPC LLVM*, 2015, p. 7.
[43] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, Jun. 2019.

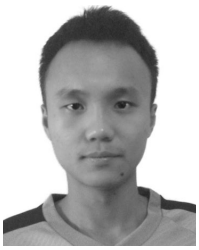


HAIFAN GONG (Student Member, IEEE) is currently a Senior Student with the School of Information Engineering, Nanchang University, Nanchang, China. His current research interests include computer vision, deep learning, and video online tracking.

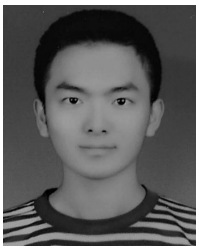
LIMIN CHEN received the Ph.D. degree from Nanchang University, China, in 2019. He has been with the School of Information Engineering, Nanchang University, since 2002, where he is currently an Associate Professor. His research interests include computer vision, intelligent sensing, and software-defined radio.

CHANGHAO LI received the bachelor's degree from the Wuhan University of Science and Technology, Wuhan, China, in 2008. He is currently pursuing the M.S. degree with the School of Information Engineering, Nanchang University. His research interests include machine learning, computer vision, and their applications.

JUN ZENG received the B.S. degree in clinical medicine from the Second Clinical Medical College, Nanchang University, in 1995. He has been with the Department of Gastroenterology, Jiangxi Provincial People's Hospital, Nanchang, China, since 1995, where he is currently an Associate Chief Physician. His research interests include gastroenterology, surgery simulation, and artificial intelligence for medicine. He is also a member of the Jiangxi Provincial Committee of the Chinese Medical Association and Digestive Pathology Committee of the Chinese Medical Association.



XICHEN TAO received the B.Eng. degree from the Jiangxi University of Science and Technology, Ganzhou, China, in 2018. He is currently pursuing the master's degree with the School of Information Engineering, Nanchang University, Nanchang, China. His current research interests include computer vision, machine learning, and deep learning.



YUE WANG (Student Member, IEEE) is currently pursuing the double bachelor's degree in clinical medicine and computer science and technology with the Second Clinical Medical College, Nanchang University, Nanchang, China. His research interests include computer vision, deep learning, and image reconstruction.