ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

Biometric recognition using wearable devices in real-life settings

Emanuela Piciucco[a,**], Elena Di Lascio[b], Emanuele Maiorana[a], Silvia Santini[b], Patrizio Campisi[a]

[a]*Università degli Studi Roma Tre, Via Vito Volterra 62, 00146 Rome, Italy*
[b]*Università della Svizzera italiana (USI), Via Giuseppe Buffi 13, 6900 Lugano, Switzerland*

ABSTRACT

The popularity of wearable devices, such as smart glasses, chestbands, and wristbands, is nowadays rapidly growing, thanks to the fact that they can be used to track physical activity and monitor users' health. Recently, researchers have proposed to exploit their capability to collect physiological signals for enabling automatic user recognition. Wearable devices inherently provide the means for detecting their unauthorized usage, or for being used as front-end in biometric recognition systems controlling the access to either physical or virtual locations and services. The present work evaluates the feasibility of performing biometric recognition using signals captured by wearable devices, considering data collected through off-the-shelf commercial wristbands, and comparing recordings taken during two distinct sessions separated by an average time of 7 days. In more detail, recognition is performed leveraging on electrodermal activity (EDA) and blood volume pulse (BVP), considering measurements taken from 17 subjects performing natural activities such as attending or teaching lectures. Several tests have been carried out to determine the most effective representation of the considered EDA and BVP signals, as well as the most suitable classifier. The best recognition performance has been achieved exploiting convolutional neural networks to extract discriminative characteristics from the combined spectrograms of the employed EDA and BVP data, guaranteeing average correct identification rate of 98.58% for test samples lasting 30 seconds.

1. Introduction

In recent years, the use of smart wearable devices (SWD) is becoming increasingly popular. Their principal use consists in monitoring the health of the user from parameters such as the heart rate, thus acting as assistants to control physical activity, and help achieving fitness goals (Hill, 2015). However, the capability of SWDs in acquiring physiological signals could be also exploited within the framework of biometric recognition systems, using the recorded data to discriminate between legitimate and unauthorized subjects (Chaki et al., 2019; Blasco et al., 2016).

The use of physiological signals for people recognition is commonly referred to as *cognitive biometrics* (Revett and de Magalhães, 2010). Exploiting these traits to recognize individuals offers several advantages compared to the exploitation of traditional physical or behavioural identifiers such as fingerprint, face, iris, or signature. First, being physiological signals not easy to be captured at a distance with conventional acquisition devices, and reflecting the mental and emotional states of an individual, they are extremely difficult to steal and replicate, making spoofing attacks almost impossible to be implemented (Revett and de Magalhães, 2010). Moreover, physiological signals are able to inherently provide liveness detection, in addition to uniqueness and universality. They also allow to perform continuous user recognition, thus preventing session hijacking, and avoiding unauthorized access to information or services after a successful recognition. Eventually, biosignals can be acquired in a non-invasive way, making the acquisition procedure convenient, and the system user friendly.

In the present study, the recognition performance achievable when exploiting two biosignals, namely electrodermal activity (EDA) and blood volume pulse (BVP), as biometric identifiers, is investigated. Electrodermal activity, also known as galvanic skin response (GSR), reflects changes in the behavior of ec-

[**] Please cite this work as: Emanuela Piciucco, Elena Di Lascio, Emanuele Maiorana, Silvia Santini, Patrizio Campisi, "Biometric recognition using wearable devices in real-life settings," Pattern Recognition Letters, Vol. 146, pp: 260-266, June 2021. Digital Object Identifier 10.1016/j.patrec.2021.03.020.

crine sweat glands, directly controlled by the sympathetic nervous system (SNS) (Boucsein, 2012). It is typically measured by placing two electrodes on the skin, and connecting them to a voltage amplifier. EDA has been widely used as measure of physiological arousal, and as a proxy for emotions and stress (Schmidt et al., 2019). EDA is characterised by peaks, also known as skin conductance responses (SCRs), in correspondence to external stimuli (Boucsein, 2012). EDA responses vary across individuals, depending also on demographic characteristics such as gender, age and culture (Fowles et al., 1981). This interpersonal variability in the EDA responses can be actually exploited for the automatic identification of different subjects (Bianco and Napoletano, 2019). The blood volume pulse describes the changes in the peripheral blood volume due to vasodilation and vasoconstriction (Peper et al., 2010). It is commonly measured by resorting to photoplethysmography (PPG) techniques, using a pulse oximeter to illuminate the skin and measure light absorption. Volume changes in the microvascular bed of tissue results in variations of the reflected light, with the recorded signals therefore providing information about cardiovascular activities such as heart rate (HR) and heart rate variability (HRV) (Peper et al., 2010). Characteristics of the BVP signal depends on the activity of the Parasympathetic and SNS (Sancho et al., 2018). Similarly to EDA, features extracted from the BVP can be used as proxy for stress, cognitive load and affect (Schmidt et al., 2019). It has been shown that a large inter-individual variability in the physiological characteristics of the heart as the heart's mass and orientation, the orientation and position of the myocardium and the shape of the torso exist (Agrafioti et al., 2011). Such differences have made the cardiovascular activity's measurements suitable for biometric recognition (Agrafioti et al., 2011; Ekiz et al., 2020).

The performed experimental evaluation has been explicitly designed to obtain a preliminary evidence that promising recognition performance can be achieved exploiting physiological data acquired using a commercial, off-the-shelf wristband. Furthermore, in order to perform a proper analysis of the discriminative capabilities of the considered biometric traits, signals recorded during two distinct acquisition sessions have been compared in the performed tests. In order to achieve such aims, we have exploited a longitudinal database comprising samples taken from 17 different subjects attending or teaching lectures, therefore taking into account a natural setting representing practical working conditions. Physiological signals have been captured, for each subject, during two different acquisition sessions, separated by an average period of 7 days. A multimodal approach, jointly using EDA and BVP to perform recognition, is here proposed, resorting to feature-level fusion to improve the performance of the proposed system. Discriminative representations of the employed signals are obtained considering representations in the time and time-frequency domains, fed to both shallow classifiers and deep-learning-based approaches to perform user identification. The effects of adopting different window sizes are also investigated, demonstrating that the proposed approach is able to guarantee good recognition performance even with test samples lasting only $10s$.

The state of the art on biometric recognition using SWDs is outlined in Section 2. The employed database is described in Section 3, while the performed signal processing is presented in Section 4. The adopted classification strategies are then detailed in Section 5, with the obtained results discussed in Section 6. Finally, conclusions, including limitations and possibilities for further advancements, are drawn in Sections 7 and 8.

2. Related Work

Although the use of wearable devices became increasingly common in recent years, the field of biometric recognition based on physiological signals recorded by SWDs is still underexplored. An overview of relevant related works on biometric recognition using SWD is given in Table 1, where the achieved performance, the kind of considered wearable devices, the experimental setting, and the time needed for recognition, are summarized. More in detail, when considering the experimental settings employed in literature, we refer to three different categories: laboratory (L), field (F), i.e., real-life scenarios in completely unconstrained environments, and field with constraint (FC), with subjects having constraints in terms of movements and environment (Schmidt et al., 2019). In the present study, we take into account an FC scenario, since signals are recorded from students and teachers participating to lectures taking place in a room.

The first relevant study on the effectiveness of biosignals recorded through wearable devices to perform biometric recognition has been proposed by Cornelius *et al.* (Cornelius et al., 2014), taking bioimpedence into account. Tests on eight subjects have achieved 98% accuracy, when comparing signals collected during a single-day usage. A deep-learning-based approach has been proposed by Everson *et al.* (Everson et al., 2018), exploiting a database of 12 subjects whose PPG signals have been recorded during physical activity. The collected temporal data have been fed into a framework consisting of two convolution neural networks (CNN), in conjunction with two long short-term memory (LSTM) units, and followed by a dense output layer, achieving a 96% recognition accuracy. Luque *et al.* (Luque et al., 2018) have exploited a dense neural network (DNN) classifier for PPG-based biometric recognition, achieving classification with an area under curve (AUC) of 0.78 and 0.83 for the two considered databases. In (Ekiz et al., 2020), the authors used wristbands to collect data from 28 subjects. Heart rate variability (HRV), derived from PPG, has been employed to authenticate users through features extracted in the frequency domain, and machine learning techniques have been used for classification. The best reported performance correspond to a correct identification rate (CIR) at 98.48%, and an equal error rate (EER) at 3.96%. All the aforementioned studies have performed recognition using a single physiological signal, and considering data recorded during a single session.

Multimodal biometric recognition has been instead proposed in Blasco *et al.* (Blasco and Peris-Lopez, 2018), where signals acquired through several SWDs, including PPG, ECG, EDA, and accelerometer (ACC), have been employed for biometric recognition. The authors have built their own low-cost wearable sensors, that have been used to capture data from 25 subjects while walking or being seated, at either resting state and after

Table 1: Overview of the state-of-the-art approaches for biometric recognition using physiological signals and wearable devices.

Reference	Signals	Database		Performance	Wristband	Experimental settings	Time needed for recognition
		Subjects	Sessions				
(Cornelius et al., 2014)	Bioimpedence	8	1	CIR = 97.8% EER = 12.7%	✓	F	15s
(Everson et al., 2018)	PPG	12	1	CIR = 96.00%	✓	L	Not specified
(Luque et al., 2018)	PPG	43 20	1 1	AUC = 78.2% AUC = 83.2%	✗	L	3s 1s
(Ekiz et al., 2020)	HRV	28	1	EER = 3.96% CIR = 98.48%	✓	FC	120s
(Blasco and Peris-Lopez, 2018)	PPG, ECG, ACC, EDA	25	1	AUC = 99.00% EER = 2.00%	✓	L	2s
(Bianco and Napoletano, 2019)	HR, BR, EDA, PER-EDA	37	1	CIR = 88.74%	✗	L	60s
(Alonso et al., 2016)	SPO2, AF, ECG, EMG, EDA,Temp	25	1	CIR = 92%	✗	L	Not specified
(Alemán-Soler et al., 2016)	EDA, EMG, ECG	18	1	CIR = 85.55%	✗	L	Not specified
(Byeon and Kwak, 2019)	ECG	211 99	2 1	CIR = 98.99% CIR = 94.03%	✗	L	Not specified
(Sancho et al., 2018)	PPG	42 56 24 24	1 2 3 3	EER = 1.0% EER = 8.0 - 21.5% EER = 6.6 - 23.2% EER = 6.0 - 20.5%	✗ ✗ ✗ ✗	L	Not specified
(Vhaduri and Poellabauer, 2019)	step count, HR, calorie burn, MET	400	17 months	CIR = 90 - 93% EER = 5%	✓	F (activity-dependent)	300s
(Vhaduri and Poellabauer, 2017)	step count, HR, calorie burn, MET	421	2 years	CIR = 92.97%	✓	F (activity-dependent)	300s

a gentle stroll. Features have been extracted exploiting the discrete Fourier and Walsh–Hadamard transforms, and then compared using Gaussian models, obtaining an EER = 2%. Multimodal physiological signals have been employed to perform biometric recognition also in (Bianco and Napoletano, 2019), considering breathing rate (BR), HR, palm electrodermal activity (P-EDA), and perinasal perspitation (PER-EDA). Classification approaches consisting of a CNN with mono-dimensional kernels, and inputs represented as windows of the raw signals stacked along the channel dimension have been employed. A database with 37 subjects, acquired during a controlled experiment on a driving simulator, has been collected and used to reach a top accuracy of 88.74%. In (Alonso et al., 2016) the authors collected signals acquired from 25 people. A combination of principal components analysis (PCA) and support vector machines (SVMs) is applied to identify people using ECG, EDA, airflow (AF), temperature (Temp), pulse oximetry (SPO2), and electromyogram (EMG). The testing results have achieved a correct identification rate of 92%. Alemán-Soler *et al.* (Alemán-Soler et al., 2016) have presented an approach to use different biomedical signals, namely EMG, ECG, and EDA, in order to perform biometric identification. Several statistical parameters have been used as features, performing classification with a neural network, achieving a CIR of 85.55%.

Although the aforementioned works have conducted interesting studies regarding the joint usage of multiple physiological signals to perform biometric recognition, all of them suffer from a notable flaw, that is, only single-session databases have been used for the experiments. Under this scenario, the estimated performance may depend more on session-specific recording conditions than on individual characteristics of the involved subjects (Maiorana et al., 2015). Furthermore, signals have been always acquired in laboratory conditions, which are unlikely to reflect real-world situations, preventing the findings to be robust to the typical noise of natural scenarios.

To the best of our knowledge, very few studies have taken into account the permanence of the considered physiological signals, properly performing tests by recording and comparing data acquired through SWDs during multiple acquisition sessions. Byeon *et al.* have evaluated electrocardiogram (ECG) biometrics using pre-configured models of convolutional neural networks, such as VGGNet, ResNet, DenseNet and Xception, with various time-frequency representations, namely spectrogram, log spectrogram, mel spectrogram, and scalogram (Byeon and Kwak, 2019). Two different databases, one composed of two sessions, and the other one with data captured during a single session, have been there considered, achieving a correct identification rate (CIR) of 98.99% and 94.04%, respectively. A long-term feasibility study on the use of PPG signals as biometric trait has been performed in (Sancho et al., 2018). Several feature extractors, based on the time domain and the Karhunen–Loève transform, and matching metrics, including Manhattan and Euclidean distances, have been tested using four different databases. The achieved equal error rates (EERs) range from 1.0% to 8.0% when a single session is used, and from 19.1% to 23.2% when signals from different sessions are compared. Despite the good results obtained in the aforementioned studies, they present some limitations. First of all, the performed experiments have been carried out in laboratory settings. More importantly, medical devices have been there employed, with such acquisition modalities being hard to be replicated in practical applications, due both to the required costs and to the user inconvenience. In this regard, it is worth remarking that the present study has been instead conducted considering the multi-session database presented in (Di Lascio et al., 2018), collected in a real-life scenario while the involved subjects attend lectures performing natural activities such as listening, talking, making gestures, and taking notes. Furthermore, a commercial off-the-shelf wristband has been employed for data collection, thus allowing to considering signals properly representing practical working conditions.

Commercial devices have been also used in the longitudinal studies performed in (Vhaduri and Poellabauer, 2019) and (Vhaduri and Poellabauer, 2017). In more detail, three types of biometric identifiers, namely step count, HR, calorie burn, and metabolic equivalent of task (MET), have been acquired through a Fitbit wearable device, and later used for user recognition, in (Vhaduri and Poellabauer, 2019). Recordings from over 400 users have been acquired in a 17-month long health study, and used to achieve an average recognition accuracy at about 93%, and an EER at 5%. A similar approach has been also investigated in (Vhaduri and Poellabauer, 2017), where an analysis on 421 Fitbit users has been carried out for two years, achieving an average recognition accuracy at 92.97%. In both studies, statistical features have been used to discriminate users, and SVMs exploited for user classification. People taking part to the experiments wore the Fitbit SWDs all day long, yet biometric recognition has been performed according to an activity-dependent modality, recognizing users only when involved in specific tasks. Moreover, the time required to perform recognition is in the order of five minutes, which seems overly long to be practically considered in real-life applications. Conversely, the proposed work focuses on keeping low the required recognition time, with promising recognition results achieved while resorting to query samples lasting only $10s$.

3. Employed Database

In this paper we investigate the feasibility of using physiological signals gathered with commercial wristbands to automatically identify several users. In more detail, the employed database, collected in natural environments, for multiple days, and with wrist-worn devices, has been presented in (Di Lascio et al., 2018), where it has been employed to evaluate students' engagement during lectures. The considered database contains EDA and BVP data from 33 healthy participants (24 students and 9 instructors), collected during 41 actual lectures in classroom from four courses over a period of three weeks. The average duration of a lecture is about 43 minutes. Physiological data have been collected using the unobtrusive off-the-shelf E4 wristband (Garbarino et al., 2014), that participants wore during the lectures. For further details regarding the data collection procedure we refer to (Di Lascio et al., 2018).

To perform the analysis presented in this paper, we have selected a subset of the available data. Specifically, only participants for which EDA and BVP samples have been recorded during at least two different days have been considered. For each subject, we have employed signals lasting at least 30 minutes, and taken from four randomly-selected lectures (two in one day, and two in another). We refer to data collected in the same day for each participant as *session*. The average distance between two sessions of the same subject is 7 days. Following the aforementioned criteria, we have employed data recorded during a total of 34 unique sessions, with samples belonging to 17 subjects, including four instructors.

4. Template generation

The employed representations of EDA and BVP signals have been obtained by first applying the pre-processing procedures commonly adopted in literature (Boucsein, 2012; Kalimeri and Saitis, 2016; Greco et al., 2015; Zhang et al., 2018).

In particular, we have filtered-out noise from the EDA traces, sampled at 4Hz, using the Butterworth low-pass filter with a 0.4Hz cut-off frequency, similarly to (Kalimeri and Saitis, 2016). We have then decomposed the EDA signal, to which we refer as *EDA-mixed*, in its *phasic* and *tonic* components, using the Python implementation[1] of the convex optimisation approach proposed by Greco *et al.* (Greco et al., 2015). The two components differ in time resolution: the *phasic* component is characterised by fluctuations in response to stimuli at time resolution of seconds, while the *tonic* component changes at scale of minutes, and provides information about the trend of the signal (Boucsein, 2012; Branković, 2012). Different combinations of the computed mixed, phasic, and tonic components of the EDA signals have been employed in the performed tests, with the configurations providing the best recognition results in each considered scenario detailed in Section 6.

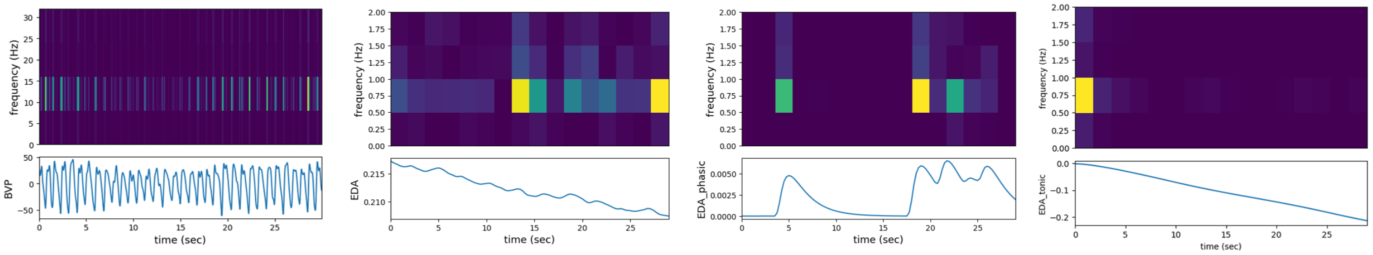
As for BVP signals, we have removed high-frequency noise from the original data, sampled at 64 Hz, using the first order Butterworth filter with a cut-off frequency of 5 Hz, similarly to (Zhang et al., 2018).

To guarantee the same amount of samples for all the subjects, we have selected the central 30 minutes of each lecture. We have then segmented the physiological signals into overlapping frames, using a sliding window approach with a window size W and an overlap O. A single frame is considered as either a training sample, or a recognition probe, in the performed tests. We have used time windows with lengths $W = \{10, 20, 30\}$ seconds. Time windows larger than $30s$ have not been considered due to the inconvenience a long recognition time can cause to the user. On the other hand, the minimum windows size has been set to $10s$ to be able to acquire enough discriminative information. An O =75% overlap factor between consecutive frames is also employed to generate a number of samples allowing to properly train the employed CNNs.

4.1. Feature Extraction and Fusion

In the performed tests, we have exploited representations of EDA and BVP physiological signals in both time and time-frequency domains. Regarding the latter one, we have resorted to spectrograms, providing two-dimensional representations of the frequency content over time. The computed spectrograms are based on the short-time Fourier transform (STFT), dividing the considered signals into continuous short segments, and applying the Fourier transform to each of them (Kehtarnavaz, 2008). The STFT is expressed mathematically as:

[1] https://github.com/lciti/cvxEDA



(a) (b) (c) (d)

Fig. 1: Examples of spectrograms of the 30 seconds segments of BVP signal (a), *EDA-mixed* (b), *phasic* (c), and *tonic* (d) component of the EDA .

$$X_{STFT}[m,n] = \sum_{k=0}^{N-1} x[k]w[k-m]e^{-j2\pi nk/N} \qquad (1)$$

where $x[k]$ represents the signal and $w[k]$ the window of N points (Kehtarnavaz, 2008). The spectrogram is an intensity plot, and each row represents the variation of the power spectral density (PSD) of the signal, corresponding to the magnitude squared of the STFT, over time. Figure 1 presents an example of the spectrogram of 30 seconds segments of the BVP, the *EDA-mixed*, and the *phasic* component of the EDA.

The considered EDA and BVP components are employed, either individually or in a combined form, as inputs to several classifiers, as detailed in the following section.

5. Employed Classifiers

In order to perform an exhaustive set of experimental tests, we have evaluated the effectiveness of both shallow classifiers and deep learning approaches. The employed standard machine learning algorithms are detailed in Section 5.1. The considered deep learning approaches are detailed in Section 5.2, where the used CNN architectures are outlined, and in Section 5.3, where the recurrent neural networks (RNNs) applied to the available temporal signals are presented.

5.1. Shallow Classifiers

Representations based on the spectrograms of the EDA and BVP signals have been used as inputs to the considered shallow classifiers. In more detail, when performing recognition using only the components derived from the EDA signals, all of them, that is, the mixed, phasic, and tonic components, are jointly used as input to the employed algorithms. The spectrograms of the three components are in fact concatenated to create a one-dimensional feature vector, representative of an EDA frame. On the other hand, in order to maximize the achievable recognition performance, on an empyrical basis we have used only the mixed and phasic components of the EDA when jointly exploiting both EDA and BVP data, performing also in this case a feature-level fusion of the available spectrogram features.

The employed feature representations have been normalized using z-score before applying the used shallow classifiers, in order to speed up the learning process. A feature selection process is also carried out before feeding the employed representations to the considered shallow classifiers. Specifically, in each considered scenario, an ANOVA test has been performed on each feature, with the resulting F-scores providing information regarding the separation of distributions belonging to different subjects. Tests have been then done for an increasing number of features, sorted for decreasing F-scores, used as inputs to the employed classifiers.

Three different shallow classifiers, relying on distinct strategies to perform classification, have been employed in the performed tests. In more detail, support vector machine (SVM) has been chosen as representative of parametric classifiers (Bishop, 2006). Its purpose is to find the optimal hyperplane that allows to correctly separate training data belonging to different classes, maximizing the margin between the decision boundaries and the samples deemed most difficult to classify, that is, the support vectors. A one-versus-all (OvA) design has been employed in the performed tests to apply the binary SVM approach to a multi-class scenario. Non-parametric standard machine learning algorithms, which do not require to make any assumption on the distributions of the treated data, have been also exploited, resorting to random forest (RF) and gradient boosting (GB) approaches (Friedman et al., 2009). Both RF and GB rely on ensembles of weak classifiers, that is, decision trees in our case. Yet, while RF exploits a bagging approach, creating an ensemble of independent decision trees trained on different subsets of the available training data, GB instead performs an incremental learning, sequentially creating decision trees based on inputs depending on the outcomes of the previously generated predictors. In the performed tests, the XGBoost implementation has been used for the employed ensemble classifiers (Chen and Guestrin, 2016).

5.2. Convolutional Neural Networks

As done when using the considered shallow classifiers, spectrograms are used as inputs of the empoyed CNNs. Specifically, also in this case all the three components of the EDA signals have been employed when using only EDA to perform recognition, while only the mixed and phasic EDA components have been used when jointly exploiting EDA and BVP data. For each considered scenario, the spectrograms of the considered components are arranged as different planes of three-dimensional tensors, thus again resorting to feature-level fusion. The created structures are then fed to well-known CNNs proposed in literature to perform object classification on RGB images, that is, the VGG-16 (Simonyan and Zisserman, 2014) and MobileNetV2 (Sandler et al., 2018) CNN architectures, after being resized to comply with the input requirements of each network.

5.2.1. *VGG-16*

VGG-Net has a simple structure, and for this reason it is widely used. In details, it consists of 13 convolutional layers, 5 pooling layers, and 3 fully-connected layers, the last one followed by a softmax classifier. All hidden layers are equipped with a rectification (ReLU) non-linearity. The input of the first convolutional layer has a size of $224\times224\times3$, and a small 3×3 filter size is employed for kernels.

We have initialized VGG-16 weights with those estimated for an image classification task over Imagenet (Russakovsky et al., 2015). The layers have been then fine-tuned using a cross-entropy loss function for back-propagation, with stochastic gradient descent (SGD) and a batch size of 16. Learning rate has been set to 0.001, with momentum at 0.9. The maximum number of training epoch is set to 100, with early stopping in case the validation loss is minimized.

5.2.2. *MobileNetV2 Architecture*

MobileNetV2 is a neural network architecture that runs very efficiently on mobile devices, making the proposed application feasible for a real scenario where wearable devices are coupled with a smartphone. MobileNetV2 builds upon the ideas from MobileNetV1 (Howard et al., 2017), using depthwise separable convolution, and a pointwise convolution replacing the full convolutional operator. However, MobileNetV2 introduces two new features to the architecture: linear bottlenecks between the layers, and shortcut connections between the bottlenecks. The architecture of MobileNetV2 contains the initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers. The input of the first convolutional layer has size $224 \times 224 \times 3$. ReLU is used as non-linearity, together with a kernel size 3×3. Dropout and batch normalization are utilized during training.

As done for VGG-16, also MobileNetV2 has been initialized with the weights estimated over Imagenet, with fine-tuning performed using a cross-entropy loss function and SGD with momentum. The maximum number of training epoch is set to 100 also for MobileNetV2.

5.3. *Recurrent Neural Networks*

Tests have been also performed using the temporal behavior of EDA and BVP signals as inputs to RNNs. Long short-term memory (LSTM) networks, the most-widely employed kind of RNN architecture, have been used for this purpose. Specifically, signals created combining, at feature level, the temporal components of EDA and BVP data, are fed to networks comprising a bi-directional LSTM with 1300 hidden states, followed by a dropout layer with dropout probability equal to 40%, and a fully connected layer with a softmax as loss function.

6. Results and Discussion

In order to estimate the achievable recognition performance, in terms of correct identification rate (CIR), we have used for each subject's enrolment the data from the first session (the first 20% for validation, and the remaining 80% for training), while samples belonging to the second session have been reserved for testing. The validation samples have been used for tuning the hyper-parameters of the shallow classifiers using a grid search approach. Specifically, we have chosen the hyper-parameters optimizing the CIR achievable on the validation set, and then used them to perform the final training process on the whole first session data. As for the hyper-parameters of the employed CNNs, we have leveraged on those of the pre-trained networks, and used the validation set for early stopping only. For each subject, a value of CIR is computed. In order to obtain the overall performance of the system, the average of the performance obtained for each subject is taken into account, with the standard deviation used as indicator of the stability of the achieved results.

Table 2 shows the results, in terms of best CIR for different sets of employed features, obtained for different combinations of employed signals, shallow classifiers, and time window durations. The obtained results show that the considered non-parametric classifiers are more efficient than the parametric one. Specifically, RF typically performs better than SVM. More importantly, the best results are generally achieved when exploiting the GB learning approach, testifying that the incremental boosting strategy adopted in GB fits the available data better than a bagging approach such as the one employed in RF. This could imply that the available data are characterized by a limited amount of noise, thus minimizing possible overfitting risks (Friedman et al., 2009). The best results obtained with shallow classifiers correspond to a CIR of 93.82% when using the fusion of the spectrograms of the EDA and BVP components, computed on window of length $W = 20$s, as input to the GB classifier. The employed hyper-parameters are a learning rate of 0.15, 120 estimators, and a maximum depth equal to 6. The GB algorithm has also shown better stability, expressed in terms of smaller standard deviation, in comparison to the other standard classifiers.

A decrease of recognition performance is typically observed with a decrease of the window size. However, when combined representations of EDA and BVP features are used as inputs to GB classifiers, the obtained recognition rates remain pretty stable, guaranteeing good performance also when exploiting very short segments.

The recognition rates achieved when training the considered shallow classifiers over combined representations of BVP and EDA features are typically better than those achieved when exploiting individual modalities. This confirms the usefulness of the proposed approach relying on multiple sources to perform recognition. Some exceptions to this general behavior can be found, when using SVM or RF classifiers, in cases of large discrepancies between the results achieved using separately EDA or BVP features.

Table 3 shows the performance, in terms of correct identification rate, obtained when exploiting the considered deep learning approaches, that is, CNNs relying on VGG-16 and MobileNetV2, and LSTM networks. While these latter achieve performance comparable to the best shallow classifier, i.e., GB, the employed CNNs outperform both the consider standard machine learning algorithms and LSTM. The CNNs are also generally more stable, in terms of performance standard deviation, compared to the shallow classifiers. In more detail, the

Table 2: Performance of the considered shallow classifiers in terms of CIR, reported as mean ± standard deviation. Best performance in bold.

Signal	SVM			RF			GB		
	$W = 30s$	$W = 20s$	$W = 10s$	$W = 30s$	$W = 20s$	$W = 10s$	$W = 30s$	$W = 20s$	$W = 10s$
EDA	28.37 ±26.37%	23.14 ±23.86%	12.60 ±12.32%	58.83 ±17.46%	48.94 ±17.87%	35.31 ±17.91%	91.01 ± 6.52%	91.83 ± 5.67%	91.65 ± 5.74%
BVP	63.03 ±17.01%	47.55 ±18.29%	16.14 ±9.26%	79.68 ±8.97%	67.70 ±10.00%	35.90 ±17.68%	86.69 ±7.30%	86.72 ±6.86%	86.62 ±7.25%
Fusion	56.89 ±21.67%	67.10 ±15.96%	45.63 ±18.47%	73.68 ±12.53%	59.94 ±16.26%	44.14 ±16.15%	93.55 ±4.29%	**93.82 ± 4.19%**	93.30 ±4.12%

Table 3: Performance of the employed deep learning approaches in terms of CIR, reported as mean ± standard deviation. Best performance in bold.

Signal	VGG-16			MobileNet v2			LSTM		
	$W = 30s$	$W = 20s$	$W = 10s$	$W = 30s$	$W = 20s$	$W = 10s$	$W = 30s$	$W = 20s$	$W = 10s$
EDA	86.18 ± 11.26%	91.71 ± 8.28%	90.14 ± 7.07%	92.69 ± 5.56%	94.91 ± 3.90 %	94.31 ± 4.43%	59.00 ± 23.15%	59.02 ± 21.35%	58.88 ± 19.78%
BVP	96.29 ± 3.10%	94.90 ± 3.45 %	80.19 ± 9.68%	96.23 ± 2.86%	95.12 ± 3.27 %	91.46 ± 4.10%	92.42 ± 5.83%	92.91 ± 6.11%	92.44 ± 5.83%
Fusion	98.13 ± 2.09%	96.83 ± 2.35%	97.62 ± 1.81 %	**98.58±1.49**%	97.66 ± 2.06%	97.28 ± 2.46%	92.84 ± 5.95%	93.30 ± 5.85%	91.03 ± 5.46%

best recognition results are obtained when resorting to time-frequency features extracted from a window of size $W = 30s$, and using them as input to CNNs. CIR at 98.13% for the VGG-16, and 98.58% in case of MobileNetv2 are obtained [2]. In general, this latter architecture guarantees the best recognition performance. It is worth remarking that MobileNetV2 has been designed to be optimized for mobile devices, thus entailing the possibility of its use in real-life scenarios, where the recognition procedure is carried out by a smartphone connected to a wearable device, or by the device itself.

Differently from what observed when using shallow classifiers, the behavior of deep learning approaches when varying the length of the employed inputs is less predictable. In fact, although recognition performance generally worsens when shortening the employed time window, it also happens that the best results for EDA are achieved for the shortest inputs. This may be due to the larger number of training samples available when resorting to smaller time windows, and to the capability of the employed networks to effectively exploit such greater amount of data for achieving improvements in recognition performance. The possibility of achieving good recognition performance with time windows as short as $10s$ implies the feasibility to design highly-performing recognition systems requiring an acceptable recognition time.

It is worth mentioning that, in the performed experimental tests, also fusion at the score level has been employed to combine the information from BVP and EDA signals. Nonetheless, the best recognition results have been achieved when resorting to feature-level fusion, which has been therefore reported. Such behavior demonstrates that the employed CNNs are not only able to provide classification accuracies better than those achievable through shallow classifiers, yet they are also able to effectively exploit joint representations of EDA and BVP signals. Better results are obtained when CNNs are trained over combined representations of the employed data rather than performing separate training over disjoint representations, and then fusing the produced output scores.

7. Limitations of the present work

Despite the presented promising results, further research is needed to perform an in-depth analysis about the effectiveness of using physiological signals collected through SWDs to perform biometric recognition.

First of all, the present study has been conducted on a limited set of subjects. It would be therefore important to collect signals using commercial SWDs from a larger number of users, and exploit such data to conduct further analysis.

Moreover, although the performed study demonstrates the existence of discriminative characteristics in EDA and BVP signals collected in two different days separated by a week, multiple recording sessions performed at increasing time distances from the first one should be considered in order to further speculate about the permanence of the employed traits. The availability of multiple acquisition sessions could be also exploited to evaluate the possibility of improving the achievable recognition rates, employing for instance samples acquired during different enrolment sessions, to be able to collect more information regarding the variability of the employed data and then designing template update strategies.

It has also to be mentioned that, although the employed data have been collected asking participants to attend, or to teach, lectures still behaving as they would have normally done, the specific classroom settings might have limited the range of possible movements, with a possible impact on collected recordings and on the obtained recognition performance. Therefore, for future developments, it would be interesting to investigate the role of different operative scenarios on the achievable recognition performance, evaluating field conditions in completely unconstrained real-life environments.

Furthermore, in this work we have focused our analysis only on user identification, while verification scenarios could be also considered in future investigations.

Lastly, it is worth mentioning that, in the performed tests, we have trained the employed models using, for each subject, data recorded during two lectures, for a total of 60 minutes. Even though the proposed approach has guaranteed good recognition performance while using very short identification probes, the amount of time needed for the user's enrolment in a real scenario might be too long. Therefore, an evaluation of the effects

[2] The trained model is available at `https://github.com/emapici/wearable-biometrics-cnn`

of shortening the enrolment acquisition duration on the recognition performance could be beneficial to assess the feasibility this system in real-life applications.

8. Conclusions

In this work we have evaluated the feasibility of identifying subjects exploiting physiological signals gathered with off-the-shelf wearable devices, collecting data in practical conditions. Using the fusion of the EDA and BVP spectrograms as input to a MobileNet-v2 network, we have achieved an average CIR at 98.58%, comparing samples taken from 17 subjects at a time distance of a week. We have also verified the superiority of deep learning models, based on CNNs, over shallow classifiers to achieve higher recognition performance. We have also analyzed the impact of the selected window length in the recognition performance, showing that our approach guarantees good recognition performance even when only 10-second identification probes are used. The methods presented in this study could be integrated into wearable devices for enabling a fast and reliable user identification, and preventing unauthorized usage of these devices. Although additional research, using databases involving a high number of subjects and comprising multiple sessions, as well as evaluating proper training strategies to derive feature representations usable within verification scenarios, is needed to further speculate on the drawn conclusions, the present study represents the first evidence that physiological signals collected through commercial SWDs could be employed to perform biometric recognition while normally carrying out real-life activities.

References

Agrafioti, F., Gao, J., Hatzinakos, D., Yang, J., 2011. Heart Biometrics: Theory, Methods and Applications, in: Biometrics. InTech Shanghai, China, pp. 199–216.

Alemán-Soler, N.M., Travieso, C.M., Guerra-Segura, E., Alonso, J.B., Dutta, M.K., Singh, A., 2016. Biometric approach based on physiological human signals, in: 2016 3rd Int. Conf. on Signal Processing and Integrated Networks (SPIN), IEEE. pp. 681–686.

Alonso, A.D.D., Travieso, C.M., Alonso, J.B., Dutta, M.K., Singh, A., 2016. Biometric personal identification system using biomedical sensors, in: 2016 2nd Int. Conf. on Communication Control and Intelligent Systems (CCIS), IEEE. pp. 104–109.

Bianco, S., Napoletano, P., 2019. Biometric recognition using multimodal physiological signals. IEEE Access 7, 83581–83588.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. springer.

Blasco, J., Chen, T.M., Tapiador, J., Peris-Lopez, P., 2016. A survey of wearable biometric recognition systems. ACM Computing Surveys (CSUR) 49, 1–35.

Blasco, J., Peris-Lopez, P., 2018. On the feasibility of low-cost wearable sensors for multi-modal biometric verification. Sensors 18, 2782.

Boucsein, W., 2012. Electrodermal activity. Springer Science & Business Media.

Branković, S., 2012. Assessment of Brain Monoaminergic Signaling Through Mathematical Modeling of Skin Conductance Response. CM Contreras (Ed.), Neuroscience–Dealing with Frontiers , 83–108.

Byeon, Y.H., Kwak, K.C., 2019. Pre-configured deep convolutional neural networks with various time-frequency representations for biometrics from ecg signals. Applied Sciences 9, 4810.

Chaki, J., Dey, N., Shi, F., Sherratt, R.S., 2019. Pattern mining approaches used in sensor-based biometric recognition: a review. IEEE Sensors Journal 19, 3569–3580.

Chen, T., Guestrin, C., 2016. Xgboost: A Scalable Tree Boosting System, in: Proceedings of the Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Cornelius, C., Peterson, R., Skinner, J., Halter, R., Kotz, D., 2014. A wearable system that knows who wears it, in: 12th annua.int. conf. on Mobile systems, applications, and services, pp. 55–67.

Di Lascio, E., Gashi, S., Santini, S., 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 1–21.

Ekiz, D., Can, Y.S., Dardagan, Y.C., Ersoy, C., 2020. Can a smartband be used for continuous implicit authentication in real life. IEEE Access .

Everson, L., Biswas, D., Panwar, M., Rodopoulos, D., Acharyya, A., Kim, C.H., Van Hoof, C., Konijnenburg, M., Van Helleputte, N., 2018. Biometricnet: Deep learning based biometric identification using wrist-worn ppg, in: 2018 IEEE Int. Symposium on Circuits and Systems, IEEE. pp. 1–5.

Fowles, D.C., Christie, M.J., Edelberg, R., Grings, W.W., Lykken, D.T., Venables, P.H., 1981. Publication Recommendations for Electrodermal Measurements. Psychophysiology 18, 232–239.

Friedman, J., Tibshirani, R., Hastie, T., 2009. The Elements of Statistical Learning. 2nd edition ed., Springer.

Garbarino, M., Lai, M., Bender, D., Picard, R.W., Tognetti, S., 2014. Empatica E3—A Wearable Wireless Multi-sensor Device for Real-time Computerized Biofeedback and Data Acquisition, in: 2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH), IEEE. pp. 39–42.

Greco, A., Valenza, G., Lanata, A., Scilingo, E.P., Citi, L., 2015. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. IEEE Transactions on Biomedical Engineering 63, 797–804.

Hill, C., 2015. Wearables–the future of biometric technology? Biometric Technology Today 2015, 5–9.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

Kalimeri, K., Saitis, C., 2016. Exploring multimodal biosignal features for stress detection during indoor mobility, in: Proceedings of the 18th ACM international conference on multimodal interaction, pp. 53–60.

Kehtarnavaz, N., 2008. Frequency Domain Processing (chapter 7).

Luque, J., Cortes, G., Segura, C., Maravilla, A., Esteban, J., Fabregat, J., 2018. End-to-end photoplethysmography (ppg) based biometric authentication by using convolutional neural networks, in: 2018 26th European Signal Processing Conference (EUSIPCO), IEEE. pp. 538–542.

Maiorana, E., La Rocca, D., Campisi, P., 2015. On the permanence of eeg signals for biometric recognition. IEEE Transactions on Information Forensics and Security 11, 163–175.

Peper, E., Shaffer, F., Lin, I.M., 2010. Garbage In; Garbage out—Identify Blood Volume Pulse (BVP) Artifacts Before Analyzing and Interpreting BVP, Blood Volume Pulse Amplitude, and Heart Rate/Respiratory Sinus Arrhythmia Data. Biofeedback 38, 19–23.

Revett, K., de Magalhães, S.T., 2010. Cognitive biometrics: Challenges for the future, in: International Conference on Global Security, Safety, and Sustainability, Springer. pp. 79–86.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. journal of computer vision 115, 211–252.

Sancho, J., Alesanco, Á., García, J., 2018. Biometric authentication using the ppg: a long-term feasibility study. Sensors 18, 1525.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: IEEE Conf. on computer vision and pattern recognition, pp. 4510–4520.

Schmidt, P., Reiss, A., Dürichen, R., Laerhoven, K.V., 2019. Wearable-based affect recognition—a review. Sensors 19, 4079.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Vhaduri, S., Poellabauer, C., 2017. Wearable device user authentication using physiological and behavioral metrics, in: IEEE 28th Annual Int. Symposium on Personal, Indoor, and Mobile Radio Communications, IEEE. pp. 1–6.

Vhaduri, S., Poellabauer, C., 2019. Multi-modal biometric-based implicit authentication of wearable device users. IEEE Transactions on Information Forensics and Security 14, 3116–3125.

Zhang, X., Lyu, Y., Luo, X., Zhang, J., Yu, C., Yin, H., Shi, Y., 2018. Touch sense: Touch screen based mental stress sense. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 1–18.