IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 December 2023, accepted 20 December 2023, date of publication 2 January 2024, date of current version 10 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3349148

RESEARCH ARTICLE

Total Variation PCA-Based Descriptors for Electrocardiography Identity Recognition

HAIYING LIU[1], HAIYAN LIN[2], AND XIANHUI WANG[1]
[1]School of Information Science and Engineering, Xinjiang University, Ürümqi 830046, China
[2]Computer Department, Xinjiang Qitai Secondary Vocational and Technical School, Qitai 831800, China

Corresponding author: Haiying Liu (1466432554@qq.com)

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2022D01C428, in part by the NSFC-Xinjiang Joint Funding under Grant U1903127, and in part by the NSFC Fund under Grant 61961039.

ABSTRACT Electrocardiographic (ECG) signals have been successfully used in biometric recognition. However, the accuracy of ECG-based biometric systems is generally lower than systems based on other physiological traits. This study introduces a local feature learning method aimed at enhancing the performance of ECG-based biometric recognition systems. Specifically, we first extracted the multi-scale differential feature (MDF) for each point in the training ECG heartbeats using the difference between each point and its neighboring points. Second, we learn feature mapping to project these MDFs into low-dimensional descriptors in an unsupervised manner, where 1) the errors between the original MDF and reconstructed MDF are minimized. 2) The total variation in the reconstructed MDFs is minimized. Third, we represented each ECG heartbeat as a histogram feature using clustering and pooling descriptors. Finally, we adopted global feature learning methods to obtain a representation of an ECG heartbeat. Experiments on the MIT-BIH Arrhythmia, ECG-ID, and Physikalisch Technische Bundesanstalt databases verified the performance of the proposed method over existing ECG biometric recognition methods using within-session analysis. Moreover, we evaluated the performance of the proposed method using an across-session analysis of the ECG-ID database.

INDEX TERMS Bag-of-words, ECG biometrics recognition, feature learning, total variation.

I. INTRODUCTION

Electrocardiogram (ECG) is an electrical signal reflecting cardiac activity, which is a weakly nonstationary and quasi-periodic signal. ECG signals, noninvasive and closely linked to human health conditions, are adopted in a variety of applications. This includes diagnosing cardiac-related diseases, detecting sleep apnea, monitoring driver drowsiness, estimating heart rate [26], and measuring blood pressure. Recently, person identity recognition using ECG signals has become a new topic in biometrics owing to the following aspects: 1) ECG signals meet all the requirements of biometrics, such as universality, uniqueness, collectability, acceptability, and permanence [7], [8]. 2) Robustness against circumvention and spoofing attacks. This lies in two facts. First, ECG signals can only be captured by human beings; thus, real-time aliveness feedback is automatically provided while capturing the ECG signals. Second, ECG signals are based on physiological characteristics inherent to the human body, making them difficult to alter or replicate. 3) Continuous authentication can be provided by an ECG signal because the ECG signal is quasiperiodic and is recorded by sensors attached to the body [11]. Therefore, person identity recognition based on ECG signals is a potential direction in biometrics.

However, using ECG signals in biometrics can be challenging. 1) The quality of the ECG signal, when sampled, is affected by several types of noise, such as power line interference, electromyogram artifacts, muscle interference, and electrode motion artifacts [25]. 2) The intra-variance of the ECG signal is large. In other words, records from the same person at different times may differ significantly.

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

This is because the waveform of the measured signal during capture is affected by numerous factors, such as the health status, physical exercise, emotion, diet, and posture of the person. Many efforts have been made to overcome these challenges from different perspectives, such as data acquisition, denoising, detection of fiducial features, feature extraction, and matching. Among them, feature extraction significantly affects the performance of the ECG identity recognition system, and many distinctive features, such as fiducial features [15], [16], [17], one-dimensional local binary patterns(1DLBP) [9], [10], [11], autocorrelation coefficients [29], [30], and discrete wavelet transforms [31], have been proposed.

Among the above methods, the 1DLBP has been successfully applied to ECG identity recognition owing to its simplicity, efficiency, and effectiveness. Two operations were required to construct the 1DLBP. One computes the difference between the values of the current point and its neighbors to obtain a difference feature (DF). The other was to threshold the DF to a fixed value, finally forming a 1DLBP at this point. The merit of the 1DLBP is that it tolerates noise and preserves the morphology of ECG waveforms [9]. However, 1DLBP has three limitations. 1) The extracted code length is limited. This was because the dimensions of the 1DLBP were equal to the length of the DF. Although rich information from ECG signals can be obtained when long codes are generated, the computational cost and storage space increase. 2) Information redundancy may exist in 1DLBP. This is because dimension reduction is not involved when the DFs are binarized. 3) Information on the amplitude of the DF is lost when a fixed value is used to transform the DF into a 1DLBP. Moreover, these shortcomings are also present in other local binary code methods, such as one-dimensional multiresolution local binary patterns [11], [45] and one-dimensional local difference patterns [10].

To overcome the above problems of handcrafted local features, several local feature learning methods [47], [48], [49] have been proposed for other recognition domains(e.g., face recognition and finger vein recognition), and satisfactory performance has been achieved. However, when we attempt to use these local feature-learning methods for ECG identity recognition, the recognition rates are considerably low. This is because the ECG signal is a time-series signal, whereas local feature-learning methods are designed for images. Therefore, applying a local feature learning method to ECG signals is the main task.

Additionally, after learning a map using these local feature learning methods, we constructed a histogram of the sample using a bag-of-words framework, and a global representation of the sample was obtained through whitened principal component analysis (WPCA) to reduce the dimensions of the histogram representation. This dimension reduction process does not use the label information of the training samples, and we assume that the global representation using these methods lacks discrimination. Fang et al. [46] proposed a regularized label relaxation linear regression method (RLRLR) for classification problems, which introduced a nonnegative label relaxation matrix into linear regression to fit the labels and construct a class graph based on manifold learning as a regularization item to avoid overfitting. Their experiments demonstrated that these novel ideas can enhance the performance of classification tasks.

Motivated by the above works, in this study, we propose a total variation principal component analysis (PCA)-based descriptor for ECG identity recognition, which we call TVPCAD because it incorporates PCA with total variation regularization when learning the local descriptor. Specifically, considering that the multi-scale differential features(MDF) are used successfully in the domain of ECG identity recognition [10], [11], [12], [13], [14], we first extract the MDF for each point of each ECG heartbeat in the training set as the original local feature. Second, to obtain compact features, we learned a map that projected these MDFs into descriptors. In the learning process, two main constraints are considered:1) the error between the original MDF and the reconstructed MDF is minimized, which is equivalent to the loss in PCA and is widely used to reduce the dimension of the features. 2) The total variation in the reconstructed signal is minimized to preserve the piecewise smoothness property, which is a data-denoising procedure that can improve the robustness of our algorithm against noise. Therefore, our local feature-learning method can extract a compact descriptor that preserves the principal components of MDF and can handle noise. Once a map is obtained, we can transform the MDF extracted from each point in the ECG heartbeat into a real descriptor.

After obtaining the descriptor for each point in the ECG heartbeat, and considering that the BoW framework has achieved promising results in ECG identity recognition [23], [24], we represent each ECG heartbeat as a histogram using the BoW framework.

However, these histogram representations have high dimensions and include a lot of redundancy, and we reduce the dimension and remove the redundancy using WPCA to obtain a global representation of the ECG heartbeat. In addition, the processes of local feature learning and reduction of histogram representation do not use the labels of the training samples; these global representations may be less discriminative. Therefore, we used the RLRLR method to fuse the label information and obtain the final representation of the ECG heartbeat.

Finally, experiments were conducted using three public ECG databases to verify the proposed method. In addition, we compared the performance of the proposed method with those of other ECG identification recognition methods.

II. BACKGROUND

In this section, we briefly summarize the necessary background of this study, including the ECG identity recognition methods, PCA, and total variation. These methods are most relevant to the proposed method.

A. ECG IDENTITY RECOGNITION

Depending on the features used, ECG identity recognition methods can be roughly grouped into two categories: fiducial and non-fiducial.

Fiducial methods first detect fiducial points and then extract features. Several studies have investigated this topic. Biel et al. [15] generated 20-dimensional features related to the onset, duration, amplitude, and morphology of the P, T, and QRS waves. Karimian et al. [16] extracted time-based and amplitude features from the detected fiducial points and adopted a hierarchical validation scheme for mobile environments. Palaniappan et al. [17] extracted the R-R interval, R amplitude, QRS interval, QR amplitude, and RS amplitude from the QRS segment and utilized a neural network as the classifier. Rezgui et al. [43] used interval and amplitude as biometric features. However, there are several problems associated with using fiducial methods for ECG identity recognition. First, it is difficult to detect fiducial points accurately because the quality of the ECG signal used in an identity recognition system is typically low. In general, when capturing an ECG signal, there is considerable noise from different sources, such as power line interference, electrode movement noise, muscle contraction noise, baseline drift, and physical noise [25]. In addition, the physiological and emotional variabilities of individuals when capturing ECG signals can affect their waves of ECG signals. Second, the fiducial features must be designed manually and require expertise in the ECG domain. To address the shortcomings of fiducial methods, non-fiducial methods for ECG identity recognition have been proposed.

Non-fiducial methods generally do not use fiducial points to generate features but typically need to detect the location of the R peak for heartbeat segmentation. Srivastva et al. [29] used the autocorrelation coefficient(AC) followed by one of three transformation techniques, that is, discrete cosine transform (DCT), discrete Fourier transform (DFT), and Walsh–Hadamard transform (WHT), and found that the DFT can achieve the best performance. Dar et al. [31] extracted wavelet coefficients using a discrete wavelet transform (DWT) and a single nearest neighbor classifier for person identification. Yu et al. [44] performed dimension reduction using PCA and optimized a neural network using resilient propagation(PRROP). Wang et al. [12] extracted multiscale differential features and used collective matrix factorization for ECG biometric recognition. Xu et al. [32] proposed a structural sparse representation algorithm, and learned a class-specific dictionary for ECG biometric recognition. Huang et al. [13] extracted multiple features and employed a unified sparse representation framework for ECG identity recognition. Li et al. [21] presented an ECG biometric method based on graph-regularized nonnegative matrix factorization and sparse representation.

Currently, deep learning methods are used for ECG identity recognition. Luz et al. [6] used convolutional neural networks (CNNs) to extract features from raw heartbeat signals and spectrograms. This work achieves state-of-the-art results in two public off-the-person databases. Chu et al. [54] proposed a multi-scale one-dimensional residual network for ECG biometrics. Labati et al. [3] presented Deep-ECG, which uses deep CNNs to extract features, produces real and binary templates for matching, and achieves better performance. Zhao et al. [4] transformed blind segments into two-dimensional images and used a CNN to learn discriminative features and representations for ECG biometric recognition. Abdeldayem and Bourlai [7] transformed the ECG segments into spectral correlation images and fed them into the CNN. Li et al. [5] first used a CNN to extract the features of ECG heartbeats and then input the extracted features into a second CNN for identification; strong generalization ability and significant performance were achieved. Srivastva et al. [50] proposed an ensemble of pre-trained deep neural networks (e.g., ResNet and DenseNet) for ECG biometric recognition. Kim and Pyun [51] proposed a real-time system using a bidirectional long short-term memory(LSTM)-based deep recurrent neural network through late fusion, achieving an accuracy of 99.8% accuracy in the MIT-BIH database for ECG biometric recognition. Alduwaile and Islam [53] input the time-frequency domain representation of a short segment of an ECG signal around the R-peak into a small CNN for biometric recognition and achieved better accuracy. Jyotishi and Dandapat [52] designed a hierarchical (HLSTM) model to capture the temporal variation of ECG signals at different abstractions and used the attention mechanism to identify ECG complexes with rich identification information. However, the heavy computational and large database requirements of deep-learning methods still need to be overcome.

B. PCA AND TOTAL VARIATION

PCA is one of the most widely used exploratory data analysis tools, and many pattern recognition tasks utilize it to reduce the dimensions of features, such as face recognition [33] and ECG biometric recognition [34]. Specifically, PCA determines a linear transformation to seek a projection that best represents the data in a least-squares manner, and can be formalized as (1).

$$\begin{aligned} &\min_{W} \ ||X - WW^TX||_F^2 \\ &s.t. \qquad W^TW = I, \end{aligned} \tag{1}$$

where $X \in R^{d\times n}$, $W \in R^{d\times k}$, and $||\cdot||_F$ is the Frobenius norm of the matrix. The objective in (1) can be solved efficiently using a repeated method [35]. Using a simple derivation, (1) can be written as (2).

$$\begin{aligned} &\max_{W} \ (W^TXX^TW) \\ &s.t. \qquad W^TW = I. \end{aligned} \tag{2}$$

In other words, (1) is equivalent to maximizing the variance of the original data and can be solved using the singular value decomposition of XX^T.

Although PCA has been widely and successfully used in image recognition and reduces the dimensions of global features, the data are assumed to be independent when PCA is used. However, many real-world data such as ECG are sequential and exhibit a strong statistical dependence among adjacent samples [18]. In this study, we used the regularization of the total variation of the reconstructed signal to address this problem.

Total variation (TV) has numerous useful applications in signal processing, such as denoising [19], [20], [28], reconstruction [22], and trend filtering [27]. TV regularization assumes that the values of neighboring points change slowly but can also exhibit abrupt level shifts. In this study, we combined PCA with TV regularization to reconstruct the original features; therefore, denoising was involved during feature learning.

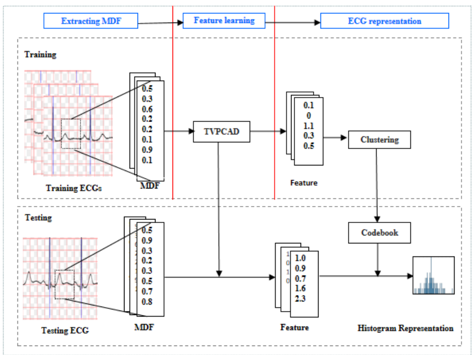


FIGURE 1. Architecture of total variation PCA-based Descriptor learning and codebook clustering for ECG signal.

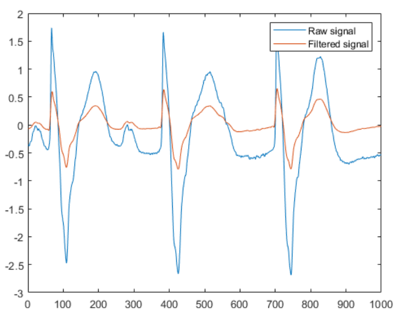


FIGURE 2. Raw signal and its corresponding filtered signal in MITDB database.

III. METHODOLOGY

A novel ECG biometric framework is designed for ECG identity recognition. First, data preprocessing was performed to obtain high-quality ECG heartbeats. Second, a base feature extraction algorithm is introduced to generate the intermediate base feature, that is, MDF. Third, TVPCAD was used to generate compact local descriptors for the ECG signals. Finally, the representation of the ECG heartbeat based on TVPCAD and the BoW framework is introduced, and a matching procedure is performed for the testing data. Fig. 1 shows the proposed framework for the identity recognition systems based on ECG signals. In the following, we introduce each component of the framework in detail, that is, data preprocessing, extraction MDF, learning TVPCAD, and representation of the ECG heartbeat based on our method and matching.

A. DATA PREPROCESSING

The ECG signals captured by the developed device contain severe noise and must be divided into ECG heartbeats. Therefore, we must apply a preprocessing step. Data preprocessing included three steps: denoising, heartbeat segmentation, and outlier removal.

During denoising, the signal first passes through the filters described in [1]. Fig. 2 shows the raw signal from the MITDB (MIT-BIH Arrhythmia) database and its filtered signals. However, the filtering process may remove important information and destroy the original signal; thus, we used the raw signal to experiment with the MITDB database. However, the other two databases, the ECG-ID database, and the Physikalisch Technische Bundesanstalt(PTB) database, contain significant noise; therefore, we used the filtered signal in the experiment.

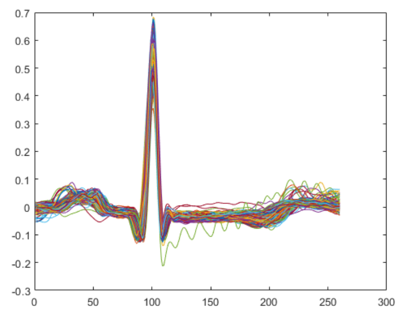


FIGURE 3. All segments of class 100 in MITDB database.

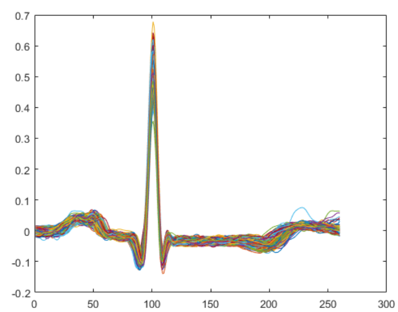


FIGURE 4. All segments of class 100 after outlier removal in MITDB database.

In the heartbeat segmentation process, first, the locations of the R peak can be detected by the Pan Tompkin algorithm [1], and then, taking 100 points forward from the R peak and 159 points backward, therefore, an ECG heartbeat has 260 points of ECG sequence. Fig. 3 shows all heartbeats extracted for recordings belonging to person 100 in the MITDB database. We found that there were many outliers.

In the outlier removal processing, we use an outlier detection method [2] to eliminate outliers. Fig. 4 shows all the heartbeats of person 100 after outlier removal for the heartbeats in Fig. 3. The outliers shown in Figs. 4 are smaller than those shown in Fig. 3.

B. EXTRACTION OF MULTI-SCALE DIFFERENTIAL FEATURE

Although one-dimensional multiresolution local binary patterns(1DMRLBP) [12] have been widely applied to ECG identity recognition, they still have some limitations. To overcome these limitations, we extracted the base features, namely MDF, in which only the amplitude of 1DMRLBP is used; that is, no binarization is performed during the extraction of 1DMRLBP. Specifically, The MDF is extracted from the time sample $y(t)$ can be defined as (3) and (4).

$$S_i = \begin{cases} y(t+i-p-d-1) - y(t) & 1 \leq i \leq p \\ y(t+i-p+d) - y(t) & p+1 \leq i \leq 2p \end{cases} \tag{3}$$

$$x_i = [S_1, S_2, \cdots, S_{2p}], \tag{4}$$

where $y(t)$ denotes the value of the heartbeat point at time t, p denotes the number of points to be calculated on each side of $y(t)$, d denotes the distance from $y(t)$ to the desired time sample. MDF preserves the amplitude of the 1DMRLBP and captures multi-scale information with different p and d. Fig. 5 shows the extracted MDF with $p = 4$ and $d = 2$ for a point at time t. However, MDFs are high-dimensional features that contain considerable noise and redundancy. Therefore, the next step is to learn a compact descriptor for MDF.

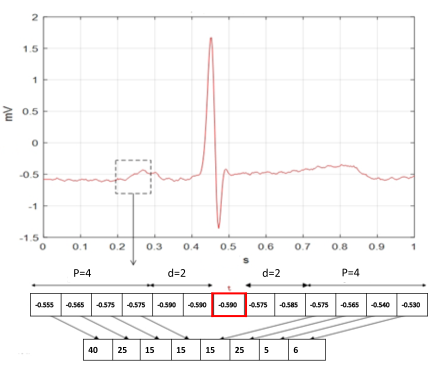


FIGURE 5. Extraction of multi-scale differential feature.

C. LEARNING TOTAL VARIATION PCA BASED DESCRIPTORS

In this section, we introduce a learning method for total variation PCA-based descriptors. This method aims to learn a linear map for projecting MDF onto lower-dimensional features.

Assume that $X = [x_1, x_2, \cdots, x_n] \in R^{d\times n}$ are the MDFs extracted from the C class ECG segment, and $W = [w_1, w_2, \cdots, w_K] \in R^{d\times K}$ is the learned map matrix that can project x_i into a low-dimensional descriptor. To learn a desirable map matrix for the original features, we believe that two principles must be considered:1) Information loss is minimized. 2) Noise was removed. With these principles in mind, we construct the objective function in (5).

$$\begin{aligned} \min_{W} \quad & \lambda_1 ||X - WW^TX||_F^2 + \lambda_2 ||BWW^TX||_1 \\ s.t. \quad & W^TW = I, \end{aligned} \tag{5}$$

where λ_1 and λ_2 are tradeoff parameters, $||\cdot||_1$ is the l_1-norm of the matrix, $B \in R^{d\times d}$ is the first-order difference matrix.

The first term in the objective function (5) is the reconstruction error, which is the same as PCA and minimizes information loss during feature learning. The second term in (5) is the one-dimensional total variation(1D-TV), which is a widely used denoising model in real applications where the neighbor features of each reconstructed MDF change are small.

Optimization: While directly solving the objective function in (5) is intractable because of the l_1 norm, we can iteratively optimize using the alternating direction multiplier method (ADMM).

First, we introduce Z and change (5) into (6).

$$\begin{aligned} \min_{W,Z} \quad & \lambda_1 ||X - WW^TX||_F^2 + \lambda_2 ||Z||_1 \\ s.t. \quad & W^TW = I, \quad Z = BWW^TX. \end{aligned} \tag{6}$$

According to the Lagrange,(6) is transformed into (7).

$$\begin{aligned} \min_{W,Z,Q} \quad & \lambda_1 ||X - WW^TX||_F^2 + \lambda_2 ||Z||_1 \\ & + 0.5\mu ||Z - BWW^TX + \frac{Q}{\mu}||_F^2 \\ s.t. \quad & W^TW = I. \end{aligned} \tag{7}$$

Eq. (7) can be easily solved using ADMM, where the variables are updated alternately by minimizing (7). Specifically, the update of the variables includes the following steps. **1) The other variables are fixed and solve W.** When the other variables are fixed, the solution to (7) can be written as (8).

$$\begin{aligned} \min_{W} \; & \lambda_1 ||X - WW^TX||_F^2 + 0.5\mu ||Z - BWW^TX + \frac{Q}{\mu}||_F^2 \\ s.t. \quad & W^TW = I. \end{aligned} \tag{8}$$

Eq. (8) can be solved efficiently using the gradient descent method with the curvilinear search algorithm proposed in [35].

2) Other variables are fixed, and update Z. When the other variables are fixed,(7) can be rewritten as(9).

$$\min_{Z} \lambda_2 ||Z||_1 + 0.5\mu ||Z - BWW^TX + \frac{Q}{\mu}||_F^2, \tag{9}$$

For convenience, we introduce the following soft-thresholding operator:

$$S_\epsilon(z) = \begin{cases} z - \epsilon & z > \epsilon, \\ z + \epsilon & z < \epsilon, \\ 0, & otherwise, \end{cases} \tag{10}$$

where $z \in R$ and $\epsilon > 0$. The soft-thresholding operator can be extended to vectors and matrices by applying it elementwise. Subsequently, (9) can be solved using(11).

$$S_\epsilon(Z) = argmin \epsilon ||Z||_1 + 0.5||Z - H||_F^2, \tag{11}$$

where $\epsilon = \frac{\lambda_2}{\mu}$, $H = BWW^TX - \frac{Q}{\mu}$.

3) Update Q by (12).

Finally, the Lagrange multiplier is updated as

$$Q^t = Q^{t-1} - \mu(Z - BWW^T X), \tag{12}$$

where Q^t is Q for the current update, and Q^{t-1} is Q for the last iteration.

Moreover, parameter μ is updated using (13).

$$\mu^t = min(\rho * \mu^{t-1}, \mu_{max}). \tag{13}$$

where ρ and μ_{max} are constants.

The detailed algorithm for solving the problem (5) is summarized in Algorithm 1.

Algorithm 1 TVPCAD Algorithm

Require:
X= training dataset; ξ= number of iterations; λ_1 and λ_2= parameters; K= length of the descriptor;

Ensure:
Optimized matrix W

Initialize W as random number, t as 1;
repeat
  Obtain W by solving (8);
  Obtain Z by solving (11);
  Update Q using (12);
  Update μ by (13)
  Set t as $t + 1$;
until $t > \xi$ or $|W^t - W^{t-1}| < \epsilon$
return W.

D. ECG HEARTBEAT REPRESENTATION BASED ON TVPCAD

Having learned the feature map and extracted the descriptors for each point in the ECG heartbeat, we represented the entire ECG heartbeat. The bag-of-words model can utilize local descriptors to capture high-level structural information and has been successfully used in ECG biometric recognition [36], [37]; thus, we adopted it to organize the learned local descriptors in this work. In addition, the heartbeat of the ECG can be divided into three main complexes, P, QRS, and T, and the different complexes have different shapes; therefore, we segmented each ECG heartbeat into several segments and learned a map for each segment. In the following section, the proposed algorithm is described in detail.

After obtaining the map matrix W, the MDFs extracted from the training ECG heartbeats were projected onto low-dimensional descriptors. The k-means method is used to cluster the descriptors into a codebook. Subsequently, given an ECG heartbeat, the descriptors obtained from the ECG heartbeat are encoded as histogram features by pooling them into a codebook. This histogram is a global representation of an ECG heartbeat. The histogram features of the other ECG heartbeats were extracted similarly. To further enhance the performance of the system, we divided each ECG heartbeat into many nonoverlapped segments and learned a map matrix and codebook for each segment. Subsequently, the histogram features extracted from the different segments were concatenated as the final representation of the entire ECG heartbeat. The ECG heartbeat representation using the proposed method is illustrated in Fig. 6. Because the histogram feature representation of each ECG heartbeat includes redundancy and does not utilize label information during local feature learning and histogram feature construction, we first used whitened PCA (WPCA) to reduce the dimensionality of the representation and then utilized the RLRLR method to fuse the category information to obtain the discriminative representation for the ECG heartbeat.

After obtaining a compact and discriminative representation of the ECG heartbeat, we calculated the matching score using the Euclidean similarity.

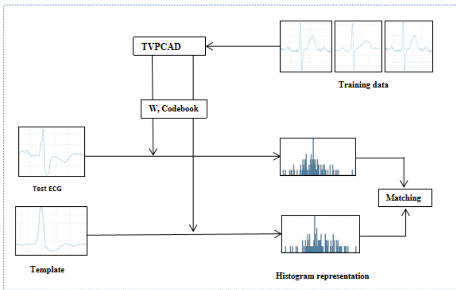


FIGURE 6. Flowchart of TVPCAD ECG representation and recognition.

IV. EXPERIMENTS AND RESULTS

In this study, we implemented the proposed method using MATLAB R2018a and conducted experiments on a common desktop PC with an i5 2.8GHZ CPU and 24 GB RAM. Three public ECG databases are used to evaluate the proposed method. To better demonstrate the advantages of our proposed method, we compared it with three benchmark methods.1) The objective function of the first benchmark method is the same as that of PCA, and singular value decomposition is used to solve it. For brevity, we refer to this descriptor as the PCA descriptor (PCAD). 2) The objective function of the second benchmark method is the same as that of PCA but uses the iterative method to optimize it. We call this the iterated PCA descriptor (IPCAD). 3) The third benchmark method uses the proposed objective function, which we call the TVPCAD-0. After obtaining the histogram representation of each ECG heartbeat, the WPCA was utilized to reduce the dimensionality of the histogram representation for the three benchmark methods. In addition, the proposed method was compared with several state-of-the-art ECG identification algorithms. The recognition rate and equal error rate (EER) were used to evaluate the recognition performance of the ECG systems. The recognition rate is the ratio of the number of correctly identified subjects to the number of subjects in the database, and EER is the value at which the false acceptance rate (FAR) is equal to the false rejection rate (FRR). Within-session and cross-session analyses were performed. The within-session analyses used heartbeats from one recording of one subject, whereas the across-session analyses used heartbeats from two recordings of one subject.

A. DATABASES AND EXPERIMENTAL SETTINGS

We conduct experiments on three public databases: MIT-BIH Arrhythmia (MITDB) dataset [39], ECG-ID database [40], and Physikalisch Technische Bundesanstalt(PTB) database [38].

MIT-BIH Arrhythmia (MITDB) database: This database contains 48 dual-channel recordings from 47 subjects (25 males and 22 females), each of which had only one recording available except for one subject that had two recordings (records 201 and 202). The recordings were digitized at 360 HZ. In our experiment, a within-session analysis was performed using the MITDB database. In other words, all heartbeats belonging to one recording were obtained from a single subject. After using the Pan Tompkin algorithm to detect the location of the R-peak, we took 100 points forward from the R-peak and 159 points backward from the raw recording to construct an ECG heartbeat. The total number of points for each ECG heartbeat was 260 in the MITDB database. Twenty-four ECG heartbeats were recorded for each recording, we took a total of 24 ECG heartbeats. Twelve heartbeats were used to construct the training set, and the remaining 12 ECG heartbeats were used to construct the testing set.

ECG-ID database: This database contains 310 ECG recordings from 90 subjects(44 males and 46 females), digitized at 500 HZ. The number of recordings for each person ranged from two (collected in one day) to 20 (collected periodically over six months). Two recordings per subject were used for the analysis because only a small subset of subjects had over two recordings. Within-session and across-session analyses were performed using this database. In the within-session analysis, one recording was used and the experimental setting was the same as that of the MITDB. In the across-session analyses, the training set came from one recording, and the testing set came from the other. The training and testing sets comprised 12 heartbeats.

Physikalisch Technische Bundesanstalt(PTB) database This database contains 549 records from 290 subjects(209 men and 81 women). The recordings were digitized at 1000 HZ. In our experiment, a within-session analysis was performed using the PTB database, and the ECG signal was downsampled at 500 HZ. Before extracting the ECG heartbeats, the recordings were denoised using filters [1]. The other experimental settings for the PTB database were the same as those for the MITDB.

B. EVALUATION OF TVPCAD

This section discusses the effectiveness of the proposed objective function. To demonstrate the effectiveness of our method, we compared it with PCAD, IPCAD, and TVPCA-0 on the MITDB and PTB databases in the within-session as well as on ECG-ID in the within-session and across-session. In the experiments, we empirically set d to four and p to 25; therefore, a 50-dimension MDF was obtained for each point in the ECG heartbeat and projected onto K-bit descriptor using the learning mapping functions. In all our experiments, K was empirically set as 16. Using the cross-validation strategy on the training sample of the MITDB database, we determined the parameters λ_1 and λ_2 to be 1,000 and 10, respectively. The codebook size was set to 1280, and we experimentally determined the number of segments for an ECG heartbeat to be seven. Consequently, each ECG heartbeat is represented as an 8,960-dimensional feature vector (8,960=1,280*7). We utilized WPCA to reduce the feature dimensions to 250 and applied the RLRLR method with Euclidean similarity for ECG heartbeat matching. In the ECG-ID and PTB databases, the parameter settings and experimental process were the same as those in the MITDB database within the session, whereas, across sessions in the ECG-ID database, we empirically set λ_1 and λ_2 to 1 and 10^4, respectively. In addition, for a fair comparison with PCAD, IPCAD, and TVPCAD-0, the training and testing sets were identical for all three databases. Performance was evaluated using the recognition rate and EER, which are the most common benchmarks in biometrics.

TABLE 1. Recognition rates of PCAD, IPCAD, TVPCAD-0 and TVPCAD on MITDB, ECG-ID and PTB databases (average recognition rate).

Databases	PCAD	IPCAD	TVPCAD-0	TVPCAD
MITDB (Within-session)	97.89%	97.99%	98.31%	100%
ECG-ID (Within-session)	95.58%	95.91%	96.12%	99.25%
ECG-ID (Across-session)	87.98%	88.82%	89.94%	93.34%
PTB (Within-session)	97.96%	97.34%	99.46%	99.80%

TABLE 2. Equal error rates of PCAD, IPCAD, TVPCAD-0, and TVPCAD on MITDB, ECG-ID and PTB databases (average recognition rate).

Databases	PCAD	IPCAD	TVPCAD-0	TVPCAD
MITDB (Within-session)	7.75%	13.32%	4.74%	0.59%
ECG-ID (Within-session)	7.89%	5.75%	5.59%	2.06%
ECG-ID (Across-session)	9.62%	7.98%	4.52%	4.55%
PTB (Within-session)	2.29%	2.73%	0.87%	0.59%

Tables 1 and 2 present the results of the comparison. The results can be analyzed in terms of the following aspects: First, PCAD was compared with IPCAD in the recognition rate and EER. The difference between PCAD and IPCAD is the process of solving the PCA problem. PCAD uses singular value decomposition and IPCAD uses the iterated method. Tables 1 and 2 show that the performance of PCAD was approximately similar to that of IPCAD for the three databases. This is because the objective functions of PCAD and IPCAD are the same. Second, TVPCAD-0 was compared with PCAD and IPCAD in terms of recognition rate and EER. The main difference between them lies in their objective functions. TVPCAD-0 contains a regulator that represents the total variation in the reconstructed MFD, whereas PCAD and IPCAD do not. Tables 1 and 2 show that TVPCAD-0 achieves better recognition performance than PCAD and IPCAD on the three databases. The main reason for these results is that we adopted the total variation of the reconstructed MFD in the objective function of the TVPCAD-0. Third, TVPCAD was compared with TVPCAD-0 in terms of recognition rate and EER. The main difference between them is that the RLRLR method was adopted in TVPCAD but not in TVPCAD-0. Tables 1 and 2 show that TVPCAD outperforms better recognition performance than TVPCAD-0 for all three databases. This occurs because the RLRLR method can combine label information into the representation of the ECG heartbeat, making our representation of the ECG heartbeat more discriminative than that of TVPCAD-0.

Therefore, these results show that using the total variation of the reconstructed MFD can enhance the discrimination of descriptors, and utilizing category information when reducing the dimensions of the ECG heartbeat representation can increase the recognition performance of the system.

TABLE 3. Comparison with different conventional recognition methods on two databases.

Methods	Databases (Subject number)	Number of heartbeats	Accuracy
[41]	ECG-ID(90)	1	97.0%
	MITDB(47)	1	98.6%
[42]	ECG-ID(89)	1	97.54%
	MITDB(47)	1	99.70%
[43]	MITDB(44)	1	98.80%
[21]	ECG-ID(89)	3	100%
	MITDB(47)	3	100%
[54]	ECG-ID(90)	1	98.24%
	MITDB(47)	1	95.99%
TVPCAD-0	ECG-ID(89)	1	96.12%
	MITDB(47)	1	98.31%
TVPCAD	ECG-ID(89)	1	99.25%
	MITDB(47)	1	100%

C. COMPARISON WITH EXISTING ECG RECOGNITION METHODS

The key objective of this set of experiments was to evaluate the ECG recognition performance of the proposed methods (TVPCAD-0 and TVPCAD) in comparison with some state-of-the-art ECG recognition methods [21], [41], [42], [43], [54]; the experimental settings were the same as those in previous settings.

Table 3 lists the ECG recognition performance of the proposed method compared to the state-of-the-art methods on the MITDB and ECG-ID databases. From these results, we can draw the following conclusions: First, the performance of TVPCAD-0 is slightly lower than that of methods in [41] and [42] for the MITDB and ECG-ID databases in terms of the recognition rate. These results originate from the methods in [41] and [42] which fused three heartbeats in the experiment, whereas we used only one heartbeat in TVPCAD-0. Second, the fiducial method in [43] achieved higher accuracy than TVPCAD-0 on the MITDB database; however, the method in [43] used 44 subjects in the experiment, whereas 47 subjects were used in our experiment. Third, compared to the no-fiducial method [54], TVPCAD-0 achieved the best results on the MITDB dataset. However, in the ECG-ID database, the method in [54] achieved a better performance. Nevertheless, the above experiments proved that the proposed local learning feature was valid. Fourth, TVPCAD achieves better performance on the MITDB and ECG-ID databases in terms of the recognition rate than the state-of-the-art ECG recognition methods [41], [42], [43], [54] in Table 3. The reason for these results lies in two aspects: 1) TVPCAD is based on learning the local features, which is more data adaptive than the other methods, and 2) TVPCAD adopts a global feature learning method, which utilizes label information during dimension reduction. However, The recognition rate of TVPCAD on the ECG-ID database is slightly lower than that of the method described in [21]. This is because the method in [21] fuses the three heartbeats as the test sample, whereas our method uses only one heartbeat. Therefore, local and global feature learning leads to a high recognition rate.

TABLE 4. Training time and matching time per heartbeat.

Databases	Training time (s)	Matching time per heartbeat (ms)
ECG-ID	736.006s	43.2ms
MITDB	125.027s	25.6ms

In an actual application scenario, the time required for ECG biometrics should not be too long. Therefore, we recorded the training and matching times to demonstrate the feasibility of the proposed method. The training time contained the time interval from the input of the training heartbeat database (after preprocessing) to obtain the projection mapping and to calculate the template representation, which was obtained from the input of a test ECG heartbeat to obtain the matching result.

A comparison of the results is presented in Table 4. The results can be analyzed in terms of three main aspects. First, the training time for the ECG-ID database was longer than that for the MITDB database. These results are attributable to the number of subjects in the ECG-ID database was 89, whereas that in the MITDB database was 47. Second, the matching time per heartbeat in the ECG-ID database is higher than that in the MITDB database. This finding occurred because the length of the heartbeat in the ECG-ID database was 360, whereas that in the MITDB database was 260. A longer heartbeat requires more time to calculate the local descriptor. Third, a significant amount of time was consumed during training, whereas the matching per heartbeat was performed considerably quickly in both databases. This satisfies the requirement in an actual application scenario that the training can be performed offline and the testing can be executed online.

V. CONCLUSION AND FUTURE WORKS

This study proposes a novel feature learning method called the total variation PCA-based descriptor (TVPCAD) for ECG biometric recognition. It enhances the performance of ECG identity recognition by introducing MDF and our objective function into feature learning. To make the proposed descriptor more effective, we utilized a global feature-learning method to integrate category information into the representation of the ECG heartbeat. The main contributions of this work are as follows:1) We adopted MDF as the base feature for the local feature-learning method. 2) We combined PCA with the TV regulator to learn the local features and experimentally showed that the TV regulator is effective when learning the local features of the ECG signal. 3) The ADMM approach was adopted to optimize the objective function. 4) We learned descriptors and dictionaries for different segments of the ECG heartbeat to enhance its discriminative ability and obtain a more precise representation of the heartbeat. 5) To further enhance the recognition performance, we adopted the WPCA and RLRLR methods to reduce the redundancy of the histogram representation of the ECG heartbeat and increase its discriminative representation. Moreover, within-session and across-session analyses show that our method outperforms state-of-the-art ECG identity recognition methods. In future work, we aim to learn the binary local features of ECG signals.

REFERENCES

[1] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[2] H. P. da Silva, A. Fred, A. Lourenço, and A. K. Jain, "Finger ECG signal for user authentication: Usability and performance," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.

[3] R. D. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ECG: Convolutional neural networks for ECG biometric recognition," *Pattern Recognit. Lett.*, vol. 126, pp. 78–85, Sep. 2019.

[4] Z. Zhao, Y. Zhang, Y. Deng, and X. Zhang, "ECG authentication system design incorporating a convolutional neural network and generalized S-transformation," *Comput. Biol. Med.*, vol. 102, pp. 168–179, Nov. 2018.

[5] Y. Li, Y. Pang, K. Wang, and X. Li, "Toward improving ECG biometric identification using cascaded convolutional neural networks," *Neurocomputing*, vol. 391, pp. 83–95, May 2020.

[6] E. J. da Silva Luz, G. J. P. Moreira, L. S. Oliveira, W. R. Schwartz, and D. Menotti, "Learning deep off-the-person heart biometrics representations," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1258–1270, May 2018.

[7] S. S. Abdeldayem and T. Bourlai, "A novel approach for ECG-based human identification using spectral correlation and deep learning," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 1, pp. 1–14, Jan. 2020.

[8] N. Belgacem, F. Bereksi-Reguig, A. Nait-Ali, and R. Fournier, "Person identification system based on electrocardiogram signal using LabVIEW," *Int. Journ. Comput. Sci. Eng.*, vol. 4, no. 6, p. 974, 2012.

[9] W. Louis, D. Hatzinakos, and A. Venetsanopoulos, "One dimensional multi-resolution local binary patterns features (1DMRLBP) for regular electrocardiogram (ECG) waveform detection," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 601–606.

[10] M. Benouis, L. Mostefai, N. Costen, and M. Regouid, "ECG based biometric identification using one-dimensional local difference pattern," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102226.

[11] W. Louis, M. Komeili, and D. Hatzinakos, "Continuous authentication using one-dimensional multi-resolution local binary patterns (1DMRLBP) in ECG biometrics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12, pp. 2818–2832, Dec. 2016.

[12] K. Wang, G. Yang, Y. Huang, and Y. Yin, "Multi-scale differential feature for ECG biometrics with collective matrix factorization," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107211.

[13] Y. Huang, G. Yang, K. Wang, H. Liu, and Y. Yin, "Learning joint and specific patterns: A unified sparse representation for off-the-Person ECG biometric recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 147–160, 2021, doi: 10.1109/TIFS.2020.3006384.

[14] K. Su, G. Yang, B. Wu, L. Yang, D. Li, P. Su, and Y. Yin, "Human identification using finger vein and ECG signals," *Neurocomputing*, vol. 332, pp. 111–118, Mar. 2019.

[15] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: A new approach in human identification," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 3, pp. 808–812, Jun. 2001.

[16] N. Karimian, Z. Guo, M. Tehranipoor, and D. Forte, "Highly reliable key generation from electrocardiogram (ECG)," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 6, pp. 1400–1411, Jun. 2017.

[17] R. Palaniappan and S. M. Krishnan, "Identifying individuals using ECG beats," in *Proc. Int. Conf. Signal Process. Commun.*, Dec. 2004, pp. 569–572.

[18] X.-L. Li, "Convolutional PCA for multiple time series," *IEEE Signal Process. Lett.*, vol. 27, pp. 1450–1454, 2020.

[19] H. Shi, R. Liu, C. Chen, M. Shu, and Y. Wang, "ECG baseline estimation and denoising with group sparse regularization," *IEEE Access*, vol. 9, pp. 23595–23607, 2021.

[20] I. W. Selesnick, H. L. Graber, D. S. Pfeil, and R. L. Barbour, "Simultaneous low-pass filtering and total variation denoising," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1109–1124, Mar. 2014.

[21] R. Li, G. Yang, K. Wang, Y. Huang, F. Yuan, and Y. Yin, "Robust ECG biometrics using GNMF and sparse representation," *Pattern Recognit. Lett.*, vol. 129, pp. 70–76, Jan. 2020.

[22] Z. Dou, B. Zhang, and X. Yu, "A new alternating minimization algorithm for total variation image reconstruction," in *Proc. Int. Conf. Wireless*, Jul. 2016, pp. 248–272.

[23] I. B. Ciocoiu, "Comparative analysis of bag-of-words models for ECG-based biometrics," *IET Biometrics*, vol. 6, no. 6, pp. 495–502, Nov. 2017.

[24] J. Wang, M. She, S. Nahavandi, and A. Kouzani, "Human identification from ECG signals via sparse representation of local segments," *IEEE Signal Process. Lett.*, vol. 20, no. 10, pp. 937–940, Oct. 2013.

[25] N. Mourad, "ECG denoising algorithm based on group sparsity and singular spectrum analysis," *Biomed. Signal Process. Control*, vol. 50, pp. 62–71, Apr. 2019.

[26] H. Gao, C. Liu, X. Wang, L. Zhao, Q. Shen, E. Y. K. Ng, and J. Li, "An open-access ECG database for algorithm evaluation of QRS detection and heart rate estimation," *J. Med. Imag. Health Informat.*, vol. 9, no. 9, pp. 1853–1858, Dec. 2019.

[27] Q. Wen, J. Gao, X. Song, L. Sun, and J. Tan, "RobustTrend: A Huber loss with a combined first and second order difference regularization for time series trend filtering," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jun. 2019, pp. 3856–3862.

[28] T. F. Chan, S. Osher, and J. Shen, "The digital TV filter and nonlinear denoising," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 231–241, Feb. 2001.

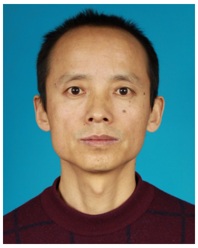
[29] R. Srivastva and Y. N. Singh, "ECG analysis for human recognition using non-fiducial methods," *IET Biometrics*, vol. 8, no. 5, pp. 295–305, Sep. 2019.

[30] S. Gutta and Q. Cheng, "Joint feature extraction and classifier design for ECG-based biometric recognition," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 2, pp. 460–468, Mar. 2016.

[31] M. N. Dar, M. U. Akram, A. Usman, and S. A. Khan, "ECG biometric identification for general population using multiresolution analysis of DWT based features," in *Proc. 2nd Int. Conf. Inf. Secur. Cyber Forensics (InfoSec)*, Nov. 2015, pp. 5–10.

[32] J. Xu, G. Yang, K. Wang, Y. Huang, H. Liu, and Y. Yin, "Structural sparse representation with class-specific dictionary for ECG biometric recognition," *Pattern Recognit. Lett.*, vol. 135, pp. 44–49, Jul. 2020.

[33] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.
[34] J. M. Irvine, S. A. Israel, W. T. Scruggs, and W. J. Worek, "EigenPulse: Robust human identification from cardiovascular function," *Pattern Recognit.*, vol. 41, no. 11, pp. 3427–3435, Nov. 2008.
[35] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, nos. 1–2, pp. 397–434, Dec. 2013.
[36] J. Wang, P. Liu, M. F. H. She, S. Nahavandi, and A. Kouzani, "Bag-of-words representation for biomedical time series classification," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 634–644, Nov. 2013.
[37] I. B. Ciocoiu, "ECG biometrics using bag-of-words models," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jul. 2015, pp. 1–4.
[38] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik/Biomed. Eng.*, vol. 40, pp. 317–318, Jul. 2009.
[39] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May 2001.
[40] T. Lugovaya, "Biometric human identification based on electrocardiogram," M.S. thesis, Faculty Comput. Technol. Inform., Electrotechnical Univ. LETI, Saint-Petersburg, Russian Federation, 2005. [Online]. Available: https://archive.physionet.org/physiobank/database/ecgiddb/images/
[41] R. Salloum and C.-C. J. Kuo, "ECG-based biometrics using recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2062–2066.
[42] B. Wu, G. Yang, L. Yang, and Y. Yin, "Robust ECG biometrics using two-stage model," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1062–1067.
[43] D. Rezgui and Z. Lachiri, "ECG biometric recognition using SVM-based approach," *IEEJ Trans. Electr. Electron. Eng.*, vol. 11, no. S1, pp. S94–S100, Jun. 2016.
[44] J. Yu, Y. Si, L. Xin, D. Wen, T. Luo, and L. Lang, "ECG identification based on PCA-RPROP," in *Proc. Int. Conf. Commun. Technol.*, Jul. 2017, pp. 2008–2013.
[45] C. Camara, P. Peris-Lopez, L. Gonzalez-Manzano, and J. Tapiador, "Real-time electrocardiogram streams for continuous authentication," *Appl. Soft Comput.*, vol. 68, pp. 784–794, Jul. 2018.
[46] X. Fang, Y. Xu, X. Li, Z. Lai, W. K. Wong, and B. Fang, "Regularized label relaxation linear regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1006–1018, Apr. 2018.
[47] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.
[48] H. Liu, G. Yang, L. Yang, and Y. Yin, "Learning personalized binary codes for finger vein recognition," *Neurocomputing*, vol. 365, pp. 62–70, Nov. 2019.
[49] H. Liu, L. Yang, G. Yang, and Y. Yin, "Discriminative binary descriptor for finger vein recognition," *IEEE Access*, vol. 6, pp. 5795–5804, 2018.
[50] R. Srivastva, A. Singh, and Y. N. Singh, "PlexNet: A fast and robust ECG biometric system for human recognition," *Inf. Sci.*, vol. 558, pp. 208–228, May 2021.
[51] B.-H. Kim and J.-Y. Pyun, "ECG identification for personal authentication using LSTM-based deep recurrent neural networks," *Sensors*, vol. 20, no. 11, p. 3069, May 2020.
[52] D. Jyotishi and S. Dandapat, "An ECG biometric system using hierarchical LSTM with attention mechanism," *IEEE Sensors J.*, vol. 22, no. 6, pp. 6052–6061, Mar. 2022.
[53] D. A. AlDuwaile and M. S. Islam, "Using convolutional neural network and a single heartbeat for ECG biometric recognition," *Entropy*, vol. 23, no. 6, p. 733, Jun. 2021.
[54] Y. Chu, H. Shen, and K. Huang, "ECG authentication method based on parallel multi-scale one-dimensional residual network with center and margin loss," *IEEE Access*, vol. 7, pp. 51598–51607, 2019.



HAIYING LIU received the Ph.D. degree from Shandong University, in 2019. He is currently a Lecturer with Xinjiang University. His main research interests include biometrics and machine learning.

HAIYAN LIN received the bachelor's degree from Shihezi University, China, in 1999. She is currently a Teacher with the Computer Department, Xinjiang Qitai Secondary Vocational and Technical School. Her research interests include biometrics and data mining.

XIANHUI WANG received the Ph.D. degree from Xi'an Jiaotong University, China, in 2005. He is currently an Associate Professor with Xinjiang University. His research interests include machine learning and data mining.