Atrial Fibrillation Detection from Face Videos by Fusing Subtle Variations

Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen and Guoying Zhao, *Senior Member, IEEE*

***Abstract*—Atrial fibrillation (AF) is one of the most common cardiac arrhythmias, which particularly occurs in the elderly individuals with heart disease. Though AF is often asymptomatic during normal activities, it has huge potential risks for stroke and other severe diseases. Thus, early detection of AF has great importance in the field of public health. Currently, electrocardiography (ECG) is the commonly used measure for the diagnosis of AF, which presents the irregular rhythm of waveform for AF patients. However, the measurement of the ECG signal requires special medical acquisition devices, which are not comfortable for practical monitoring in daily life. In this paper, we explore a very promising algorithm to detect AF from remote face videos by analyzing the color variations of face skin. The main challenge is that the current remote photoplethysmography (rPPG) technique is rather immature, which causes difficulty in extracting accurate pulse signals for describing the cardiac rhythm. To solve this problem, we first utilize various rPPG algorithms to capture pulse rhythms from different regions on the face video. We then investigate biomedical statistical methods to extract suitable features from each pulse signal. Due to the imprecision of video-extracted pulse signals, some traditional physiological features may lose their utility since they were originally proposed for ECG signals. Furthermore, some of them are very susceptible to the influence of noise. Thus, we propose a feature fusion algorithm to select and combine reasonable information from multiple physiological features, which aims to preserve the discriminability of detecting AF in the presence of the noise and outlier disturbances. The experimental results on a real-world database demonstrate the effectiveness of the proposed method in providing useful information for AF detection.**

***Index Terms*—Atrial fibrillation, heart rate variability (HRV), cardiac disease diagnosis, feature fusion, classification.**

I. Introduction

ATRIAL fibrillation (AF) is the most common type of arrhythmia, and has been reported to significantly increase the risk for heart failure, stroke and mortality. Although AF treatment strategies [1]–[4] have achieved great development over the last decade, early detection of AF is very important for preventing the occurrence of serious cardiac disease. Unfortunately, AF is often asymptomatic in the early phase, which induces the difficulty in diagnosing AF in time. Currently, ECG signals [5]–[8] have been successfully utilized in AF diagnosis for clinical applications. However, the acquisition of ECG signals requires specific biomedical equipment, which limits the application of monitoring AF risk in daily life. Recently, many researchers [9]–[11] have tried to capture the cardiac pulse signal with the wearable device and smart phone for predicting AF. Nevertheless, the measurement needs to be conducted in a skin-contact manner that is inconvenience and uncomfortable for the examinee. In this paper, we will discuss the possibility of detecting AF from short clips of face videos in a contactless manner as illustrated in Fig. 1. In such case, AF can be simply examined by a common camera or webcam, which is very promising for the early intervention of potential AF patients.

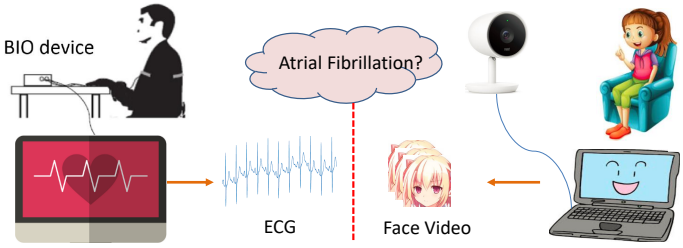


Fig. 1. An illustration for the proposed remote AF detection method. The traditional AF detection approach is presented in the left part of the picture, which requires to put specific sensors on the body and utilize biomedical devices for capturing the ECG signal. As shown in the right part of the picture, the proposed remote AF detection model aims to detect AF from face videos in a contactless manner, which is very convenient to monitor the AF risk in daily life since it only needs a common RGB camera.

The diagnosis of AF from ECG signals can be roughly categorized into two classes [7], e.g., (1) P-wave detection and (2) R-R interval (RRI) variability. The methods based on P-wave detection aim to monitor the absence of the P-wave in the ECG signal of AF patients, which would be replaced by rapid oscillations in actual conditions. However, the P-wave is very sensitive to motion and noisy artifacts in the measurement, which causes challenges in acquiring accurate P-wave information. In contrast, the R-peak is the most prominent characteristic feature of ECG signals and is robust to different kinds of noise. Generally, the RRI-based algorithms diagnose patients with AF by spotting irregularities in the extracted RRI variability series, which means that the precise detection of the R-peak is very important to the final

Manuscript received March 5, 2019; revised June 6, 2019; accepted June 24, 2019. This work was supported by the National Natural Science Foundation of China (No. 61772419), Academy of Finland, Tekes Fidipro Program (No. 1849/31/2015), Business Finland Project (No. 3116/31/2017), Academy of Finland 6Genesis Flagship (No. 318927), Tekniikan Edistämissäätiö and Infotech Oulu. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jonathan Wu. *(Corresponding author: Guoying Zhao.)*

G. Zhao is with the School of Information and Technology, Northwest University, Xi'an 710069, China, and also with the Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014 Oulu, Finland. (e-mail:guoying.zhao@oulu.fi).

J. Shi, I. Alikhani, X. Li, Z. Yu and T. Seppänen are with the Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014 Oulu, Finland. (e-mail: jingang.shi@oulu.fi; iman.alikhani@oulu.fi; xiaobai.li@oulu.fi; zitong.yu@oulu.fi; tapio.seppanen@oulu.fi).

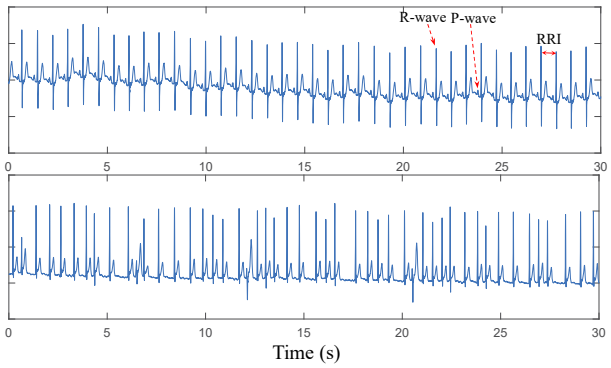


Fig. 2. ECG signals for a healthy individual (top) and an AF patient (bottom) respectively. The irregularity of beat-to-beat variability occurs in the ECG signal of AF patient.

AF diagnosis. In Fig. 2, we illustrate the distinction of ECG signals for healthy and AF cases.

Recently, some interesting studies [12]–[18] have focused on the contactless monitoring of cardiac activity by detecting the pulse-induced subtle color variations on the human face with a common RGB camera, which inspired us to further explore cardiac disease detection from remote videos. More comprehensive studies are presented in [19] [20] to monitor human heart activity by the PPG sensor, wearable device and smartphone. These algorithms have the ability to extract the pulse signals from the face or skin, which can approximate the real heart rate variability (HRV) of humans. Thus, we can detect the local maximum peaks from the extracted signals and measure the specific changes between successive pulses for estimating the RRI variability. As shown in Fig. 3 and Fig. 4, we respectively extract the pulse rhythms from the face videos of one healthy individual and one AF patient via three different algorithms, while the real ECG signals are also provided for comparison. Though the video-extracted pulse rhythms lose the typical morphology of ECG signals, the R-peaks are successfully captured in the waveforms, which can be utilized to detect AF. However, the pulse extraction techniques have not been fully developed so far, which can cause artificial effects in the video-extracted pulse signals. Thus, the approximation of RRI series from video-extracted pulse rhythms induces a certain bias in the application of AF detection. To better illustrate the above problem, we present an example in Fig. 5. Due to the influence of noises in the real scene, all the video-extracted pulse rhythms are contaminated by artifacts at different levels, which introduces difficulties in R-peak analysis for AF detection. Couderc et al. [21] conducted a preliminary study for detecting AF using contactless video monitoring. However, they achieved a relatively high error rate due to the imprecision of video-extracted pulse rhythms.

In this paper, we discuss the possibility of detecting AF risk from short clips of face videos by exploiting the special characteristics of face images. In the entire human face, some regions contain abundant vascular tissues, which are more suitable for monitoring the variations of skin color and extracting the pulse rhythms. Meanwhile, other regions are more vulnerable to the influence of nonrigid movements

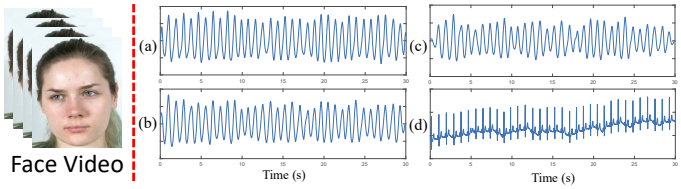


Fig. 3. The pulse rhythms extracted from the face video of one healthy individual by three various algorithms on ROI 1 (refer to Table I for details). The peak points in the waveforms are exactly the same as the R-wave peaks in ECG signal. (a) Method [12]. (b) Method [13]. (c) Method [14]. (d) ECG signal.

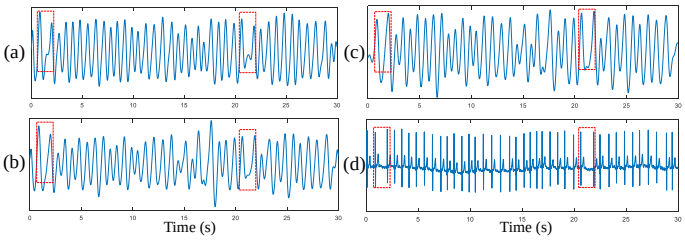


Fig. 4. The pulse rhythms extracted from the face video of one AF patient by three various algorithms on ROI 1 (refer to Table I for details). The irregularity of heart variability can be seen from the rhythms. (a) Method [12]. (b) Method [13]. (c) Method [14]. (d) ECG signal.

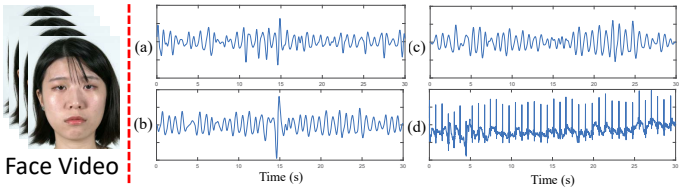


Fig. 5. Another example on one healthy individual. Apparently, the pulse rhythms suffer from artificial effects, which induce difficulties to obtain accurate R-wave peak points as ECG signal. The waveforms are extracted by three different approaches on ROI 1 (refer to Table I for details). (a) Method [12]. (b) Method [13]. (c) Method [14]. (d) ECG signal.

such as expression variations, which tend to produce noisy pulse rhythms. Due to the diversity of human individuals, it is truly difficult to determine a universal face region of interest (ROI) that produces the most robust video-extracted pulse rhythm. Thus, it is better to extract pulse rhythms on multiple plausible ROIs and fuse these signals for superior results. Meanwhile, in order to enhance the robustness of video-extracted pulse rhythms, we propose to utilize various pulse extraction methods for acquiring multiple pulse signals on each ROI. Thus, the problem turns into mining the potential HRV information in multiple pulse rhythms extracted from multiple ROIs, which enables us to utilize the knowledge of multiple tasks to facilitate final classification.

An intuitive method is to extract HRV features from each video-extracted pulse rhythm and concatenate all the features together for AF detection. However, as discussed in the above section, some ROIs on the face image may be affected by voluntary movements or other disturbances, while current video-based pulse extraction methods also produce rhythms with artifacts and noises. In this case, some selected HRV features may lose effectiveness for handling video-extracted

rhythms since they are not robust to the influence of artificial effects on the pulse signals. The above fact means that the contributions of various HRV features are rather different in the AF detection problem, while some of the HRV features may suffer from noisy artifacts. Thus, it would not be a good choice to utilize all the features together because some of them may play a negative role in classification. Recently, dimension reduction techniques [22]–[30] have attracted great attention for further refining the extracted features. These methods aim to learn low-dimensional representations from the original high-dimensional data samples by removing the redundancy in the raw features and enhancing the discriminability of data according to the label information. Considering the local manifold structure of HRV features, we seek a linear projection matrix to learn a low-dimensional embedding space, which enhances the compactness of neighboring samples according to label information. For the binary classification problem of AF detection, this approach maintains the intrinsic geometric structure of features from the same category, while simultaneously preserving the discriminative information across healthy/AF categories.

One major challenge of the remote AF detection task is the problem of outliers, i.e., noisy samples in the training set and redundant dimensions in each training sample. As shown in the sample case in Fig. 5, the extracted pulse signals of one healthy individual are contaminated by noises and thus show the irregular beat rhythm, which may incorrectly imply potential AF contrary to the fact. Thus, the defects of current pulse extraction methods cause the contradiction between the training sample and label information, which induces outlier samples in the whole training set. To improve the robustness of the proposed method, we utilize $\ell_{2,1}$ norm metric to measure the scatter of data samples. The $\ell_{2,1}$ norm metric is less sensitive than the traditional Euclidean distance and reduces the disturbance of noisy training samples (outliers). To construct the discriminative embedding space by local manifold structure, we utilize the strategy of adaptively adjusting the neighborhood connectivity between data samples. The above strategy further optimizes the relationship of neighboring samples in the embedding space, which avoids the influence of noises in the original HRV features. To alleviate the inappropriate characteristics in the extracted HRV features, we also consider the $\ell_{2,1}$ norm regularization of the projection matrix for conducting the feature selection procedure, which aims to exclude the redundant dimensions and enhance the discriminability. Note that the target of the proposed algorithm is not to eliminate the noises and artifacts in the video-extracted signals, but to achieve superior AF detection performance based on the fact that noises and artifacts exist in the pulse signals.

Finally, the unified feature fusion framework can be optimized by an iterative procedure to obtain the projection matrix. The embedding HRV features of the testing samples can be conveniently extracted by linear projection, which also avoids the out-of-sample problem. The final AF detection result can be simply predicted using the SVM in the learned embedding space. In [12], we collected the OBF database for AF detection and evaluated the performance as a baseline. In this paper, we further improve the performance of AF detection in two aspects: (1) Since video-extracted pulse rhythms are susceptible to the influence of noise, we extract multiple pulse rhythms from multiple ROIs and fuse the HRV features together to enhance the robustness. (2) We propose a novel learning-based method to obtain more discriminative feature for facilitating the task of AF detection. The experimental results on the OBF database [12] present the capability of the proposed method, which efficiently improves the accuracy from 77.89% to 92.56%. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first study to improve the capability of AF detection using a learning-based method by analyzing the specific characteristics in human face videos.
- Current methods for measuring heart beat signals from face videos are all not robust enough to achieve accurate pulse rhythms. Given this fact, the proposed method partitions the whole face into several ROIs and extracts multiple pulse signals from each ROI by various methods. We fuse the HRV features in multiple pulse rhythms extracted from multiple ROIs to facilitate AF detection.
- By analyzing the property of AF detection, we propose a robust feature fusion method for extracting suitable HRV features, which simultaneously improves the problem of outliers and enhances the discriminability between healthy/AF individuals.
- The extracted HRV features for the testing samples can be conveniently obtained by the learned projection matrix. The experimental results demonstrate the possibility of detecting AF risk from remote face videos in a contactless manner.

II. Methodology

In this section, we describe the proposed approach for detecting AF risk from remote face videos. The whole procedure can be roughly divided into three steps, which are presented in the following subsections.

A. ROI detection and pulse signal extraction

As we previously discussed, it is essential to capture multiple pulse signals from various ROIs to further compensate the disadvantages of current pulse extraction methods. To partition the entire face image into several ROIs, we first utilize OpenFace [31] to localize and track 68 facial landmarks. To avoid the effect of nonrigid motions, we only use 14 landmarks located on the contour of a face image to construct ROIs, which have relatively stable coordinates during the movement and expression variation. An illustration of the 14 landmarks is presented in Fig. 6. The potential ROIs are obtained by the connection of specific landmarks. As shown in Table I, we utilize a set of facial keypoints to define ROIs on the face image, where the selected landmarks serve as the polygon vertices of each ROI. Thus, we select 21 ROIs in total for the proposed approach.

For each ROI, we record the mean RGB values of pixels inside it and further eliminate the noises by a moving-average filter on the nearest 15 points. Then, we employ three various

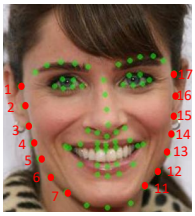


Fig. 6. The selected 14 facial landmarks on the contour of a face image for defining the ROIs. The above picture is cited from [32].

TABLE I
THE SELECTED POLYGON VERTICES FOR TOTAL 21 ROIS

Index	Landmarks	Index	Landmarks	Index	Landmarks
1	(1,7,11,17)	**8**	(2,5,13,16)	**15**	(5,7,11,13)
2	(1,6,12,17)	**9**	(3,6,12,15)	**16**	(1,2,16,17)
3	(2,7,11,16)	**10**	(4,7,11,14)	**17**	(2,3,15,16)
4	(1,5,13,17)	**11**	(1,3,15,17)	**18**	(3,4,14,15)
5	(2,6,12,16)	**12**	(2,4,14,16)	**19**	(4,5,13,14)
6	(3,7,11,15)	**13**	(3,5,13,15)	**20**	(5,6,12,13)
7	(1,4,14,17)	**14**	(4,6,12,14)	**21**	(6,7,11,12)

pulse extraction methods [12]–[14] to acquire the video-extracted pulse rhythms. An ideal example of the final pulse rhythm signals can be seen in Fig. 3. Subsequently, we develop a customized peak detection function for detecting pulse peaks from the video-extracted signals and compute the RR interval (RRI). As shown in Fig. 2, the intervals between adjacent heart beats are mostly stable for the healthy individual, while the intervals can have dramatic changes for the AF patient. We extract the following HRV standard features [33] from RRI to distinguish healthy/AF individuals:

• Time-domain: mean RRI, standard deviation of RRI, root mean square of successive differences (RMSSD), square root of the sum of the squares of differences of individual values compared to the mean value, divided by the number of RRI in a period (RMSM) and percentage of samples with more than 50 ms difference from the consecutive beat (pNN50).

• Geometrical-domain: Poincare plot standard deviations (SD1, SD2).

• Spectral-domain: LF, HF and their ratio in normalized units.

B. Feature fusion and selection

Suppose that we have in total N face video clips as the training set. As shown in the above subsection, we extract the HRV features from 21 ROIs by three different algorithms. Thus, the HRV descriptors of each face video clip can be represented by 63 views. We define the vector $\boldsymbol{x}_v^n$ as the feature of the vth view for the nth sample and $\mathbf{X}_v = \left[\boldsymbol{x}_v^1, \boldsymbol{x}_v^2, \ldots, \boldsymbol{x}_v^N\right]$ as the feature of the vth view for all the training samples. Now, the problem transforms into utilizing the features of all the 63 views for enhancing the discriminability across healthy/AF samples. The simplest way to employ multiview data is to concatenate all the features together for AF classification. For example, let $\boldsymbol{x}^n = \left[(\boldsymbol{x}_1^n)^T, (\boldsymbol{x}_2^n)^T, \ldots, (\boldsymbol{x}_{63}^n)^T\right]^T \in R^d$ represent the feature for the nth sample and $\mathbf{X} = \left[(\mathbf{X}_1)^T, (\mathbf{X}_2)^T, \ldots, (\mathbf{X}_{63})^T\right]^T \in R^{d\times N}$ describe the set for the whole training samples. We can directly use some classification algorithms (e.g., SVM) to train the related model for AF detection. However, due to the voluntary movement or disturbance on the face image, the video-extracted pulse rhythms often suffer from artificial effects and noise. Some HRV features may not be robust to the effects of artifacts on the pulse signals and therefore can hardly describe the characteristics of AF. In this case, the corresponding features lose discriminative ability and become less important in the AF detection. Thus, it is better to combine these HRV features according to their importance rather than employ simple concatenation. To further refine the features from multiple views, we aim to construct the discriminative features by the linear projection of the original features:

$$\mathbf{Y} = \mathbf{P}^T\mathbf{X} \tag{1}$$

where $\mathbf{Y} = [\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^N] \in R^{m\times N}(m < d)$ represents the low-dimensional embedding for the HRV features, and $\mathbf{P} \in R^{d\times m}$ is the linear transformation matrix. The problem converts to training a suitable projection matrix $\mathbf{P}$ to preserve the discriminability of embedding features $\mathbf{Y}$.

The AF detection problem is a typical binary classification problem. The label of each data sample belongs to the category of healthy or AF. Given the training samples in $\mathbf{X}$, we aim to optimize the data graph $\{\mathbf{Y}, \mathbf{W}\}$ with the embedding features $\mathbf{Y}$ and the similarity matrix $\mathbf{W}$, which further enhances the compactness of samples from the same category. For each training sample $\boldsymbol{x}^n$, we define the affiliated data group $\mathbf{X}^n = [\boldsymbol{x}^n, \boldsymbol{x}^{n_1}, \boldsymbol{x}^{n_2}, \ldots \boldsymbol{x}^{n_k}] \in R^{d\times(k+1)}$ to further investigate the local relationship, where $\boldsymbol{x}^{n_1}, \boldsymbol{x}^{n_2}, \ldots \boldsymbol{x}^{n_k}$ represent the k nearest neighboring samples from the same category of $\boldsymbol{x}^n$. Thus, we can obtain the corresponding low-dimensional embedding results $\mathbf{Y}^n = [\boldsymbol{y}^n, \boldsymbol{y}^{n_1}, \boldsymbol{y}^{n_2}, \ldots \boldsymbol{y}^{n_k}] \in R^{m\times(k+1)}$ by the linear projection matrix $\mathbf{P}$. To preserve the intraclass compactness in the low-dimensional subspace, we minimize the distances between each training sample and the affiliated neighbors in the data group, which encourages enhancing the discriminability of features in the classification. According to spectral analysis theory, the objective function can be formulated as:

$$\arg\min_{\mathbf{P}} \sum_{q=1}^{k} \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^{i_q}\right\|_2^2 \omega_{i,i_q}, s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{2}$$

where the weight ω_{i,i_q} is employed to measure the similarity between each sample and its neighbors in the data graph and the constraint preserves the orthogonality of the projection matrix. Commonly, the weight ω_{i,i_q} can be simply defined as $1/k$, while the number of neighboring samples is set to 5 in this paper. By considering all the samples in the training set, we can further represent the optimization function as:

$$\arg\min_{\mathbf{P}} \sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2^2 \omega_{i,j}, s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{3}$$

where the corresponding weight for the sample $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ can be represented as:

$$\omega_{i,j} = \begin{cases} 1/k & j \in N_i = \{i^1, \ldots, i^k\} \\ 0 & otherwise \end{cases} \tag{4}$$

As discussed in the introduction, AF detection mainly encounters two problems due to the effect of outliers. The first problem is the existence of noisy samples in the training set. For the example in Fig. 5, though these samples are actually captured from a healthy individual, the defects of current pulse extraction methods induce artificial effects on the pulse rhythm signals. The artifacts cause irregular RRI series between adjacent pulse peaks, which could imply HRV characteristics that are similar to the AF case. Such samples can be considered as outliers in the training set since the representative features lose their consistency with the corresponding label information. Actually, the feature extraction method in Eq. (3) is prone to suffering from outliers due to the exaggeration of the error by the Euclidean norm metric. In contrast, the $\ell_{2,1}$ norm metric is less sensitive to the effect of outliers, which may improve the robustness of the AF detection problem. Thus, we rewrite the optimization problem (3) as follows:

$$\arg\min_{\mathbf{P}} \sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2 \omega_{i,j}, s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad (5)$$

Since the low-dimensional embedding feature $\boldsymbol{y}^i$ has superior discriminability than the original data point $\boldsymbol{x}^i$, we can adaptively optimize the data graph in the embedding space rather than simply predefine it by experience, as in Eq. (4). As we know, each data point $\boldsymbol{y}^i$ belongs to one of the two categories (i.e., healthy or AF), and it can be connected by all the neighboring data points of the same category with the probability $\omega_{i,j}$. The probability on the graph measures the similarity between two neighboring samples. Thus, it is natural to give a larger probability $\omega_{i,j}$ for the neighbors with a small distance $\left\|\boldsymbol{y}^i - \boldsymbol{y}^j\right\|_2$. By modifying the formulation of (5), we can assign the probabilities of neighboring samples for graph construction by the following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{P},\mathbf{W}} \sum_{i=1}^{N}\left\{\sum_{j=1}^{N}\left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2\omega_{i,j} + \alpha\left\|\boldsymbol{\omega}_i\right\|_2^2\right\} \\ &s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I}, \boldsymbol{\omega}_i^T\mathbf{1} = 1, \\ &0 \le \omega_{i,j} \le 1(j \in N_i), \omega_{i,j} = 0(j \notin N_i) \end{aligned} \quad (6)$$

where $\mathbf{W}$ is the similarity matrix with each element as $\omega_{i,j}$, the vector $\boldsymbol{\omega}_i$ describes the similarity connections for the ith sample with the jth element as $\omega_{i,j}$, and N_i indicates the neighborhood of the ith sample. The constraints ensure that the whole data graph is partitioned into two clusters, i.e., healthy and AF. The regularization term is utilized to smooth the elements in the matrix $\mathbf{W}$. It avoids the trivial solution in which the probability of the nearest neighbor is set to 1, while all the other similarity probabilities are set to zero. By optimizing Eq. (6), we can obtain the data connection graph automatically, which alleviates the effect of noises in the training samples.

Another problem is the influence of redundant dimensions in each training sample $\boldsymbol{x}^i$. The corresponding features in these dimensions are affected by inappropriate HRV characteristics or noises, which induce disadvantages to the discriminablity of samples. To further eliminate the redundant dimensions in the HRV feature, it is reasonable to enforce the row sparsity for the projection matrix $\mathbf{P}$, which means that some rows of $\mathbf{P}$ are filled with elements of all zeros. Such a projection matrix ignores the corresponding redundant dimensions and conducts feature selection in the embedding procedure. The row sparsity property can be achieved by the minimization of the $\ell_{2,1}$ norm regularization term [34]. It can be defined as:

$$\|\mathbf{P}\|_{2,1} = \sum_{i=1}^{d}\|\boldsymbol{p}_i\|_2 = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{m} p_{ij}^2} \quad (7)$$

where $\boldsymbol{p}_i$ represents the ith row of matrix $\mathbf{P}$.

Considering feature fusion and feature selection in a unified framework, we can finally obtain the following objective function:

$$\begin{aligned} &\arg\min_{\mathbf{P},\mathbf{W}} \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\left\|\mathbf{P}^T\mathbf{x}^i - \mathbf{P}^T\mathbf{x}^j\right\|_2\omega_{i,j} + \alpha\left\|\boldsymbol{\omega}_i\right\|_2^2\right) + \beta\|\mathbf{P}\|_{2,1} \\ &s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I}, \boldsymbol{\omega}_i^T\mathbf{1} = 1, \\ &0 \le \omega_{i,j} \le 1(j \in N_i), \omega_{i,j} = 0(j \notin N_i) \end{aligned} \quad (8)$$

where β is the regularization parameter. The optimization of Eq. (8) will be shown in Section III.

C. AF detection

Once we obtain the optimal projection matrix $\mathbf{P}$, the discriminative feature for each sample can be extracted in the embedding space by linear projection according to Eq. (1). We utilize the radial basis function (RBF) kernel support vector machine (SVM) [35] to train the classifier, which considers the AF detection as a typical binary classification problem.

III. Optimization

In this section, we discuss the optimization of Eq. (8) to obtain the projection matrix $\mathbf{P}$. The objective function in (8) is a non-convex optimization problem, which means that it is difficult to acquire the optimal solution. To solve the tough problem, we can expect to get a local optimal solution by an iterative procedure. Specifically, we alternatively optimize over matrices $\mathbf{P}$ and $\mathbf{W}$, while keep the other one fixed. To find a reasonable solution, we also give a warm start to the variables. The details are presented in the follows.

A. Initialization

Rather than randomly initialize the variables $\mathbf{P}$ and $\mathbf{W}$, we give them a more reasonable start in order to obtain superior results after convergence. For convenience, we initialize the data similarity matrix by Eq. (4), which also defines the elements in $\mathbf{W}^0$. The projection matrix $\mathbf{P}^0$ is then initialized by Eq. (3). The objective function can be rewritten as:

$$\arg\min_{\mathbf{P}} tr(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}), s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad (9)$$

where $\mathbf{L} = \mathbf{D} - (\mathbf{W} + \mathbf{W}^T)/2$ is the graph Laplacian matrix and the diagonal degree matrix $\mathbf{D}$ is defined as $\mathbf{D}_{ii} = \sum_j (\mathbf{W}_{ij} + \mathbf{W}_{ji})/2$. According to Ky-Fan theorem [36], the initial projection matrix $\mathbf{P}^0$ can be produced by the eigenvectors corresponding to the m smallest eigenvalues of $\mathbf{X}\mathbf{L}\mathbf{X}^T$.

B. The iterative procedure

With the initialization of variables $\mathbf{W}^0$ and $\mathbf{P}^0$, we can apply the alternative optimization method to solve Eq. (8). When the value of $\mathbf{W}$ is fixed, the optimal $\mathbf{P}$ can be obtained by minimizing the following function:

$$\begin{aligned} &J(\mathbf{P}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2 \omega_{i,j} + \beta\|\mathbf{P}\|_{2,1} \\ &s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I} \end{aligned} \tag{10}$$

The objective function in Eq. (10) can be deduced as:

$$\begin{aligned} J(\mathbf{P}) &= \sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{P}^T(\boldsymbol{x}^i - \boldsymbol{x}^j)\right\|_2 \omega_{i,j} + \beta\|\mathbf{P}\|_{2,1} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \left\|\mathbf{P}^T(\boldsymbol{x}^i - \boldsymbol{x}^j)\omega_{i,j}\right\|_2 + \beta\|\mathbf{P}\|_{2,1} \\ &= \left\|\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}\right\|_{2,1} + \beta\|\mathbf{P}\|_{2,1} \end{aligned} \tag{11}$$

where the matrix $\tilde{\mathbf{X}} = [(\boldsymbol{x}^1 - \boldsymbol{x}^2), \ldots, (\boldsymbol{x}^1 - \boldsymbol{x}^N), \ldots, (\boldsymbol{x}^N - \boldsymbol{x}^1), \ldots, (\boldsymbol{x}^N - \boldsymbol{x}^{N-1})]$ and the corresponding coefficient matrix $\tilde{\mathbf{W}} = diag(\omega_{1,2}, \ldots, \omega_{1,N}, \ldots, \omega_{N,1}, \ldots, \omega_{N,N-1})$. Eq. (11) can be further rewritten as:

$$J(\mathbf{P}) = tr(\mathbf{P}^T\tilde{\mathbf{X}}\tilde{\mathbf{W}}^T\mathbf{M}_d\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}) + \beta tr(\mathbf{P}^T\mathbf{M}_r\mathbf{P}) \tag{12}$$

where the affiliated matrices $\mathbf{M}_d$ and $\mathbf{M}_r$ are diagonal matrices with the ith element on the diagonal to be $\frac{1}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i\right\|_2}$ and $\frac{1}{2\|\mathbf{P}_i\|_2}$, respectively. In practice, we add a very small constant ε on the denominator, which avoids the value to be zero. Thus, Eq. (10) can be represented as the following optimization problem:

$$\arg\min_{\mathbf{P}} tr(\mathbf{P}^T(\tilde{\mathbf{X}}\tilde{\mathbf{W}}^T\mathbf{M}_d\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T + \beta\mathbf{M}_r)\mathbf{P}), s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{13}$$

The projection matrix $\mathbf{P}$ can be obtained by solving:

$$(\tilde{\mathbf{X}}\tilde{\mathbf{W}}^T\mathbf{M}_d\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T + \beta\mathbf{M}_r)\boldsymbol{p} = \lambda\boldsymbol{p} \tag{14}$$

where λ is the eigenvalue and $\boldsymbol{p}$ is the corresponding eigenvector. The optimal solution of Eq. (14) is the eigenvectors associated with the first m smallest eigenvalues. The optimal projection matrix is then represented by: $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_m]$. Note that the affiliated matrices $\mathbf{M}_d$ and $\mathbf{M}_r$ are dependent to $\mathbf{P}$ and thus Eq. (14) contains unknown variables. To deal with this problem, we alternatively update the affiliated matrices $\{\mathbf{M}_d, \mathbf{M}_r\}$ by Eq. (12) and calculate the solution $\mathbf{P}$ according to Eq. (14). The convergence of such an iterative algorithm will be proved in the next section.

When the value of $\mathbf{P}$ is fixed, the optimization of matrix $\mathbf{W}$ reduces to minimize:

$$\begin{aligned} &J(\mathbf{W}) = \sum_{i=1}^{N}\left(\sum_{j=1}^{N} \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2 \omega_{i,j} + \alpha\left\|\boldsymbol{\omega}_i\right\|_2^2\right) \\ &s.t. \boldsymbol{\omega}_i^T\mathbf{1} = 1, 0 \le \omega_{i,j} \le 1 (j \in N_i), \omega_{i,j} = 0 (j \notin N_i) \end{aligned} \tag{15}$$

Note that the above problem is independent between each index i. Thus, it can be solved respectively according to various i:

$$\begin{aligned} &J(\boldsymbol{\omega}_i) = \sum_{j=1}^{N} \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2 \omega_{i,j} + \alpha\left\|\boldsymbol{\omega}_i\right\|_2^2 \\ &s.t. \boldsymbol{\omega}_i^T\mathbf{1} = 1, 0 \le \omega_{i,j} \le 1 (j \in N_i), \omega_{i,j} = 0 (j \notin N_i) \end{aligned} \tag{16}$$

We define the vector $\tilde{\boldsymbol{\omega}}_i \in R^k$ that contains all the weights of $\omega_{i,j}$ $(j \in N_i)$ and set the values of other elements $\omega_{i,j}$ in $\boldsymbol{\omega}_i$ to be zero. The problem is converted to the optimization of $\tilde{\boldsymbol{\omega}}_i$, while the objective function can be rewritten as:

$$\begin{aligned} &J(\tilde{\boldsymbol{\omega}}_i) = \sum_{j=1}^{k} (\left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2 \tilde{\omega}_{i,j} + \alpha\tilde{\omega}_{i,j}^2) \\ &s.t. \tilde{\boldsymbol{\omega}}_i^T\mathbf{1} = 1, 0 \le \tilde{\omega}_{i,j} \le 1 \end{aligned} \tag{17}$$

Denote $d_{ij} = \left\|\mathbf{P}^T\boldsymbol{x}^i - \mathbf{P}^T\boldsymbol{x}^j\right\|_2$ in Eq. (17) and vector $\boldsymbol{d}_i \in R^k$ with the jth element as d_{ij}. It can be further represented by the following optimization problem:

$$\arg\min_{\tilde{\boldsymbol{\omega}}_i} \left\|\tilde{\boldsymbol{\omega}}_i + \frac{1}{2\alpha}\boldsymbol{d}_i\right\|_2^2, s.t. \tilde{\boldsymbol{\omega}}_i^T\mathbf{1} = 1, 0 \le \tilde{\omega}_{i,j} \le 1 \tag{18}$$

The Lagrangian function of Eq. (18) can be given as:

$$L(\tilde{\boldsymbol{\omega}}_i, \eta, \boldsymbol{\rho}) = \frac{1}{2}\left\|\tilde{\boldsymbol{\omega}}_i + \frac{1}{2\alpha}\boldsymbol{d}_i\right\|_2^2 - \eta(\tilde{\boldsymbol{\omega}}_i^T\mathbf{1} - 1) - \boldsymbol{\rho}^T\tilde{\boldsymbol{\omega}}_i \tag{19}$$

where η and $\boldsymbol{\rho}$ are the Lagrangian multipliers. We calculate the derivative of (19) with respect to $\tilde{\boldsymbol{\omega}}_i$, and set to be 0:

$$\frac{\partial L(\tilde{\boldsymbol{\omega}}_i, \eta, \boldsymbol{\rho})}{\partial\tilde{\boldsymbol{\omega}}_i} = \tilde{\boldsymbol{\omega}}_i + \frac{1}{2\alpha}\boldsymbol{d}_i - \eta\mathbf{1} - \boldsymbol{\rho} = \mathbf{0} \tag{20}$$

From (20), we can represent the jth element $\tilde{\omega}_{ij}$ in the vector $\tilde{\boldsymbol{\omega}}_i$ as:

$$\tilde{\omega}_{ij} + \frac{1}{2\alpha}d_{ij} - \eta - \rho_j = 0 \tag{21}$$

According to the Karush-Kuhn-Tucker condition [36], we have:

$$\tilde{\omega}_{ij}\rho_j = 0 \tag{22}$$

Combining (21) and (22) together, we can deduce the solution of $\tilde{\omega}_{ij}$ as:

$$\tilde{\omega}_{ij} = \left(\eta - \frac{1}{2\alpha}d_{ij}\right)_+ \tag{23}$$

Utilizing the constraint $\tilde{\boldsymbol{\omega}}_i^T\mathbf{1} = 1$, we can get the following equation:

$$\sum_j \tilde{\omega}_{ij} = \sum_j \left(\eta - \frac{1}{2\alpha}d_{ij}\right)_+ = 1 \tag{24}$$

Define the following function:

$$f(\eta) = \sum_j \left(\eta - \frac{1}{2\alpha}d_{ij}\right)_+ - 1 \tag{25}$$

The optimal solution of Lagrangian multiplier η is the root of $f(\eta) = 0$. Thus, the above problem can be solved by Newton-Raphson method iteratively:

$$\eta^t = \eta^{t-1} - \frac{f(\eta^{t-1})}{f'(\eta^{t-1})} \tag{26}$$

Substituting the solution of η into (23), we can get the optimal solution of weight $\tilde{\omega}_{ij}$. The matrix $\mathbf{W}$ can be further obtained by constituting all the similarity weights. Notice that the value of α steers the adaptive selection of neighboring samples and assigns the corresponding weights automatically.

With the initial value of the two variables, we can optimize the projection matrix $\mathbf{P}$ by minimizing (10). Once this is done, the similarity matrix $\mathbf{W}$ can be optimized by solving (15). We

repeat the iterative procedure to update the above variables until the optimization problem (8) converges to the local minimum. The details of the complete optimization procedure are summarized in Algorithm 1.

Algorithm 1 The complete optimization procedure.

Input: The training data samples $\mathbf{X} \in R^{d\times N}$.

Initialization: Set t=0. Estimate the initial similarity matrix $\mathbf{W}^{(0)}$ by (4) and projection matrix $\mathbf{P}^{(0)}$ by (9).
repeat
  Update t=t+1.
  / fix matrix $\mathbf{W}^{(t-1)}$, update matrix $\mathbf{P}^{(t)}$ */*
  repeat
    Calculate the affiliated matrices $\mathbf{M}_d^{(t)}$ and $\mathbf{M}_r^{(t)}$ in (12).
    Update the projection matrix $\mathbf{P}^{(t)}$ by solving (13).
  until convergence
  / fix matrix $\mathbf{P}^{(t)}$, update matrix $\mathbf{W}^{(t)}$ */*
  for i= 1 to N **do**
    Update the ith row of similarity matrix $\mathbf{W}^{(t)}$ by solving (18).
  end for
  Constitute the similarity weight $\mathbf{W}^{(t)}$.
until convergence

Output: The final projection matrix $\mathbf{P}^*$.

IV. Convergence

In this section, we analyze the convergence of the proposed approach, which demonstrates that the algorithm finally converges to a local optimal solution. In advance, we give the following lemma, which has been proved in [34].

Lemma 1: For any nonzero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, the following inequality holds:

$$\|\boldsymbol{u}\|_2 - \frac{\|\boldsymbol{u}\|_2^2}{2\|\boldsymbol{v}\|_2} \le \|\boldsymbol{v}\|_2 - \frac{\|\boldsymbol{v}\|_2^2}{2\|\boldsymbol{v}\|_2} \tag{27}$$

Theorem 1: The iterative strategy for solving (10) monotonically decreases the value of $J(\mathbf{P})$ during optimization.

Proof: According to Eq. (12), the optimization of matrix (10) can be represented as the following formulation in the lth iteration:

$$\begin{aligned}\mathbf{P}^{(l)} = \arg\min_{\mathbf{P}^T\mathbf{P}=\mathbf{I}} & tr(\mathbf{P}^T\tilde{\mathbf{X}}\tilde{\mathbf{W}}^T\mathbf{M}_d^{(l-1)}\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}) \\ & + \beta tr(\mathbf{P}^T\mathbf{M}_r^{(l-1)}\mathbf{P})\end{aligned} \tag{28}$$

Thus, we have:

$$\begin{aligned}& tr((\mathbf{P}^{(l)})^T\tilde{\mathbf{X}}\tilde{\mathbf{W}}^T\mathbf{M}_d^{(l-1)}\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}^{(l)}) + \beta tr((\mathbf{P}^{(l)})^T\mathbf{M}_r^{(l-1)}\mathbf{P}^{(l)}) \\ & \le tr((\mathbf{P}^{(l-1)})^T\tilde{\mathbf{X}}\tilde{\mathbf{W}}^T\mathbf{M}_d^{(l-1)}\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}^{(l-1)}) \\ & \quad + \beta tr((\mathbf{P}^{(l-1)})^T\mathbf{M}_r^{(l-1)}\mathbf{P}^{(l-1)})\end{aligned} \tag{29}$$

Considering the specific format of affiliated matrices $\mathbf{M}_d^{(l-1)}$ and $\mathbf{M}_r^{(l-1)}$, the above inequality can be rewritten as:

$$\begin{aligned}& \sum_i \frac{\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l)}\right\|_2^2}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2} + \beta\sum_i \frac{\left\|\mathbf{P}_i^{(l)}\right\|_2^2}{2\left\|\mathbf{P}_i^{(l-1)}\right\|_2} \\ & \le \sum_i \frac{\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2^2}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2} + \beta\sum_i \frac{\left\|\mathbf{P}_i^{(l-1)}\right\|_2^2}{2\left\|\mathbf{P}_i^{(l-1)}\right\|_2}\end{aligned} \tag{30}$$

According to Lemma 1, we have:

$$\begin{aligned}& \left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l)}\right\|_2 - \frac{\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l)}\right\|_2^2}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2} \\ & \le \left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2 - \frac{\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2^2}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2}\end{aligned} \tag{31}$$

$$\left\|\mathbf{P}_i^{(l)}\right\|_2 - \frac{\left\|\mathbf{P}_i^{(l)}\right\|_2^2}{2\left\|\mathbf{P}_i^{(l-1)}\right\|_2} \le \left\|\mathbf{P}_i^{(l-1)}\right\|_2 - \frac{\left\|\mathbf{P}_i^{(l-1)}\right\|_2^2}{2\left\|\mathbf{P}_i^{(l-1)}\right\|_2} \tag{32}$$

Considering all the row vectors in the corresponding matrices, we can further deduce:

$$\begin{aligned}& \sum_i \left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l)}\right\|_2 - \sum_i \frac{\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l)}\right\|_2^2}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2} \\ & \le \sum_i \left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2 - \sum_i \frac{\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2^2}{2\left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2}\end{aligned} \tag{33}$$

$$\begin{aligned}& \beta\sum_i \left\|\mathbf{P}_i^{(l)}\right\|_2 - \beta\sum_i \frac{\left\|\mathbf{P}_i^{(l)}\right\|_2^2}{2\left\|\mathbf{P}_i^{(l-1)}\right\|_2} \\ & \le \beta\sum_i \left\|\mathbf{P}_i^{(l-1)}\right\|_2 - \beta\sum_i \frac{\left\|\mathbf{P}_i^{(l-1)}\right\|_2^2}{2\left\|\mathbf{P}_i^{(l-1)}\right\|_2}\end{aligned} \tag{34}$$

By summing inequalities (30), (33) and (34), we get:

$$\begin{aligned}& \sum_i \left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l)}\right\|_2 + \beta\sum_i \left\|\mathbf{P}_i^{(l)}\right\|_2 \\ & \le \sum_i \left\|(\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P})_i^{(l-1)}\right\|_2 + \beta\sum_i \left\|\mathbf{P}_i^{(l-1)}\right\|_2\end{aligned} \tag{35}$$

It can be rewritten as:

$$\left\|\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}^{(l)}\right\|_{2,1} + \beta\left\|\mathbf{P}^{(l)}\right\|_{2,1} \le \left\|\tilde{\mathbf{W}}\tilde{\mathbf{X}}^T\mathbf{P}^{(l-1)}\right\|_{2,1} + \beta\left\|\mathbf{P}^{(l-1)}\right\|_{2,1} \tag{36}$$

According to (11), we can get the conclusion:

$$J(\mathbf{P}^{(l)}) \le J(\mathbf{P}^{(l-1)}) \tag{37}$$

Thus, the convergence of the iterative strategy has been proved.

Theorem 2: Denote the objective function in (8) as $J(\mathbf{P}, \mathbf{W})$. Algorithm 1 monotonically decreases the value of $J(\mathbf{P}, \mathbf{W})$ in each iteration.

Proof: When we fix matrix $\mathbf{W}$ and update $\mathbf{P}$, we have the conclusion $J(\mathbf{P}^{(l)}, \mathbf{W}^{(l-1)}) \le J(\mathbf{P}^{(l-1)}, \mathbf{W}^{(l-1)})$ according to *Theorem 1*. When we fix matrix $\mathbf{P}$ and update $\mathbf{W}$, the optimal solution can be deduced from the convex problem (15). Apparently, we have $J(\mathbf{P}^{(l)}, \mathbf{W}^{(l)}) \le J(\mathbf{P}^{(l)}, \mathbf{W}^{(l-1)})$. Thus, we can get the conclusion $J(\mathbf{P}^{(l)}, \mathbf{W}^{(l)}) \le J(\mathbf{P}^{(l-1)}, \mathbf{W}^{(l-1)})$, which indicates that the objective function will monotonically decrease to the local minimum by conducting Algorithm 1.

V. Experimental results

In this section, we conduct experiments to explore the possibility of detecting AF cases from short clips of face videos. The experiments are performed on the Oulu Bio-Face (OBF) database [12], which was captured at the University of Oulu for healthy participants and Oulu University Hospital for clinical patients[1]. The OBF database contains approximately five-minute video recordings of participants in various statuses, including resting-state/post-exercise sessions for healthy

[1]https://sites.google.com/site/jshiwebpage/af

individuals and prior-treatment/after-treatment sessions for AF patients. In the experiments, we utilize the resting-state recordings of healthy individuals and prior-treatment recordings of patients in the database, which indicate the healthy and AF samples, respectively. For each video sample, we divide it into nonoverlapped short clips of 30 seconds in length. Since the duration of a few videos is slightly less than 5 min, we only employ the first 9 clips of each video to avoid bias. To perform the experiments, we randomly select the video clips from 20 healthy subjects and 20 AF patients to constitute the training set, while collect the clips from other 10 healthy subjects and 10 AF patients as the testing set. The training and testing sets are independently divided by subject. Thus, there is no overlapping subject in the training and testing sets. In total, we have 180 video clips in the testing set and employ 360 video clips as the training set. The experiments are independently repeated 10 times, while the average results are reported as the output.

A. Parameter settings

Several parameters are required to be fixed in the proposed method. We first randomly select one independent validation set from the database to adjust the parameters. To acquire the best performance, we choose the dimension of the embedding feature space as $m = 490$. The regularization parameters α and β in (8) are set to 0.1 and 80, respectively. We will further analyze the influence of these parameters in Section V-I. Actually, the performance is also rather stable when the above parameters vary within a certain range around the optimal values.

B. The pipeline for AF detection

To better illustrate the problem of AF detection, we further present a flowchart in Fig. 7 that explains the whole procedure. For an input video clip, we first detect the 21 ROIs by the connection of specific landmarks. We then utilize three pulse extraction methods to generate the heart beat rhythm from each ROI. Thus, we have 63 pulse rhythm signals in total extracted from various regions by multiple methods. The peak detection procedure is conducted on each pulse signal to produce the RRI signal, which describes the intervals between adjacent heart beats. We extract the basic HRV features from each RRI signal and normalize them as a feature vector. Furthermore, we concatenate all of the 63 HRV feature vectors to represent the characteristic of AF for the input video clip. The trained linear projection matrix (i.e., matrix $\mathbf{P}$) is utilized to conduct the feature fusion and selection procedure, which transforms the extracted HRV features into an embedding space to enhance the discriminability. Based on the discriminative feature, we can finally obtain the binary classification results for the problem of AF detection.

We further provide the running time of each component on an Intel Core i5 3.2 GHz CPU to analyze the efficiency of the proposed method. The procedure of ROIs detection is conducted by the landmark detection algorithm [31]. In the proposed approach, the processing speed is approximately 10 fps. Notice that the speed is highly related to the attributes of the video (e.g., resolution). Additionally, it can be further improved by using more efficient landmark detection methods [32]. The following three procedures are rather efficient. For a 30-second video clip, it takes approximately 3.66 s to extract the HRV feature, 0.6 ms to complete feature fusion and selection from the trained projection matrix (i.e., matrix $\mathbf{P}$), and 2 ms to obtain the final AF detection result from the SVM.

C. Experiments on the OBF database

In this subsection, we evaluate the performance of the proposed approach on the OBF database to illustrate the advantages of the proposed method. To demonstrate the effectiveness for AF detection, we compare the proposed discriminative features with seven other HRV features, which are presented in detail as follows:

- Single ROI feature I (SRF1) [12]: It first extracts the pulse rhythm by [12] from ROI 1 in Table I (ROI 1 contains the whole cheek). Then, it generates the RRI signal and produces the HRV feature. This is also the baseline method presented in [12].
- Single ROI feature II (SRF2): It conducts the same procedure as SRF1 except that it utilizes [13] to extract the pulse rhythm.
- Single ROI feature III (SRF3): It conducts the same procedure as SRF1 except that it utilizes [14] to extract the pulse rhythm.
- All ROIs feature I (ARF1): Different from SRF1 which only extracts pulse rhythm from ROI 1, it extracts signals from all of the 21 ROIs and concatenates all the corresponding HRV feature vectors as the final output.
- All ROIs feature II (ARF2): It conducts the same procedure as ARF1 except that it utilizes [13] to extract the pulse rhythm.
- All ROIs feature III (ARF3): It conducts the same procedure as ARF1 except that it utilizes [14] to extract the pulse rhythm.
- Multi-feature: It concatenates the features of ARF1, ARF2 and ARF3 together.

To better present the advantage of the proposed framework, we also conduct the above feature fusion algorithm to refine ARF1, ARF2 and ARF3. We adjust the optimal parameters to adapt these features while defining the refined results as Fused-ARF1, Fused-ARF2 and Fused-ARF3 for comparison. Finally, we compare the AF detection ability of these features by employing the SVM classifier with RBF kernel. The parameters of the RBF-SVM classifier are tuned in the range of $\left\{2^{-6}, 2^{-5}, \ldots, 2^{5}, 2^{6}\right\}$. To evaluate the experimental results, we calculate the number of True Positives (TP), True Negatives (TN), False Positives (TP) and False Negatives (FN). Then, we utilize the sensitivity ($\frac{TP}{TP+FN}$), specificity ($\frac{TN}{TN+FP}$) and accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) as the validation metrics. Table II further shows the quantitative indexes of all the compared methods, together with the results from one preliminary approach [21]. The advantage of the proposed method is remarkable. As shown in Table II, the sensitivity, specificity and accuracy results of the proposed approach are

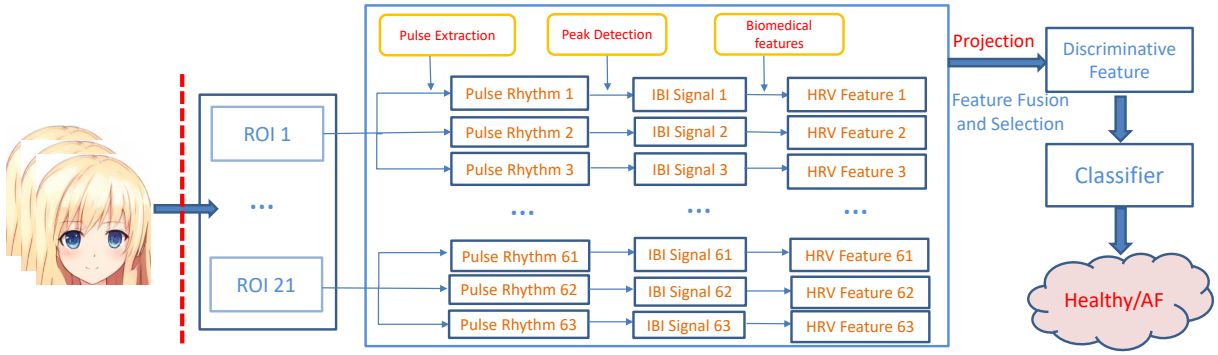


Fig. 7. The flowchart for explaining the problem of AF detection, which presents the whole procedure from the input video clip to the final classification result.

TABLE II
THE COMPARISON OF VARIOUS ALGORITHMS FOR AF DETECTION.

	Sensitivity(%)	Specificity(%)	Accuracy(%)
Couderc et al. [21]	83.22	77.89	80.56
SRF1 [12]	83.56	72.22	77.89
SRF2	87.11	83.56	85.34
SRF3	89.78	81.11	85.45
ARF1	83.22	89.78	86.50
ARF2	89.44	89.22	89.33
ARF3	85.67	92.11	88.89
Multi-feature	87.78	92.78	90.28
Fused-ARF1	88.00	93.22	90.61
Fused-ARF2	89.56	93.22	91.39
Fused-ARF3	89.44	92.67	91.06
Proposed	**91.00**	**94.11**	**92.56**

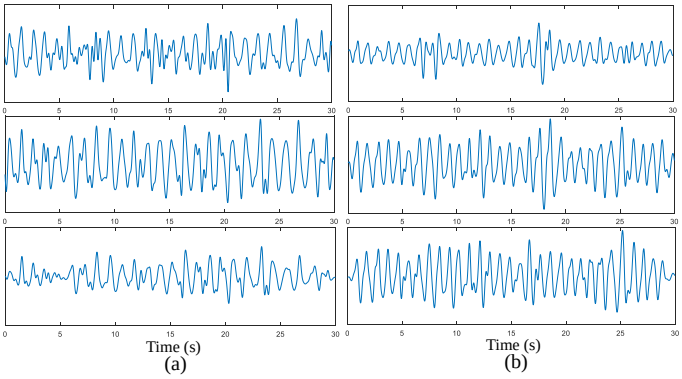


Fig. 8. The extracted pulse rhythms from ROI 1 in some failure cases. From top to bottom: Rhythms extracted by [12], [13] and [14], respectively. (a) One healthy sample, but was incorrectly classified as an AF patient. (b) One AF sample, but was incorrectly classified as a healthy individual.

91.00%, 94.11% and 92.56%, respectively. Compared with the baseline method (SRF1) [12], the proposed method achieves more promising results with an improvement of 14.67% in terms of accuracy value. It also outperforms 2.28% in AF detection accuracy when compared with the second best method (Multi-feature). According to the results in Table II, we can obtain the following conclusions:

• The pulse signals extracted by [13] [14] are more suitable to be utilized in AF detection than the signal obtained by [12].

• The combination of HRV features in multiple ROIs achieves better performance than utilizing the feature from single ROI for AF detection.

• The employment of multiple pulse extraction algorithms improves the results of AF detection since they can complement each other's defects caused by the artificial effects.

• The proposed method conducts a training phase to learn more discriminative features, which achieves the best performance in the compared approaches.

In Fig. 8, we present one failure case for a healthy individual and another failure case for an AF patient. For the case of the healthy individual, the main reason is that the pulse extraction approaches are disturbed by the nonrigid movements or expression variations during recording. Thus, the pulse signals suffer from noises and artificial effects (e.g., the signals extracted by [12] and [14] in Fig. 8), which resemble the characteristics of an AF patient. The failure case of the AF patient occurs when the symptom is not very serious. In this situation, the irregularity of the beat-to-beat variability does not occur or rarely occurs in the 30-second recording, which causes the misclassification of this sample.

D. Influence of gender, age and race

In this subsection, we further conduct experiments to discuss the influence of individual factors (e.g., gender, age and race) on the performance of the proposed algorithm. To analyze the effect of gender, we randomly select the video clips from five male healthy individuals and five male AF patients as the testing samples while utilizing two different training sets (i.e., a specific training set and a general training set) to compare the classification results. Both training sets contain 20 healthy subjects and 20 AF patients. The difference is that the specific training set is only selected from the remaining male samples, while the general training set includes both male and female samples. Thus, the composition of the specific training set excludes the factor of gender that could influence the final results. The above experiments are independently conducted 10 times, which ensures the diversity of the testing sets. Fig. 9(a) presents the average accuracy for samples in each testing set on the specific and general training sets. Although the results obtained in the two training sets are somewhat different, the changes remain within a reasonable range for most cases. Similarly, we also conduct the same strategy to

investigate the influence of race and age. We utilize the same amount of healthy and AF samples in both testing and training sets, but select different kinds of samples according to the factor we want to analyze. To study the effect of race, we only utilize Caucasian individuals to constitute the testing set and specific training set, while the general training set contains volunteers of both Caucasian and other races. For the factor of age, we randomly select the individuals from 35 to 70 years old to construct the testing set and specific training set. The samples in the general training set are chosen from volunteers of all ages. Fig. 9(b) and Fig. 9(c) show the experimental results for the factors of race and age, respectively. According to the average accuracy of each testing set, the variation of the two training sets does not induce considerable change in the final results. Thus, we can conclude that the impact of the above factors is not apparent in the experiments.

E. Influence of various recording conditions

To evaluate the robustness of the proposed approach, we capture additional face videos for 10 healthy individuals and 10 AF patients with a 10-min duration of each video. The 10-min recording is divided into two periods. In both periods, we utilize one Blackmagic camera (Cam 1) and one GoPro camera (Cam 2) to simultaneously capture videos. In the first 5-min recording, the cameras are placed in front of volunteers at a distance of approximately one meter. The videos are recorded under a cold lighting condition. In the second 5-min recording, Cam 1 is still set at a one-meter distance, but Cam 2 is arranged at the distance of approximately two meters. We also adjust the illumination to warm lighting condition. Briefly, we can divide the above captured videos into four testing sets and utilize them to evaluate the proposed algorithm under different recording conditions:

Set 1: Cam 1, cold light, and one meter distance.
Set 2: Cam 2, cold light, and one meter distance.
Set 3: Cam 1, warm light, and one meter distance.
Set 4: Cam 2, warm light, and two meter distance.

We also employ the protocol described previously, which divides the videos into short clips of 30-second lengths. The videos of Set 1 are captured under the same environment as the training set, while the videos from the other three sets are recorded with alternative cameras, various lighting conditions and different distances. In Table III, we present the quantitative indicators for the performance on the four testing sets. The results of Set 1 and Set 2 compare the performance of the proposed method when the videos are captured by different cameras. Different cameras have various imaging characteristics, so the accuracy drops slightly for the AF detection task. A similar situation also exists for the videos captured under different lighting conditions, which can be concluded from the results of Set 1 and Set 3. However, the decline in performance is not considerable under the variations of the camera and lighting, which illustrates that the proposed approach has the capability to obtain stable results. The videos of Set 4 are recorded from a longer distance. The performance decreases when compared with the results of the other three sets. Thus, it is essential to maintain a valid recording distance for remote AF detection.

TABLE III
THE QUANTITATIVE INDICATORS FOR THE TESTING SETS CAPTURED UNDER VARIOUS ENVIRONMENTS

	Sensitivity(%)	Specificity(%)	Accuracy(%)
Set 1	94.44	100.00	97.22
Set 2	92.22	94.44	93.33
Set 3	90.00	100.00	95.00
Set 4	92.22	71.11	81.67

F. Influence of facial ROIs

In this subsection, we discuss the impact of local facial ROI feature detectors on the final AF detection results. To achieve this objective, we respectively remove a couple of landmarks in Fig. 6 (i.e., {1,17}, {2,16},{3,15},{4,14},{5,13},{6,12} and {7,11}) and further evaluate the AF detection accuracy. According to Table I, there are 15 remaining ROIs after the abovementioned landmarks have been discarded. The pulse signals are then extracted on the 15 ROIs and utilized to predict AF by the proposed algorithm. Table IV presents the AF detection results when the experiments are conducted without the corresponding landmarks. The accuracy will be slightly decreased after some landmarks are removed. However, the difference is not so obvious when compared with the original results in Table II, which illustrates that the proposed method is robust to the ROI detector.

We further investigate another situation in which only one ROI is available on the face image. Fig. 10 shows the corresponding accuracy when the experiment is performed on various ROIs (i.e., from ROI 1 to ROI 21). It achieves relatively high accuracy on the ROI 1, ROI 2, ROI 4, ROI 5, ROI 7, ROI 11 and ROI 12 while obtaining the lowest accuracy on ROI 21. According to the experimental results, we can arrive at the following two conclusions. First, it is more suitable to extract the pulse signals from the region of the cheek, while the region of the chin is susceptible to additional disturbance. Second, a relatively large ROI is more robust in achieving a better AF detection result. It alleviates the negative impact of facial landmark misalignment on the tracking procedure through the whole video.

A reduction in the overall performance undoubtedly occurs in the case of degrading facial ROI feature detectors (such as recording the video in a very dark environment). In this case, the facial landmarks cannot be precisely located, which causes the misalignment of the ROI on each frame of the video. Accordingly, the variation in color on the facial skin cannot truly describe the pulse of heart activity. Given that this paper is only a primary study on the possibility of AF detection from face videos, we capture the frontal face video under a specific lighting condition to explore the AF detection problem without a medical acquisition device. Moreover, it is also interesting to consider conducting further studies on applications in the natural environment.

G. Comparison with deep learning-based method

We further conduct an additional experiment to compare the proposed approach with the deep learning-based method. The extracted pulse signals can be considered as time-based

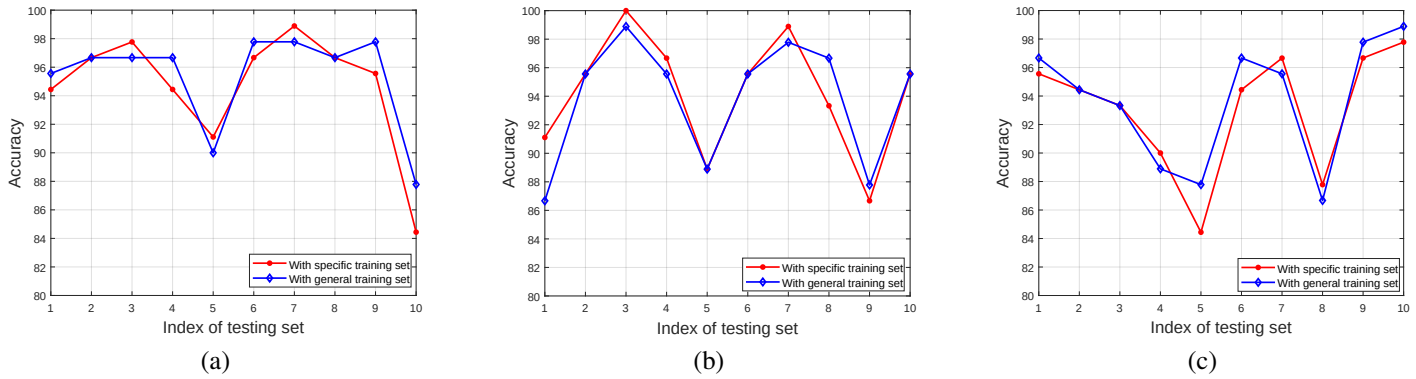


Fig. 9. Comparison on the factors of gender, race and age. (a) Gender. (b) Race. (c) Age.

TABLE IV
THE AF DETECTION RESULTS WHEN SOME FACIAL LANDMARKS ARE REMOVED

w/o landmarks	{1,17}	{2,16}	{3,15}	{4,14}	{5,13}	{6,12}	{7,11}
Sensitivity(%)	90.67	90.44	90.89	91.00	90.56	90.56	90.44
Specificity(%)	92.00	92.22	93.67	93.11	92.44	92.44	93.33
Accuracy(%)	91.33	91.33	92.27	92.05	91.50	91.50	91.89

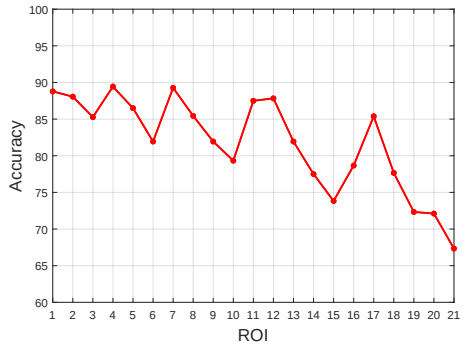


Fig. 10. The average AF detection accuracy on each ROI.

sequences. We can feed the sequences into an LSTM-based deep network to predict the final results. The deep architecture is presented in Fig. 11. It packages up three LSTM layers with 256 hidden neurons to form a stacked LSTM, which is followed by one fully-connected layer with an ReLU activation and one fully-connected layer with the sigmoid. The binary cross entropy is utilized as the loss function. As discussed, the pulse rhythms are extracted by three different methods on 21 ROIs. Therefore, we can obtain a 63-dimensional feature from one frame to describe the pulse status and consider it as a time-based sequence for the entire video. We feed the sequence into the LSTM and predict the category of a healthy individual or an AF patient. We utilize the same training/testing protocols as the experiments in Section V-C. We employ PyTorch for implementation and use Adam as the optimizer. The batch size is set to 64. The learning rate starts from 0.001 and is divided by 10 after 20 epochs until the loss is steady. The average sensitivity, specificity and accuracy results of the LSTM-based method are 71.12%, 71.55% and 71.34%, respectively. The performance is inferior to the quantitative indicators of the compared approaches in Table II. One reason might be that noisy samples exist in the training set, which induces bias since the deep network has a superior capability to fit the training data. The small scale of the dataset also limits the learning capacity of the deep network.

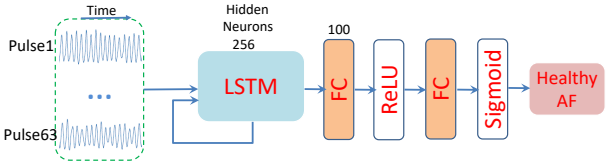


Fig. 11. The architecture of the compared deep network.

H. Performance on the live streaming videos

Previously, we considered each 30-second video clip as the testing sample while obtaining the experimental results with the proposed method. In this subsection, we further evaluate the AF detection performance on live streaming videos. To achieve this objective, we maintain the same training protocol to learn the projection matrix $\mathbf{P}$, but consistently evaluate the testing video every 5 s during playing. The experimental results are presented in Fig. 12. As shown in the figure, the AF detection results become more stable along with the increase in the video length. The performance of the average accuracy tends to be improved from 56.28% to 92.56% during the playing of the whole video clip. The experiment also demonstrates that it is important to employ face videos with enough length to achieve the AF detection task.

I. Parameter Sensitivity Analysis

In this subsection, we study the robustness of important parameters in the proposed method, i.e., the dimension m of the embedding space, the regularization parameters α and β. We evaluate the sensitivity of parameters by analyzing the average accuracy on the 10-fold experiments. As shown in Section V-A, the optimal values for parameters are set to $m = 490$, $\alpha = 0.1$ and $\beta = 80$, respectively.

Fig. 13 presents the results on the OBF database with the variation of parameter m. As shown in Fig. 13(a), the label of the x-axis represents the dimension of the low-dimensional

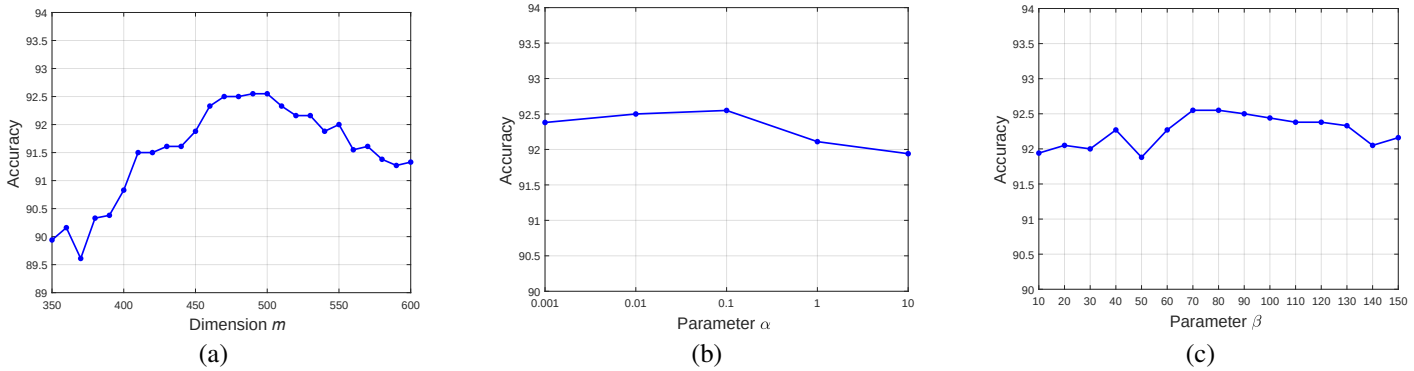


Fig. 13. The performance of the proposed AF detection algorithm with different values of dimension m, parameter α and β. (a) Dimension m. (b) Parameter α. (c) Parameter β.

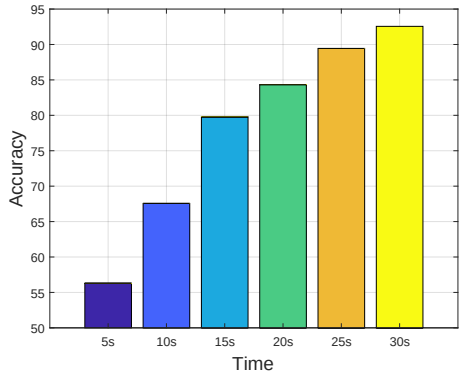


Fig. 12. AF detection performance during the playing of the video clip.

space, and the y-axis denotes the accuracy with respect to different dimensions. We can observe that the proposed method obtains the best performance when the parameter m is around 490. When the dimension is too high, the discriminative feature contains considerable redundancy and noise, which can cause the decrease of accuracy for AF detection. When the dimension is too low, we cannot obtain enough discriminative features to describe the characteristics of AF. Thus, the performance of AF detection will drop dramatically. The proposed algorithm achieves stable results when parameter m varies within a certain range. The experimental results show that the proposed algorithm is insensitive to parameter m.

Next, we present the result of AF detection accuracy with the variation of parameter α. As discussed in Section III-B, parameter α automatically controls the number of adaptive neighbors in the construction of the local similarity graph. For each training sample, a small value of α enforces it to be connected by less neighbors in the graph, while a large value of α assigns more neighbors with nonzero weights. The experimental results are shown in Fig. 13(b). The proposed method obtains the highest accuracy when parameter α is set to 0.1. However, we can also obtain reasonable performance with other selections of α, which shows that the result is not sensitive to the parameter. To further demonstrate the efficiency for the strategy of adaptively adjusting the local similarity graph, we revise the proposed method with a fixed similarity graph as the initialization in Eq. (4) for comparison. The experimental results are shown in Table V. The proposed method achieves an improvement of 0.73% in accuracy, which illustrates the advantage in the usage of the adaptive similarity graph.

TABLE V
THE COMPARISON OF AF DETECTION RESULTS FOR UTILIZING THE FIXED SIMILARITY GRAPH AND THE ADAPTIVE STRATEGY.

Graph	Sensitivity(%)	Specificity(%)	Accuracy(%)
Fixed	90.78	92.89	91.83
Adaptive	**91.00**	**94.11**	**92.56**

We further investigate the effect of parameter β on the AF detection accuracy. The results are shown in Fig. 13(c). The regularization parameter β determines the row sparsity property for the projection matrix $\mathbf{P}$. A large value of β induces more rows in matrix $\mathbf{P}$ to be filled with elements of all zeros, which means that more HRV features will be considered as redundant dimensions and eliminated in the feature embedding procedure. Combining the requirement of discarding redundant features and reserving discriminative features, we should determine the value of β as a moderate value for practical applications. According to the experimental results, the proposed method obtains superior performance when the value of parameter β is around 80. We also observe that the experimental results remain stable during the variation in parameter β within a large range. It indicates that the feature selection term is efficient in improving the performance of the proposed AF detection approach.

J. Discussion

To further present the advantage of the proposed approach, we compare the accuracy of AF detection with or without the proposed feature fusion and selection procedure. The results for all the 10 trials are shown in Fig. 14, which indicate that the feature fusion and selection procedure consistently helps to improve the performance of the AF detection task. Once the projection matrix $\mathbf{P}$ is obtained, we can observe that some rows of matrix $\mathbf{P}$ are filled with elements of all zeros. It means that the corresponding redundant features will

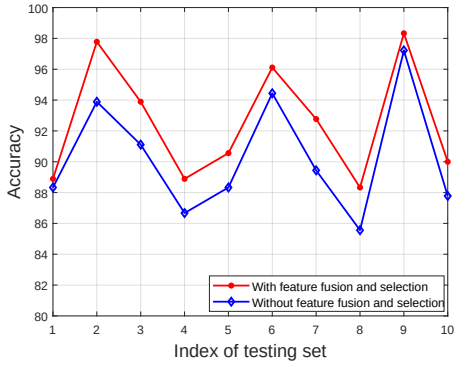
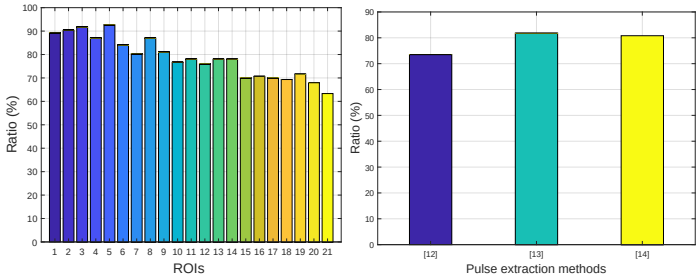


Fig. 14. Comparison for the advantage of feature fusion and selection.



(a) (b)

Fig. 15. Discussion about the importance of the following attributes in the AF detection. (a) ROIs. (b) Pulse extraction methods.

be ignored in the feature fusion and selection procedure. In the proposed method, we capture HRV pulse signals from 21 ROIs by three different pulse extraction methods. Thus, we want to further investigate the importance of these attributes (i.e., various ROIs and pulse extraction algorithms) in the AF detection task. We utilize the property of projection matrix $\mathbf{P}$ to further illustrate this problem. For each attribute, we count the number of selected features according to $\mathbf{P}$ and calculate the corresponding ratio to represent the importance. The average results of all the experiments are shown in Fig. 15. For example, each ROI completely corresponds to 30-dimensional HRV features that are extracted from three pulse signals. The result in Fig. 15(a) shows that 89.33% of features from ROI 1 are selected to construct the discriminative feature, which is much higher than the ratio of 63.33% from ROI 21. According to Fig. 15, we can obtain the following conclusions:

• The ROIs with a larger area are more suitable for extracting pulse signals in AF detection since such ROIs are less sensitive to the influence of misalignment and noise.

• The pulse extraction algorithms [13] [14] exhibit superior ability compared with the method [12]. Similar experimental results are also demonstrated in Table II.

In the above sections, we have shown many experimental results and discussions to analyze the performance of the proposed algorithm. Since this paper presents a preliminary study to achieve AF detection using a learning-based method by fusing multiple subtle changes, there are still some open issues that could be improved to promote the study. We list three points to facilitate future work as follows:

• We conduct the experiments on the OBF dataset to evaluate the performance of AF detection. We also change the recording conditions (e.g., different cameras, lighting conditions and recording distances) to verify the robustness. It is expected that more challenging datasets under various environments will be collected to promote the development of AF detection.

• In the work, we mainly aim to extract and refine discriminative features from multiple subtle signals for AF detection. It is also expected that a more sophisticated classifier will be designed to further improve the performance.

• It is also desirable to analyze subtle facial changes for performing micro-expression recognition and lie detection in the future.

VI. Conclusion

In this paper, we aim to perform an innovative study that exploits the capability of AF detection by fusing specific characteristics in human face videos. Recently, the development of the video-extracted pulse extraction approach is ongoing in a preliminary stage, which cannot measure cardiac activity as accurately as ECG signals do. To improve this problem, the proposed method divides the whole face into multiple ROIs and extracts pulse rhythms by various methods from each ROI. We further combine the HRV features in multiple pulse rhythms by conducting a robust feature fusion method that simultaneously eliminates the influence of outliers and enhances the discriminability between healthy/AF samples. The experimental results illustrate that the proposed method produces superior results in comparison with other baselines. Overall, we demonstrate the possibility of detecting AF risk from remote face videos, which is a promising research topic in real-world applications. In the future, it is also anticipated that other diseases associated with arrhythmia can be similarly diagnosed.

References

[1] M. Baumert, P. Sanders, and A. Ganesan, "Quantitative-electrogram-based methods for guiding catheter ablation in atrial fibrillation," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 416–431, 2016.

[2] A. Brost, A. Wimmer, R. Liao, and et al., "Constrained registration for motion compensation in atrial fibrillation ablation procedures," *IEEE Trans. on medical imaging*, vol. 31, no. 4, pp. 870–881, 2012.

[3] M. W. Krueger, G. Seemann, K. Rhode, and et al., "Personalization of atrial anatomy and electrophysiology as a basis for clinical modeling of radio-frequency ablation of atrial fibrillation," *IEEE Trans. on medical imaging*, vol. 32, no. 1, pp. 73–84, 2013.

[4] Y. Zheng, D. Yang, M. John, and D. Comaniciu, "Multi-part modeling and segmentation of left atrium in c-arm ct for image-guided ablation of atrial fibrillation," *IEEE Trans. on medical imaging*, vol. 33, no. 2, pp. 318–331, 2014.

[5] G. D. Clifford, C. Liu, B. Moody, and et al., "AF classification from a short single lead ecg recording: The physionet computing in cardiology challenge 2017," *Computing in Cardiology*, vol. 44, 2017.

[6] S. Asgari, A. Mehrnia, and M. Moussavi, "Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine," *Computers in biology and medicine*, vol. 60, pp. 132–142, 2015.

[7] J. Lee, Y. Nam, D. D. McManus, and K. H. Chon, "Time-varying coherence function for atrial fibrillation detection." *IEEE Trans. Biomed. Engineering*, vol. 60, no. 10, pp. 2783–2793, 2013.

[8] F. Andreotti, O. Carr, M. A. Pimentel, A. Mahdi, and M. De Vos, "Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ecg," *Computing in Cardiology*, pp. 1–4, 2017.

[9] A. G. Bonomi, F. Schipper, L. M. Eerikäinen, and et al., "Atrial fibrillation detection using photo-plethysmography and acceleration data at the wrist," in *Proceedings of Computing in Cardiology*. IEEE, 2016, pp. 277–280.
[10] J. Lee, B. A. Reyes, D. D. McManus, and et al., "Atrial fibrillation detection using an iphone 4s," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 203–206, 2013.
[11] O. Lahdenoja, T. Hurnanen, Z. Iftikhar, S. Nieminen, T. Knuutila, A. Saraste, T. Kiviniemi, T. Vasankari, J. Airaksinen, M. Pänkäälä *et al.*, "Atrial fibrillation detection via accelerometer and gyroscope of a smartphone," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 108–118, 2017.
[12] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, and G. Zhao, "The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 242–249.
[13] L. Feng, L.-M. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 879–891, 2015.
[14] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation." *IEEE Trans. Biomed. Engineering*, vol. 63, no. 9, pp. 1974–1984, 2016.
[15] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in non-contact, multiparameter physiological measurements using a webcam," *IEEE Trans. on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2011.
[16] W. Wang, S. Stuijk, and G. De Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg." *IEEE Trans. Biomed. Engineering*, vol. 62, no. 2, pp. 415–425, 2015.
[17] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.
[18] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4264–4271.
[19] S. Fallet, "Signal processing techniques for cardiovascular monitoring applications using conventional and video-based photoplethysmography," *EPFL thesis*, 2018.
[20] B. Yan, W. Lai, C. Chan, and et al., "Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals," *Journal of the American Heart Association*, vol. 7, no. 8, 2018.
[21] J.-P. Couderc, S. Kyal, L. K. Mestha, and et al., "Detection of atrial fibrillation using contactless facial video monitoring," *Heart Rhythm*, vol. 12, no. 1, pp. 195–201, 2015.
[22] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1299–1313, 2009.
[23] Y. Huang, D. Xu, and F. Nie, "Patch distribution compatible semisupervised dimension reduction for face and human gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 479–488, 2012.
[24] L. An, Z. Qin, X. Chen, and S. Yang, "Multi-level common space learning for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1777–1787, 2018.
[25] W. Wang, Y. Yan, F. Nie, S. Yan, and N. Sebe, "Flexible manifold learning with optimal graph for image and video representation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2664–2675, 2018.
[26] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on grassmann manifold with application to video based face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 140–149.
[27] Z. Huang, R. Wang, X. Li, W. Liu, S. Shan, L. Van Gool, and X. Chen, "Geometry-aware similarity learning on spd manifolds for visual recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2513–2523, 2018.
[28] X. Zhu, H. Suk, S.-W. Lee, and D. Shen, "Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification," *IEEE Trans. on Biomedical Engineering*, vol. 63, no. 3, pp. 607 – 618, 2016.
[29] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1921–1932, 2010.
[30] R. Zhang, F. Nie, and X. Li, "Regularized class-specific subspace classifier," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2738–2747, 2017.
[31] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.
[32] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.
[33] A. Camm, M. Malik, J. Bigger, and et al., "Heart rate variability: standards of measurement, physiological interpretation and clinical use." *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.
[34] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.
[35] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM trans. on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[36] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

Jingang Shi received the B.S. degree and Ph.D. degree both from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China. Since 2017, he has been a postdoctoral researcher at the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His current research interests mainly include image restoration, face analysis and biomedical signal processing.

Iman Alikhani is currently working toward the Ph.D. degree in the Center for Machine Vision and Signal Analysis, University of Oulu. His research interests focus on biomedical signal processing and analysis.

Xiaobai Li received her B.Sc degree in Psychology from Peking University, M.Sc degree in Biophysics from the Chinese Academy of Science, and Ph.D. degree in Computer Science from University of Oulu. She is currently a postdoctoral researcher in the Center for Machine Vision and Signal Analysis of University of Oulu. Her research of interests include spontaneous vs. posed facial expression comparison, micro-expression and deceitful behaviors, and heart rate measurement from facial videos.

Zitong Yu is currently a Ph.D. candidate in the Center for Machine Vision and Signal Analysis, University of Oulu. His research interests focus on remote photoplethysmograph measurement and face anti-spoofing.

Tapio Seppänen received the MSc degree in electrical engineering and the DSc degree in computer engineering from the University of Oulu, Finland, in 1985 and 1990, respectively. He is currently a professor of biomedical engineering at the University of Oulu, where he is involved in teaching and conducting research on biomedical signal processing and multimedia signal processing. His research interests include cardiovascular signal processing, respiratory signal processing, EEG signal processing, and affective computing.

Guoying Zhao is currently a professor with the School of Information and Technology, Northwest University, China and a professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She has authored or co-authored more than 200 papers in journals and conferences. Her papers have currently over 9900 citations in Google Scholar (h-index 44). Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, dynamic-texture recognition, human motion analysis, and person identification. Dr. Zhao was a Co-Chair of many International Workshops at ECCV, ICCV, CVPR, ACCV and BMVC. She is co-publicity chair for FG2018, has served as Area Chairs for several conferences. Currently, she is Associate Editor for *Pattern Recognition*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *Image and Vision Computing* Journals.