GAIT-BASED PERSON IDENTIFICATION BY SPECTRAL, CEPSTRAL AND ENERGY-RELATED AUDIO FEATURES

Jürgen T. Geiger, Martin Hofmann, Björn Schuller and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Germany

{geiger,martin.hofmann,schuller,rigoll}@tum.de,

ABSTRACT

With this work, we address the problem of acoustic gait-based person identification, which is the task of identifying humans by the sounds they make while walking. We examine several acoustic features from speech processing tasks for their suitability for acoustic gait recognition. Using a wrapper-based feature selection technique, we reduce the feature set while at the same time increasing the identification accuracy by 10 % (relative). For classification, Support Vector Machines (SVMs) are employed. Experiments are conducted using the TUM GAID database, which is a large gait recognition database containing 3 050 recordings of 305 subjects in three variations.

Index Terms— Acoustic gait-based person identification, gait recognition, feature selection

1. INTRODUCTION

Recognizing people by the way they walk (also known as gait recognition) has been an active field of research in the last decade. While most of previous studies focussed on visual information, acoustic information can also be used for gait recognition. Even though the focus on this modality has so far been significantly less, results are promising. The characteristics of the sounds of walking persons are mainly dependent on the gait, shoes (and other characteristics like trousers) and the floor type. In a user study [1], it was shown that humans are able to distinguish other people by their walking sounds. After a training phase, twelve subjects were able to identify their co-workers by their walking sounds with an accuracy of 66 %. This study shows that walking sounds convey characteristic information about the walking person and can be used for person identification. Using audio information for gait-based person identification has potential applications in indoor surveillance-scenarios, to enhance visual surveillance and facilitate multimodal approaches. As compared to video-based person identification, acoustic systems will also work in the darkness, require less expensive hardware and are less obtrusive. Acoustic gait-based person identification is also known as *acoustic gait recognition*.

Until now, only few works have been addressing the problem of acoustic gait-based person identification. In [2], the task was to detect footstep sounds in a corpus of various different environmental sounds. A system for person identification using footstep detection was introduced in [3]. Mel-cepstrum analysis, walking intervals and the degree of similarity of spectrum envelope are used as features. For classification, a method based on k-means clustering is used. The system was tested with a database of five persons. This work was extended in [4] by adding psychoacoustic features like loudness, sharpness, fluctuation strength and roughness. Finally, in [5], Dynamic Time Warping (DTW) was used for classification and the database was extended to contain ten persons.

In [6], a system for person identification based on walking sounds is presented. From the audio signal, the gait frequency, spectral envelope, Linear Predictive Coding (LPC) coefficients, Mel-frequency Cepstral Coefficients (MFCCs) and loudness are computed. A subset of the features is selected using Fisher's linear discriminant analysis. For classification, k-nearest neighbours (k-NN) is compared with k-means. Using a database with 15 individuals with six different shoe types, classification rates range from 33.5 % to 97.5 %. The weakness of all previous studies about acoustic gait-based person identification that are mentioned here is the fact that only small databases (mostly no more than ten subjects) have been employed.

In this contribution, we describe experiments for acoustic gait-based person identification using the TUM GAID corpus, which contains a large number of subjects. Furthermore, the database features recordings with three different variations (normal walking, walking with a backpack and walking with shoe covers), which allows for realistic experiments. We employ a large candidate feature set adopted from speech processing and systematically select features that are relevant for acoustic gait-based person identification. Support Vector Machines (SVMs) are used for classification, and features are ranked and selected using a wrapper approach. On an independent test set, we achieve a 10 % relative improvement in identification accuracy with the selected features compared to using all features.

2. THE TUM GAID DATABASE

For our experiments, we use our freely available[1] TUM Gait from Audio, Image and Depth (GAID) database [7]. The motivation behind the TUM GAID database is to foster multimodal gait recognition. Therefore, data was recorded with an RGB-D sensor, as well as with a four-channel microphone array. Thus, a typical color video stream, a depth stream and an audio stream are simultaneously available. The database contains recordings of 305 subjects walking perpendicular to the recording device in a 3.5 m wide hallway corridor with a solid floor. In each recorded sequence, the subject walks for roughly 4 m, typically performing between 1.5 and 2.5 gait cycles. The sequences each have a length of approximately 2 – 3 s. Three variations are recorded for each subject: Normal walking (N), walking with a backpack (B), and walking with shoe covers (S). The backpack constitutes a significant variation in gait pattern and sound (sounds are created by the backpack itself), and the shoe covers pose

This research was supported by the ALIAS project (AAL-2009-2-049) co-funded by the EC, the French ANR and the German BMBF.

[1] www.mmk.ei.tum.de/tumgaid

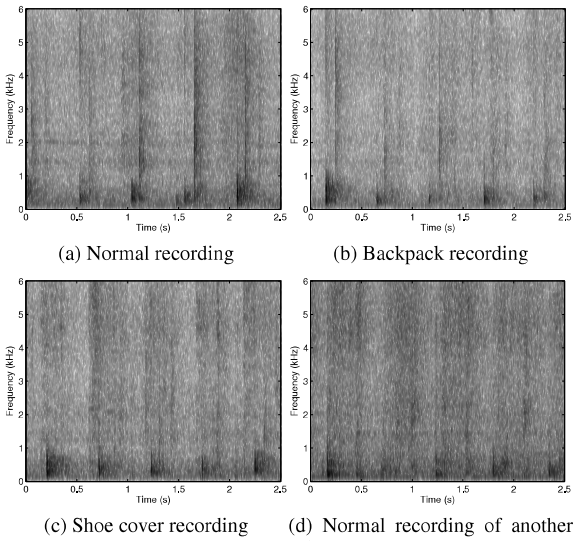
	Development (150 subj.)	**Test** (155 subj.)
N1 – N4	Enrollment	Enrollment
N5 – N6	Identification	Identification
B1 – B2	Identification	Identification
S1 – S2	Identification	Identification

Table 1: Setup of the TUM GAID database

a considerable change in acoustic condition. For each subject, there are six recordings of the N setup, and two each of the B and S setups. This sums to a total number of 3 050 recordings. The metadata distribution of the database is well-balanced with a female proportion of 39 % and ages from 18 to 55 years (average 24.8 years and standard deviation 6.3 years). More than half of the subjects are wearing sneakers while other commonly-used shoe types are boots and loafers.

To allow for a proper scientific evaluation and to prevent overfitting on the test data, the database is divided into a *development set* and a *test set*. The two sets are person-disjunct and contain 150 and 155 subjects, respectively. Both for the development and for the test set, the first four N recordings of each subject are used for the enrollment process. The other two N recordings as well as the B and S recordings are used to perform the identification experiments. This means that models are learned only using the N recordings, while the B and S setups constitute previously unseen variations during the identification experiments and will therefore deteriorate the identification performance. The partition of the database is shown in Table 1.

Figure 1 shows the spectrograms of four exemplary recordings in the used database. The spectrograms reveal a lot of static background noise, which is due to the recording environment. Figure 1 a shows a normal recording of one subject, where each walking step is characterized by two successive sounds, where the first sound has stronger low-frequency components and the second sound has stronger high-frequency components. With a backpack (Figure 1 b), the steps get softer and audible sounds are added by the backpack. When wearing shoe covers (Figure 1 c), more and longer high-frequency components are introduced, which are the rustle-like sounds of the shoe covers. For reference, Figure 1 d shows the spectrogram of another subject, with sounds between the steps, which result from the legs of the trousers rubbing against each other.



(a) Normal recording (b) Backpack recording

(c) Shoe cover recording (d) Normal recording of another subj.

Fig. 1: Spectrograms of four recordings in the TUM GAID database.

3. BASELINE SYSTEM

To address the problem of person identification based on gait sounds, we use SVMs for classification and examine a number of acoustic features. We use features which are established in audio processing tasks like speech recognition, emotion recognition or acoustic event classification. Our candidate feature set also includes features which have been used in previous studies on acoustic gait-based person identification.

3.1. Candidate Features

The database provides audio signals with four audio channels recorded with a sampling rate of 16 kHz. Before the feature extraction step, the recordings are converted to mono by averaging over the four individual channels. In order to provide a first well reproducible and transparent baseline system, we use a brute-force large-scale feature extraction approach, employing our open-source toolkit openSMILE [8].

The employed audio feature set is based on the baseline audio features we had provided for the Audio/Visual Emotion Challenge 2011 (AVEC 2011) [9] and contains a number of energy, spectral and cepstral features. Compared to the AVEC 2011 feature set, the voicing related features were omitted, as we found out that they are not relevant for our problem. The employed features are of supra-segmental nature. This means that the acoustic descriptors such as energy and spectral entropy (which are sampled at a fixed rate) are summarized over a recording (of variable length) into a single feature vector of constant length. This is achieved by applying statistical functionals to the acoustic low-level descriptors (LLD). Thereby, each functional maps each LLD signal into a single value for the given segment. Examples for functionals are mean, standard deviation, higher order statistical moments, quartiles, etc.

The set of LLDs and the functionals are listed in Table 2 and Table 3, respectively. All LLDs are computed every 10 ms, where a window size of 60 ms is applied for the MFCCs and loudness features while all other features are computed based on windows with a length of 25 ms. Features which have been analyzed in previous studies about acoustic gait-based person identification [4, 5] such as the loudness, psychoacoustic sharpness or Mel-Frequency Cepstral Coefficients (MFCCs) are included in our feature set. In addition, our feature set provides a substantial number of further acoustic feature information. Furthermore, for each LLD, first order delta coefficients (equivalent to the first derivative) are computed. The final feature set is then made up of 25 LLDs $\times$ 42 functionals and 25 delta coefficients $\times$ 23 functionals, summing up to 1 625 features in total per recording.

3.2. Classification

To foster reproducibility, as a classifier, SVMs with a linear Kernel function (as implemented in the WEKA toolkit [10]) are applied. Sequential minimal optimization (SMO) (complexity 1.0) is used for training. The multi-class classification problem is handled by constructing pairwise SVMs. SVMs are discriminative classifiers which do not require large amounts of training data. This makes them especially suited for our task.

Energy-related features (3)
loudness (auditory model based), energy in bands from 250 Hz – 650 Hz, 1 kHz – 4 kHz,
Spectral features (12)
zero crossing rate, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity,
Cepstral featurea (10)
MFCCs 1 – 10

Table 2: 25 energy and spectral-related acoustic low-level descriptors (LLDs).

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 %-99 %, percentage of frames contour is above: minimum + 25 %, 50 %, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length, standard deviation of segment length
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1] (6)
Linear Predictive Coding gain/coefficients 1 – 5

Table 3: Set of all 42 functionals used for audio feature extraction. [1]Not applied to delta coefficient contours. [2]For delta coefficients, the mean of only positive values is applied, otherwise the arithmetic mean is applied.

3.3. Baseline Results

Results for the candidate feature groups and their combinations (using the development set) are shown in Table 4. Generally, the best results are obtained in the normal (N) setup. Carrying a backpack (B setup) leads to a different walking pattern as well as to additional sounds and therefore to a decrease in identification performance. Using shoe covers (S setup) completely changes the characteristics of the footstep sounds. However, all results in the S setup are still better than the chance level (0.7 % accuracy).

The best single feature group, consisting of the MFCCs leads to an average accuracy of 23.1 %. Energy features (which constitute the smallest feature group with only three features) lead to the worst performance. Looking at the different combinations of the feature groups, it can be concluded that MFCCs and spectral features are somewhat redundant, since there is no significant improvement when combining those two feature groups (significance was evaluated using a one-sided z-test). The best result is obtained by combining MFCCs with energy features, with an averagy accuracy of 28.1 %. This result is significantly better than with only MFCCs (significant at a 0.01 level) and even better than the combination of all three feature groups.

Features	**N**	**B**	**S**	**avg.**
MFCC	41.7	22.3	5.3	23.1
Energy	29.3	17.0	5.0	17.1
Spectral	41.0	20.7	4.0	21.9
MFCC + Energy	49.7	28.7	**6.0**	28.1
MFCC + Spectral	44.0	26.3	4.7	25.0
Energy + Spectral	43.0	24.7	4.3	24.0
MFCC + Energy + Spectral	48.3	29.0	4.3	27.2
Best 400 features	**57.7**	**30.7**	3.3	**30.6**
Best LLDs	43.7	29.7	4.3	25.9

Table 4: Results on the development set (150 subjects), using different combinations of feature groups, the best 400 features as determined by the described feature selection technique and the best acoustic low-level descriptors (the three energy features, spectral kurtosis, flux and skewness and MFCC 1). Shown is the identification accuracy for the three setups N (normal walking), B (backpack) and S (shoe covers) and the average. The chance level is 0.7 %.

4. FEATURE ANALYSIS

In order to reduce our feature set and to identify the relevant features, we apply an automatic wrapper-based [11] feature selection technique.

4.1. Feature Selection

We use a simplified version of Sequential Forward Selection [12] for feature analysis. For each of the $N = 1\,625$ candidate features (including all the delta coefficients and functionals) the classifier is trained and evaluated on the N setup of the development set. For the whole feature set $F = f_1, f_2, ..., f_N$, this accuracy $a(f_n)$ is computed as

$$a(f_n) = \frac{1}{T}\sum_{t=1}^{T}\delta\Big(\arg\max_{c} P(x_t|m(c, f_n)) - l(t)\Big) \qquad (1)$$

where $X = x_1, x_2, ..., x_T$ are the instances of the development set, $l(t)$ denotes the true label for instance t, $\delta(\cdot)$ is the Kronecker delta function and $m(c, f_n)$ is the model of the classifier for class c, built using only feature f_n. In this experiment, accuracies $a(f_n)$ between 0 % and 7 % are achieved, with a mean of 1.8 %. The features are then sorted according to their classification accuracy $a(f_n)$. Then, starting with the single best feature, more and more features are added to the feature set according to their ranking until the whole feature set is used. Figure 2 shows the results of this experiment on the development set. The best result is obtained using 400 out of the total 1 625 features. This result is also shown in Table 4. Out of these 400 features, 89 are derived from MFCCs, 90 from energy features and 221 from spectral features. This composition suggests that in all of the three feature groups, there are relevant features. While with all 1 625 features, an accuracy of 48.3 % is achieved in the N setup, with the 400 best features, this accuracy is raised by 19 % relative to 57.7 %. This improvement is significant at the 0.05 level. For the B setup, there is a non-significant improvement to 30.7 %, while the performance in the S setup undergoes a slight (non-significant) decrease to 3.3 %, compared to using all features.

We then further analyzed the employed low-level features and functionals. In order to understand the relevance of each feature, we

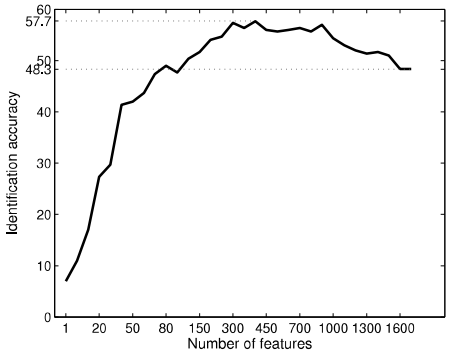


Fig. 2: Identification accuracy on the N (normal) setup of the development set for different numbers of features. More features are added according to their identification accuracy as a single feature.

obtained a single score for each low-level descriptor and for each of the functionals. For each low-level descriptor, we compute the average of $a(f_n)$ over the top 50 % performing functionals. Analogously, for each of the functionals, the average accuracy over the top low-level descriptors for this functional is computed. The features which get the highest scores are the three energy features, spectral kurtosis, flux and skewness and the first MFCC coefficient. These are also the features which are among the most common features in the feature set of the best 400 features as determined by the simplified sequential forward selection technique. When only these LLDs are used (together with all delta coefficients and functionals, summing up to 455 features in total), an average accuracy of 25.9 % is obtained with the development set (see Table 4). Since these features have been ranked and selected independent of each other, it is understandable that the result is slightly worse than the baseline result of 27.2 % when using all features. Among the functionals, the best scores are achieved by the means (arithmetic and root quadratic), the standard deviation, the quartiles and quartile ranges and the percentiles and percentile ranges. Similarly as with the best scored features, these functionals are also among the most common functionals in the best 400 features.

4.2. Test Set Results

Table 5 shows results for experiments using the test set containing 155 subjects. In general, the same trends can be observed as on the development set. The combination of MFCCs and energy features is better than the combination of all three feature groups. Using the best 400 features which have been selected as described in Section 4.1, the best result is achieved with an average identification accuracy of 28.2 %. This is a 10 % relative improvement compared to using the whole feature set.

5. CONCLUSIONS

In this paper, we presented an extensive feature analysis for acoustic gait-based person identification. Using the development set of the TUM GAID database, suitable features have been analyzed and selected from a large candidate feature set. Out of all 1 625 features, a subset of 400 features was selected with a wrapper-based feature selection technique, which led to the best average results on the test set. This feature set contains features from all three feature groups (energy, spectral and cepstral). When features are examined independent of each other, the three energy features, spectral kurtosis,

Features	**N**	**B**	**S**	**avg.**
MFCC	42.3	21.9	**7.4**	23.9
Energy	24.2	17.7	3.6	15.2
Spectral	33.2	10.0	1.9	15.1
MFCC + Energy	46.5	25.5	**7.4**	26.5
MFCC + Spectral	43.6	24.8	3.6	24.0
Energy + Spectral	37.1	22.9	3.2	21.1
MFCC + Energy + Spectral	44.5	27.4	4.8	25.6
Best 400 features	**51.9**	**28.4**	4.2	**28.2**
Best LLDs	38.1	21.6	4.5	21.4

Table 5: Results on the test set (155 subjects), using different combinations of feature groups, the best 400 features as determined by the described feature selection technique and the best acoustic low-level descriptors (the three energy features, spectral kurtosis, flux and skewness and MFCC 1). Shown is the identification accuracy for the three setups N (normal walking), B (backpack) and S (shoe covers) and the average. The chance level is 0.7 %.

flux and skewness and the first MFCC coefficient are found to be relevant for acoustic gait-based person identification. The best results were achieved on the normal recordings (N experiments), while wearing a backpack (B) or shoe covers (S) influenced the achieved identification accuracy in a negative way.

Future work includes investigation of features which are specifically taylored to gait sounds. Furthermore, fusion of our acoustic approach with vision-based approaches (as in [13], where speaker recognition was combined with face recognition) should lead to improved performance. In [14], it is investigated how new classes can be added to the set of already known classes of acoustic events. Given the fact that relatively small amounts of training data are available in acoustic gait recognition tasks, such adaptation techniques could lead to improved performance. Additionally, adopting adaptation approaches from the speaker recognition domain [15] could also address this problem. To increase the robustness of the systems, we plan to apply signal separation techniques and noise-robust recognizers as in [16].

6. RELATION TO PRIOR WORK

The most-widespread approach for video-based gait recognition is the Gait Energy Image [17], which is a simple silhouette-based approach. It can be combined with face recognition [18] or with depth information [19]. Furthermore, model-based approaches have been proposed for visual gait recognition [20]. Besides using video or audio information, other methods to identify walking persons include using acoustic Doppler sonar [21] or pressure sensors [22].

Compared to previous works on acoustic gait recognition [3, 4, 5, 6], we investigated a larger number of features and used a much larger database. In these studies, the employed audio features include gait frequency, spectral envelope, LPC coefficients, MFCCs or loudness. In [23], inter-peak distances and peak height were used as audio features.

In [24], acoustic features from the speech domain are used for the classification of acoustic events, which is similar to our work, since they try to adopt features which were developed for speech processing for another audio recognition task. Another study about feature selection for acoustic event detection is [25], where the discriminant capability of each feature (candidate features are MFCCs and log frequency filter bank parameters) is quantified according to the approximated Bayesian accuracy.

7. REFERENCES

[1] K. Mäkelä, J. Hakulinen, and M. Turunen, "The use of walking sounds in supporting awareness," in *Proc. of International Conference on Auditory Display*, Boston, MA, USA, 2003, pp. 144–147.

[2] B. She, "Framework of footstep detection in in-door environment," Kyoto, Japan, 2004, pp. 715–718.

[3] Y. Shoji, T. Takasuka, and H. Yasukawa, "Personal identification using footstep detection," in *Proc. of IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Seoul, South Korea, 2004, pp. 43–47.

[4] A. Itai and H. Yasukawa, "Footstep recognition with psyco-acoustics parameter," in *Proc. of IEEE Asia Pacific Conference on Circuits and Systems*, Singapore, 2006, pp. 992–995.

[5] A. Itai and H. Yasukawa, "Footstep classification using simple speech recognition technique," in *Proc. of IEEE International Symposium on Circuits and Systems*, Seattle, WA, USA, 2008, pp. 3234–3237.

[6] R.L. de Carvalho and P.F.F. Rosa, "Identification system for smart homes using footstep sounds," in *Proc. of IEEE International Symposium on Industrial Electronics*, Bari, Italy, 2010, pp. 1639–1644.

[7] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and Rigoll G., "The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits," *Journal of Visual Communication and Image Representation (JVCI), Special Issue on Visual Understanding and Applications with RGB-D Cameras, Elsevier*, 2013.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Florence, Italy, 2010, pp. 1459–1462.

[9] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–The first international audio/visual emotion challenge," in *Proc. of International Audio/Visual Emotion Challenge and Workshop*, Memphis, TN, USA, 2011, pp. 415–424.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The weka data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[11] R. Kohavi and G.H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[12] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[13] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 10, pp. 955–966, 1995.

[14] J. Geiger, M. Lakhal, B. Schuller, and G. Rigoll, "Learning new acoustic events in an hmm-based system using map adaptation," in *Proc. Interspeech 2011, Florence, Italy*, 2011, pp. 293–296.

[15] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[16] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The munich 2011 chime challenge contribution: Nmf-blstm speech enhancement and recognition for reverberated multisource environments," in *Proc. Machine Listening in Multisource Environments (CHiME 2011), satellite workshop of Interspeech 2011, ISCA, Florence, Italy*, 2011, pp. 24–29.

[17] Ju Han and Bir Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 316–322, 2006.

[18] M. Hofmann, S.M. Schmidt, AN Rajagopalan, and G. Rigoll, "Combined face and gait recognition using alpha matte preprocessing," in *IAPR/IEEE International Conference on Biometrics*, New Delhi, India, 2012, pp. 1–6.

[19] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5d gait biometrics using the depth gradient histogram energy image," in *Proc. of IEEE International Conference on Biometrics: Theory, Applications and Systems*, Washington, DC, USA, 2012.

[20] C.Y. Yam, M.S. Nixon, and J.N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, vol. 37, no. 5, pp. 1057–1072, 2004.

[21] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recogination," in *Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance*, London, UK, 2007, pp. 27–32.

[22] J.S. Yun, S.H. Lee, W.T. Woo, and J.H. Ryu, "The user identification system using walking pattern over the ubifloor," in *Proc. International Conference on Control, Automation, and Systems*, Gyeongju, Korea, 2003, pp. 1046–1050.

[23] D.T. Alpert and M. Allen, "Acoustic gait recognition on a staircase," in *Proc. IEEE World Automation Congress (WAC)*, 2010, pp. 1–6.

[24] A. Temko and C. Nadeu, "Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering," in *Proc. IEEE ICASSP'05*, Philadelphia, PA, USA, 2005, pp. 502–505.

[25] X. Zhuang, X. Zhou, T.S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Proc. IEEE ICASSP'08*, Las Vegas, NV, USA, 2008, pp. 17–20.