Person identification using biometric markers from footsteps sound

M. Umair Bin Altaf, Taras Butko, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

umair, taras.butko, juang@ece.gatech.edu

Abstract

Human identification using unobtrusive methods is a challenging problem that has many applications in surveillance tasks. In this work we propose a set of biometric features extracted from a footsteps audio signal that can be used to identify a person. Instead of using short-time spectral domain, Teager-Kaiser energy operator is employed to transform a time-domain signal into a representational domain where the intrinsic properties of the walking style of each individual become apparent. We show that these features are biometrically significant as they are physical correlates of human gait. The experimental results obtained using a recently recorded publicly available database show prominent results in a human identification task.

Index Terms: Temporal modeling, Teager-Kaiser operator, Biometric in footsteps sound, Acoustic gait recognition.

1. Introduction

A biometric is a physical characteristic of an individual that can be used for person identification and verification. A person possesses many biometrics and a few have been used to identify an individual such as finger prints, facial characteristics, palm prints, retinal images, speech, etc. Gait is one such characteristic which is unique to a person [1] and it is attractive as a biometric as it does not need the cooperation of the person and is, arguably, not easy to fake in the standard surveillance scenarios. It is also less intrusive. Nevertheless, it is a challenging problem in signal processing to use human gait as a biometric, even though human subjects have been shown to be capable in identifying people using gait only [2].

The biometric gait characteristics can be modeled with features derived from spectral, spatio-temporal, kinematic, or kinetic domains. Spectral flux, harmonicity and cepstrum are the spectral features, which are mainly derived from the short-time Fourier power analysis (STFA) of the signal. Spatio-temporal features include average walking velocity, stride length, step length, step time, cadence, stance phase time, swing phase time, single support (when only one foot is in contact with the floor), double support (when both feet are in contact with the floor), and stride width. Kinematic features are derived from the study of angles between the ankle, hip and knee joints. Finally, kinetic analysis examines moments, energy and power at these joints.

These features can be extracted from various modalities. By far the most widely used modality has been video [3, 4, 5], especially under the auspices of DARPA’s HumanID project [6]. In some scenarios seismic signals are used to model and identify a person’s gait [7, 8]. Audio has also shown promise in this task [9], which is called acoustic gait recognition (AGR), and has the advantage of being less intrusive, less expensive and more amenable to nocturnal surveillance tasks. Tanaka and Inoue [10] use the STFA to extract the frequency and pitch of a footstep while [11, 12] use MFCCs, short-time spectral features designed for speech recognition, spectral envelope and append walking intervals. Spectral features designed for speech and music processing are also used in [13, 14] for AGR. An acoustic Doppler radar [15] has been used to identify kinematic features of the gait. In [16], the authors use the spectro-temporal modulation features while in [17] the structural information, i.e., footsteps frequency and footsteps signal energy is used to identify the person and detect whether that person is going up or down on a staircase.

As one might notice, the emphasis in AGR has been on using short-time spectral features. This is not surprising as conventional modeling of audio processing is dominated by the short-time Fourier-spectral perspective in which one views the significant audio features as arising out of a short-time linear Fourier power spectrum analyzer. In audio signal processing this framework endures, in part, because the human ear is considered as a frequency analyzer [18].

In the acoustic domain, a gait comprises the sound of successive steps of a person. In addition to identifying a person, acoustic information can be used to derive high-level intelligence such as the type of walk–slow, fast or random, its direction. A change in a person’s gait may indicate a pathological condition. It also encodes the information about the number of persons walking simultaneously. Information needed to derive these intelligent inferences is not captured by Short-time spectral features.

In this paper we identify novel features from the structure of a footsteps waveform which are physically correlated with the individual walking style. They are derived from temporal and energy measurements. The idea of using temporal and energy measurements is not new and the walking intervals and energy has been used before [11, 17, 12] but walking intervals also depend on the walking speed, which reduces their usefulness as a biometric feature. The features that we derive from these measurements show good discriminative ability for identifying persons, are independent of walking speed by construction, and are better suited to extracting information from footsteps sound.

2. Database

The databases available for AGR are sparse and sporadic. Almost all studies on AGR that we have cited in the previous section use proprietary databases. In the databases for acoustic events detection, a task related to AGR, such as UPC [19], BBC sound series [20], and FBK [21], the footsteps sound is treated as one of many acoustic events. Under DARPA’s humanID project, many research groups recorded more than dozen databases for gait recognition but all of them were video only [5]. A recent database that specifically addresses the footsteps sound for person identification is TUM Gait from Audio, Image and Depth...
(GAID) database [14] which includes the audio modality, in addition to video and depth. The audio includes four channel recordings. The annotations for each individual footstep are not provided.

For the current study, we recorded a new publicly available database of footsteps sounds. The recordings were performed in a low-reverberant room schematically shown in Figure 1. The room reverberation time, $RT_{60}$, is about 200ms. In total, sixteen microphones were employed for recordings: eight cardioid and eight omni-directional attached in pairs to the walls of the room. Additionally, one video camera with fish-eye lens was used to facilitate the annotation process. The audio signals are provided in pcm format, 24 bit resolution, and 48 kHz sampling frequency. The difference between the peaks of footsteps and the ambient noise in the recordings is around 20dB.

Ten sessions corresponding to ten participating persons (1 female and 9 males) were recorded. During each session the person was asked to walk around the recording laboratory at his/her natural speed with each person making 15 rounds clockwise and 15 rounds anti-clockwise. For convenience, as shown in Figure 1, the room was divided into 12 zones so each individual footstep of the person may take place in a particular zone.

The database was annotated by two annotators using the ELAN annotation tool [22] that allows creating, editing, visualizing and searching annotations for video and audio data. The annotation file includes the starting time of each footstep sound together with one of 24 possible text labels: L01, L02, . . . , L12, R01, R02, . . . , R12. Each label corresponds to the zone in which the footstep occurred and the foot (left or right) which produced the sound. The text labels were manually assigned based on information from the video camera.

3. Biometrics in a footstep sound

A sample footstep acoustic event is shown in Figure 2a. It consists of repeated impulses and each impulse is followed by a gradual decay, collectively called the peak. Each peak is produced by an interaction of a foot or shoe with the ground surface. Within this interaction, the structure and shape of different parts of the sole of the foot—the heel, the toe and the middle—will affect this peak. Moreover, it is the series of such roughly similar peaks, irregularly separated and interspersed with background sounds that gives rise to the perceptual quality of a footstep. These successive footsteps are produced by either the left or right foot, with subtle differences between the two.

Short-time spectral features derived from STFA would fail to account for these perceptual cues in at least two ways. The rapid temporal changes cannot be represented accurately due to the time-frequency uncertainty principle. Secondly, the short time interval is not long enough to account for the long-term (> 100 msec) repeatability. Increasing the frame-size to account for this will lead to unreliable estimates due to non-stationarity of the signal at such time-scales. Using the states of hidden Markov model (HMM) to model the temporal sequence of the events does not adequately model the duration of such events as the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed independent of the data distribution [23].

3.1. Teager-Kaiser energy operator

We want to find a representation domain of acoustic signal where the footsteps of each person preserve the prominent characteristics that we have outlined before. Since these characteristics require fine temporal precision and measurements of energy over long-term, we propose to extract the temporal envelope of time waveform. A temporal envelope is a smoothed instantaneous energy representation of the signal.

The Teager-Kaiser energy operator (TKEO) [24] provides an unconventional perspective on the instantaneous energy of a signal. It relates energy to square of the signal amplitude and the square of its frequency. The discrete instantaneous energy,
$x_{\text{TKEO}}[n]$ of a signal $x[n]$ given by TKEO is:

$$x_{\text{TKEO}}[n] = x^2[n] - x[n + 1]x[n - 1] \quad (1)$$

For comparison, we extract the Hilbert envelope of the signal and smooth it with a 20Hz low pass filter to obtain the smoothed envelope [25]. In Figures 2b and 2c, smoothed Hilbert and smoothed TKEO energy envelope, respectively, of the footsteps sound in Figure 2a. TKEO provides the most crisply representative of the sound, as the peaks are clearly delineated. This is because it includes instantaneous frequency, in addition to signal amplitude, in the energy estimate — combining the local and long-term measurement in a single estimate.

### 3.2. Biometric Features

In the TKEO domain in Figure 2c, we can notice several prominent characteristics of a footsteps acoustic signal which are easy to identify when compared to Figures 2a and 2b. These characteristics are correlated with the physical properties of the walking person. We identify them below:

**Similarity of the successive footsteps.** We can observe that subsequent footsteps look similar. The same conclusion can not be done by just looking at time domain signal.

**The presence of more than one hump in a peak.** Each footprint has a prominent shape that includes one, two or even more humps. Based on number of these humps, we can designate a footprint as “single”, “double” or “triple” footprint. These humps have different energy distribution and are temporally separated from each other within a peak. The energy distribution and number of humps is correlated with a particular walking style of an individual, which arises due to the gradual process of interaction of the heel, the middle and toe with the floor [7].

**Distribution of energy between left and right foot.** Another characteristic of a series of footsteps is that acoustic energy in a footprint produced by left foot is systematically equal, lower or higher than the acoustic energy from the footprint produced by the right foot. This can be explained as intrinsic characteristic of the walking style, i.e., different people distribute pressure differently between their legs to balance themselves as they walk. This pressure variation appears as energy asymmetry in the sound of the right and left footstep. In most people, this asymmetry expresses itself in the form of a dominant leg and foot [6]. Thus, the identification of the dominant foot and the degree of dominance is a biometric feature.

Taking into account these observations, we propose to extract the biometric features from footsteps audio signal. Schematically, we show these features in Figure 3. From this signal we can extract the following features:

1. Number of humps in a peak from a footprint sound. This corresponds to a “single”, “double” and “triple” footprint.
2. Temporal intervals between humps ($D_1$ and $D_2$). These values are estimated for left and right foot.
3. TK energy of the footprint $E$. It is defined as the energy of the highest peak. We distinguish TK energy from left and right foot, $E_L$ and $E_R$ respectively.

In Figure 4 we show the probability distribution of $E_L / E_R$ for different people. A value of less than 1 corresponds to the right dominant foot while a value a higher than 1 correspond to the left dominant foot. One can observe that person 2 has left dominant foot, person 4 has right dominant foot, while person 7 does not show dominance in either foot.

**4. Experiments**

The recorded database was divided into two parts: 80% were used for training footsteps models for each individual person.
The task of person identification is defined as finding a person \( m \in \{1 \ldots M\} \) that most likely produced a test sequence of \( N \) footsteps \( \{f_1, f_2, \ldots f_N\} \). \( M = 10 \) is the total number of persons in the database. Note that each individual footstep \( f_j \) corresponds to either left or right foot and may include one, two or three humps (in rare cases even more than three humps may appear; in this case we take into consideration just the first three). Thus for each person \( m \) the likelihood, \( P_m(x, H, L) \), of a footstep consisting of \( H \) humps and produced by the foot \( L \) with the interval between humps \( x = (D_1; D_2) \) is given by equation:

\[
P_m(x, H, L) = \sum_{h=1}^{3} \sum_{l=1}^{2} \delta_{H,L}(h = H, l = L)p_m(h, l)N_m(x, h, l)
\]

(2)

Where \( \delta_{H,L}(\cdot) \) is an indicator function equal to 1 when \( h = H, l = L \) and 0 otherwise; \( p_m(h, l) \) is the prior probability that the footstep produced by the person \( m \) and the foot \( l \) consists of \( h \) humps; \( L = 1 \) for left foot and \( L = 2 \) for right foot. \( N_m(x, h, l) \) is the posterior probability distribution from the GMM that models the interval, \( x \), between the humps. Note that for \( h = 1 \) ("single" footstep) both values \( D_1 \) and \( D_2 \) are assumed to be equal to 0 and \( N_m = 1 \). The log-likelihood of a sequence of \( N \) footsteps is computed as a sum of log-likelihoods from each individual footstep \( f_j \).

Given a set of \( N \) footsteps we also model the energy ratio \( R \) between left and right foot for each person \( m \). We group \( N \) footsteps into \( K \) chunks in such a way that each chunk include footsteps from the same zone in the room. This way the left and right footstep from the same chunk occur at similar distance to the microphone. Assuming each chunk \( \Omega_k \) includes \( V_k \) left and \( W_k \) right footsteps with the corresponding energies \( \{E_{V_k,\text{left}}, \ldots, E_{V_k,\text{right}}, E_{W_k,\text{left}}, \ldots, E_{W_k,\text{right}}\} \), the energy ratio can be defined as:

\[
R_k = \frac{\sum_{j=1}^{V_k} E_{i,j,\text{left}}}{\sum_{j=1}^{W_k} E_{j,\text{right}}}
\]

(3)

Given a set of features \( R_k \), where \( k = 1, \ldots, K \) we train a GMM for each person \( m \). During testing the log-likelihood from this model is combined with the log-likelihood obtained from eq. (2).

The human identification results are presented in Figure 7. We show the recognition rate as the function of number of individual footsteps used for identification during testing.

As one may notice, 45% of correct identification is achieved just using 3 footsteps, and the maximum identification rate, 95%, is achieved by using more than 50 footsteps from each individual during testing.

5. Discussions and Conclusions

In this work a set of biometric markers extracted from a footstep sound is proposed. These markers have shown to be particularly useful: 95% of a maximum accuracy is achieved in a human identification task. Note that 45% of accuracy is achieved by using just 3 footsteps from a person. These experiments were performed on a newly recorded database with about 500 casual footsteps per person. This database is annotated by hand and is made available from authors for research purposes.

The proposed approach assumes that footsteps are accurately detected, roughly localized to a region and that the left and right foot assignment is made. In the study, the annotations provided this information. This is necessary in light of the relatively unexplored nature of the AGR problem and the preliminary nature of the study. A more complete system should include footsteps detection and localization from the audio signal itself. These tasks can clearly be performed from the database—former, for example using [26] and the latter can be extracted from the microphone array. The lateral assignment of a footstep to left or right can only be provided by video, though in some limited scenarios, such as going around in a circle, it is possible to extract this information from audio.

Our results, although not definitive, are nonetheless an interesting step in the identification of person using footsteps sound without using short-time spectral features. A much harder problem is that of recognizing individuals performing more complex walks such as random walk and running. However, the acoustic footstep features proposed represent only a part of the information that could be obtained. The identified novel features are constructed to be room and shoe invariant since they are based on timing and energy ratio information from audio signal. Future work will test this empirically with a more diverse database.
6. References


