Scattering features for multimodal gait recognition

Srđan Kitić
Technicolor R&I
Home Experience Lab - Data Analytics

GlobalSIP 2017
# Table of contents

1. Introduction

2. Normalized scattering for gait signals

3. Performance and wrap-up
Identification is a core component in many applications:

- Recommender systems,
- Online banking and commerce,
- Surveillance,
- Gaming,
- Administration etc.

Different biometrics: fingerprint, face, speech, retinal scan, gait (*this work)*...

Each comes with advantages and drawbacks, e.g. accuracy or intrusiveness.
Gait-based identification

Prior art - various modalities exploited:

- Video (silhouette) (1, 2): high accuracy, privacy issues.
- Mechanical force sensors (3, 4): high setup cost.
- Wearables (5, 6): intrusive.
- WiFi (7): limited accuracy and range.
- Sound (8, 9, 10, 11): (assuming VAD) privacy-preserving, wideband, widespread availability.
- Seismic (12): privacy-preserving, robust, secure, narrowband.

Complementary properties of sound and seismic cues indicate that a bimodal approach may be effective.
Gait-based identification

Prior art - various modalities exploited:

- **Video (silhouette) (1, 2):** high accuracy, privacy issues.
- **Mechanical force sensors (3, 4):** high setup cost.
- **Wearables (5, 6):** intrusive.
- **WiFi (7):** limited accuracy and range.
- **Sound (8, 9, 10, 11):** (assuming VAD) privacy-preserving, wideband, widespread availability.
- **Seismic (12):** privacy-preserving, robust, secure, narrowband.

Complementary properties of sound and seismic cues indicate that a *bimodal* approach may be effective.
Gait-based identification

Open set identification:

1. Identify a person, if coming from a known set.
2. Otherwise, decide that the person is unknown.

Addressed through *GMM-UBM framework* (13).

Remaining challenges:

- No publicly available bimodal data.
- No generally acclaimed feature type.
- Seamless feature fusion?
Gait-based identification

Open set identification:

1. Identify a person, if coming from a known set.
2. Otherwise, decide that the person is unknown.

Addressed through *GMM-UBM framework* [13].

Remaining challenges:

- No publicly available bimodal data.
  - We recorded a small scale dataset (size precludes deep learning).
- No generally acclaimed feature type.
  - Tailored *scattering transform* [14] based features.
- Seamless feature fusion?
  - Surprisingly simple - stay tuned.
Gait signals

Particle velocity:

\[ \hat{v}(\omega) = \mathcal{F}(v(t)) \propto \mathcal{F}\left(\int \vec{F}_{GRF} dt\right) \]

Footfall \( \approx 0.15 \text{s.} \)
Period \( \approx 2 \times 0.61 \text{s.} \) \( (15) \)

Acquired signals are band-passed and convoluted:

- Sound, for \( 200 \text{Hz} \lesssim \omega \lesssim 20 \text{kHz} \):

  \[ \hat{x}_a(\omega, \vec{r}(t)) = \hat{h}_a(\omega, \vec{r}(t)) \hat{v}(\omega) + \hat{e}_a(\omega) = \hat{g}_a(\omega, \vec{r}(t)) \frac{\hat{v}(\omega)}{\hat{\omega}(\omega)} + \hat{e}_a(\omega) \]

- Seismic, for \( 20 \text{Hz} \lesssim \omega \lesssim 300 \text{Hz} \):

  \[ \hat{x}_g(\omega, \vec{r}(t)) = \hat{h}_g(\omega, \vec{r}(t)) \hat{v}(\omega) + \hat{e}_g(\omega) = S_g \hat{g}_g(\omega, \vec{r}(t)) \hat{v}(\omega) + \hat{e}_g(\omega). \]

Local stationarity assumption (LSA)

Within (short) temporal segment of duration \( \tau \):

\[ \hat{g}.(\omega, \vec{r}(t + t')) \approx \hat{g}.(\omega, \vec{r}(t)), \text{ analogously } \hat{h}.(\omega, \vec{r}(t + t')) \approx \hat{h}.(\omega, \vec{r}(t)). \]
Gait signals

Particle velocity:

\[ \hat{v}(\omega) = \mathcal{F}(v(t)) \propto \mathcal{F} \left( \int \vec{F}_{GRF} dt \right) \]

Footfall \( \approx 0.15 \)s.
Period \( \approx 2 \times 0.61 \)s. (15)

Acquired signals are band-passed and convoluted:

- **Sound**, for \( 200 \text{Hz} \lesssim \omega \lesssim 20 \text{kHz} \):
  \[
  \hat{x}_a(\omega, \vec{r}(t)) = \hat{h}_a(\omega, \vec{r}(t))\hat{v}(\omega) + \hat{e}_a(\omega) = \hat{g}_a(\omega, \vec{r}(t)) \frac{\hat{v}(\omega)}{\hat{z}(\omega)} + \hat{e}_a(\omega)
  \]

- **Seismic**, for \( 20 \text{Hz} \lesssim \omega \lesssim 300 \text{Hz} \):
  \[
  \hat{x}_g(\omega, \vec{r}(t)) = \hat{h}_g(\omega, \vec{r}(t))\hat{v}(\omega) + \hat{e}_g(\omega) = S_g\hat{g}_g(\omega, \vec{r}(t))\hat{v}(\omega) + \hat{e}_g(\omega).
  \]

**Local stationarity assumption (LSA)**

Within (short) temporal segment of duration \( \tau \):

\[ \hat{g}_s(\omega, \vec{r}(t + t')) \approx \hat{g}_s(\omega, \vec{r}(t)), \text{ analogously } \hat{h}_s(\omega, \vec{r}(t + t')) \approx \hat{h}_s(\omega, \vec{r}(t)). \]
Gait signals

Particle velocity:
\[ \hat{v}(\omega) = \mathcal{F}(v(t)) \propto \mathcal{F}\left(\int \vec{F}_{\text{GRF}} dt\right) \]

Footfall \( \approx 0.15\) s.
Period \( \approx 2 \times 0.61\) s. (15)

Acquired signals are band-passed and convoluted:
- Sound, for \( 200\,\text{Hz} \lesssim \omega \lesssim 20\,\text{kHz} \):
  \[ \hat{x}_a(\omega, \vec{r}(t)) = \hat{h}_a(\omega, \vec{r}(t))\hat{v}(\omega) + \hat{e}_a(\omega) = \hat{g}_a(\omega, \vec{r}(t)) \frac{\hat{v}(\omega)}{\hat{z}(\omega)} + \hat{e}_a(\omega) \]

- Seismic, for \( 20\,\text{Hz} \lesssim \omega \lesssim 300\,\text{Hz} \):
  \[ \hat{x}_g(\omega, \vec{r}(t)) = \hat{h}_g(\omega, \vec{r}(t))\hat{v}(\omega) + \hat{e}_g(\omega) = S_g \hat{g}_g(\omega, \vec{r}(t))\hat{v}(\omega) + \hat{e}_g(\omega) . \]

**Local stationarity assumption (LSA)**

Within (short) temporal segment of duration \( \tau \):
\[ \hat{g}.(\omega, \vec{r}(t + t')) \approx \hat{g}.(\omega, \vec{r}(t)), \text{ analogously } \hat{h}.(\omega, \vec{r}(t + t')) \approx \hat{h}.(\omega, \vec{r}(t)). \]
Feature extraction

- Signals depend on impact velocity 😊 and relative position 😐
- Sound and seismic signals represent different physical quantities.
- To cope, we rely on a “CNN-like” scattering transform (16).

Feature extraction up to the order \(p\):

0: \(S_0(x) = \phi_T * x\),
1: \(S_1(x) = \phi_T * |\psi_{\lambda_1} * x|\),
2: \(S_2(x) = \phi_T * |\psi_{\lambda_2} * |\psi_{\lambda_1} * x||\),
... \(p: S_p(x) = \phi_T * |\psi_p * ... |\psi_{\lambda_2} * |\psi_{\lambda_1} * x|| ... |.

\(\phi_T := \phi_T(t)\) - a lowpass \((2\pi/T)\) filter, \(\psi_\lambda := \psi_\lambda(t)\) - a complex wavelet at scale \(\lambda\)

**Rule of thumb**

1. Computational cost increases with \(T\) (“time-invariance”).
2. \(T \propto \) duration of a classified event (crucial for performance!).
Feature extraction

- Signals depend on impact velocity 😊 and relative position 😞
- Sound and seismic signals represent different physical quantities.
- To cope, we rely on a “CNN-like” scattering transform (16).

Feature extraction up to the order $p$:

0: $S_0(x) = \phi_T * x$,
1: $S_1^{\lambda_1}(x) = \phi_T * |\psi_{\lambda_1} * x|$,
2: $S_2^{\lambda_1,\lambda_2}(x) = \phi_T * |\psi_{\lambda_2} * |\psi_{\lambda_1} * x||$,
...$p$: $S_p^{\lambda_1,\ldots,\lambda_p}(x) = \phi_T * |\psi_p * \ldots |\psi_{\lambda_2} * |\psi_{\lambda_1} * x|| \ldots |$.

$\phi_T := \phi_T(t)$ - a low-pass ($2\pi/T$) filter, $\psi_{\lambda} := \psi_{\lambda}(t)$ - a complex wavelet at scale $\lambda$

Rule of thumb

1. Computational cost increases with $T$ (“time-invariance”).
2. $T \propto$ duration of a classified event (crucial for performance!)
Feature extraction

- Signals depend on impact velocity 🙂 and relative position 😞
- Sound and seismic signals represent different physical quantities.
- To cope, we rely on a “CNN-like” scattering transform (16).

Feature extraction up to the order $p$:

0: $S_0(x) = \phi_T \ast x,$
1: $S_1^{\lambda_1}(x) = \phi_T \ast |\psi_{\lambda_1} \ast x|,$
2: $S_2^{\lambda_1,\lambda_2}(x) = \phi_T \ast |\psi_{\lambda_2} \ast |\psi_{\lambda_1} \ast x||,$
... $p$: $S_p^{\lambda_1,\ldots,\lambda_p}(x) = \phi_T \ast |\psi_p \ast \ldots \psi_{\lambda_2} \ast |\psi_{\lambda_1} \ast x|| \ldots |.$

$\phi_T := \phi_T(t)$ - a lowpass $(2\pi/T)$ filter, $\psi_{\lambda} := \psi_{\lambda}(t)$ - a complex wavelet at scale $\lambda$

Rule of thumb

1. Computational cost increases with $T$ (“time-invariance”).
2. $T \propto$ duration of a classified event (crucial for performance!).
Feature extraction

Competing requirements for $T$:

1. Short ($T \sim 0.15\text{s}$): characterizes only the footfall event, requires $p = 1$.
2. Large ($T \sim 1.22\text{s}$): captures also the temporal dynamics, but violates LSA and increases cost.

Can we avoid this tradeoff?

Visual comparison - two $p = 1$ scattering matrices (audio):
Feature extraction

Competing requirements for $T$:

1. Short ($T \sim 0.15s$): characterizes only the footfall event, requires $p = 1$.
2. Large ($T \sim 1.22s$): captures also the temporal dynamics, but violates LSA and increases cost.

Can we avoid this tradeoff?

Visual comparison - two $p = 1$ scattering matrices (audio):

Invariances mostly due to a global temporal offset!
Feature extraction

Competing requirements for $T$:

1. Short ($T \sim 0.15 \text{s}$): characterizes only the footfall event, requires $p = 1$.
2. Large ($T \sim 1.22 \text{s}$): captures also the temporal dynamics, but violates LSA and increases cost.

Can we avoid this tradeoff?

Visual comparison - two $p = 1$ scattering matrices (audio):

Remedy - compute Fourier modulus across rows (time).
Robust scattering features: normalized scattering

What about feature dependency on $\mathbf{r}$?

<table>
<thead>
<tr>
<th>Normalized scattering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under certain assumptions on $h := h(t)$, it can be shown:</td>
</tr>
<tr>
<td>$S_{p}^{\lambda_{1},\ldots,\lambda_{p}}(h \ast x) \approx</td>
</tr>
<tr>
<td>then:</td>
</tr>
<tr>
<td>$\tilde{S}<em>{p}^{\lambda</em>{1},\ldots,\lambda_{p}}(h \ast x) := \frac{S_{p}^{\lambda_{1},\ldots,\lambda_{p}}(h \ast x)}{S_{p}^{\lambda_{1},\ldots,\lambda_{p-1}}(h \ast x)} \approx \tilde{S}<em>{p}^{\lambda</em>{1},\ldots,\lambda_{p}}(x)$.</td>
</tr>
</tbody>
</table>

Consequence: if LSA holds, normalized scattering features depend only on $v(t)$!

A cheap channel normalization technique - “scattering CMS“.
Robust scattering features: normalized scattering

What about feature dependency on $\vec{r}$?

**Normalized scattering**

Under certain assumptions on $h := h(t)$, it can be shown:

$$S_{p}^{\lambda_1,\ldots,\lambda_p}(h \ast x) \approx |\hat{h}(\lambda_1)|S_{p}^{\lambda_1,\ldots,\lambda_p}(x),$$

then:

$$\tilde{S}_{p}^{\lambda_1,\ldots,\lambda_p}(h \ast x) := \frac{S_{p}^{\lambda_1,\ldots,\lambda_p}(h \ast x)}{S_{p}^{\lambda_1,\ldots,\lambda_{p-1}}(h \ast x)} \approx \tilde{S}_{p}^{\lambda_1,\ldots,\lambda_p}(x).$$

Consequence: if LSA holds, normalized scattering features depend only on $\nu(t)$!

A cheap channel normalization technique - “scattering CMS”.
Robust scattering features: normalized scattering

What about feature dependency on $r$?

**Normalized scattering**

Under certain assumptions on $h := h(t)$, it can be shown:

$$S_{p}^{\lambda_{1},\ldots,\lambda_{p}}(h * x) \approx |\hat{h}(\lambda_{1})|S_{p}^{\lambda_{1},\ldots,\lambda_{p}}(x),$$

then:

$$\tilde{S}_{p}^{\lambda_{1},\ldots,\lambda_{p}}(h * x) := \frac{S_{p}^{\lambda_{1},\ldots,\lambda_{p}}(h * x)}{S_{p}^{\lambda_{1},\ldots,\lambda_{p-1}}(h * x)} \approx \tilde{S}_{p}^{\lambda_{1},\ldots,\lambda_{p}}(x).$$

Consequence: if LSA holds, normalized scattering features depend only on $v(t)$!

A cheap channel normalization technique - “scattering CMS”.
Feature fusion

What about fusion?

- Recall that $\hat{x}_a$ and $\hat{x}_g$ have (approx) complementary frequency range.
- Hence, $\tilde{S}_{1}^{\lambda_1}(x_a) > 0$ and $\tilde{S}_{1}^{\lambda_1}(x_g) > 0$ should be complementary as well.

- Due to channel normalization, $\tilde{S}_{1}^{\lambda_1}(x_a)$ and $\tilde{S}_{1}^{\lambda_1}(x_g)$ “live” in the same feature space, we can simply sum them up:\footnote{$\alpha$ is a normalization constant}

$$\tilde{S}_{fused}^{\lambda_1} = \alpha_a \tilde{S}_{1}^{\lambda_1}(x_a) + \alpha_g \tilde{S}_{1}^{\lambda_1}(x_g)$$

$\tilde{S}_0(x_a)$ and $\tilde{S}_0(x_g)$ are concatenated to $\tilde{S}_{fused}^{\lambda_1}$. 
Feature fusion

What about fusion?

- Recall that \( \hat{x}_a \) and \( \hat{x}_g \) have (approx) complementary frequency range.
- Hence, \( \tilde{S}_1^{\lambda_1}(x_a) > 0 \) and \( \tilde{S}_1^{\lambda_1}(x_g) > 0 \) should be complementary as well.

- Due to channel normalization, \( \tilde{S}_1^{\lambda_1}(x_a) \) and \( \tilde{S}_1^{\lambda_1}(x_g) \) “live” in the same feature space, we can simply sum them up\(^1\):

\[
\tilde{S}_{fused}^{\lambda_1} = \alpha_a \tilde{S}_1^{\lambda_1}(x_a) + \alpha_g \tilde{S}_1^{\lambda_1}(x_g)
\]

\( \tilde{S}_0(x_a) \) and \( \tilde{S}_0(x_g) \) are concatenated to \( \tilde{S}_{fused}^{\lambda_1} \).

\(^1\alpha \) is a normalization constant
Experiments

Experimental setup (17):

- Data collected internally, on a prototype dual sensor setup.
- 12 participants (8m and 4f), up to two types of shoes per person.
- (Low noise) recordings in a carpet-covered room, on 3 different days\(^2\).
- 6 persons randomly chosen for training the UBM.
- From the remaining, randomly chosen 3 targets and 3 unknowns.
- Hyperparameters: \( \tau, T, N \) (the number of retained coefficients after PCA).

\(^2\)To avoid environmental effects: 2 days for training, 3rd day for evaluation.
Results

- Performance metric: *Equal Error Rate (EER)*, lower is better.
- Median results for the best-performing N, after 100 random partitions.

```
\[
\tau = 1.5s
\]
```

- “Optimal” hyperparameters agree with predictions:
  1. $T$ on the order of the footfall impact duration.
  2. Larger $\tau$ degrades performance (violates LSA).
  3. “Richer” representations (i.e. audio and fused) favor larger N.
Results

- Performance metric: *Equal Error Rate (EER)*, lower is better.
- Median results for the best-performing $N$, after 100 random partitions.

- “Optimal” hyperparameters agree with predictions:
  1. $T$ on the order of the footfall impact duration.
  2. Larger $\tau$ degrades performance (violates LSA).
  3. “Richer” representations (*i.e.* audio and fused) favor larger $N$. 
Results

Best setting for each modality

Classification with fused features:
- exhibits the smallest variance,
- is the most robust wrt parameterization.
Bimodal gait-based identification wrap-up:

- Confirmed identification by both sound and seismic observations.
- Performance gradation: fused > sound > seismic.
- Further research directions:
  - Recognition in noisy conditions and using cheap MEMS sensors.
  - “Walker diarization”?
  - Relevance of the shoe type, gender and/or environment.
  - A better way to fuse / extract features (new datasets), etc.
THANK YOU!
References I


References II


