IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 11, 2020, accepted June 17, 2020, date of publication June 22, 2020, date of current version July 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003993

Foldover Features for Dynamic Object Behaviour Description in Microscopic Videos

XIALIN LI[1], CHEN LI[1], FRANK KULWA[1], MD MAMUNUR RAHAMAN[1], WENWEI ZHAO[1], XUE WANG[1], DAN XUE[1], YUDONG YAO[2], (Fellow, IEEE), YILIN CHENG[1], JINDONG LI[1], SHOULIANG QI[1], (Member, IEEE), AND TAO JIANG[3]

[1]Microscopic Image and Medical Image Analysis Group, MBIE College, Northeastern University, Shenyang 110169, China
[2]Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[3]Control Engineering College, Chengdu University of Information Technology, Chengdu 610103, China

Corresponding author: Chen Li (lichen201096@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61806047, in part by the Fundamental Research Funds for the Central Universities under Grant N2019003, and in part by the China Scholarship Council under Grant 2017GXZ026396 and Grant 2018GBJ001757.

ABSTRACT A behavior description helps analyze tiny objects, similar objects, objects with weak visual information, and objects with similar visual information. It plays a fundamental role in the identification and classification of dynamic objects in microscopic videos. To this end, we propose foldover features to describe the behavior of dynamic objects. Foldover is defined as: Each frame of an object's motion is superimposed on the same spatial plane in the spacetime order of the motion, the result of the superposition is the foldover of the object's motion. Foldover of an object contains temporal information, spatial information, behavior features and static features. Therefore, the features extracted based on the foldover of the object are the foldover features. In this work, we first generate foldover for each object in microscopic videos in X, Y and Z directions, respectively. Then, we extract foldover features from the X, Y and Z directions with statistical methods, respectively. The core content of this paper is to construct the foldovers and extract the foldover features. Through these two steps, the temporal information, spatial information, behavior features and static features of the object are enhanced and included in the foldover features. Furthermore, the description of the behavior of dynamic objects by the foldover features is strengthened. Finally, we use four different classifiers to test the effectiveness of the proposed foldover features. In the experiment, we use a microscopic sperm video dataset to evaluate the proposed foldover features, including three types of 1374 sperms, and obtain the highest classification accuracy of 96.5%.

INDEX TERMS Foldover feature extraction, content-based microscopic image analysis, microscopic videos, dynamic object behavior.

I. INTRODUCTION

In computer vision, a video is made up of many frames and video analysis is basically image analysis [1]. In addition, we tend to focus on a specific target object or class of video rather than the whole video. Therefore, image feature extraction is very important for video analysis [2]. Currently, static features [2] and dynamic features [3] are mainly used to identify or classify different objects in images as shown in TABLE 1.

From TABLE 1 we can see that when facing the following three conditions, it is easy to describe the objects with existing static and dynamic features (similarity corresponds to distinction, if the similarity of two objects is high, the distinction between them is low; similarly, if two objects are very similar, the distinction between them must be high): (1) the distinction

TABLE 1. A comparison table of static and dynamic features.

Two objects	Static features similarity	Dynamic features similarity	Whether existing studies solve
A and B	low	low	Yes
C and D	low	high	Yes
E and F	high	low	Yes
G and H	high	high	Our work

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

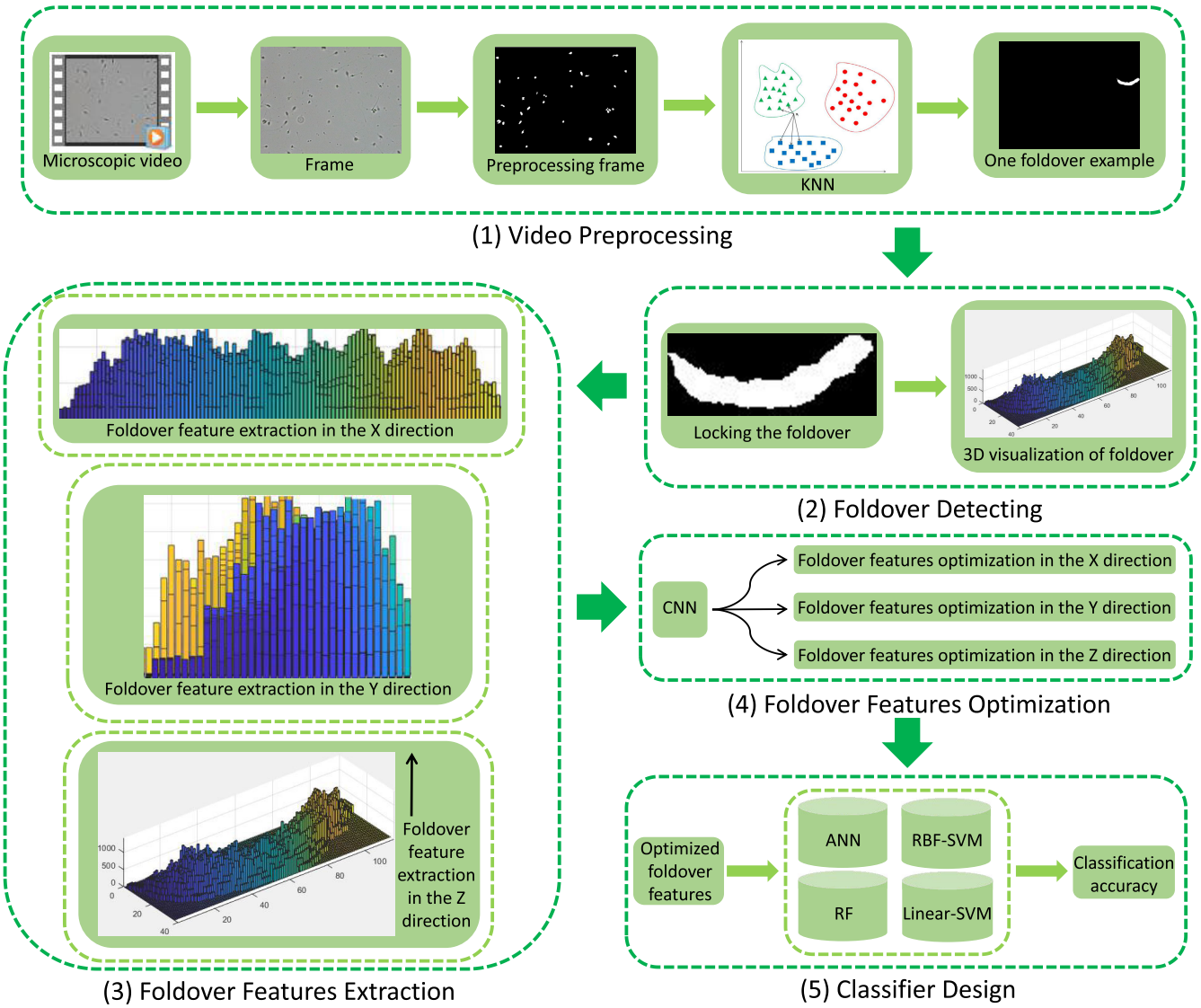

FIGURE 1. Work flow diagram of the foldover feature extraction method.

between static features and dynamic features is both high. (2) there are obvious differences in static features and little differences in dynamic features. (3) the difference between static features is little, while the difference between dynamic features is large. However, objects in microscopic videos are hard to identify or classify in the following two cases: (1) two objects with very similar static and dynamic features. (2) two objects with very weak static features and very similar dynamic features. To this end, we propose new foldover features to describe the behavior of objects in microscopic videos.

In microscopic videos, the following difficulties usually exist in identifying or classifying different individuals of the same class of tiny objects. Firstly, because most of the tiny objects are colorless or transparent, they have little color or texture information. Secondly, when tiny objects have similar morphological characteristics, it is difficult to distinguish them by shape features. Thirdly, if the size of the objects are only several pixels, it is tough to obtain available information. Fourthly, if two objects have both similar static and dynamic features, it is hard to identify or classify them. Hence, we select the microscopic sperm videos as the experimental material, where sperms have little color information, weak shape information, tiny sizes, and similar static and dynamic features.

Foldover is defined as: Each frame of an object's motion is superimposed on the same spatial plane in the space-time order of the motion, the result of the superposition is the foldover of the object's motion. Foldover of an object contains temporal information, spatial information, behavior features and static features. Foldover features are a kind of behavior feature that is based on dynamic targets. The workflow diagram of the proposed algorithm is presented in FIGURE 1.

There are five steps in FIGURE 1: (1) Video preprocessing, the purpose of which is to obtain the motion foldover of each tiny object in the video. (2) Foldover detection, which is used to detect the foldover of each object. (3) Foldover features extraction, extracting the foldover features from the X, Y, Z, three directions. (4) Foldover features optimization, the Convolutional Neural Network [4] (CNN) removes the redundant information in the foldover features and further enhances the foldover features. (5) Classifier design, four different classifiers are used to verify the superiority of foldover features.

The core content of this paper is to construct the foldovers and extract the foldover features. The contributions of the foldover features are as follows: (1) The foldover features provide a feature extraction method for the behavior classification of tiny objects. (2) The foldover features provide a method to extract feature information for objects with little feature information. (3) In the behavior classification of similar objects, the application of foldover features makes the classification to obtain good results.

II. RELATED WORK

This section summarizes the existing works that is related to our study. II-A summarizes the static feature extraction methods, including various classical feature extraction methods and deep learning feature extraction methods. II-B summarizes the dynamic feature extraction methods, including several common dynamic feature extraction methods and deep learning feature extraction methods. II-C summarizes the technologies of target detection and feature engineering, including several common feature engineering and target detection methods. II-D summarizes the classifier design, including some well-known algorithms.

A. STATIC FEATURES

Static features usually include color, texture and shape features. Color features describe the surface properties of the scene corresponding to an image region based on pixel information [5]. However, when images have little color information (microscopic sperm videos), the color features are almost identical. For example, in the article [6], a brightness histogram is used to retrieve images with good results. However, when multiple objects have similar color brightness, this method is secure to lose effectiveness.

Texture features reflect the properties of surface structure organization and arrangement with slow or periodic change [7]. For example, the proposal of the Histogram of Oriented Gradient (HOG) feature [8], the advantage of HOG is that the geometric deformation and optical deformation of images have little influence on HOG. However, HOG is difficult to deal with the occlusion. For example, in the microscopic sperm videos, when two sperms collide and overlap, the extraction result of HOG feature will have errors. Another example, the application of Gray-Level Co-occurrence Matrix (GLCM) [9]. GLCM is used to calculate uniformity and strength values to identify candidate areas of Ground Glass Opacity (GGO) nodules. However, GLCM cannot identify two very similar objects by describing the gray relationship between a certain pixel and a pixel within a certain distance.

There are many effective shape features, such as geometric features, Hu moments [10], shape signature [11] and Scale-invariant Feature Transform (SIFT) features [12]. The geometric features mainly include perimeter, area, long axis, short axis, length-width ratio and complexity, which can be used for motion analysis. However, in the analysis and recognition of similar targets (such as microscopic sperm videos), it is not effective to use geometric features. Hu moments are higher-order geometric features used to reflect the distribution of random variables in statistics. Translation, scale expansion, rotation these changes will not affect the invariant moment. It has good invariance. However, Hu moments depend on image segmentation a lot, and their application fields are limited. SIFT feature [12] is a local feature of images, which is invariant to rotation, scale scaling and brightness change, and has excellent stability to angle change, affine transformation and noise influence. However, the detection of critical points is an essential step in SIFT feature extraction but the features extracted from tiny targets are limited. Shape signature is a boundary - based shape descriptor formed by a set of one-dimensional signals called shape signatures [11], which is robust to environmental conditions (partial occlusion) and image transformation (scaling, rotation, translation). But, the point of shape signature is to identify objects based on their shape, which is not effective at recognizing object (such as sperm) with similar shape.

With the development of deep learning technology, we can adopt different neural network frameworks to extract the target objects' in-depth features. *Convolutional Neural Network* [4] (CNN) is an efficient identification method because it avoids the complicated pre-processing steps and can directly input the original images. VGG16 [13] network is a classical CNN explores the relationship between the depth of the convolutional neural network and its performance. The error rate is significantly reduced. Therefore, we can use VGG16 network to directly extract the in-depth features of static targets. Deep learning features can be used for further data statistics at the pixel level. However, when objects are tiny (such as sperms in microscopic videos), the feature extraction ability of CNN is minimal.

B. DYNAMIC FEATURES

With the development of pattern recognition and intelligent video processing technology, there is much research on dynamic target analysis. The dynamic texture is an extension of static texture in the time domain, which includes both static and dynamic information [14]. For example, in the Motion Energy Model [15], a video sequence is regarded as the direction in the three-dimensional space-time, and a directionally selective filter is used to extract the motion information on each position. In [16], based on the expansion of separable guided filtering theory, a 3D filter is

decomposed into three independent one-dimensional filters, which are filtering along the horizontal, vertical and time directions, and the filtering efficiency is significantly improved. Gaussian Mixture Model (GMM) [17] is widely used to model the background of complex dynamic scenes, especially on the occasions of periodic movement, such as shaking branches, turbulent water, snowstorms and fountains. GMM can steadily and quickly detect suspected motion prospects. Mixtures of Dynamic Textures (MDT) [18] is used for video frame sequence modeling. MDT can use dynamic textures to generate a series of video sequences into specific samples, which has excellent performance in motion clustering and segmentation. The above four examples are target motion analysis in video based on dynamic texture. However, the basis of dynamic texture is static texture, which is an extension of static texture in the time domain. In microscopic video analysis, we encounter the following difficulties: (1) Multi-objective analysis, there are many objects of our analysis in each frame. (2) All the objects are very tiny, and there is no significant difference in the appearance of different objects (such as sperms). (3) Little texture information of tiny objects. (4) Interference of impurities, some impurities are similar to our analysis objects in appearance. The above difficulties cannot be solved by dynamic texture features.

The acquisition of motion parameters is also useful for object motion analysis. For example, a series of motion parameters of each sperm are continuously collected to analyse sperm motion [19] and achieve good results. However, in the case that there are many sperm targets in the camera lens, different sperm targets have similar motion patterns and little difference in motion parameters. Therefore, it is not enough to rely on motion parameters alone.

In recent years, deep learning method has been successfully applied in object tracking field, and gradually surpasses the traditional method in performance. A typical strategy is that first obtaining the feature representation of a target by using CNNs, then the CNNs are trained on a large-scale classification database like ImageNet [20], and the trained CNNs are finally used to classify and track the objects. This approach not only avoids the problem of insufficient samples of large-scaleCNN, but also makes full use of the strong representation ability of deep learning features.

FCNT mainly analyses the conv4-3 and conv5-3 output feature maps of VGG-16 [21]. FCNT constructes a feature screening network and two complementary heat-map prediction networks based on the analysis of features of different CNN layers. FCNT makes the targets more robust during deformation. The work of [22] uses the output of conv3-4, conv4-4 and conv5-4 in a pre-trained VGG-19 [13] as the feature extraction layer. The Features extracted from these three layers are respectively studied through relevant filters to obtain different templates, and then the obtained three results are fused to obtain the final target position. However, the above method is not applicable to the identification and analysis of multi-target motion in microscopic sperm videos. The difficulties in using deep learning in the field of target tracking and recognition are appearance deformation, light change, fast movement, motion blur, interference from similar objects, scale change, occlusion and target movement out of the field of view. These difficulties are also the problems that we encounter in the microscopic sperm videos. In addition, the five difficulties proposed in this paper in the section on dynamic texture are still not well solved by using the above methods. These five difficulties are also the key problems to be solved in this paper.

C. FEATURE ENGINEERING AND TARGET DETECTION

In the recognition and analysis of dynamic objects, image processing and object detection are very important, because the accuracies of image processing and object detection affect the results of recognition and analysis. Specifically, image segmentation and feature extraction are two important steps in image processing.

Image segmentation is critical to the effectiveness of feature extraction. Mask-Refined R-CNN (MR R-CNN) [23] adjusts the stride of ROIAlign (region of interest align), and the feature fusion is realized by replacing the full convolutional layer with a new semantic segmentation layer. Combining with the feature layer of global and detail information, the segmentation accuracy is greatly improved. Article [24] presents an automated data augmentation method for synthesizing labeled medical images, learning a model of transformations from the images, and using the model along with the labeled example to synthesize additional labeled examples. Each transformation is comprised of a spatial deformation field and an intensity change, enabling the synthesis of complex effects.

Noise removal and contrast enhancement constitute important topics in image processing, which can improve the accuracy of image segmentation. Article [25] proposes a noise-level estimation method, whereby the noise level is estimated by computing the standard deviation and variance in a local block. The obtained noise level is then used as an input parameter for the block-matching and 3D filtering (BM3D) algorithm, and the denoising process is then performed, the method converts low contrast data into high contrast data and reduces high noise level. Article [26] remove both impulse and Gaussian noise, and enhance contrast. To enhance image contrast, low contrast pixels become even lower, and high contrast pixels become even higher.

Correspondingly, the quality of feature extraction is based on the result of image segmentation, for example, two-dimensional discrete cosine transform (2DDCT) is used to extract the features of left and right palmprints to constitute a double-source space [27]. More discriminant coefficients can be preserved and retrieved with discrimination power analysis (DPA) from dual-source space, the accuracy performance is improved. Another example, PalmHash Code and PalmPhasor Code, as two cancelable palmprint coding schemes, are proposed to balance the conflict between security and verification performance [28].

In addition, we can manipulate the features to improve the quality of the features, such as, select, weight and combine. In [29], dynamic weighted discrimination power analysis (DWDPA) enhances the discrimination power (DP) of the selected discrete cosine transform coefficients (DCTCs) without premasking window, in other words, it does not need to optimize the shape and size of premasking window. Dynamic weighting gives larger weights to the DCTCs with larger discrimination power values (DPVs) which optimizes and enhances the recognition performance. Conjugate 2DPalmHash Code (CTDPHC) [30] is constructed by 2DPalmHash Codes (2DPHCs) of palmprint and palmvein, it is proposed as a cancelable multi-modal biometric. CTDPHC enjoys higher verification accuracy and stronger anti-counterfeit ability, while trades neither computational complexity nor storage cost.

Object tracking is a component of dynamic object detection, because the accuracy of object tracking affects the result of dynamic object recognition and analysis. The difficulty of object tracking is background interference and object occlusion. In the case of occlusion and scale variation, article [31] proposes a scale adaptive target tracking method with good performance. This article proposes an update strategy based on occlusion detection, which provides an effective method for object detection with occlusion. A double-channel object tracking (DCOT) is proposed in [32]. The discriminative correlation filter (DCF), which has strong discriminative power of low-level features, is employed for the position deviation suppress of the samples generated from MDNet. This method guarantees the accuracy of tracked positions effectively.

In target recognition, saliency detection has important application value, which can bring a series of significant help and improvement to visual information processing. Article [33] proposes a new salient property of part-object relationships provided by the Capsule Network (CapsNet) for salient object detection, and presents a deep Two-Stream Part-Object Assignment Network (TSPOANet). The proposed model requires less computation budgets while obtaining better wholeness and uniformity of the segmented salient object. The proposal of the Deep Conditional Random Field network (DCRF) [34] takes into account both the depth features and the neighbor information. DCRF is a good combination of low-level internal context and high-level semantic information, keeping object boundaries clear and suppressing background noise. Another example, article [35] proposes a novel end-to-end network for multi-modal salient object detection, which turns the challenge of RGB-T saliency detection to a CNN feature fusion problem. Under challenging conditions, such as poor illumination, complex background and low contrast, The network performs the saliency detection task well. Article [36] proposes an approach that considers the internal color and saliency properties of the image. It changes the saliency map via an optimization framework that relies on patch-based manipulation using only patches from within the same image to maintain its appearance characteristics. This method has significant results in both the saliency manipulation and the realistic appearance of the resulting images. Article [37] proposes a framework to learn deep salient object detectors without requiring any human annotation. It is a good solution to the problem that it is expensive and time-consuming to provide pixel-level ground-truth masks for each training image. Another example, article [38] proposes a two-stage mechanism for robust unsupervised object saliency prediction, it refines the pseudo-labels from different unsupervised handcraft saliency methods in isolation, and improves the supervisory signal for training the saliency detection network. The two-stage mechanism is crucial to improve the quality of pseudo-labels and hence achieve competitive performance on the object saliency detection tasks.

D. CLASSIFIER DESIGN

There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines [39].

A kind of well-known algorithms are based on the notion of perceptron, such as multilayered perceptrons (Artificial Neural Networks) [39]. The advantages of Artificial Neural Networks (ANNs) are: Strong parallel distributed processing ability, strong distributed storage and learning ability, strong robustness and fault tolerance to noise nerves [40]. Another well-known algorithms are based on the ensemble learning, such as Random Forests (RFs) [41]. The advantages of random forests are: It has a strong ability to process high-dimensional data, the generalization ability of the model is strong, it is fast to train the model, and the model can handle unbalanced data [42]. Another well-known algorithms are based on the Support Vector Machines (SVM) [39]. The advantages of SVM are: It can solve machine learning problems in small samples, improve generalization performance, solve nonlinear problems, and the problem of neural network structure selection and local minima can be avoided [43].

III. FOLDOVER FEATURES

In this section, we introduce the proposed foldover feature extraction method, referring to III-A foldover construction, III-B foldover feature extraction.

For the convenience of narration, the variables are used in this paper as follows: (1) We define a data set of videos as $\chi = \{X_1, X_2, \ldots, X_i, \ldots, X_n\}$, $i = 1, 2, 3, \ldots, n$, where X_i is the video variable, i is the video number, and n is the total number of videos in χ. Furthermore, $X_i = \{x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,j)}, \ldots, x_{(i,m)}\}$ ($j = 1, 2, 3, \ldots, m$) is a set of frames (static images), where $x_{(i,j)}$ is the frame variable, j is the frame number, m is the total number of frames in X_i. In addition, $x_{(i,j)} = \{x_{(i,j,1)}, x_{(i,j,2)}, \ldots, x_{(i,j,k)}, \ldots, x_{(i,j,h)}\}$ ($k = 1, 2, 3, \ldots, h$) is a set of pixels, where $x_{(i,j,k)}$ denotes

the image pixel, k is the pixel number, h is the total number of pixels in a frame, $h = h_1 \times h_2$, $h_1 = 1, 2, 3, \ldots$ is the number of pixels in a row, and $h_2 = 1, 2, 3, \ldots$ is the number of pixels in a column. (2) We define the intensity (pixel value) at pixel $x_{(i,j,k)}$ as $p\left(x_{(i,j,k)}\right) \in [0, 255]$. (3) We define a set of sperms in each frame as $\varsigma_{(i,j)} = \left\{s_{(i,j,1)}, s_{(i,j,2)}, \ldots, s_{(i,j,l)}, \ldots, s_{(i,j,q)}\right\}$, $l = 1, 2, 3, \ldots, q$, where $s_{(i,j,l)}$ is one sperm, l is the sperm number, and q is the total number of sperms in this frame.

A. CONSTRUCTION OF FOLDOVERS

There are many sperms in a semen microscopic video, we construct a foldover for each sperm by the following six steps. The work flow of the constrction of foldover is shown in FIGURE 2.

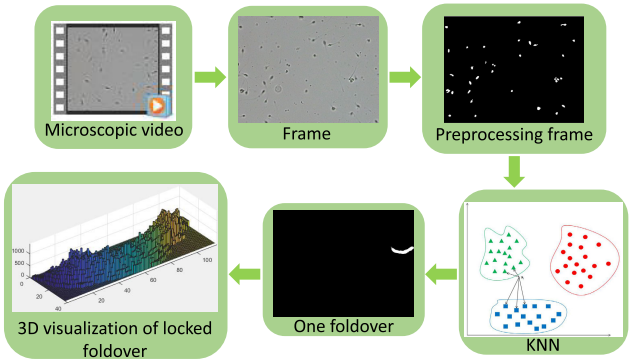

FIGURE 2. An example of the constrction of one foldover.

As the work flow is shown in FIGURE 2, Each frame of an object's motion is superimposed on the same spatial plane in the space-time order of the motion. The result of the superposition is the foldover of the object's motion. Besides, we can extract the temporal information, spatial information, behavioral features and static features of the object from the foldover.

1) VIDEO DECOMPOSITION

We decompose a semen microscopic video X_i into frames $x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,j)}, \ldots, x_{(i,m)}$. Each frame (such as $x_{(i,j)}$) is a static gray-scale image, and an example is shown in FIGURE 3.

2) IMAGE SEGMENTATION

We define the threshold value of the image $x_{(i,j)}$ as $T\left(x_{(i,j)}\right)$, the segmentation result of $x_{(i,j)}$ as $x^{\text{seg}}_{(i,j)}$, and the value of the k-th pixel in $x^{\text{seg}}_{(i,j)}$ as $p\left(x^{\text{seg}}_{(i,j,k)}\right)$ in Eq. (1).

$$p\left(x^{\text{seg}}{}_{(i,j,k)}\right) = \begin{cases} 0 & p\left(x_{(i,j,k)}\right) \le T\left(x_{(i,j)}\right) \\ 1 & otherwise \end{cases} \tag{1}$$

In Eq. (1), When the pixel value $p\left(x_{(i,j,k)}\right)$ is lower than the threshold $T\left(x_{(i,j)}\right)$, the result of threshold segmentation $p\left(x^{\text{seg}}_{(i,j,k)}\right)$ is 0 (black); otherwise, the result of threshold segmentation $p\left(x^{\text{seg}}_{(i,j,k)}\right)$ is 1 (white). Finally, we get the image segmentation result $x^{\text{seg}}_{(i,j)}$, and all the sperms $\varsigma_{(i,j)}$ in each frame $x_{(i,j)}$ are obtained. An example of the threshold segmentation result is shown in FIGURE 4.

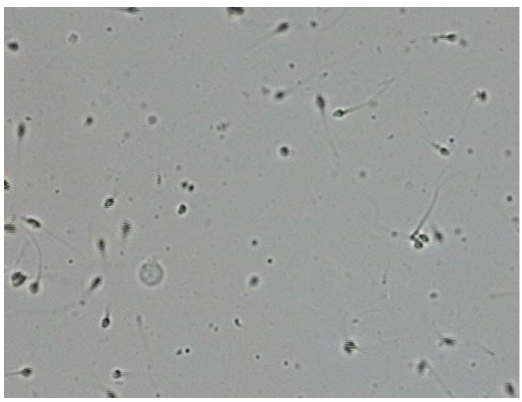

FIGURE 3. An example of a semen microscopic video frame (a static gray-scale image) $x_{(i,j)}$.

FIGURE 4. An example of the threshold segmentation result $x^{s}_{(i,j)}$.

3) BARYCENTER COORDINATES EXTRACTION

Based on the image segmentation results $x^{s}_{(i,j)}$, we define a barycenter coordinates set of all sperms for total frames in the video X_i as $\psi_{(i)} = \left\{C_{(i,1)}, C_{(i,2)}, \ldots, C_{(i,j)}, \ldots, C_{(i,m)}\right\}$, where $C_{(i,j)}$ is the barycenter coordinate variable, i is the video number, j is the frame number, and m is the total number of frames. Furthermore, $C_{(i,j)} = \left\{c\left(s_{(i,j,1)}\right), c\left(s_{(i,j,2)}\right), \ldots, c\left(s_{(i,j,l)}\right), \ldots, c\left(s_{(i,j,q)}\right)\right\}$ is a set of barycenter coordinates for all sperms $\varsigma_{(i,j)}$ in the frame $x_{(i,j)}$, where $c\left(s_{(i,j,l)}\right)$ is the barycenter coordinates of l-th sperm in the j-th frame of i-th video. In conclusion, we extract all barycenter coordinates $\psi_{(i)}$ from all sperms in the video X_i.

4) TARGET MATCHING

Currently, the commonly used sperm quality test method is computer-assisted sperm analysis (CASA) [44], CASA applies computer technology and advanced image processing technology to the analysis of sperm dynamics.

The quantitative data of sperm dynamics are provided by analyzing the sperm motility images. Nearly all commerical CASA instruments use the nearest neighbor (NN) tracking scheme [19], in which the initial image processing provides a centroid for each spermatozoon in the first frame of a scene, for each cell location of the most probable centroid in successive frames is deduced, and connecting the centroids for a spermatozoon provides its actual trajectory [44].

Our challenge is to match the same target from the current frame to the next frame, and we choose a classical k-nearest neighbor (k-NN) [45] algorithm to solve this problem, and an example of k-NN is shown in FIGURE 5.

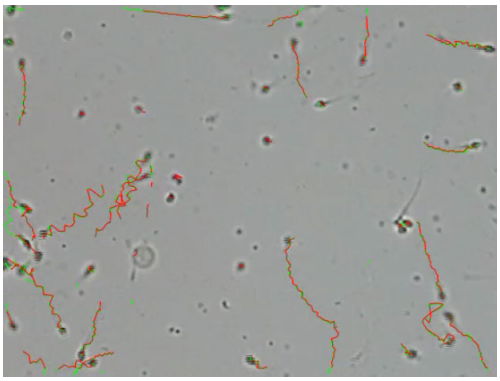

FIGURE 5. An example of the k-NN classification result for sperms. The green represents the actual trajectory of the sperm, and the red represents the trajectory calculated by k-NN.

Based on the results of $\psi_{(i)}$, we obtain the barycenter coordinates of all sperms $\varsigma_{(i,j)}$ in the video X_i. Then, we use k-NN algorithm to calculate Euclidean distance: There is a barycenter coordinate $c\left(s_{(i,j,l)}\right)$ in frame $x_{(i,j)}$, and next frame $x_{(i,j+1)}$, where all the barycenter coordinates are $C_{(i,j+1)} = \left\{c\left(s_{(i,j+1,1)}\right), c\left(s_{(i,j+1,2)}\right), \ldots, c\left(s_{(i,j+1,l)}\right), \ldots, c\left(s_{(i,j+1,q)}\right)\right\}$. We calculate the Euclidean distance between $c\left(s_{(i,j,l)}\right)$ and all the barycenter coordinates in $C_{(i,j+1)}$, and figure out a set of Euclidean distance $D_{(i,j)} = \left\{d_{(i,j,1)}, d_{(i,j,2)}, \ldots, d_{(i,j,l)}, \ldots, d_{(i,j,q)}\right\}$, where $d_{(i,j,l)}$ is the Euclidean distance between $c\left(s_{(i,j,l)}\right)$ and $c\left(s_{(i,j+1,l)}\right)$ in Eq. (2).

$$d_{(i,j,l)} = \sqrt{\left[c\left(s_{(i,j+1,l)}\right) - c\left(s_{(i,j,l)}\right)\right]^2} \tag{2}$$

We find the minimum in $D_{(i,j)}$, and define this minimum as $d_{\min}\left(D_{(i,j)}\right)$. We use k-NN algorithm to classify all the barycentric coordinates to their corresponding coordinates in the former frame of the video. The result of classification is that all barycentric coordinates $\psi_{(i)} = \left\{C_{(i,1)}, C_{(i,2)}, \ldots, C_{(i,j)}, \ldots, C_{(i,m)}\right\}$ of the same sperm target in the video X_i are classified into one category. An example of a classification is shown in FIGURE 6.

As the example shown in FIGURE 6, we define a set of classification result as $\phi_{(i)} = \left\{S_{(i,1)}, S_{(i,2)}, \ldots, S_{(i,g)}, \ldots, S_{(i,\tau)}\right\}$, where $S_{(i,g)} = \left\{I_{(i,j,g)}, I_{(i,j+1,g)}, \ldots\right\}$ is a set of all the barycentric coordinates of one sperm in this video X_i, I is the barycentric coordinate variable, j is the frame number, g is the index number of classification result, τ is the total number of classification result, and i is the video number.

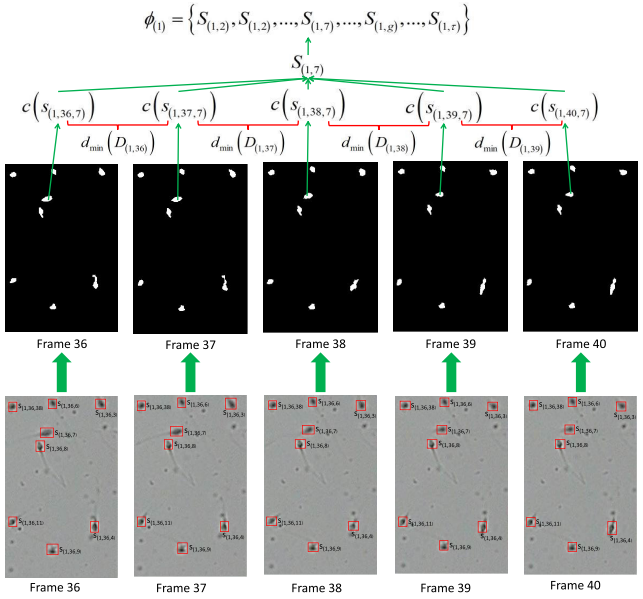

FIGURE 6. An example of the k-NN classification result for sperms.

In the video X_i, there are sperms constantly swimming into or out of the visual field, therefore, sperm counts are inequality in different frames. According to this practical situation, we give a solution strategy as follow:

- **Case-I:** If there is a sperm swimming into the visual field, we define this sperm as a new target, and it will have a new classification result for its own with the k-NN classifier.
- **Case-II:** If there is a sperm swimming out of the visual field, we stipulate that the motion of this sperm is over.

Based on **Case-I** and **Case-II**, we can conclude that the number of classification result $\phi_{(i)} = \left\{S_{(i,1)}, S_{(i,2)}, \ldots, S_{(i,g)}, \ldots, S_{(i,\tau)}\right\}$ is the total number of sperms in the video X_i, where τ is the total number of sperms. FIGURE 7 shows an example of the sperm count statistics from frame 36 to frame 80 in video X_i.

5) CONSTRUCTION OF THE FOLDOVER

According to the result of k-NN classification, we get the barycentric coordinates $\phi_{(i)} = \left\{S_{(i,1)}, S_{(i,2)}, \ldots, S_{(i,g)}, \ldots, S_{(i,\tau)}\right\}$ of all the sperms in the video X_i. The following operations are performed for each k-NN classification result $\phi_{(i)}$. First, according to k-NN classification result $S_{(i,g)}$, we determine the range of the frames in which the sperm moves. Then, we extract these frames from the segmentation results $X_i^{\text{seg}} = \left\{x_{(i,1)}^{\text{seg}}, x_{(i,2)}^{\text{seg}}, \ldots, x_{(i,j)}^{\text{seg}}, \ldots, x_{(i,m)}^{\text{seg}}\right\}$ according to the range of frames. In these extracted frames, setting the barycentric coordinates $S_{(i,g)} = \left\{I_{(i,j,g)}, I_{(i,j+1,g)}, \ldots\right\}$ as the center, setting r pixels as a standard radius. We calculate the distance between the barycentric coordinates

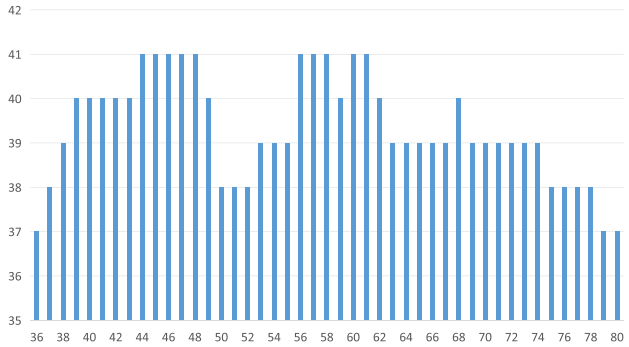

FIGURE 7. The total number of sperms from frame 36 to frame 80 in video X_i.

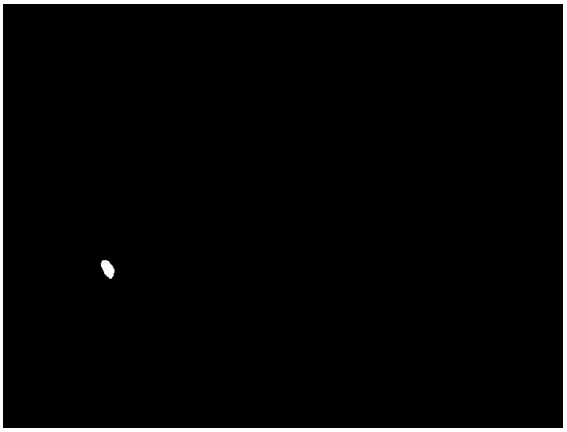

FIGURE 8. An example of a $L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right)$ result.

$S_{(i,g)} = \{I_{(i,j,g)}, I_{(i,j+1,g)}, \ldots\}$ and all other pixels in the $X_i^{\text{seg}} = \left\{x^{\text{seg}}_{(i,1)}, x^{\text{seg}}_{(i,2)}, \ldots, x^{\text{seg}}_{(i,j)}, \ldots, x^{seg}_{(i,m)}\right\}$, so the pixel value $p\left[L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j,k)}\right)\right]$ is defined as Eq. (3).

$$p\left[L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j,k)}\right)\right] = \begin{cases} p\left(x^{\text{seg}}_{(i,j,k)}\right) & \sqrt{\left(x^{\text{seg}}_{(i,j,k)} - I_{(i,j,g)}\right)^2} \le r \\ 0 & otherwise \end{cases} \quad (3)$$

The pixel value $p\left[L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j,k)}\right)\right]$ is 0 which is more than r pixels away from the barycentric coordinate $I_{(i,j,g)}$; otherwise, the pixel value $p\left[L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j,k)}\right)\right]$ is $p\left(x^{\text{seg}}_{(i,j,k)}\right)$. In this way, we target each sperm in the segmentation results, and we define the result as $L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right)$. $L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right)$ is the image segmentation result of the g-th sperm in the j-th frame of the i-th video, and an example of the $L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right)$ result is shown in FIGURE 8.

Second, according to the $L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right)$ result, we can get a set $\theta_{(i,g)} = \left\{L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right), L_{(i,j+1,g)}\left(x^{\text{seg}}_{(i,j+1)}\right), \ldots\right\}$ of the same sperm. We use $L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j)}\right)$ to localize the sperm region from the original frame (image) according to Eq. (4).

$$p\left[o_{(i,j,g)}\left(x_{(i,j,k)}\right)\right] = \begin{cases} 0 & p\left[L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j,k)}\right)\right] = 0 \\ p\left(x_{(i,j,k)}\right) & otherwise \end{cases} \quad (4)$$

In Eq. (4), we define the extracted result as $o_{(i,j,g)}\left(x_{(i,j)}\right)$. If the pixel value $p\left[L_{(i,j,g)}\left(x^{\text{seg}}_{(i,j,k)}\right)\right]$ is equal to 0, the pixel value $p\left[o_{(i,j,g)}\left(x_{(i,j,k)}\right)\right]$ is 0; otherwise, the pixel value $p\left[o_{(i,j,g)}\left(x_{(i,j,k)}\right)\right]$ is the pixel value $p\left(x_{(i,j,k)}\right)$ corresponding to the original image. According to Eq. (4), we hold on the image of each sperm, and we define the result as $o_{(i,j,g)}\left(x_{(i,j)}\right)$. $o_{(i,j,g)}\left(x_{(i,j)}\right)$ is the image of the g-th sperm in the j-th frame of the i-th video, in which the background is black, and FIGURE 9 is an example of $o_{(i,j,g)}\left(x_{(i,j)}\right)$.

Thirdly, a set of the $o_{(i,j,g)}\left(x_{(i,j)}\right)$ is denoted as $O_{(i,g)}$, where we define $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$ as the total number of extracted results in $O_{(i,g)}$, and $O_{(i,g)}$ is defined as Eq. (5).

$$O_{(i,g)} = \left\{o_{(i,j,g)}\left(x_{(i,j)}\right), o_{(i,j+1,g)}\left(x_{(i,j+1)}\right), \ldots, o_{(i,\Gamma_{(i,g)}(O_{(i,g)}),g)}\left(x_{(i,\Gamma_{(i,g)}(O_{(i,g)}))}\right)\right\} \quad (5)$$

According to Eq. (5), we can obtain a set $O_{(i,g)}$ of images of the g-th sperm in the i-th video. $O_{(i,g)}$ contains all the images of the g-th sperm in the i-th video, these images have a black background such as the example in FIGURE 9.

FIGURE 9. An example of a $o_{(i,j,g)}\left(x_{(i,j)}\right)$ result.

Based on the Eq. (5), we define $F_{(i,g)}$ as the foldover of the g-th sperm in the i-th video, and $p\left[F_{(i,g)}\left(x_{(i,j,k)}\right)\right]$ is expressed by Eq. (6).

$$p\left[F_{(i,g)}\left(x_{(i,j,k)}\right)\right] = \sum_{j}^{\Gamma_{(i,g)}(O_{(i,g)})} p\left[o_{(i,j,g)}\left(x_{(i,j,k)}\right)\right] \quad (6)$$

As the definition in Eq. (6), we add up the k-th pixel of each frame in $O_{(i,g)}$, and the sum is $p\left[F_{(i,g)}\left(x_{(i,j,k)}\right)\right]$, $k = 1, 2, 3, \ldots, h$, k is the pixel number, h is the total number of pixels in a frame, $h = h_1 \times h_2$, $h_1 = 1, 2, 3, \ldots$ is the number of pixels in a row, and $h_2 = 1, 2, 3, \ldots$ is the number

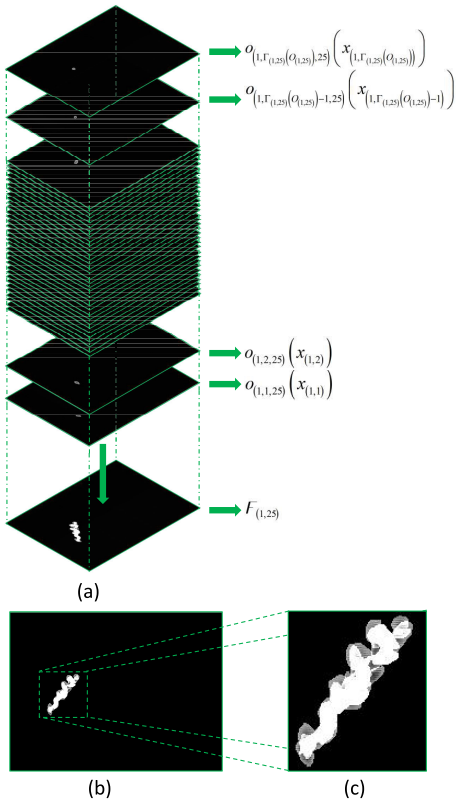

FIGURE 10. An example of foldover $F_{(i,g)}$ construction. (a) is the process of $O_{(i,g)}$ accumulation. We add up corresponding pixels in $O_{(i,g)}$, and the cumulative result is the foldover $F_{(i,g)}$. (b) is the cumulative result of the $O_{(i,g)}$. (c) is the two-dimensional visualization of the foldover $F_{(i,g)}$.

of pixels in a column. In this way, we add the corresponding pixel values in different frames to obtain $F_{(i,g)}$. $F_{(i,g)}$ is the foldover of the g-th sperm in the i-th video, $p\left[F_{(i,g)}\left(x_{(i,j,k)}\right)\right]$ is the pixel value of the k-th position in $F_{(i,g)}$, and the $F_{(i,g)}$ is shown in FIGURE 10.

By method of accumulation in FIGURE 10, images of the same sperm in different frames are placed on the same spatial plane. In this spatial plane, the images in $O_{(i,g)}$ construct the foldover of the g-th sperm in the i-th video.

6) CONSTRUCTION OF 3D IMAGES

In the video X_i, the swimming directions of sperms are uncertain. Therefore, we need to unify the swimming directions of sperms to facilitate our experimental analysis. We define the direction in which the starting barycentric coordinate of sperms to their ending coordinate as the positive direction (forward direction), and the horizontal direction is defined as the X direction. In order to unify the swimming directions, we rotate the foldover $F_{(i,g)}$ into this positive direction to the X direction, and we define the rotated foldover $F_{(i,g)}$ as $F^R_{(i,g)}$. An example of $F^R_{(i,g)}$ is shown in FIGURE 11.

FIGURE 11 is only the two-dimensional visualization result of the $F^R_{(i,g)}$, it cannot contain all the information of the $F^R_{(i,g)}$. Therefore, we show a 3D vision of the $F^R_{(i,g)}$ to reflect all the information of the foldover in FIGURE 12.

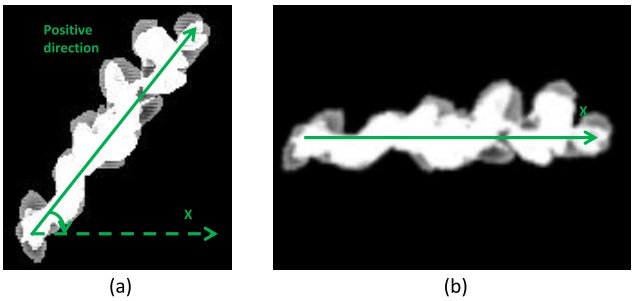

FIGURE 11. An example of $F^R_{(i,g)}$ in 2D vision. (a) is the foldover $F_{(i,g)}$ before the rotation. (b) is the rotated foldover $F^R_{(i,g)}$.

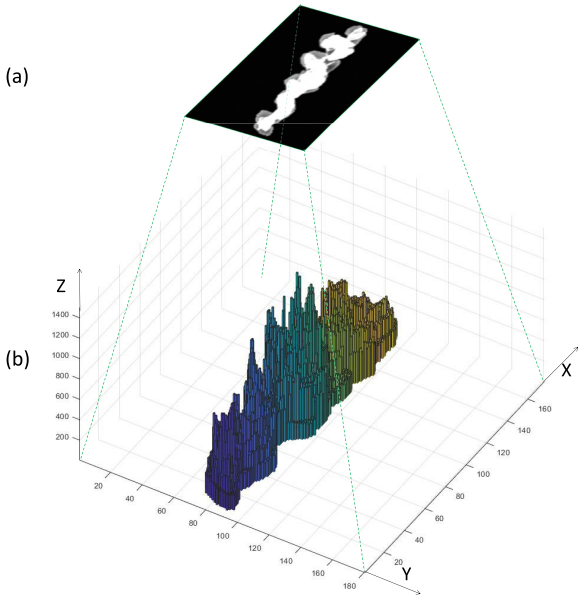

FIGURE 12. An example of the $F^R_{(i,g)}$ in 3D vision. (a) is the two-dimensional visualization of the foldover $F^R_{(i,g)}$. (b) is a 3D vision of the foldover $F^R_{(i,g)}$.

B. FOLDOVER FEATURES EXTRACTION

Foldover feature extraction is the statistics of the information in the $F^R_{(i,g)}$, which is also the focus of our whole method, and the method of foldover feature extraction consists of the following four steps.

1) FOLDOVER PROCESSING IN THE X, Y, AND Z DIRECTIONS

Foldover processing in the X, Y and Z directions is shown in FIGURE 13. First, we define the length of $F^R_{(i,g)}$ on X, Y and Z three directions as $F^R_{(i,g)}(\Theta)$, where Θ is defined in Eq. (7).

$$\Theta = \begin{cases} \text{X} & \text{along the X axis} \\ \text{Y} & \text{along the Y axis} \\ \text{Z} & \text{along the Z axis} \end{cases} \tag{7}$$

Second, we cut the foldover $F^R_{(i,g)}$ along the direction of Θ with a step length of v_Θ, and we can receive a set of slices which is defined as $F^{(R,\Theta)}_{(i,g)}$ in Eq. (8), u is the number, and

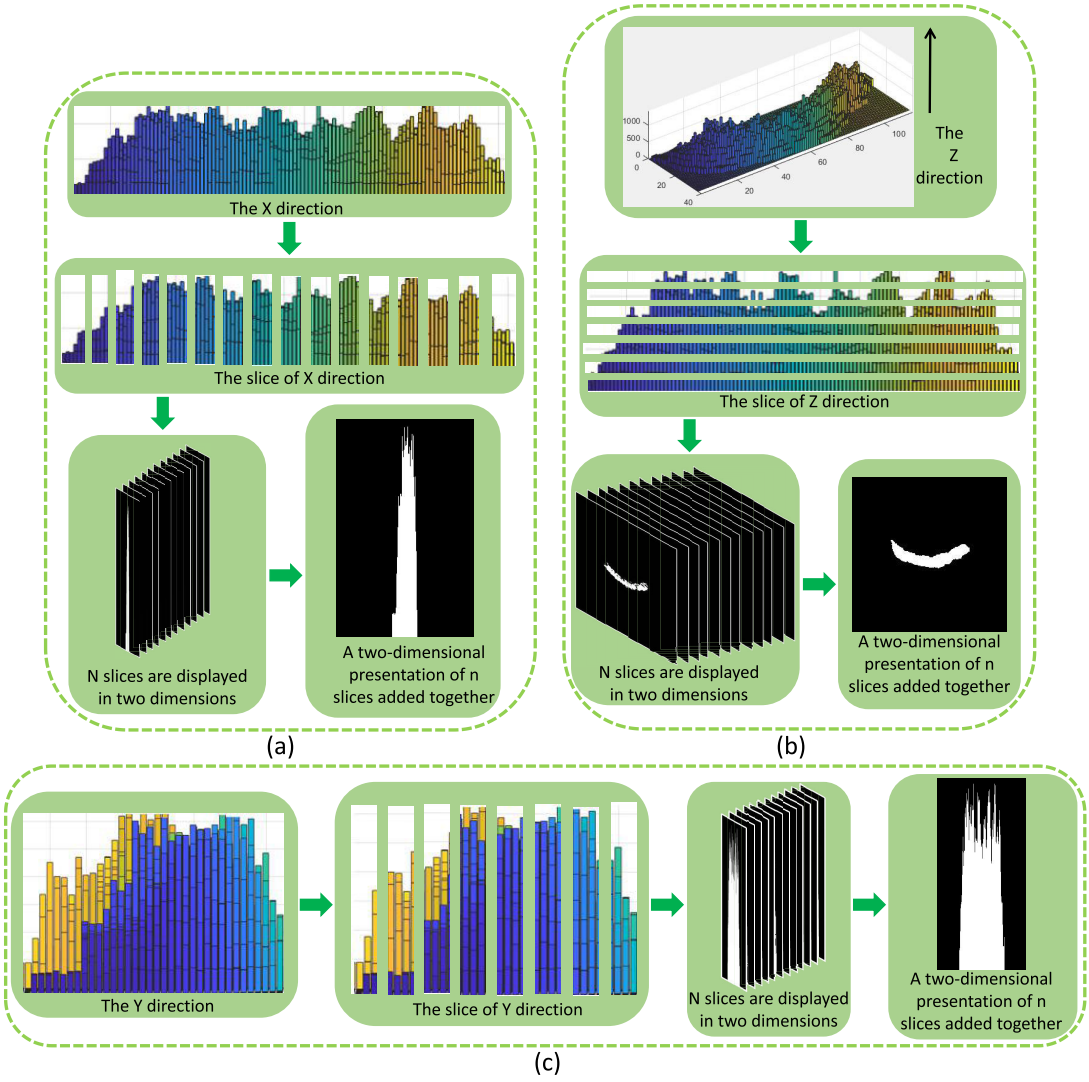

FIGURE 13. An example of foldover processing in the X, Y and Z directions. (a) is the foldover processing in the X direction, (b) is the foldover processing in the Z direction and (c) is the foldover processing in the Y direction.

$\frac{F^{R}_{(i,g)}(\Theta)}{v_\Theta}$ is the total number.

$$F^{(R,\Theta)}_{(i,g)} = \left\{ F^{(R,\Theta)}_{(i,g,1)}, F^{(R,\Theta)}_{(i,g,2)}, \ldots, F^{(R,\Theta)}_{(i,g,u)}, \ldots, F^{(R,\Theta)}_{\left(i,g,\frac{F^{R}_{(i,g)}(\Theta)}{v_\Theta}\right)} \right\} \tag{8}$$

Third, in X and Y directions, $F^{R}_{(i,g)}$ can reflect time information and movement information of sperms, but $F^{R}_{(i,g)}$ cannot reflect the information of pixel accumulation. For slices $F^{(R,\Theta)}_{(i,g)}$ (Θ = X or Y), we set the pixel values of the areas where the foldover exists to 1 and other areas to 0. We add $F^{(R,\Theta)}_{(i,g)}$ (Θ = X or Y) together as the result of $F^{R}_{(i,g)}$ in the X and Y directions. Unlike the foldover slices in the X and Y directions, the foldover slices in Z direction truly reflect the effect of pixel accumulation. Therefore, we have no necessary to set the pixel values, so the pixel values of the areas where the foldover slices in the Z direction exists are added directly. We define the cumulative result of foldover slices $F^{(R,\Theta)}_{(i,g)}$ as $U\left(F^{(R,\Theta)}_{(i,g)}\right)$ (Θ = X, Y, or Z) in Eq. (9).

$$U\left(F^{(R,\Theta)}_{(i,g)}\right) = \sum_{u=1}^{\frac{F^{(R)}_{(i,g)}(\Theta)}{v_\Theta}} F^{(R,\Theta)}_{(i,g,u)} \tag{9}$$

Finally, we get three cumulative results $U\left(F^{(R,\Theta)}_{(i,g)}\right)$ of the foldover slices $F^{(R,\Theta)}_{(i,g)}$ in X, Y, and Z directions as shown in FIGURE 14.

2) FOLDOVER FEATURES EXTRACTION

Foldovers contain different behavior information in different directions. As the example shown in FIGURE 15, two sperms with different behaviors have completely different foldovers.

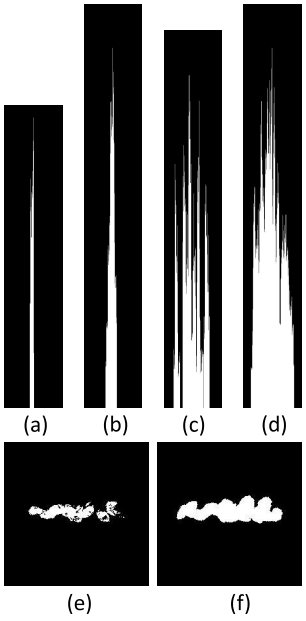

FIGURE 14. A slicing example and cumulative results of $U\left(F^{(R,\Theta)}_{(i,g)}\right)$. One of the slices in X,Y and Z directions ($F^{(R,X)}_{(i,g,u)}$, $F^{(R,Y)}_{(i,g,u)}$ and $F^{(R,Z)}_{(i,g,u)}$) is shown in (a), (c) and (e). The cumulative result of slices ($U\left(F^{(R,X)}_{(i,g)}\right)$, $U\left(F^{(R,Y)}_{(i,g)}\right)$ and $U\left(F^{(R,Z)}_{(i,g)}\right)$) are shown in (b), (d) and (f).

TABLE 2. Human sperm motility grade by World Health Organization (WHO) [46].

No.	Grades	Name	Movements μm/s
1	A	Rapid progressive motility	25
2	B	Slow progressive motility	5 <speed <25
3	C	non-progressive motility	<5
4	D	Immotility	No Movements

According to the human sperm quality assessment proposed by the World Health Organization (WHO) [46], sperm motility is grouped into four categories as shown in TABLE 2.

So, the grade of FIGURE 15 (a) is D (immotility), and the grade of FIGURE 15 (b) is A (rapid progressive motility). In X direction, the foldover contains the range of moving direction, which is the length of the foldover along the X direction $F^{R}_{(i,g)}(\mathrm{X})$. The total number of frames $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$ that make up the foldover $F_{(i,g)}$ is the time information, which is the movement time of sperm, and by $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$ we calculate the frame rate of $F_{(i,g)}$ in the X direction. We define the frame rate as $v^{(\mathrm{FPS,X})}_{(i,g)}$ in Eq. (10), and 2D visualization of two foldovers in the X direction are shown in (a) and (b) of FIGURE 16.

$$v^{(\mathrm{FPS,X})}_{(i,g)} = \frac{F^{R}_{(i,g)}(\mathrm{X})}{\Gamma_{(i,g)}\left(O_{(i,g)}\right)} \tag{10}$$

In Y direction, the foldover contains the range of the orthogonal direction of moving direction, which is the length of the foldover along the Y direction $F^{R}_{(i,g)}(\mathrm{Y})$. Similar to X direction, we calculate the frame rate of $F_{(i,g)}$ in the

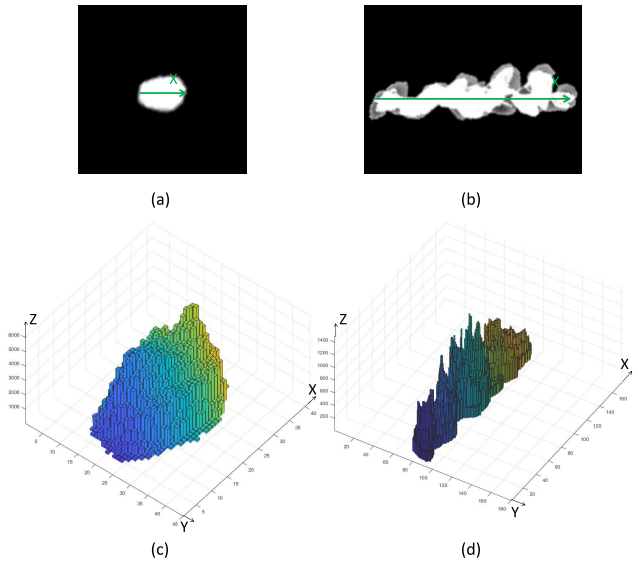

FIGURE 15. A comparison of foldovers of two different sperms. (a) (b) are 2D visualizations of foldovers. (c) (d) are 3D of the foldovers.

Y direction by $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$, we define the frame rate as $v^{(\mathrm{FPS,Y})}_{(i,g)}$ in Eq. (11), and 2D visualization of two foldovers in the Y direction are shown in (c) and (d) of FIGURE 16.

$$v^{(\mathrm{FPS,Y})}_{(i,g)} = \frac{F^{R}_{(i,g)}(\mathrm{Y})}{\Gamma_{(i,g)}\left(O_{(i,g)}\right)} \tag{11}$$

In Z direction, the foldover contains trajectory, shape, and brightness information. By the trajectory of the foldover we calculate the motion distance, the motion displacement and the average path length. Furthermore, we calculate the motion distance and the motion displacement by $\phi_{(i)} = \left\{S_{(i,1)}, S_{(i,2)}, \ldots, S_{(i,g)}, \ldots, S_{(i,\tau)}\right\}$ ($S_{(i,g)} = \left\{I_{(i,j,g)}, I_{(i,j+1,g)}, \ldots\right\}$), and by fitting $\phi_{(i)} = \left\{S_{(i,1)}, S_{(i,2)}, \ldots, S_{(i,g)}, \ldots, S_{(i,\tau)}\right\}$ to the third power, an equation can be calculated based on the motion path, then the average path length of sperm is calculated by combining this equation. We define the motion distance as $A_{(i,g)}$, the motion displacement as $B_{(i,g)}$, the fitted equation as $\varrho\left(I_{(i,j,g)}\right)$ and the average path length as $M_{(i,g)}$, the formula of $A_{(i,g)}$, $B_{(i,g)}$ and $M_{(i,g)}$ are expressed by Eq. (12), Eq. (13) and Eq. (14).

$$A_{(i,g)} = \sum_{j}^{\Gamma_{(i,g)}(O_{(i,g)})-1} \left[I_{(i,j+1,g)} - I_{(i,j+1,g)}\right] \tag{12}$$

In Eq. (12), we add up all the barycentric coordinates $S_{(i,g)} = \left\{I_{(i,j,g)}, I_{(i,j+1,g)}, \ldots\right\}$ contained in the foldover as the distance of motion $A_{(i,g)}$.

$$B_{(i,g)} = I_{(i,\Gamma_{(i,g)}(O_{(i,g)}),g)} - I_{(i,j,g)} \tag{13}$$

In Eq. (13), we calculate the distance between the first position $I_{(i,j,g)}$ and the last position $I_{(i,\Gamma_{(i,g)}(O_{(i,g)}),g)}$ of the

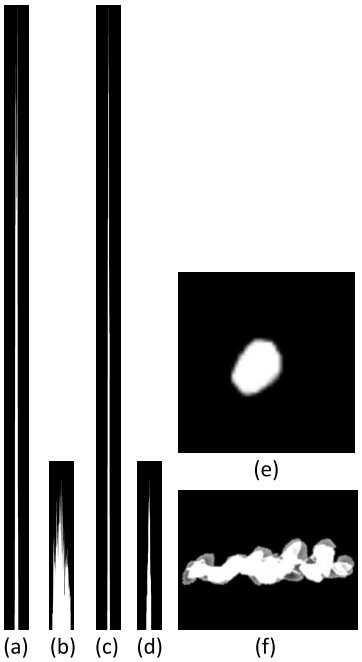

FIGURE 16. 2D visualization of two foldovers in the X, Y, and Z directions. (a) (c) (e) are the foldovers of a nearly stationary sperm in the three directions, therefore, the foldovers of this sperm in the X and Y directions are significantly higher than the others. (b) (d) (f) are the foldovers of the swimming sperm.

foldover as the motion displacement $B_{(i,g)}$.

$$M_{(i,g)} = \sum_{j}^{\Gamma_{(i,g)}\left(O_{(i,g)}\right)-1} \left[\varrho\left(I_{(i,j+1,g)}\right) - \varrho\left(I_{(i,j,g)}\right)\right] \quad (14)$$

In Eq. (14), by fitting the equation $\varrho\left(I_{(i,j,g)}\right)$, we can calculate the new coordinates corresponding to $S_{(i,g)} = \left\{I_{(i,j,g)}, I_{(i,j+1,g)}, \ldots\right\}$, and add the distance between these new coordinates we can obtain the average path length $M_{(i,g)}$.

According to the motion distance $A_{(i,g)}$, motion displacement $B_{(i,g)}$ and average path length $M_{(i,g)}$, we can further calculate curvilinear velocity (VCL), straight line velocity (VSL) and average path velocity (VAP) in Eq. (15).

$$\begin{aligned} v_{(i,g)}^{(\mathrm{VCL})} &= \frac{A_{(i,g)}}{\Gamma_{(i,g)}\left(O_{(i,g)}\right)} \\ v_{(i,g)}^{(\mathrm{VSL})} &= \frac{B_{(i,g)}}{\Gamma_{(i,g)}\left(O_{(i,g)}\right)} \\ v_{(i,g)}^{(\mathrm{VAP})} &= \frac{M_{(i,g)}}{\Gamma_{(i,g)}\left(O_{(i,g)}\right)} \end{aligned} \quad (15)$$

We define the VCL as $v_{(i,g)}^{(\mathrm{VCL})}$, and we obtain VCL based on $A_{(i,g)}$ and $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$. The VSL is defined as $v_{(i,g)}^{(\mathrm{VSL})}$, we calculate the VSL based on $B_{(i,g)}$ and $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$. The VAP is defined as $v_{(i,g)}^{(\mathrm{VAP})}$, we calculate the VAP based on $M_{(i,g)}$ and $\Gamma_{(i,g)}\left(O_{(i,g)}\right)$.

Furthermore, using $v_{(i,g)}^{(\mathrm{VCL})}$, $v_{(i,g)}^{(\mathrm{VSL})}$ and $v_{(i,g)}^{(\mathrm{VAP})}$, we can calculate linearity (LIN), Straightness (STR) and Wobble (WOB) in Eq. (16).

$$\begin{aligned} \mathrm{LIN}_{(i,g)} &= \frac{v_{(i,g)}^{(\mathrm{VSL})}}{v_{(i,g)}^{(\mathrm{VCL})}} \\ \mathrm{STR}_{(i,g)} &= \frac{v_{(i,g)}^{(\mathrm{VSL})}}{v_{(i,g)}^{(\mathrm{VAP})}} \\ \mathrm{WOB}_{(i,g)} &= \frac{v_{(i,g)}^{(\mathrm{VAP})}}{v_{(i,g)}^{(\mathrm{VCL})}} \end{aligned} \quad (16)$$

$\mathrm{LIN}_{(i,g)}$ is the ratio of $v_{(i,g)}^{(\mathrm{VSL})}$ to $v_{(i,g)}^{(\mathrm{VCL})}$, $\mathrm{STR}_{(i,g)}$ is the ratio of $v_{(i,g)}^{(\mathrm{VSL})}$ to $v_{(i,g)}^{(\mathrm{VAP})}$, and $\mathrm{WOB}_{(i,g)}$ is the ratio of $v_{(i,g)}^{(\mathrm{VAP})}$ to $v_{(i,g)}^{(\mathrm{VCL})}$.

Regarding the shape information, foldovers can detect the deformation of sperm during the movement. The brightness information mainly includes the pixel accumulation process, the higher brightness area indicates that the sperm stay in this area for the longer time. 2D visualization of two foldovers in the Z direction are shown in (e) and (f) of FIGURE 16.

Although $U\left(F_{(i,g)}^{(R,X)}\right)$, $U\left(F_{(i,g)}^{(R,Y)}\right)$ and $U\left(F_{(i,g)}^{(R,Z)}\right)$ include the information of foldovers, they are three matrices of an object (such as a sperm) with a lot of redundant information. Therefore, we make statistics on all the information of $U\left(F_{(i,g)}^{(R,X)}\right)$, $U\left(F_{(i,g)}^{(R,Y)}\right)$ and $U\left(F_{(i,g)}^{(R,Z)}\right)$ to optimize them. Especially, we apply convolutional operations to achieve the optimization, where we define the process of convolution optimization as H^{Θ} (Θ = X, Y and Z), $H_{(\mathrm{i,g,k})}^{\Theta}$ is the k-th pixel of the g-th foldover in the i-th video, and H is defined in Eq. (17).

$$H_{(\mathrm{i,g,k})}^{\Theta} = U * G = \sum_{e} p\left[U_{(i,g,k)}\left(F_{(i,g)}^{(R,\Theta)}\right)\right] p\left(G_{(k-e)}\right) \quad (17)$$

In Eq. (17), G is the convolution kernel, and e is the dimension of the G. Here, because we cannot consolidate all the useful information and get rid of all the redundant information by just once convolution, we need to do multiple convolutions.

Furthermore, $v_{(i,g)}^{(\mathrm{FPS,X})}$, $v_{(i,g)}^{(\mathrm{FPS,Y})}$, $A_{(i,g)}$, $B_{(i,g)}$, $M_{(i,g)}$, $v_{(i,g)}^{(\mathrm{VCL})}$, $v_{(i,g)}^{(\mathrm{VSL})}$, $v_{(i,g)}^{(\mathrm{VAP})}$, $\mathrm{LIN}_{(i,g)}$, $\mathrm{STR}_{(i,g)}$, $\mathrm{WOB}_{(i,g)}$ and $H_{(\mathrm{i,g,k})}^{\Theta}$ are joined together to form three foldover feature vectors, where $v_{(i,g)}^{(\mathrm{FPS,X})}$ and $H_{(\mathrm{i,g,k})}^{\mathrm{X}}$ are concatenated to form the foldover feature $F_{(i,g)}^{\mathrm{X}}$ of the X direction; $v_{(i,g)}^{(\mathrm{FPS,Y})}$ and $H_{(\mathrm{i,g,k})}^{\mathrm{Y}}$ are concatenated to form the foldover feature $F_{(i,g)}^{\mathrm{Y}}$ of the Y direction; $A_{(i,g)}$, $B_{(i,g)}$, $M_{(i,g)}$, $v_{(i,g)}^{(\mathrm{VCL})}$, $v_{(i,g)}^{(\mathrm{VSL})}$, $v_{(i,g)}^{(\mathrm{VAP})}$, $\mathrm{LIN}_{(i,g)}$, $\mathrm{STR}_{(i,g)}$, $\mathrm{WOB}_{(i,g)}$ and $H_{(\mathrm{i,g,k})}^{\mathrm{Z}}$ are concatenated to form the foldover feature $F_{(i,g)}^{\mathrm{Z}}$ of the Z direction. The algorithm of the foldover features are shown in Algorithm 1.

Algorithm 1 Generation of $H^{\Theta}_{(\mathrm{i,g,k})}$

Input: Videos χ
preprocessed video X_i
Output: $H^{\Theta}_{(\mathrm{i,g,k})}$, $\Theta = \mathrm{X}, \mathrm{Y}$ and Z

1: video decomposition:
$X_i = \{x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,j)}, \ldots, x_{(i,m)}\}$
2: image segmentation: $x^{seg}_{(i,j)} \Leftarrow x_{(i,j)}$

$$p\left(x^{\mathrm{seg}}{}_{(i,j,k)}\right) = \begin{cases} 0 & p\left(x_{(i,j,k)}\right) \le T\left(x_{(i,j)}\right) \\ 1 & otherwise \end{cases}$$

3: barycenter coordinates extraction:
$\psi_{(i)} = \{C_{(i,1)}, C_{(i,2)}, \ldots, C_{(i,j)}, \ldots, C_{(i,m)}\}$
4: target matching:

$$d_{(i,j,l)} = \sqrt{\left[c\left(s_{(i,j+1,l)}\right) - c\left(s_{(i,j,l)}\right)\right]^2}$$

5: construction of the foldover:

$$p\left[F_{(i,g)}\left(x_{(i,j,k)}\right)\right] = \sum_{j}^{\Gamma_{(i,g)}(O_{(i,g)})} p\left[o_{(i,j,g)}\left(x_{(i,j,k)}\right)\right]$$

6: rotate the foldover: $F^{R}_{(i,g)} \Leftarrow F_{(i,g)}$
7: foldover processing: $F^{(R,\Theta)}_{(i,g)} \Leftarrow F^{R}_{(i,g)}$

$$U\left(F^{(R,\Theta)}_{(i,g)}\right) = \sum_{u=1}^{\frac{F^{(R)}_{(i,g)}(\Theta)}{v_\Theta}} F^{(R,\Theta)}_{(i,g,u)}$$

8: the optimization of $U\left(F^{(R,\Theta)}_{(i,g)}\right)$:
$H^{\Theta}_{(\mathrm{i,g,k})} = U * G$
9: the generation of foldover features: $F^{\mathrm{X}}_{(i,g)}, F^{\mathrm{Y}}_{(i,g)}, F^{\mathrm{Z}}_{(i,g)}$

Finally, We obtain the foldover feature vectors, $F^{\mathrm{X}}_{(i,g)}, F^{\mathrm{Y}}_{(i,g)}$ and $F^{\mathrm{Z}}_{(i,g)}$. $F^{\mathrm{X}}_{(i,g)}, F^{\mathrm{Y}}_{(i,g)}$ and $F^{\mathrm{Z}}_{(i,g)}$ are extracted from the single foldover $F^{R}_{(i,g)}$ of the same sperm, and they represent the foldover features from the three directions of X, Y and Z respectively. The visual information of the foldover $F^{R}_{(i,g)}$ is different in the X, Y and Z directions, while $F^{\mathrm{X}}_{(i,g)}, F^{\mathrm{Y}}_{(i,g)}$ and $F^{\mathrm{Z}}_{(i,g)}$ represent the visual information in each direction. $F^{\mathrm{X}}_{(i,g)}, F^{\mathrm{Y}}_{(i,g)}$ and $F^{\mathrm{Z}}_{(i,g)}$ contain temporal information, spatial information, behavior features and static features, and foldover features are a kind of behavior feature based on foldover for dynamic targets. According to the foldover features, we can solve the following difficulties we encounter in the microscopic videos: (1) Multi-object recognition, (2) Similar object recognition, (3) Tiny object recognition, (4) Impurity interference and (5) Little feature information.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, experimental results and analysis are discussed, including IV-A experimental setting, IV-B experimental results.

A. EXPERIMENTAL SETTING

1) EXPERIMENTAL DATA

In this paper, a practical microscopic video set $\chi = \{X_1, X_2, \ldots, X_i, \ldots, X_{59}\}$ with 59 semen videos is applied to test our method. The format of the videos is grey-scale mp4, the size of each frame is 698 × 528 × 3 pixels and the frame rate is 30 frames per second (FPS). There are 1,374 sperms in set χ. For all the sperms, ground truth (GT) images are prepared manually by four biomedical engineers and two medical doctors, where the sperms are labeled as foreground object with 1 (white) and other regions are labeled as background with 0 (black). We mark the number of each sperm in the video and propose the following strategy for sperm numbering:

- **Case-I:** All the sperms in the video (moving or stationary) are numbered, the numbers increased from 1, and each sperm is numbered horizontally from the top of the visual field.
- **Case-II:** If there is a sperm swimming out of the visual field, we stipulate that the motion of this sperm is over.
- **Case-III:** If there is a sperm swimming into the visual field, we assume we have a new sample and give it a new number.
- **Case-IV:** A malformed sperm is considered a sample, for example a sperm with two heads or two tails.

Furthermore, based on the diagnosis of the medical doctors, all sperms are grouped into three classes, including poor motion state, good motion state and excellent motion state. There are 950 samples of poor motion state, 262 samples of good motion state and 162 samples of excellent motion state. Also, the number of samples in the training set is equal to that in the testing set, 687 samples are used for the training set and 687 samples are for testing. In the training set, the sample number of poor motion state is 462, the sample number of good motion state is 138, and the sample number of excellent motion state is 87. In the testing set, the sample number of poor motion state is 488, the sample number of good motion state is 124, and the sample number of excellent motion state is 75. An example of the video frames and their GT images is shown in FIGURE 17.

2) EVALUATION INDEX

We use classifiers to evaluate foldover features with a three-class classification task of sperms, and the classification evaluation indicators are shown in TABLE 2 [46]. Specifically, four classifiers are tested in this paper, including Artificial Neural Networks [40] (ANNs), Random Forests [42] (RFs) and Support Vector Machines [43] (linear-SVM and RBF-SVM). Because there are more motionless, slow-swimming sperms and fewer fast-swimming sperms in the videos, we calculate multiple indexes to evaluate the proposed foldover features. Firstly, we calculate the confusion matrix of all classification results. Then, based on the confusion matrices, we can further calculate the accuracy, precision, recall, specificity and F1-measure as shown in TABLE 3.

The negative number of the actual sample is N = TN + FP, the number of positive is P = FN + TP, and the total sample size is C = N + P, where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. Recall (also known as sensitivity) can measure the

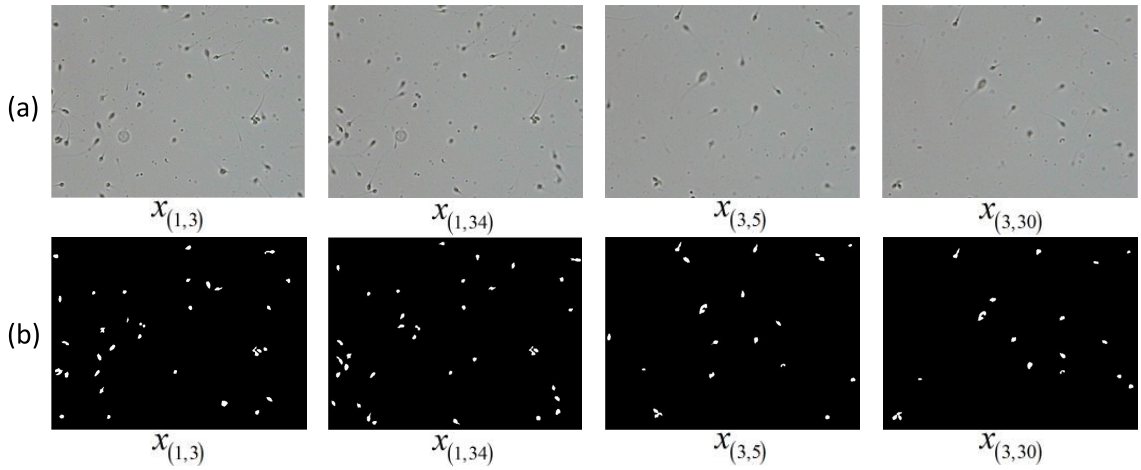

FIGURE 17. An example of frames and their GT images in a semen microscopic video. (a) shows the frames and (b) shows the GT images.

TABLE 3. The evaluation of confusion matrix.

Indicators	Formulas
Accuracy	$(\mathrm{TP}+\mathrm{TN})/(\mathrm{P}+\mathrm{N})$
Precision	$\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
Recall	$\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$
Specificity	$\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$
F1-measure	$2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$

reliability of the model's prediction with a positive sample, a higher recall means that an algorithm returns more relevant results. Precision can measure the accuracy of the model in predicting positive samples, a higher precision means that an algorithm returns substantially more relevant results than irrelevant ones. Specificity (also called the TN rate) measures the proportion of actual negatives that are correctly identified as such. F1-measure is a measure of the accuracy of a test, considering both the precision and the recall of the test to compute the score.

Thirdly, because our experiment is used for three categories, the precision has three values, and each class has its corresponding precision, we define the three values of precision as Precision1, Precision2 and Precision3. In the same way, there are also three values for recall defined as Recall1, Recall2, Recall3. Based on the confusion matrices, we can calculate the macro precision, the macro recall and the macro F1-measure as shown in TABLE 4.

Since our experiment is a triage experiment, therefore, when we calculate Macro_P, we need to calculate the mean of Precision1, Precision2 and Precision3, and the calculation of Macro_R is the same. Finally, based on the accuracy of each category, we calculate the varianceas shown in TABLE 4.

B. EXPERIMENTAL RESULTS

1) EVALUATION FOR FOLDOVER FEATURES

Artificial Neural Networks [40] (ANNs), Random Forests [42] (RFs) and Support Vector Machine [43] (linear-SVM and RBF-SVM) are used to test the effectiveness of the foldover features. Specifically, the parameters of the ANNs are set as follows: The number of network layers is 2, the number of hidden nodes is 10, and the activation function is log-sigmoid; The parameter of the RFs is set as follows: The number of decision tree is 200; The parameters of the Support Vector Machine are set as follows: Kernel function of linear-SVM is linear kernel, kernel function of RBF-SVM is radial basis function.

The foldover features, $F^{\mathrm{X}}_{(i,g)}$, $F^{\mathrm{Y}}_{(i,g)}$ and $F^{\mathrm{Z}}_{(i,g)}$ are classified by ANNs, RFs, linear-SVM and RBF-SVM, and the confusion matrices of classification results are shown in FIGURE 18.

$F^{\mathrm{Z}}_{(i,g)}$ obtains the best results in four classifiers, especially in ANNs, the accuracy is 91.8%, and the classification accuracy of each category is also excellent, 93.5%, 87.1% and 89.2%, respectively.

2) COMPARISON WITH STATIC FEATURES

Firstly, according to the $\phi_{(i)} = \{S_{(i,1)}, S_{(i,2)}, \ldots, S_{(i,g)}, \ldots, S_{(i,\tau)}\}$, each sperm is detected to a size of 26 by 26 pixels in the corresponding frame, the pixel size of 26 by 26 is an ideal size after repeated experiments to ensure which is the only sperm we want in the detected image, and an example of some detected sperms is shown in FIGURE 19.

Secondly, we extract the static features of sperms after detection, including Histogram of Oriented Gradient [8] (HOG), Grey-Level Co-occurrence Matrix [9] (GLCM), the geometric invariant moment proposed by Hu [10], Scale-Invariant Feature Transform [12] (SIFT) and gray histogram [6]. All static features are extracted from detected sperm images, but the movement of a sperm exists in multiple frames, therefore, we adopt the method of multiple extraction, and randomly select one sperm image from all the images of this sperm at a time to extract the static features. Thirdly, the number of times to extract the static feature is ten,

TABLE 4. The evaluation of three categories.

Indicators	Formulas
Macro_P	$1/3\ (\text{Precision1} + \text{Precision2} + \text{Precision3})$
Macro_R	$1/3\ (\text{Recall1} + \text{Recall2} + \text{Recall3})$
Macro_F1	$(2 \times \text{Macro\_P} \times \text{Macro\_R})/(\text{Macro\_P+Macro\_R})$
Variance	$1/3\ \left[(\text{Recall1} - \text{Macro\_R})^2 + (\text{Recall2} - \text{Macro\_R})^2 + (\text{Recall3} - \text{Macro\_R})^2\right]$

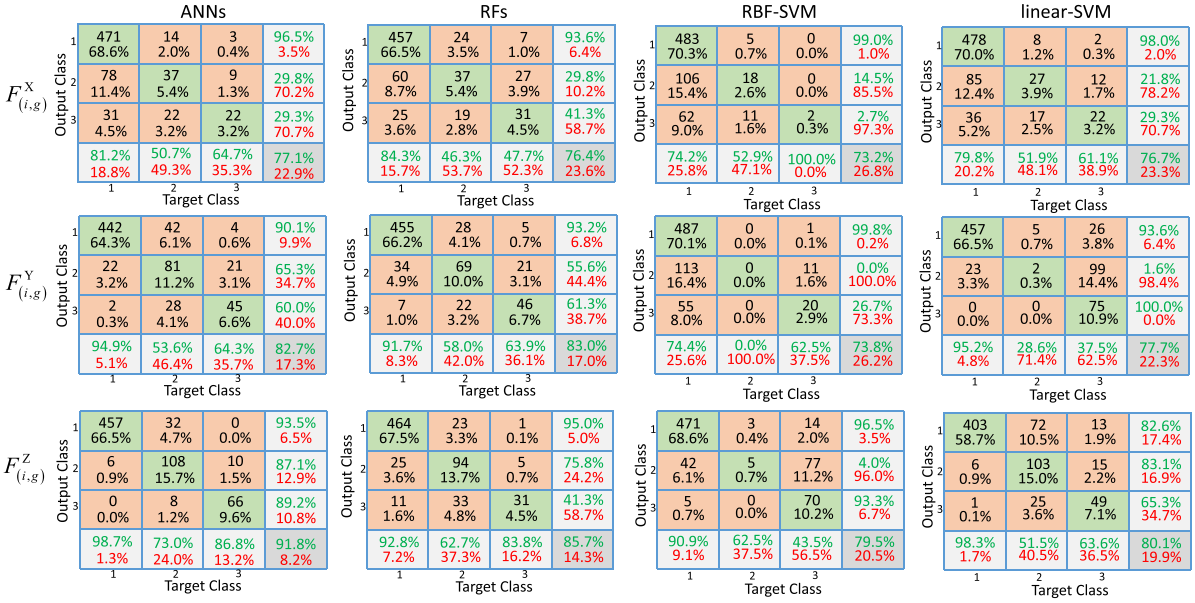

FIGURE 18. The confusion matrices of $F^{X}_{(i,g)}$, $F^{Y}_{(i,g)}$ and $F^{Z}_{(i,g)}$. Rows represent the foldover features, and columns represent the classifiers.

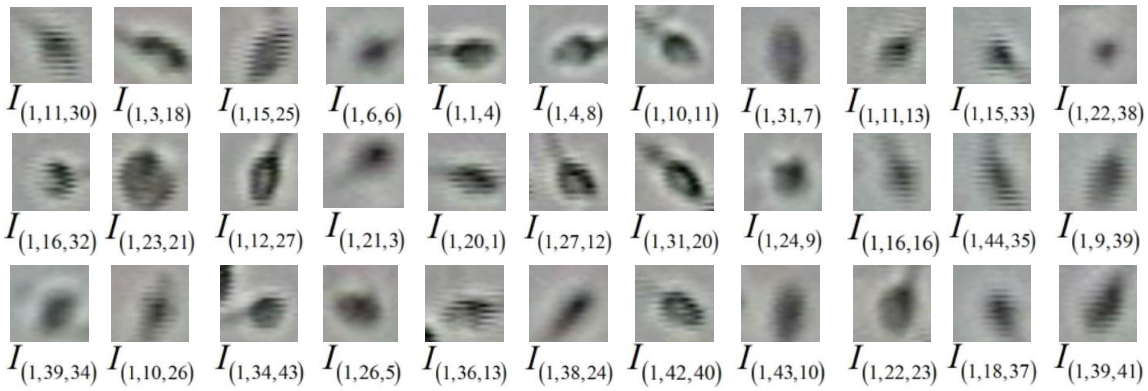

FIGURE 19. An example of detected sperms.

obviously, the number of times to classify the static feature is ten. We use Artificial Neural Networks [40] (ANNs), Random Forests [42] (RFs) and Support Vector Machine [43] (linear-SVM and RBF-SVM)) classifiers to classify static features and construct the total confusion matrices of ten experiments to represent the classification results. The classification results of static features in four classifiers are shown in in FIGURE 20.

According to the confusion matrices of static features in FIGURE 20, we calculate the evaluations, and the comparison evaluations between static features and foldover features are shown in TABLE 5.

Considering the comparison in TABLE 5, the accuracy of $F^{X}_{(i,g)}$, $F^{Y}_{(i,g)}$ and $F^{Z}_{(i,g)}$ are significantly higher than that of static features. The reason for the low accuracy of static features is: Static features are extracted from detected sperms

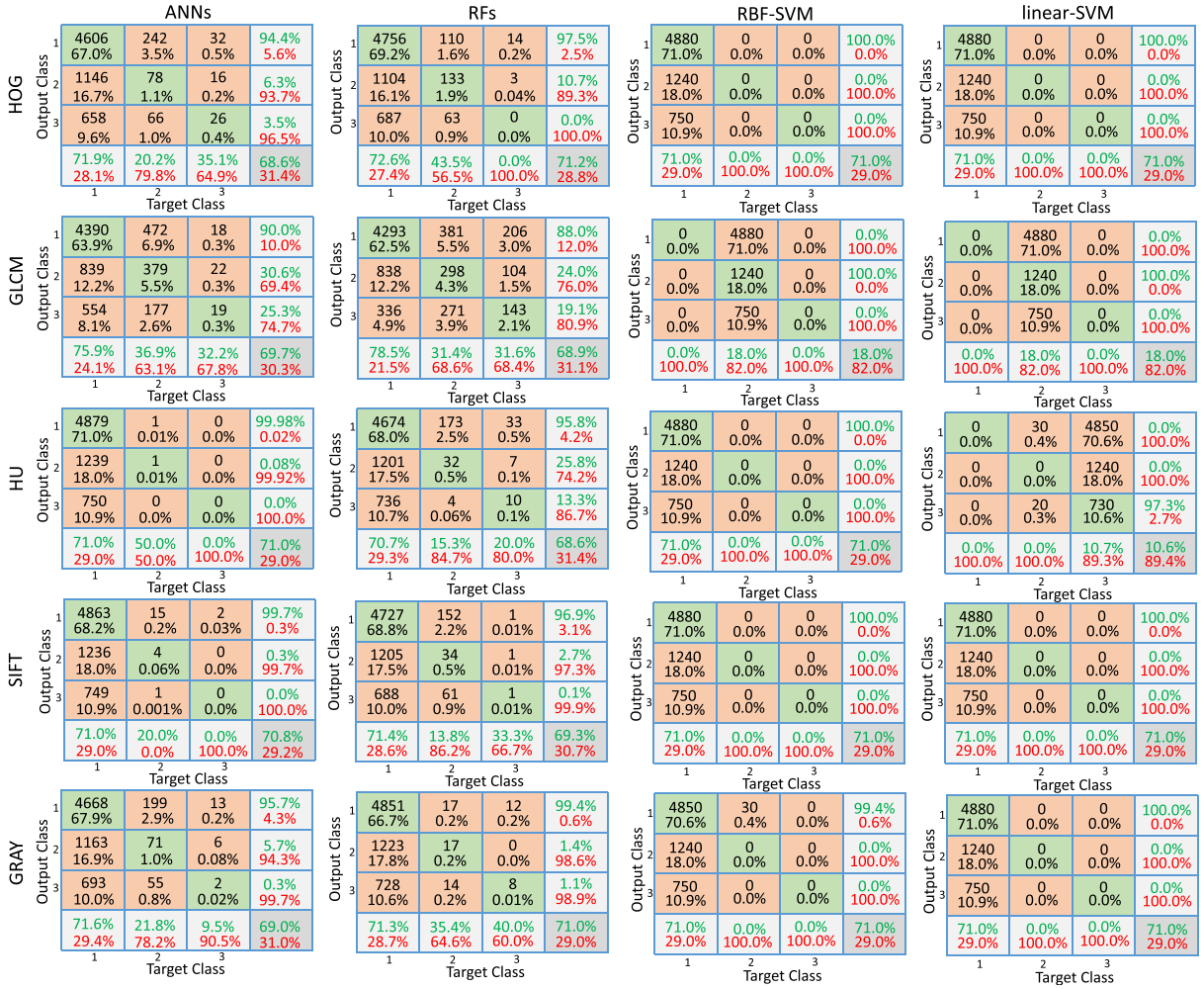

FIGURE 20. Confusion matrices of the static features. Rows represent feature types, and columns represent classifier types.

TABLE 5. Evaluation of static features with four classifiers. The first column shows the types of static features, the second column shows the types of classifiers, the third to the last columns show the calculated evaluations. We use the first three letters of each evaluation to indicate the evaluation metric, such as Acc is accuracy, Pre is precision, Mac_P is Macro_P, Rec is recall, Mac_R is Macro_R, Spe is specificity, F1-mea1 is F1-measure, Mac_F1 is Macro_F1 and Var is variance. The red font value means that the value is the maximum value in the column (Unit: %).

Feature	Classifier	Acc	Pre1	Pre2	Pre3	Mac_P	Rec1	Rec2	Rec3	Mac_R	Spe1	Spe2	Spe3	F1-mea1	F1-mea2	F1-mea3	Mac_F1	Var
HOG	ANNs	68.6	71.9	20.2	35.1	42.4	94.4	6.3	3.5	34.7	5.3	82.3	76.5	81.6	9.6	6.4	38.2	0.51
	RFs	71.2	72.6	43.5	0.0	38.7	97.5	10.7	0.0	36.1	6.7	85.0	79.9	83.2	17.2	0.0	37.4	0.53
	RBF-SVM	71.0	71.0	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.7	79.7	83.0	0.0	0.0	27.8	0.58
	Linear-SVM	71.0	71.0	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.7	79.7	83.0	0.0	0.0	27.8	0.58
GLCM	ANNs	69.7	75.9	36.9	32.2	48.3	90.0	30.6	25.3	48.6	20.0	78.3	77.9	82.4	33.5	28.3	48.4	0.36
	RFs	68.9	78.5	31.4	31.6	47.2	88.0	24.0	19.1	43.7	22.3	78.8	75.0	83.0	27.2	23.8	45.4	0.38
	RBF-SVM	18.0	0.0	18.0	0.0	6.0	0.0	100.0	0.0	33.3	62.3	0.0	20.3	0.0	30.5	0.0	10.2	0.58
	Linear-SVM	18.0	0.0	18.0	0.0	6.0	0.0	100.0	0.0	33.3	62.3	0.0	20.3	0.0	30.5	0.0	10.2	0.58
HU	ANNs	71.0	71.0	50.0	0.0	40.3	99.98	0.08	0.0	33.4	0.05	86.7	79.7	83.0	0.6	0.0	36.5	0.58
	RFs	68.6	70.7	15.3	20.0	35.3	95.8	25.8	13.3	45.0	2.1	83.2	76.9	81.4	4.4	16.0	39.6	0.45
	RBF-SVM	71.0	71.0	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.7	79.7	83.0	0.0	0.0	27.7	0.58
	Linear-SVM	10.6	0.0	0.0	10.7	3.6	0.0	0.0	97.3	32.4	36.7	13.0	0.0	0.0	0.0	19.3	6.5	0.56
SIFT	ANNs	70.8	71.0	20.0	0.0	30.3	99.7	0.3	0.0	33.3	0.2	86.4	79.5	82.9	0.6	0.0	31.7	0.58
	RFs	69.3	71.4	13.8	33.3	39.5	96.9	2.7	0.1	33.2	1.8	84.0	77.8	82.2	4.5	0.2	36.0	0.55
	RBF-SVM	71.0	71.0	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.7	79.7	83.0	0.0	0.0	27.7	0.58
	Linear-SVM	71.0	71.0	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.7	79.7	83.0	0.0	0.0	27.7	0.58
GRAY	ANNs	69.0	71.6	21.8	9.5	34.3	95.7	5.7	0.3	33.9	3.8	83.0	77.4	81.9	9.0	0.6	34.1	0.54
	RFs	71.0	71.3	35.4	40.0	48.9	99.4	1.4	1.1	34.0	1.3	86.3	79.5	83.0	2.7	2.1	40.1	0.57
	RBF-SVM	71.0	71.0	0.0	0.0	23.7	99.4	0.0	0.0	33.1	0.0	86.1	79.3	82.9	0.0	0.0	27.6	0.57
	Linear-SVM	71.0	71.0	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.7	79.7	83.0	0.0	0.0	27.7	0.57
$F^{X}_{(i,g)}$	ANNs	77.1	81.2	50.7	64.7	65.5	96.5	29.8	29.3	51.9	29.7	87.6	83.0	88.2	37.5	40.3	57.9	0.39
	RFs	76.4	84.3	46.3	47.7	59.4	93.6	29.8	41.3	54.9	34.2	86.7	80.7	88.7	36.3	44.3	7.0	0.34
	RBF-SVM	73.2	74.2	52.9	100.0	75.7	99.0	14.5	2.7	38.7	10.0	86.4	81.9	84.8	22.8	5.3	51.2	0.52
	Linear-SVM	76.7	79.8	51.9	61.1	64.3	98.0	21.8	29.3	49.7	24.6	88.8	82.5	88.0	30.7	39.6	56.0	0.42
$F^{Y}_{(i,g)}$	ANNs	82.7	94.9	53.6	64.3	70.9	90.1	65.3	60.0	71.8	63.3	86.5	85.5	92.4	58.9	62.1	71.3	0.16
	RFs	83.0	91.7	58.0	63.9	71.2	93.2	55.6	61.3	70.0	57.8	89.0	85.6	92.4	56.8	62.6	70.1	0.20
	RBF-SVM	73.8	74.4	0.0	62.5	45.6	99.8	0.0	26.7	42.2	10.0	90.0	79.6	85.2	0.0	37.4	43.8	0.52
	Linear-SVM	77.7	95.2	28.6	37.5	53.8	93.6	1.6	100.0	65.1	38.7	94.5	75.0	94.4	3.0	54.5	58.9	0.55
$F^{Z}_{(i,g)}$	ANNs	91.8	98.7	73.0	86.8	86.2	93.5	87.1	89.2	89.9	87.9	92.9	92.3	96.0	79.4	88.0	88.0	0.04
	RFs	85.7	92.8	62.7	83.8	80.0	95.0	75.8	41.3	70.7	62.8	87.9	91.2	93.4	68.6	55.3	75.0	0.27
	RBF-SVM	79.5	90.9	62.5	43.5	65.6	96.5	4.0	93.3	64.6	37.7	96.1	77.8	93.6	7.5	59.3	65.1	0.53
	Linear-SVM	80.1	98.3	51.5	63.6	71.1	82.6	83.1	65.3	77.0	76.4	80.3	82.7	89.8	63.6	64.4	73.9	0.10

images, in which there is few difference between stationary sperms and moving sperms, therefore it is difficult to distinguish different categories of sperms by static features. Because there are not many differences between static sperms and moving sperms, it is easy to miss-classify all sperms into one category by using static features to classify sperm

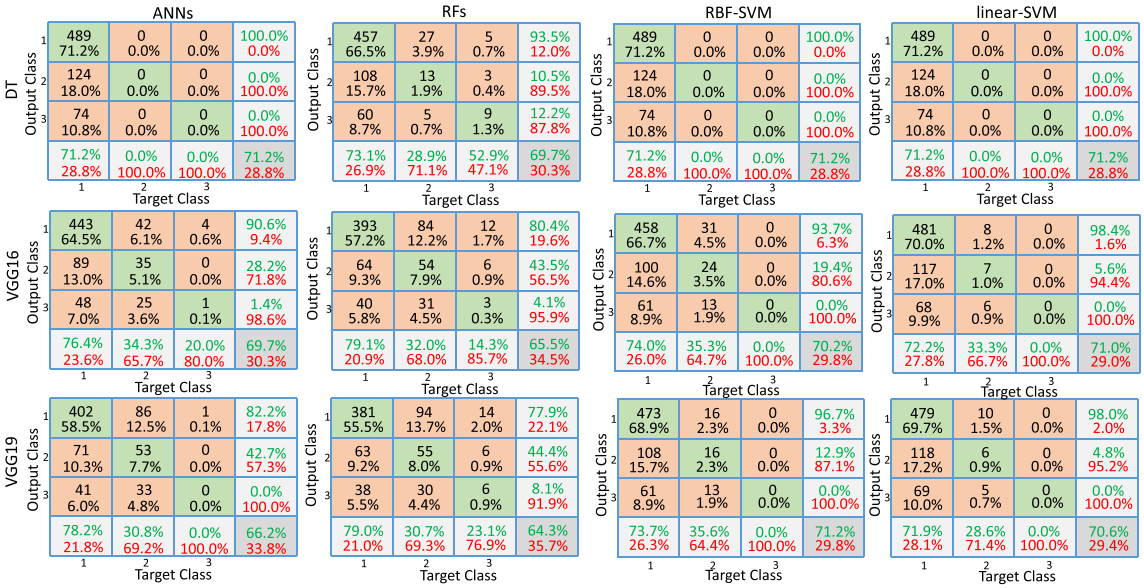

FIGURE 21. The confusion matrices of dynamic features. Rows represent feature types, columns represent classifier types, and DT represents dynamic texture features.

in different motion states, consequently, one precision (precision 1, precision 2 and precision 3) for one static feature is very high and the others are very low. The case for recall values are totally similar to that of the precision. The difference of values between precision 1, precision 2 and precision 3 further affect the macro of precision, recall and F1-measure.

$F^{X}_{(i,g)}$, $F^{Y}_{(i,g)}$ and $F^{Z}_{(i,g)}$ can distinguish three categories well by the advantage of foldover in classification, especially the information of foldover in the Z direction is very beneficial to distinguish sperms in different motion states. Therefore, the value of precision and recall are higher than which of the static features. Furthermore, foldover features perform well in the macro of precision, recall and F1-measure.

3) COMPARISON WITH DYNAMIC FEATURES

Three dynamic features are selected for the comparative experiment, including dynamic texture features and features extracted based on the CNNs (VGG-16 and VGG-19 networks). The first step is the same as the operation of static features, where each sperm is detected to a size of 26 by 26 pixels in the corresponding frame. The difference is what we need is the entire movement of the detected sperm. Therefore, the detected sperm images are combined into a video of the corresponding sperm. Secondly, we refer to the articles [14], [21], [22] to extract dynamic texture and deep learning (VGG-16 and VGG-19 networks) features in the detected semen videos. The third step is the same as the operation of static features, where we use ANNs [40], RFs [42] and SVM [43] (Linear- and RBF-SVM) classifiers to distinguish dynamic features, and the classification results are shown in FIGURE 21.

According to the classification results of dynamic features in FIGURE 21, we compare evaluations between dynamic features and foldover features in TABLE 6.

Considering the comparison in TABLE 6, the accuracy of $F^{X}_{(i,g)}$, $F^{Y}_{(i,g)}$ and $F^{Z}_{(i,g)}$ are significantly higher than that of dynamic features. The reason for the low accuracy of dynamic features is: Sperms are tiny and there is very little dynamic information. Therefore, it is difficult to distinguish different categories of sperms by dynamic features. It is easy to classify most of sperms into one category by using dynamic features to classify sperm in different motion states, consequently, the value of the true positive (TP) further affect the calculation results of all evaluations.

4) ADDITIONAL EXPERIMENT PART A: COMPARISON WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

Currently, deep convolutional neural networks (DCNNs) are applied successfully to various applications, in which depth plays a major factor in increasing efficiency of the network. Especially, in the field of image classification, DCNNs has excellent advantages [47]. Therefore, we compare two well-known DCNNs (VGG-16 and VGG-19 networks) in the classification task of 1374 sperms, and the experimental results are shown in FIGURE 22.

There are several reasons for the poor results of DCNNs: (1) The high similarity of different sperms makes it difficult to extract effective features for DCNNs classification. (2) DCNNs are difficult to operate on the selection of visual information. In the process of deep convolution, DCNNs discard some visual information judged as redundant, which is terrible for sperms with little visual information. (3) Due to the lack of sperm visual information, the depth of DCNNs is required to be relatively high. A larger depth may cause

TABLE 6. Evaluation of dynamic features with four classifiers. The first column shows the types of dynamic features, the second column shows the types of classifiers, the third to the last columns show the calculated evaluations. We use the first three letters of each evaluation to indicate the evaluation metric, such as Acc is accuracy, Pre is precision, Mac_P is Macro_P, Rec is recall, Mac_R is Macro_R, Spe is specificity, F1-mea1 is F1-measure, Mac_F1 is Macro_F1 and Var is variance. The red font value means that the value is the maximum value in the column (Unit: %).

Feature	Classifier	Acc	Pre1	Pre2	Pre3	Mac_P	Rec1	Rec2	Rec3	Mac_R	Spe1	Spe2	Spe3	F1-mea1	F1-mea2	F1-mea3	Mac_F1	Var
DT	ANNs	71.2	71.2	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.9	79.8	83.2	0.0	0.0	27.7	0.58
	RFs	69.7	73.1	28.9	52.9	51.6	93.5	10.5	12.2	38.7	10.0	82.3	76.7	82.1	15.4	19.8	44.2	0.48
	RBF-SVM	71.2	71.2	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.9	79.8	83.2	0.0	0.0	27.7	0.58
	Linear-SVM	71.2	71.2	0.0	0.0	23.7	100.0	0.0	0.0	33.3	0.0	86.9	79.8	83.2	0.0	0.0	27.7	0.58
VGG16	ANNs	69.7	76.4	34.3	20.0	43.6	90.6	28.2	1.4	40.1	18.2	78.9	78.0	82.9	31.0	2.6	41.8	0.46
	RFs	65.5	79.1	32.0	14.3	41.8	80.4	43.5	4.1	42.7	28.8	70.3	72.9	79.7	36.9	6.4	42.2	0.38
	RBF-SVM	70.2	74.0	35.3	0.0	36.4	93.7	19.4	0.0	37.7	12.1	81.4	78.6	82.7	25.0	0.0	37.0	0.50
	Linear-SVM	71.0	72.2	33.3	0.0	35.2	98.4	5.6	0.0	34.7	35.4	76.1	79.6	83.3	9.6	0.0	34.9	0.55
VGG19	ANNs	66.2	78.2	30.8	0.0	36.3	82.2	42.7	0.0	41.5	26.8	71.4	74.2	80.2	35.8	0.0	38.7	0.41
	RFs	64.3	79.0	30.7	23.1	44.3	77.9	44.4	8.1	43.5	30.8	68.7	71.1	78.4	36.3	12.0	43.9	0.35
	RBF-SVM	71.2	73.7	35.6	0.0	36.4	96.7	12.9	0.0	36.5	8.1	84.0	79.8	83.6	18.9	0.0	36.4	0.53
	Linear-SVM	70.6	71.9	28.6	0.0	33.5	98.0	4.8	0.0	34.3	3.0	85.1	79.1	82.9	8.2	0.0	33.9	0.55
$F^{X}_{(i,g)}$	ANNs	77.1	81.2	50.7	64.7	65.5	96.5	29.8	29.3	51.9	29.7	87.6	83.0	88.2	37.5	40.3	57.9	0.39
	RFs	76.4	84.3	46.3	47.7	59.4	93.6	29.8	41.3	54.9	34.2	86.7	80.7	88.7	36.3	44.3	7.0	0.34
	RBF-SVM	73.2	74.2	52.9	100.0	75.7	99.0	14.5	2.7	38.7	10.0	86.4	81.9	84.8	22.8	5.3	51.2	0.52
	Linear-SVM	76.7	79.8	51.9	61.1	64.3	98.0	21.8	29.3	49.7	24.6	88.8	82.5	88.0	30.7	39.6	56.0	0.42
$F^{Y}_{(i,g)}$	ANNs	82.7	94.9	53.6	64.3	70.9	90.1	65.3	60.0	71.8	63.3	86.5	85.5	92.4	58.9	62.1	71.3	0.16
	RFs	83.0	91.7	58.0	63.9	71.2	93.2	55.6	61.3	70.0	57.8	89.0	85.6	92.4	56.8	62.6	70.1	0.20
	RBF-SVM	73.8	74.4	0.0	62.5	45.6	99.8	0.0	26.7	42.2	10.0	90.0	79.6	85.2	0.0	37.4	43.8	0.52
	Linear-SVM	77.7	95.2	28.6	37.5	53.8	93.6	1.6	100.0	65.1	38.7	94.5	75.0	94.4	3.0	54.5	58.9	0.55
$F^{Z}_{(i,g)}$	ANNs	91.8	98.7	73.0	86.8	86.2	93.5	87.1	89.2	89.9	87.9	92.9	92.3	96.0	79.4	88.0	88.0	0.04
	RFs	85.7	92.8	62.7	83.8	80.0	95.0	75.8	41.3	70.7	62.8	87.9	91.2	93.4	68.6	55.3	75.0	0.27
	RBF-SVM	79.5	90.9	62.5	43.5	65.6	96.5	4.0	93.3	64.6	37.7	96.1	77.8	93.6	7.5	59.3	65.1	0.53
	Linear-SVM	80.1	98.3	51.5	63.6	71.1	82.6	83.1	65.3	77.0	76.4	80.3	82.7	89.8	63.6	64.4	73.9	0.10

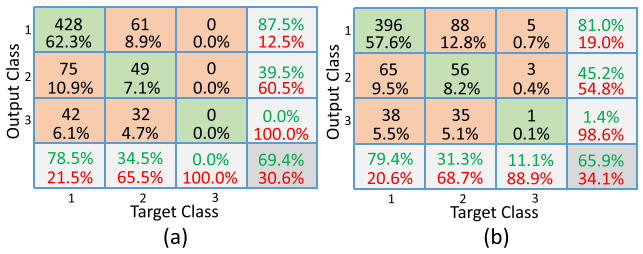

FIGURE 22. The sperm classification results using two DCNNs. (a) is the result of VGG-16 and (b) is the result of VGG-19.

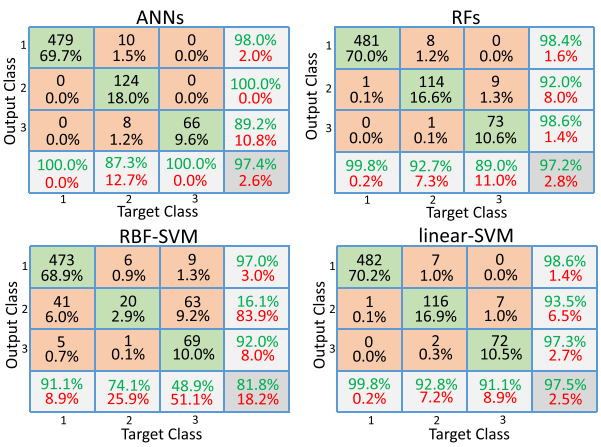

FIGURE 23. The sperm classification results using early fusion.

sperms to have no visual information to extract, while a smaller depth may cause the extracted features to have no differentiation.

5) ADDITIONAL EXPERIMENT PART B: FOLDOVER FEATURE FUSION

In order to enhance the discriminative ability of features, one important method is feature fusion, including early fusion and late fusion [48]. Early fusion is defined as the integrates unimodal features before learning concepts, and late fusion is defined as that first reduces unimodal features to separately learned concept scores, then these scores are integrated to learn concepts. Especially, because early fusion is easy to operate and requires only one learning phase, it is wildly used in video analysis tasks [48]. Hence, for foldover features $F^{X}_{(i,g)}$, $F^{Y}_{(i,g)}$ and $F^{Z}_{(i,g)}$, we adopt the early fusion method for feature fusion. We classify the foldover features of 1374 sperms by the results of early fusion, and the experimental results are shown in FIGURE 23.

According to FIGURE 23, the results on the four classifiers are excellent after the early fusion of the foldover features ($F^{X}_{(i,g)}$, $F^{Y}_{(i,g)}$ and $F^{Z}_{(i,g)}$). After early fusion, the accuracy of the three classifiers (ANNs, RFs and linear-SVM) reaches more than 97%, nearly 6% higher than the highest accuracy of 91.8% without early fusion in FIGURE 18. The accuracy of the RBF-SVM is 81.8% higher than that of the RBF-SVM in FIGURE 18 (79.5%). The recall of the three classifiers (ANNs, RFs and linear-SVM) is excellent, most of them above 90%, among which the highest reaches 100% in ANNs, and the recall of the RBF-SVM is also much better than that of the RBF-SVM in FIGURE 18.

6) EXPERIMENTAL ANALYSIS

There are two main reasons why the classification results of foldover features are superior to classical static and dynamic features. First, there is a high degree of similarity between different sperms. When two sperms are very similar in shape, size, color and texture, static features cannot effectively distinguish two sperms. However, the foldover features can solve this problem well, because the differences of foldovers between two sperms are very obvious as the example shown in FIGURE 24.

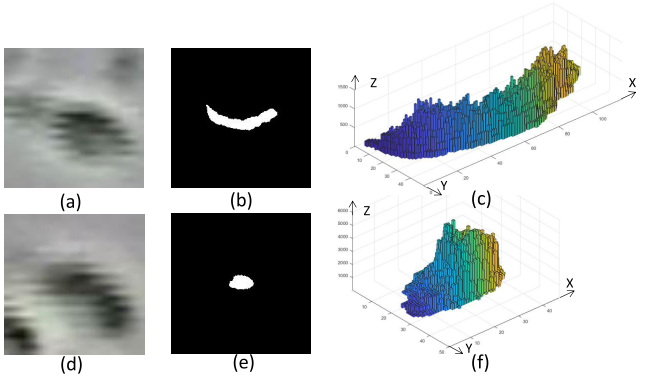

FIGURE 24. An example of different sperms. (a) and (d) represent two different sperms, (b) is the foldover of (a) in the Z direction, (e) is the foldover of (d) in the Z direction, (c) is the foldover of (a) in the 3D visualization, and (f) is the foldover of (d) in the 3D visualization.

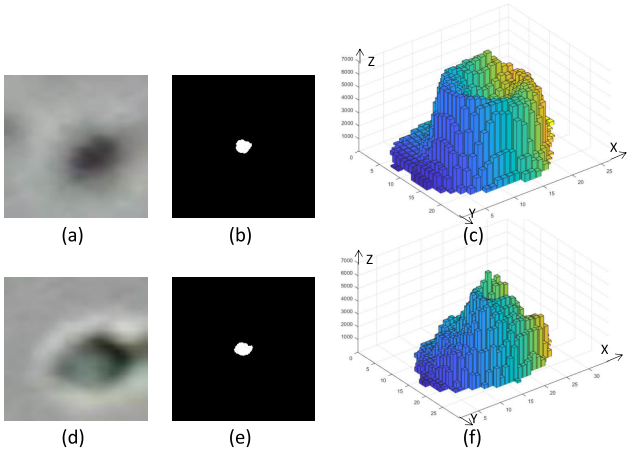

FIGURE 25. Different features of two sperms. (a) and (d) are two different sperms, (b) and (e) are 2D visualization of the two foldovers in the Z direction, and (c) and (f) are 3D visualization of the two foldovers.

According to FIGURE 24, we can find that because the sperms in FIGURE 24 (a) and (d) are very similar in shape, size and color, it is difficult to distinguish them by static features. However, the differences between FIGURE 24 (a) and (d) on the foldovers are very obvious.

Second, because sperms are very tiny and there is very little visual information, it is very difficult to distinguish two different sperms. However, due to the foldover features contain not only the original shape and texture information of sperms, but also the movement information of sperms, they can discover more useful visual information. Furthermore, we analyse the X, Y and Z directions of the sperm foldovers, and expand the sperms movement information, the 3D visualizations of two sperms are shown in FIGURE 24 (c) and (f).

According to FIGURE 24 (c) and (f), although FIGURE 24 (a) and (d) contain little visual information, the information contained in their foldovers is abundant, and the differences between the foldovers are obvious. In addition, even the static and dynamic features of two sperms are very similar, foldover features contain a lot of visual information to distinguish the sperms as shown in FIGURE 25.

The two sperms in FIGURE 25 (a) and (d) are very similar in shape, color, size and texture. In FIGURE 25 (b) and (e), there is a high similarity between the two sperms motility states. However, according to the FIGURE 25 (c) and (f), when both static and dynamic features are similar, the information contained in the foldovers is significantly different. It proves that the foldover features are superior in distinguishing tiny objects, similar objects, objects with little visual information and objects with similar visual information.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose novel foldover features, which are applied to dynamic object behavior description in microscopic videos. Compared with classical static and dynamic features, the foldover features show obvious advantages in distinguishing tiny objects, similar objects, objects with little visual information and objects with similar visual information. In the experiment, we use four different classifiers (ANN, RF, linear-SVM and RBF-SVM) to test the effectiveness of the foldover features, and an overall outstanding classification accuracy is obtained, indicating the effectiveness and potential of the proposed foldover features.

In the future, we plan to increase the amount of data in a single category, allowing the same doctors to expand the data and address the imbalance in our experimental data. Then, although we have tested the foldover features on the semen microscopic videos, we will test it on more highly similar objects to improve the generalization of the foldover features.

ACKNOWLEDGMENT

The authors would like to thank M. D. Peng Xu and M. D. Hong Yan from Dongfang Jinghua Hospital, Shenyang, China, for their great data preparation work. They would also like to thank Miss Zixian Li and Mr. Guoxian Li for their important discussion. They would also like to thank the 2001 Hongkong movie " Mist in Judge" and 2013 American TV Season "Agents of S.H.I.E.L.D.," which brought the inspiration of "foldover" to us in this research work.

REFERENCES

[1] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 996–1011, Jun. 1996.
[2] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction*. Berlin, Germany: Springer, 2006.
[3] A. Hadid, "Analyzing facial behavioral features from videos," in *Proc. Int. Workshop Hum. Behav. Understand.* Amsterdam, The Netherlands: Springer, 2011, pp. 52–61.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, no. 2, pp. 1097–1105.
[5] F. Long, H. Zhang, and D. Feng, "Fundamentals of content-based image retrieval," in *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*. Berlin, Germany: Springer, 2003, pp. 1–26, doi: 10.1007/978-3-662-05300-3_1.
[6] R. Brunelli and O. Mich, "On the use of histograms for image retrieval," in *Proc. IEEE Int. Conf. Multimedia Comput. Syst. (ICMCS)*, Jun. 2002, pp. 143–147.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[9] S. Park, B. Kim, J. Lee, J. Mo Goo, and Y.-G. Shin, "GGO nodule volume-preserving nonrigid lung registration using GLCM texture analysis," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 10, pp. 2885–2894, Oct. 2011.

[10] M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Trans. Inf. Theory*, vol. IT-8, no. 2, pp. 179–187, Feb. 1962.

[11] S. Giannarou and T. Stathaki, "Shape signature matching for object identification invariant to image transformations and occlusion," in *Computer Analysis of Images and Patterns*, W. Kropatsch, M. Kampel, and A. Hanbury, Eds. Berlin, Germany: Springer, 2007, pp. 710–717.

[12] E. N. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 184–190.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," vol. 3, 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[14] S. Soatto, G. Doretto, and N. Ying, "Dynamic textures," in *Proc. ICCV*, vol. 2, 2001, pp. 439–446.

[15] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 2, pp. 284–299, 1985.

[16] K. G. Derpa and R. P. Wildes, "Dynamic texture recognition based on distributions of spacetime oriented structure," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 191–198.

[17] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 2, 2004, pp. 28–31.

[18] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[19] L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam, "Automatic tracking and motility analysis of human sperm in time-lapse images," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 792–801, Mar. 2017.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[21] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2016, pp. 3119–3127.

[22] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[23] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020.

[24] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8543–8553.

[25] G. Jeon, "Denoising in contrast-enhanced X-ray images," *Sens. Imag.*, vol. 17, no. 1, p. 14, Dec. 2016.

[26] G. Jeon, "Computational intelligence approach for medical images by suppressing noise," *J. Ambient Intell. Humanized Comput.*, vol. 8, pp. 1–11, Nov. 2017.

[27] L. Leng, M. Li, C. Kim, and X. Bi, "Dual-source discrimination power analysis for multi-instance contactless palmprint recognition," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 333–354, Jan. 2017.

[28] L. Leng and J. Zhang, "Palmhash code vs. palmphasor code," *Neurocomputing*, vol. 108, no. 2, pp. 1–12, May 2013.

[29] L. Leng, J. Zhang, J. Xu, M. K. Khan, and K. Alghathbar, "Dynamic weighted discrimination power analysis in DCT domain for face and palmprint recognition," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, vol. 5, Nov. 2010, pp. 467–471.

[30] L. Leng, M. Li, L. Leng, and A. B. J. Teoh, "Conjugate 2DPalmHash code for secure palm-print-vein verification," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, Dec. 2013, pp. 1705–1710.

[31] Y. Yuan, J. Chu, L. Leng, J. Miao, and B.-G. Kim, "A scale-adaptive object-tracking algorithm with occlusion detection," *EURASIP J. Image Video Process.*, vol. 2020, no. 1, pp. 1–15, Dec. 2020.

[32] J. Chu, X. Tu, L. Leng, and J. Miao, "Double-channel object tracking with position deviation suppression," *IEEE Access*, vol. 8, pp. 856–866, 2020.

[33] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1232–1241.

[34] W. Qiu, X. Gao, and B. Han, "Saliency detection using a deep conditional random field network," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107266.

[35] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.

[36] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Saliency driven image manipulation," *Mach. Vis. Appl.*, vol. 30, no. 2, pp. 189–202, Mar. 2019.

[37] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[38] T. Nguyen, M. Dax, K. Mummadi, N. Ngo, P. Nguyen, Z. Lou, and T. Brox, "Deepusps: Deep robust unsupervised saliency prediction via self-supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds. Montreal, QC, Canada: Curran Associates, 2019, pp. 204–214.

[39] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 1, pp. 249–268, 2007.

[40] E. Judith and J. Deleo, "Artificial neural networks," *Cancer*, vol. 91, no. S8, pp. 1615–1635, 2001.

[41] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 161–168.

[42] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1, 1995, pp. 278–282.

[43] C. Saunders, M. Stitson, J. Weston, R. Holloway, L. Bottou, B. Scholkopf, and A. Smola, "Support vector machine," *Comput. Sci.*, vol. 1, no. 4, pp. 1–28, 2002.

[44] R. P. Amann and D. Waberski, "Computer-assisted sperm analysis (CASA): Capabilities and potential developments," *Theriogenology*, vol. 81, no. 1, pp. 5.e3–17.e3, Jan. 2014.

[45] M. Goldstein, "k_n-nearest neighbor classification," *IEEE Trans. Inf. Theory*, vol. 18, no. 5, pp. 627–630, Sep. 2003.

[46] D. Lamb, "World health organization laboratory manual for the examination of human semen and sperm-cervical mucus interaction, 4th ed," *J. Androl.*, vol. 21, no. 1, p. 32, 2000.

[47] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[48] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia (MULTIMEDIA)*, 2005, pp. 399–402.

XIALIN LI received the B.E. degree from the Hebei Normal University of Science and Technology, China, in 2017.He is curently pursuing the master's degree with the Research Group for Microscopic Image and Medical Image Analysis, College of Medicine and Biological Information Engineering, Northeastern University, China. His research interests include microscopic video analysis, object detection, and classification.

CHEN LI received the B.E. degree from the University of Science and Technology Beijing, China, in 2008, the M.Sc. degree from Northeast Normal University, China, in 2011, and the Dr.-Ing. degree from the University of Siegen, Germany, in 2016. From 2016 to 2017, he worked as a Postdoctoral Researcher with the Johannes Gutenberg University of Mainz, Germany. He is currently working as an Associate Professor with Northeastern University, China. He is also the Head of the Research Group for Microscopic Image and Medical Image Analysis, College of Medicine and Biological Information Engineering, Northeastern University. His research interests include microscopic image analysis, medical image analysis, machine learning, pattern recognition, machine vision, and multimedia retrieval. He is a reviewer for several journals and conferences, including *Pattern Recognition*, *Future Generation Computer Systems*, *Artificial Intelligence in Medicine*, *Chemometrics and Intelligent Laboratory Systems*, IEEE Access, *Neurocomputing*, *Journal of X-Ray Science and Technology*, AAAI-20, and ITIB-2020.

FRANK KULWA was born in Tanzania, in 1986. He received the B.E. degree from the Dar es salaam Institute of Technology, Tanzania, in 2013. He is currently pursuing the master's degree with the Research Group for Microscopic Image and Medical Image Analysis, Northeastern University, China. Since 2017, he has been working as a Tutorial Assistant with the Dar es salaam Institute of Technology. His research interests include microscopic image analysis and deep learning.

MD MAMUNUR RAHAMAN received the B.Sc. degree from BRAC University, Dhaka, Bangladesh, in 2017. He is currently pursuing the master's degree with the Research Group for Microscopic Image and Medical Image Analysis, Biomedical and Information Engineering School, Northeastern University. His research interests include microscopic image analysis, medical image analysis, machine learning, pattern recognition, and machine vision.

WENWEI ZHAO was born in 1997. He received the B.E. degree from Dalian Minzu University, China, in 2018. He is currently pursuing the master's degree with the Research Group for Microscopic Image and Medical Image Analysis, Northeastern University, China. His research interests include microscopic video analysis, object detection, and classification.

XUE WANG received the bachelor's and master's degrees from Northeastern University, China, in 2014 and 2017, respectively. She is currently an Engineer with Northeastern University. Her research interest includes microscopic machine design.

DAN XUE received the B.E. degree from Shenyang Ligong University, China, in 2017, and the M.E. degree from Northeastern University, China, in 2019. She is currently working at NeuSoft Medical Company as an Algorithm Engineer and also an External Researcher with the Research Group for Microscopic Image and Medical Image Analysis, Northeastern University. Her research interests include microscopic image analysis, medical image analysis, machine learning, pattern recognition, and machine vision. She is a reviewer of IEEE Access.

YUDONG YAO (Fellow, IEEE) received the B.Eng. and M.Eng. degrees in electrical engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 1988. From 1987 to 1988, he was a Visiting Student with Carleton University, Ottawa, ON, Canada. From 1989 to 2000, he was with Carleton University, Spar Aerospace Ltd., Montreal, QC, Canada, and Qualcomm Inc., San Diego, CA, USA. Since 2000, he has been with the Stevens Institute of Technology, Hoboken, NJ, USA, where he is currently a Professor and the Chair of the Department of Electrical and Computer Engineering. He holds one Chinese patent and 13 U.S. patents. His research interests include wireless communications, machine learning and deep learning techniques, and healthcare and medical applications. For his contributions to wireless communications systems, he was elected as a Fellow of the National Academy of Inventors, in 2015, and the Canadian Academy of Engineering, in 2017. He has served as an Associate Editor for the IEEE Communications Letters, from 2000 to 2008, and the IEEE Transactions on Vehicular Technology, from 2001 to 2006. He has served as an Editor for the IEEE Transactions on Wireless Communications, from 2001 to 2005.

YILIN CHENG was born in 1996. He received the B.Eng. degree from Northeastern University, China, in 2019. He is currently an External Researcher with the Research Group for Microscopic Image and Medical Image Analysis, Northeastern University. His research interests include microscopic image analysis and medical image analysis.

JINDONG LI was born in 1998. He is currently pursuing the bachelor's degree with Northeastern University, China. He is an Experimental Assistant with the Research Group for Microscopic Image and Medical Image Analysis, Northeastern University. His research interests include microscopic image analysis and medical image analysis.

SHOULIANG QI (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, in 2007. He is currently an Associate Professor with the Sino-Dutch Biomedical and Information Engineering School, Northeastern University, China. He joined the GE Global Research Center, where he was responsible for designing innovative magnetic resonance imaging (MRI) system. From 2014 to 2015, he was a Visiting Scholar with the Eindhoven University of Technology and the Kempenhaeghe Epilepsy Center, The Netherlands. In the recent years, he has been conducting productive studies in intelligent medical imaging computing and modeling, machine learning, brain networks, and brain models. He has published more than 80 papers in peer-reviewed journals and international conferences. He has won many academic awards, such as the Chinese Excellent Ph.D. Dissertation Nomination Award and the Award for Outstanding Achievement in Scientific Research from the Ministry of Education.

TAO JIANG was born in 1975. He received the Ph.D. degree from the University of Siegen, Germany, in 2013. He is currently a Professor with the Chengdu University of Information Technology (CUIT), China, where he is also the Dean of the Control Engineering College. His research interests include machine vision, artificial intelligence, robot control, self-driving auto, and membrane computing.