The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits

Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Arcisstr. 21, 80333 Munich, Germany

Abstract

Recognizing people by the way they walk – also known as gait recognition – has been studied extensively in the recent past. Recent gait recognition methods solely focus on data extracted from an RGB video stream. With this work, we provide a means for multimodal gait recognition, by introducing the freely available TUM Gait from Audio, Image and Depth (GAID) database. This database simultaneously contains RGB video, depth and audio. With 305 people in three variations, it is one of the largest to-date. To further investigate challenges of time variation, a subset of 32 people is recorded a second time. We define standardized experimental setups for both person identification and for the assessment of the soft biometrics age, gender, height, and shoe type. For all defined experiments, we present several baseline results on all available modalities. These effectively demonstrate multimodal fusion being beneficial to gait recognition.

Keywords: Multimodal Gait Recognition, Depth Gradient Histogram Energy Image, Acoustic Gait Recognition

1. Introduction

Recognizing people by the way they walk has been an active field of research in the last decade. Over the years, it has been shown in various studies, that gait motion together with static gait posture can be efficiently used to identify humans. Gait recognition has unique advantages over traditional biometrics such as fingerprint, iris, retina, DNA and face. Most notably, gait features can be obtained from people at larger distances and at low resolution, when other features such as face are obscured. In addition, capturing gait features is non-invasive and does not require the cooperation of the subject as it is usually necessary for example for fingerprint recognition. These properties make gait recognition an ideal biometric modality in many applications such as intelligent video surveillance, access control, forensics, as well as in tracking and monitoring. Furthermore, gait recognition methods can be used to estimate characteristics like gender, shoe type, age or height of a person.

Email addresses: martin.hofmann@tum.de (Martin Hofmann), geiger@tum.de (Jürgen Geiger), sebastian.bachmann@mytum.de (Sebastian Bachmann), schuller@tum.de (Björn Schuller), rigoll@tum.de (Gerhard Rigoll)

Many recent gait recognition approaches rely solely on visual data [1, 2, 3, 4, 5, 6, 7]. A multitude of methods and techniques in feature extraction from a visual image stream have been developed. However, due to the recent advances in depth imaging devices, interest in depth-based gait recognition is growing [8, 9, 10]. In addition, it has been shown that gait recognition can also be performed from audio data [11]. 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 [12], 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 as well as for soft biometrics tasks. The field of person identification as well as soft biometrics based on audio features is typically referred to as *acoustic gait recognition*.

In this work, we strive towards multimodal gait recognition combining information from (1) the visual RGB image sequence, (2) the depth image sequence and (3) the four channel audio stream. In the recent past, RGB-D sensors have received remarkable attention, paving the way for the envisioned multimodal gait recognition. Especially the emergence of low-priced consumer depth cameras such as Kinect and Xtion PRO have made the simultaneous acquisition of all three above mentioned modalities conveniently easy. We present the freely available[1] TUM Gait from Audio, Image and Depth (GAID) database, which to our knowledge is the first database which allows to simultaneously address the problem of recognizing humans and selected traits using multiple modalities.

In addition to person identification, gait recognition methods can also be used for soft biometrics tasks like gender, age, height or shoe type recognition. These can all be helpful in characterizing persons, e. g. in forensic applications. Thus, the TUM GAID database comes with a set of labels which allows to carry out such soft biometrics recognition. Experimental setups for both person identification and soft biometrics, as well as baseline algorithms for both setups are presented in this article.

The remainder of this article is organized as follows: First, we present related work, especially related gait databases in Section 3. Then, the three main contributions are presented: For the first part, a new large database for gait recognition featuring video, depth and audio has been recorded and is presented here in full detail (Section 4). Second, in Section 5, a set of experiments is defined for both the human identification task, as well as for several soft biometric tasks, i. e., gender, shoe type, age and height recognition. The third contribution is a set of baseline algorithms including feature extraction for the visual domain (Section 6) and the audio domain (Section 7), as well as classification and fusion (Section 8). Comparative results for all algorithms as well as for multimodal fusion are presented and analyzed in Section 9, followed by concluding

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

remarks in Section 10.

2. Related Work

Major approaches for gait recognition include model-based [13, 14] and model-free (appearance-based) methods [1, 2, 3, 4, 5, 15]. In model-based methods, a human pose model is extracted at each frame and the underlying kinematics are used for individual identification. While this is conceptionally solid, in practice pose estimation proves highly difficult and results show limited performance. In contrast, model-free methods bypass the model fitting and extract a variety of features directly on the input data. This way, a correspondence of the person's appearance to its identity is created. In most experiments so far, model-free methods greatly outperform model-based approaches, because they prove to be more robust in practice.

Many model-free methods build on silhouette extraction for each frame in a gait cycle. Silhouettes are either averaged, as in the prominent Gait Energy Image (GEI) [1, 4], or all silhouettes are used simultaneously [3, 5, 16]. Different classifiers ranging from nearest neighbor [1], Support Vector Machines (SVMs) [3], and Hidden Markov Models (HMMs) [16] have been applied with similarly good results.

A majority of current model-free methods use only 2D visual data. One of the most prominent and most widely used methods for 2D visual gait recognition is probably the silhouette averaging method as used for example in GEI. With the availability of depth information, several new types of feature extraction have so far emerged. For example in [8], a multi-camera system together with a structure from motion algorithm is used to build binary 3D voxel representations of the human. The voxel set is then back-projected to the side, front and top view, where 2D gait recognition methods are applied. In [9], the authors use a similar voxel reconstruction and in addition they use the Kinect sensor to obtain depth data. They define the Gait Energy Volume (GEV) as a 3D extension to the Gait Energy Image (GEI). The Depth Gradient Histogram Energy Image (DGHEI)[10] is another successful feature extraction method, which specifically makes use of depth information and outperforms the other baseline methods.

Several approaches have investigated the effects of fusion of face and gait features, which is an important step towards a practical wide area surveillance system. In such scenarios, faces are typically visible at relatively low resolution and from various viewing angles. Profile and side-view approaches as well as multi-view approaches have been used in combination with gait recognition [7, 17, 18, 19, 20, 21]. Temporal super-resolution [22] can be applied to enhance side-view profiles for better face recognition at a distance. The database presented in this article will potentially allow for further investigation of combined face and gait recognition, as it provides both, gait features as well as good face profiles.

Until now, only few works have been addressing the problem of acoustic gait recognition. In [23], 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 [11]. Mel-cepstrum analysis, walking intervals and

the degree of similarity of spectrum envelope are used as features and classified with a method based on k-means clustering. The system was tested with a database of five persons. This work was extended in [24] by adding psychoacoustic features like loudness, sharpness, fluctuation strength and roughness. Finally, in [25], Dynamic Time Warping (DTW) was used for classification and the database was extended to contain ten persons.

In [26], 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 recognition which are described here is that only small databases have been employed and no session variability experiments are performed. In this contribution, we present the first corpus for acoustic gait recognition which contains a large number of persons and where session variability experiments can be performed.

Besides using video or audio information, other methods to identify walking persons include using acoustic Doppler sonar [27] or pressure sensors [28].

3. Related Gait Databases

Since the field of gait recognition has been in existence for roughly a decade, the research community has long utilized publicly available databases for comparative performance evaluation.

Table 1 summarizes the most prominent publicly available gait recognition corpora, most of which focus on the video modality. This table also shows the important features of the particular databases. These are the number of subjects , as well as a good set of person variations. Such variations include, but are not limited to: Viewing angle, clothing, shoe types, surface types, indoor/outdoor variation, carrying condition, illumination, and time.

The first available dataset was the 1998 UCSD Dataset [29], which contains merely six subjects. Most of the following early gait recognition databases were published in 2001 from various institutions [30, 31, 32, 33, 34, 35]. Those datasets feature a medium number (about 25) of subjects. It was then found that, for meaningful evaluation, datasets should contain at least 30 subjects and possibly more.

The most comprehensive database to date, which features a large set of subjects as well as a substantial set of variations, is probably the HumanID Gait Challenge database [5].

Other databases such as CASIA (Dataset B) [36] also contain high numbers of subjects and a significant number of variations. CASIA additionally features an exhaustive number of views, which allows for precise 3D reconstruction.

Database, Ref.	**# subj. / # sequ.**	**Environment**	**Year**	**Variations**
UCSD ID [29]	6 / 42	outdoor, wall background	1998	-
CMU Mobo [30]	25 / 600	indoor, treadmill	2001	viewpoint, walking speeds, carrying conditions, surface incline
Georgia Tech [31]	15 / 268	outdoor	2001	time (six months), viewpoint
	18 / 20	magnetic tracker	2001	time (six months)
HID-UMD Dataset 1 [32]	25 / 100	outdoor		
HID-UMD Dataset 2 [33]	55 / 222	outdoor, top mounted	2001	viewpoints (front, side), time
MIT, 2001 [34]	24 / 194	indoor	2001	view
SOTON Small Database [35]	12 / –	indoor, green background	2001	carrying condition, clothing, shoe, view
SOTON Large Database [35]	115 / 2 128	indoor, outdoor, treadmill	2001	view
HumanID Gait Challenge [5]	122 / 1 870	outdoor	2002	viewpoint, surface, shoe, carrying condition, time (months)
CASIA Database A [36]	20 / 240	outdoor	2001	three viewpoints
CASIA Database B [36]	124 / 13 640	indoor	2005	11 viewpoints, clothing, carrying condition
CASIA Database C [36]	153 / 1 530	outdoor, night, thermal camera	2005	speed, carrying condition
OU-ISIR A [37]	34/ 408	indoor, treadmill	2007 - 2012	speed
OU-ISIR B [37]	68/ 1 350	indoor, treadmill	2007 - 2012	clothing
OU-ISIR D [37]	185/ 370	indoor, treadmill	2007 - 2012	gait fluctuations
SOTON temporal [38]	25/ 2 280	indoor	2012	time, view
TUM-IITKGP [39]	35 / 840	indoor, hallway	2010	carrying condition, occlusions
TUM GAID	305 / 3 370	indoor, hallway, + depth + audio	2012	time (months), carrying condition, shoe variation

Table 1: Overview of related publicly available databases for gait recognition

The TUM GAID database presented in this article is – to the best knowledge of the authors – the only database to date which allows for multimodal gait recognition using video, depth and audio features. With a total of 305 subjects and 3370 sequences it is one of the largest publicly available datasets. The time variation (where clothing, lighting, and other recording properties are significantly different) has proven to be extremely challenging in the field of gait recognition. Besides the Georgia Tech, HumanID and the SOTON temporal database[38], the TUM GAID database is the only database to address such time variation.

4. Database Description

The central motivation behind the TUM GAID database is to foster multimodal gait recognition. To meet this goal, 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 audio stream are simultaneously available.

4.1. Recording time

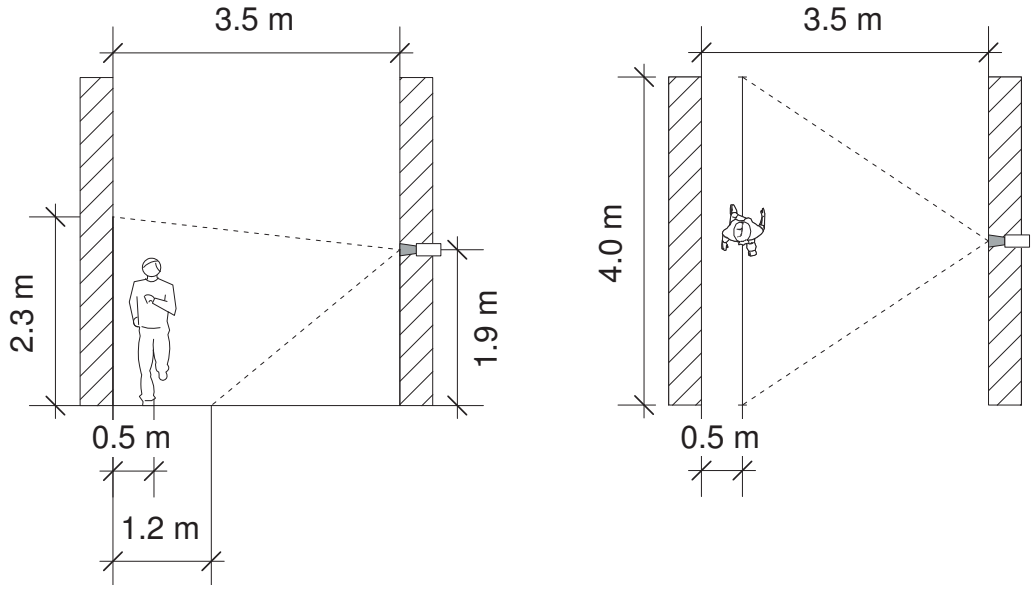
The TUM GAID database was recorded in two sessions in Munich, Germany. The first session was held in January 2012, which happened to have some of the coldest days in the year (-15° C). Thus, the subjects are wearing heavy jackets and mostly winter boots. A total of 176 subjects were recorded in this session. The second session was recorded in April 2012. Temperatures were substantially higher (+15°C), thus, subjects were wearing significantly different clothes. In this recording, 161 subjects participated. A total of 32 subjects were recorded in both the first and the second session, thus the database contains a total of 305 individuals. The subset of 32 subjects allows research in time and clothing invariant gait recognition.

4.2. Sensor

For the recording, the Microsoft Kinect sensor [40] was used. This sensor provides a video stream, a depth stream and four-channel audio. Both video and depth are recorded at a resolution of 640×480 pixels at a frame rate of approximately 30 fps (slightly varying). The depth resolution is on the order of 1 cm. For depth acquisition, the sensor sends beams of infrared light and infers the depth from reflections on the objects. Therefore, placing the sensor outside is not possible, since infrared light from the sun can interfere with the depth sensor. According to the manufacturer, the optimal distance between the subject and the camera should be between 1.8 m and 3.6 m [40]. The four-channel audio is sampled with 24 bit at 16 kHz. The four microphones are spread horizontally at equal distances along the sensor.

4.3. Recording site

In order to simultaneously allow for video, depth and audio recording, a compromise in terms of location and the recording site had to be found. Ideally, for gait recognition, a larger distance of the subjects to the



(a) side view (b) top down view

Figure 1: Schematic of the recording site. (a) side view and (b) top-down view.

sensor is favorable, because it allows to capture multiple gait cycles in the field of view. A larger distance would also better resemble a wide area surveillance scenario, which is one of the primary applications for gait recognition. This, however, is in contrast to the technical restrictions of the sensor which allows a maximum of 3.6 m distance and favors an indoor environment. Also, for audio-based gait recognition a quiet environment and a good surface are preferred (when no additional enhancement methods are applied) to make the footsteps audible at high quality. These recording conditions, while mainly necessary for technical reasons, still closely relate to practical application scenarios such as access control in a narrow corridor.

To comply with the above requirements, a 3.5 m wide hallway corridor at the TUM university campus in downtown Munich was chosen as the recording site. Figure 1 illustrates the setup. The hallway has a constant background and a solid surface for good audio quality. The sensor was placed at 1.9 m height and is facing downwards at an angle of roughly 13°. The subjects are walking perpendicular to the line of sight at a distance of roughly 3 m close to the opposite wall. Thus, the subjects cover a distance of roughly 4 m when walking through the field of view. With the visible walking distance of roughly 4 m, typically each person makes between 1.5 and 2.5 gait cycles in each recorded sequence. The sequences each have a length of approximately 2 – 3 s.

(a) N (b) B (c) S (d) TN (e) TB (f) TS

Figure 2: Thumbnails of three male (top rows) and three female (bottom rows) participants in six variations: a) normal (N), b) backpack (B), c) coating shoes (S), d) time (TN), e) time + backpack (TB), f) time + coating shoes (TS).

4.4. Recording procedure

For recording, people walking in the hallway were randomly chosen and asked for their willingness to participate. Two markers were placed on the floor roughly one meter to the left and to the right of the field of view, respectively. Participating subjects were then asked to start at the left marker and walk to the right marker, then turn around (outside the field of view) and come back. This way, both sides of the person were recorded (essentially a 180° change in view angle). Placing the markers one meter to the left and the right of the visible area has proven to be a sufficient distance to allow for the participants to accelerate to a stable walking speed.

In order to allow for some kind of diversification in the video and depth stream as well as in the audio stream, we defined the following three variations:

- Normal walking (N1 – N6): People were asked to walk the distance a total of six times (three times to the right and three times back to the left) in a normal way.
- Backpack variation (B1 – B2): Carrying a backpack (of approximately 5 kg), people were asked to walk once to the right and once back to the left. The backpack constitutes a significant variation in visual appearance, as well as in gait pattern and sound.
- Shoes variation (S1 – S2): Coating shoes (as used in clean rooms for hygiene conditions) were put on the test subjects' shoes. This variation poses a considerable change in acoustic condition.

In summary, ten sequences were recorded for each of the 305 persons: Six normal walking (N1-N6), two backpack variation (B1 – B2) and two shoe variation (S1 – S2). Furthermore, the 32 people who participated in both recordings underwent ten additional recordings, namely TN1 – TN6, TB1 – TB2, TS1 – TS2. Figure 2 depicts images for three male and three female participants in all six configurations. It can be seen that the time variation shows significant changes in clothing, shoes and hair style.

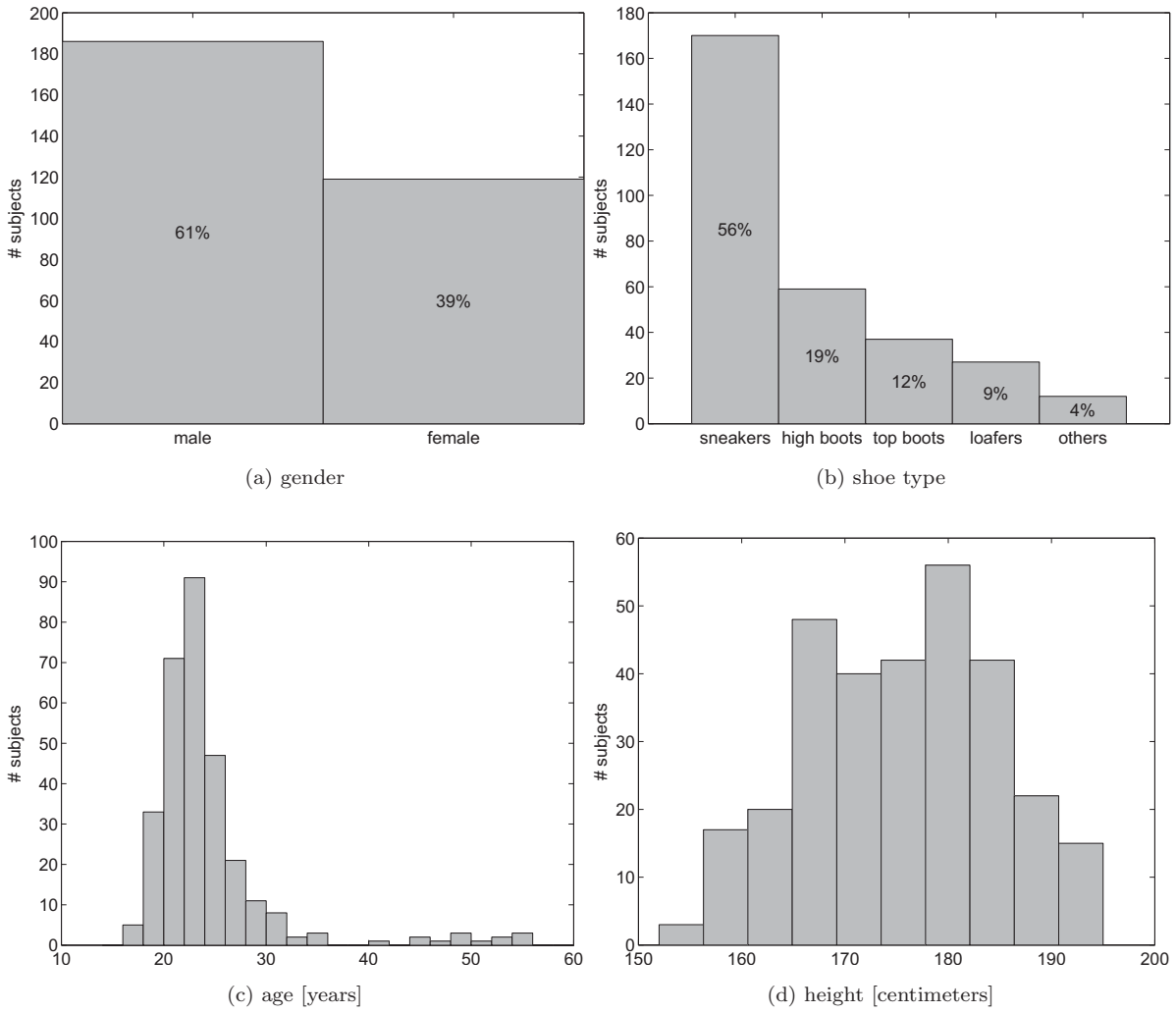
After the recording, participants were asked to sign a consent form to allow for research usage of their recordings.

The data was first captured within a raw *.oni* container (which is the raw format in the OpenNI framework[2]). In a pre-processing step, video frames were extracted frame-by-frame as jpeg images and depth data was extracted in a raw 16 bit binary format. The audio is captured in uncompressed (4-channel) wav format using 24 bit quantization at 16 kHz.

4.5. Metadata and database statistics

For each subject, the following metadata was recorded: (1) gender, (2) age, (3) height and (4) shoe type. The distribution of the metadata is shown in Figure 3. Of the 305 participating subjects, 186 (61 %) are

[2]http://openni.org



(a) gender

(b) shoe type

(c) age [years]

(d) height [centimeters]

Figure 3: Distribution of the subject metadata: gender (a), shoe type (b), age (c) and height (d)

male and 119 (39 %) are female. Ages of participants range from 18 to 55 years with an average of 24.8 years and a standard deviation of 6.3 years. The person height ranges from 152 cm to 195 cm with an average of 175.5 cm and a standard deviation of 9.5 cm. For shoe type, five classes of shoes were defined to classify the observed types : sneakers (56 %), high-boots (19 %), top-boots (12 %), loafers (9 %) and others (4 %) including sandals, ballerina and rubber boots.

5. Experiment Description

In order to facilitate and unify the evaluation process, we propose a set of experiments on the TUM GAID database. These experiments are meant to address a variety of challenges and will provide the basis

for competitive performance comparison of various algorithms. In order to define development data and prevent overfitting on the test data, the database (with 305 subjects) is divided into a *training*, *validation* and *test* set, which contain 100, 50 and 155 individuals, respectively. Each of these three subsets is roughly balanced in the available metadata. That means that each set contains roughly the same ratio of male/female, and similar distributions for age, height and shoe type. The data is used slightly differently for identification and soft biometrics experiments. Table 2 illustrates the setup for identification experiments and Table 3 for soft biometrics experiments.

5.1. Exp. I: Identification

For identification experiments, the training (100 subjects) and the validation set (50 subjects) are combined and together form the *development* set for person identification. This leaves a development set of 150 subjects and a test set of 155 subjects. Of the 32 persons who took part in both recording sessions, 16 are contained in the development set and 16 in the test set. The development set may be used for model building and for learning of covariates, as well as for parameter tuning. In the baseline experiments presented in Section 9.1, this data is only used for parameter tuning. Having such a dedicated development set (which is often ignored in previous gait recognition setups) is of central importance for building robust covariate-invariant recognition systems.

Final recognition experiments are carried out solely on the test set. Consequently, all results reported in this article are based on the test set. As depicted in Table 2, the test set is utilized as follows to get gallery and probe samples: The recordings N1, N2, N3, N4 are used as the gallery samples (for enrollment) and the recordings N5, N6, B1, B2, S1, S2 are used as probe data, separated in three experiments, namely the N ('normal walking'), B ('backpack variation') and S ('shoes variation') experiments.

In order to test the identification system for variances in appearance and background, a separate experiment is performed using only those persons who participated in both recording sessions. In these experiments, the enrollment (gallery) recordings (N1 – N4) are taken from the first recording session (the same 155 subjects as for experiment N, B, S), while the recordings from the second session (TN1 – TN6, TB1 – TB2, TS1 – TS2) are used for the identification experiments. In the experimental section, these experiments are denoted as TN, TB and TS. Note that

5.2. Exp. II: Soft Biometrics

Here, the term soft biometrics encompasses classification of gender, shoe type, height and age. For gender and shoe type, the task is a classification problem with a fixed number of classes (two for gender and five for shoe type), while height and age recognition are handled as a regression problem. For these experiments the test set is used to report results, while the training and validation sets are used for model creation and system optimization. The different subsets are person-disjunct, to perform person-independent experiments.

	test (155 subj.)
N1 – N4	Gallery
N5 – N6	Probe N
B1 – B2	Probe B
S1 – S2	Probe S
TN1 – TN4	-
TN5 – TN6	Probe TN
TB1 – TB2	Probe TB
TS1 – TS2	Probe TS

Table 2: Setup of the database for identification experiments

	train. (100 subj.)	val. (50 subj.)	test (155 subj.)
N1 – N4	✓	✓	✓
N5 – N6	✓	✓	✓
B1 – B2	-	✓	✓
S1 – S2	-	✓	✓
TN1 – TN4	-	-	-
TN5 – TN6	-	-	-
TB1 – TB2	-	-	-
TS1 – TS2	-	-	-

Table 3: Setup of the database for soft biometrics experiments

As shown in Table 3, for training, only the N recordings are used, while three validation and three test experiments are performed each using the N, B and S recordings. For validation experiments, training is done on the training set, while parameter tuning is done using the validation set. When performing test experiments, the training and validation set are merged to provide more data for model building.

6. Feature Extraction on RGB-D Data

To demonstrate the use of both visual and depth information, we applied the following algorithms to show baseline results on the newly created database: (1) The Gait Energy Image (GEI), (2) Gait Energy Image on Depth Data (depth-GEI), (3) Gait Energy Volume (GEV) and (4) Depth Gradient Histogram Energy Image (DGHEI). As depicted in Figure 4, only the standard GEI takes data from the RGB channel, while the other three methods extract features solely on the depth channel. The final features are visualized in Figure 5.

6.1. Gait Energy Image

The idea of Gait Energy Image (GEI) [1] is simple, yet has proven highly efficient. Assuming that all gait information is captured in a full gait cycle, the information of each frame within this gait cycle is averaged. This averaging seemingly discards information; however, assuming that the noise in each frame is independent, averaging removes a substantial part of the degradation.

The simplest feature (which is assumed to capture the gait information) is the silhouette. Thus, for GEI, in a first step, binary silhouettes are extracted at each frame, for example using Gaussian mixture

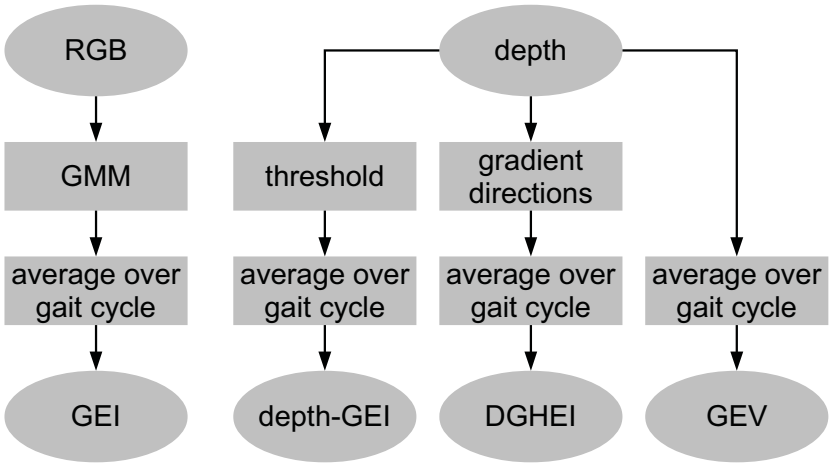
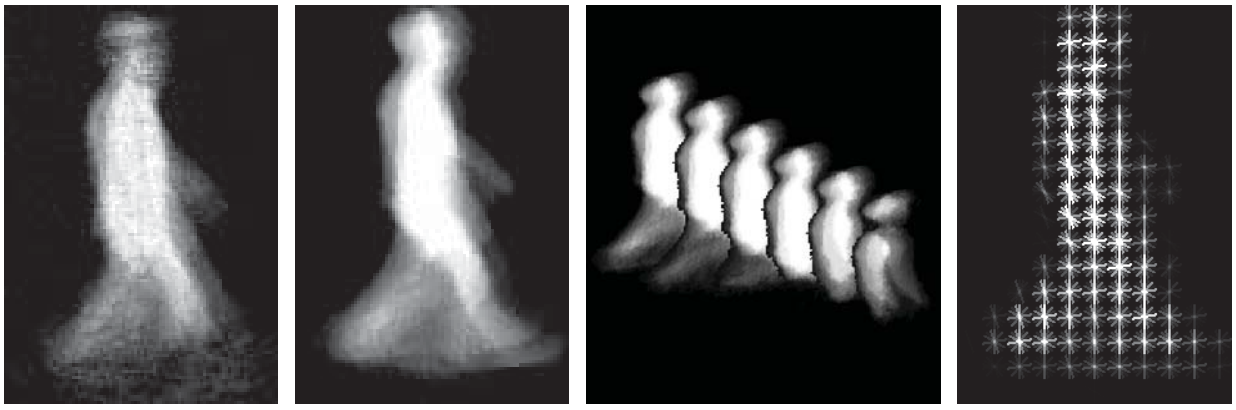


Figure 4: Feature extraction: The GEI is calculated using GMM background modeling on the RGB stream. Depth-GEI, DGHEI and GEV are extracted on the depth data.



(a) GEI (b) depth-GEI (c) GEV (d) DGHEI

Figure 5: Visual representation of the four used feature extraction techniques: a) GEI, b) depth-GEI, c) GEV, d) DGHEI. (All features are for the person in the top row of Figure 2)

models (GMMs) [41]. After processing with morphologic operations, the person is tracked by finding the largest blob. This found blob is extracted from the binary image and is resized (to 128×88 pixels in our experiments), such that all blobs in a sequence have the same size. In addition, the blobs are horizontally aligned by centering the blob around the center of the upper half of the blob. Thus, the horizontal position is normalized such that the torso in each frame is roughly at the same location.

Finally, the aligned silhouettes $S_t(x, y)$ are averaged over a gait cycle of T frames yielding the Gait Energy Image $GEI(x, y)$:

$$GEI(x, y) = \frac{1}{T} \sum_{t=1}^{T} S_t(x, y). \tag{1}$$

Extracting the GEI assumes on the one hand that the binary silhouettes actually capture the gait information, and second that errors in silhouette extraction result from independent noise at each frame. However, in practice the silhouette extraction is performed using background modeling methods which are run on the color images. Due to difficulties in the segmentation process, silhouettes can be of quite low quality and do not reliably capture the boundary of the subject. In addition, errors often result from local similarities of the person to the background. In these regions, the error is not independent at each frame.

6.2. Gait Energy Image on Depth Data

To overcome the limitations of GEI as explained above, depth information can be used instead of color information to obtain binary silhouettes $S_t(x, y)$. With depth information, the object can be reliably segmented from the background. Here, the background model is defined by the depth values of the empty scene. Since the depth-distance between object and background is relatively large, a large margin exists and simple thresholding of the depth difference results in good segmentation. With this modified extraction of silhouettes, the depth-GEI results from silhouette averaging analogously to Equation (1). The depth data used for silhouette extraction as explained above may be noisy and can contain (small) holes and missing data. Several methods, e. g. based on inpainting [42] or based on filter techniques [43] can be used to efficiently fill the holes. However, due to the temporal averaging in the GEI concept, such small degradations have little effect on the final features. As can be seen in Figure 5, the quality of the resulting depth-GEI is superior to that of the traditional GEI.

6.3. Gait Energy Volume

The Gait Energy Volume (GEV), as presented in [9], is the three dimensional extension of the Gait Energy Image. Instead of averaging binary 2D silhouettes, in GEV, 3D binary voxel volumes are averaged:

$$GEV(x, y, z) = \frac{1}{T} \sum_{t=1}^{T} V_t(x, y, z). \tag{2}$$

The binary voxel representations $V_t(x, y, z)$ is the volume of voxels, which are behind the surface reconstruction obtained using the depth channel.

The voxel data has to be carefully aligned (similarly to GEI) for a meaningful representation. It is important to note that, in the orginal publication [9], depth data from the frontal view is used, while in the TUM GAID database, profile side views are used.

6.4. Depth Gradient Histogram Energy Images

The Depth Gradient Histogram Energy Image (DGHEI) was first introduced in [10]. The noise reducing property of GEI by averaging feature vectors of each frame within a full gait cycle has proven highly efficient. The DGHEI also makes use of this concept. It is interesting to note that, in the standard GEI, all information is reduced to binary silhouettes. With the newly available depth information, the edges and depth gradients within the person's silhouettes can also be used. In order to capture all gradients and edges in a robust and efficient manner, we propose the use of histogram binning. This idea is motivated by the concept of 'histograms of oriented gradients' (HOG) as they are frequently used for object detection [44].

Extraction of DGHEI therefore in a first step consists of calculation histograms of oriented gradients at each frame t. While HOG uses grayscale images, DGHEI uses depth data. To this end, magnitude r and orientation θ of the gradient of the depth data D are computed in a first step:

$$r(x, y) = \sqrt{u(x, y)^2 + v(x, y)^2} \tag{3}$$

$$\theta(x, y) = \text{atan2}(v(x, y), u(x, y)) + \pi \tag{4}$$

with $u(x, y) = D(x-1, y) - D(x+1, y)$ and $v(x, y) = D(x, y-1) - D(x, y+1)$. Then, gradient orientations at each pixel are discretized into nine orientations:

$$\hat{\theta}(x, y) = \left\lfloor \frac{9 \cdot \theta(x, y)}{2\pi} \right\rfloor \tag{5}$$

These discretized gradient orientations θ are weighted by r and then aggregated into a dense grid of non-overlapping square image regions, the so called 'cells' (each containing typically 8×8 pixels). Each of these cells is thus represented by a 9-bin histogram of oriented gradients. Finally, each cell is normalized four times (by blocks of four surrounding cells each) leading to $9 \cdot 4 = 36$ values for each cell.

Next, following the averaging concept of GEI, the calculated gradient histograms are finally averaged over a full gait cycle consisting of T frames and result in the DGHEI:

$$H(i, j, f) = \frac{1}{T} \sum_{t=1}^{T} h_t(i, j, f) \tag{6}$$

Here, i and j are pointing to the histogram cell h at position (i, j) and $f = \{1 \dots 36\}$ is the index to the histogram bin. Each gait cycle is finally represented by a multidimensional feature vector $H(i, j, f)$.

In summary, DGHEI can be considered an extension to GEI which uses depth information instead of purely silhouette boundaries.

7. Feature Extraction on Audio Data

The Kinect sensor 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 [45].

The employed audio feature set is based on the baseline audio features we had provided for the Audio/Visual Emotion Challenge 2011 (AVEC 2011) [46] and contains a number of energy and spectral features. Compared to the AVEC 2011 feature set, the voicing related features were omitted. The employed features are supra-segmental features. This means that the acoustic descriptor signals 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 Tables 4 and 5, 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 on acoustic gait recognition [24, 25] such as the loudness, psychoacoustic sharpness or Mel-Frequency Cepstral Coefficients (MFCCs) are included in our feature set, which provides a substantial number of further acoustic feature information. 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.

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

Table 4: 25 energy and spectral-related acoustic low-level descriptors

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 5: 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.

8. Learning algorithms

Different learning algorithms are used for identification experiments and soft biometrics experiments. In this section we present the applied methods. We also explain the score level fusion applied for multi-modal fusion in case of identification experiments.

8.1. Exp. I: Identification

For person identification using RGB-D data, we apply Principal Component Analysis (PCA) followed by Linear Discriminant Analysis (LDA). While PCA seeks a projection that best represents the data in the direction of the highest covariance, LDA seeks a projection that best separates the data according to the class affiliation. The number of PCA components is set to the number of classes (thus 155 for test experiments). Final person identification is done using a 1-nearest neighbor classifier with a cosine distance measure. This combination of dimension reduction and classifier has proven highly effective [1] for problems with small amount of training data, such as it is typical for gait recognition.

When using audio features for person identification, SVMs with a linear Kernel function are applied. Sequential minimal optimization (SMO) (complexity 1.0) is used for training.

Fusion of video, depth and audio features is carried out on the score level. The best depth method (i. e., DGHEI) is fused with the video method (GEI) and with audio. Scores for video and depth are normalized from the cosine distance output. Normalized scores for audio are obtained from the pairwise voting of the multiclass SVM. For score level fusion, sum, product, max and min rules were applied, with the sum rule giving the best results. In addition, weighting of audio, depth and video scores led to further improvement. On the development set, it was found that weighting depth scores by a factor of two and weighting audio features with a factor of one half gave the best results.

8.2. Exp. II: Soft Biometrics

For the gender and shoe type classification experiments, both for audio video and depth features, SVMs are applied with the same settings as for the identification experiments. Two fusion strategies are tested for shoe type classification: Feature level fusion, where features of both modalities are combined before being fed to the classifier, and score-level fusion, where the output scores of the SVMs are combined additively (i. e. sum rule).

For video and audio features, the regression problems of age and height are addressed with unpruned REPTrees (25 cycles) with Random Subspace meta-learning, as implemented in the WEKA toolkit [47, 48]. Parameters are optimized using the validation set and led to the following values: A subspace size of 1 % of the features is used for age regression, while for height regression, 1 % is employed for the audio features and 5 % for the video features. For height regression, 500 iterations were employed and for age regression, 500 iterations for the video features and 200 iterations for the audio features.

9. Results

For all experiments, results are reported separately for the three different recording conditions (N: normal, B: backpack, S: coating shoes). As a measure, average identification accuracy (in percent) is used in the person identification experiments. This corresponds to the rank 1 identification rate. In addition we report rank 5 identification rate, which is the rate of correctly found subjects within the top 5 retrieved results.

For gender and shoe type classification, unweighted and weighted average recall (UAR/WAR) in percent are reported. WAR thereby resembles accuracy. UAR is the sum of recalls divided by the number of classes. As such, it can be more informative in case of class-imbalance as typically given in person trait assessment. In particular, the chance level is intuitive: 50 % UAR in the (two-class) case of gender assessment and 20 % UAR in the (five-class) case of shoe types. When doing age and height estimation, the cross correlation (CC), which is equal to the pearson correlation coefficient and mean absolute error (MAE) are reported. These measure follow the standard set by AVEC [46].

For identification as well as for soft biometrics experiments, all reported significance comparisons have been obtained using a one-sided z-test.

9.1. Exp. I: Identification

Test set results: Person identification							
Rank 1	N	B	S	TN	TB	TS	Ø
GEI	**99.4**	27.1	52.6	44	**6**	9	56.0
depth-GEI	96.8	3.9	88.7	28	0	22	58.8
GEV	94.2	13.9	87.7	41	0	31	61.4
DGHEI	99.0	40.3	**96.1**	50	0	44	74.1
Audio	44.5	27.4	4.8	3	0	3	23.4
Fusion (DGHEI + GEI)	**99.4**	51.3	94.8	**66**	3	**50**	77.9
Fusion (Audio + GEI)	**99.4**	45.2	44.2	41	**6**	9	58.8
Fusion (Audio + DGHEI)	99.0	52.3	95.5	50	0	47	77.6
Fusion (Audio + DGHEI + GEI)	**99.4**	**59.4**	94.5	**66**	3	**50**	**80.2**
Rank 5	N	B	S	TN	TB	TS	Ø
GEI	**100.0**	55.8	69.0	66	**31**	44	72.3
depth-GEI	99.0	16.8	96.5	50	9	41	67.3
GEV	98.1	37.7	95.5	59	9	50	73.6
DGHEI	**100.0**	73.2	**99.4**	69	16	**75**	87.3
Audio	70.3	44.2	11.0	3	6	12	38.6
Fusion (DGHEI + GEI)	**100.0**	80.3	99.0	**81**	28	72	90.1
Fusion (Audio + GEI)	**100.0**	68.4	70.0	66	**31**	41	76.3
Fusion (Audio + DGHEI)	**100.0**	79.0	**99.4**	69	16	**75**	89.1
Fusion (Audio + DGHEI + GEI)	**100.0**	**84.8**	99.0	**81**	28	72	**91.4**

Table 6: Rank 1 and rank 5 identification rates in percent for person identification for six different recording variations and their weighted average (Ø): normal (N), backpack (B), coating shoes (S), time (TN), time + backpack (TB), time + coating shoes (TS). In the N, B and S experiments, for each of 155 individuals, two recordings are evaluated while the TN, TB and TS sets each contain 16 persons with two recordings. For all experiments, the same gallery set with 155 classes is used. Thus, the chance level for all experiments is 0.6 %. Results are shown for video-based (GEI), depth-based (depth-GEI, GEV, DGHEI) and audio-based features as well as for four score-level fusion schemes.

Results for person identification are shown in Table 6. As described in Section 5.1, experiments are

carried out on the test set containing 155 subjects. Person IDs are trained on the gallery set (i. e. recording N1 – N4), containing 155×4 samples. The experiments N, B, S each contain 155×2 test samples. Of the 155 subjects, 16 have the additional time covariate, thus for the TN, TB, TS experiments, 16×2 test samples are available. For all six experiments, recognition is equivalent to a 155-class classification task. Thus, the chance level for all six experiments is 0.6 %.

First, we compare the results of the vision-based feature GEI and the three depth-based methods depth-GEI, GEV and DGHEI. With 99.4 %, the best results can be achieved for the normal setup (N) and using the vision-based Gait Energy Image (GEI). Thus, GEI performs best in cases of no change in appearance, while the depth-based methods reach lightly inferior results in the N setup with 94.2 % - 99.0 %.

The setup with the backpack (B) represents a significant degradation for the visual appearance (due to the strong change in silhouette shape) and consequently observed results are substantially lower than in the N and S setups. Out of the presented algorithms, DGHEI achieves the best accuracy of 40.3 %. The DGHEI seems to have the best generalization capability, which allows handling such significant changes in visual appearance. This degradation with the backpack becomes even more evident in the TB setup, where recognition rates drop to zero. In the other time setups TN and TS, creditable results can be achieved, given the difficulty of these experimental setups.

The Gait Energy Image (GEI) is based on the color video stream and uses a foreground/backgound segmentation technique, which can be erroneous. One hypothesis was that depth-GEI which uses the depth channel and has visually much better silhouettes would outperform the traditional GEI. Looking at the results, this is only true in the shoe setups S and TS. Similar results can be reported for the Gait Energy Volume which outperforms traditional GEI mainly in the shoe setups. Overall, the Depth Gradient Histogram Energy Image (DGHEI) gives the best average performance results, significantly better than GEI, depth-GEI and GEV (at a 0.001 level). DGHEI performs worse than GEI in setup N and TB, which, however, is not significant even at a 0.05 level. The used gradient histogram representation thus seems to able to reliably handle variations such as backpack, coating shoes and time. Similar tendencies can be observed in rank 5 recognition results.

Interesting results can be obtained in case of multimodal fusion. For this, score level fusion as detailed in Section 8.1 is applied. We combine GEI from the color video channel with our best depth-based method, namely DGHEI. Table 6 shows that this fusion takes the best out of the two modalities. In difficult set-ups such as the backpack scenario, fusion can exceed a mere 27.1 % for GEI and 40.3 % for DGHEI in case of single modality and reaches up to 51.3 %. When fusion methods within the same modality are considered, for example GEV and DGHEI, no such improvements can be observed. This shows the significance of using separate modalities in the fusion process.

As expected, recognizing humans by acoustic features turns out to be much more challenging than with visual features. Nevertheless, the results show that it is in fact possible to recognize people by their acoustic

walking characteristics, achieving 44.5 % accuracy in the N setup. The backpack naturally poses the least downgrade to acoustic gait recognition, while coating shoes completely degrade recognition results. In the time setups TN, TB, TS, where acoustics are significantly different than in the gallery set, audio-based gait recognition does not work with the proposed baseline algorithm, which does not contain any session variability handling.

However, audio-based gait recognition has potential to improve vision-based algorithms, especially in those variations in which vision-based methods struggle. This can be seen in case of fusion of audio with vision-based features (GEI). In setup (B), a 27.1 % recognition rate in GEI combined with 27.4 % in audio leads to 45.2 % in case of fusion. Significant fusion gain can be observed in case of fusing audio and depth-based features (DGHEI). Here, in case of the backpack setup (B), fusion of multiple modalities gives a relative performance gain of 29.8 % (from 40.3 % to 52.3 %; significant at a 0.002 level), while fusion does not decrease recognition rates in other setups (except setups S, which is not significant). Thus, when considering the weighted average score, a performance gain from 74.1 % to 77.6 % (significant at the 0.05 level) can be observed.

Finally, experimental results for multimodal fusion across audio, video (GEI) and the best depth-based feature (DGHEI) have been calculated. While for most categories, the recognition results stay the same with this fusion scheme, fusion can help in the case of challenging categories such as in case of backpack variation. While the best single modality reaches a mere 40.3 % recognition rate for setup B, fusion can lift recognition rates to 59.4 % (significant improvement at a 0.001 level). Thus, all modalities show contribution, which demonstrates the effectiveness of simultaneously using multiple modalities.

It is important to note that, in all our person identification experiments, we only make use of the 155 persons of the test set (as detailed in Section 5.1). Here, learning of person identities is solely done on the gallery samples, which only contain normal walking. Consequently recognition rates plummet in case of variations such as backpack and shoes. It can therefore be assumed that recognition rates could greatly benefit from separately learning the covariates (backpack, coating shoes, time) on the other 150 people in the *development set*, which is in principle possible using the presented database.

9.2. Exp. II: Soft Biometrics

Table 7 shows results for soft biometrics experiments on the test set. In gender recognition, with our video features, the best result in UAR is 95.6 % in the N setup. Results for the S setup are only slightly worse while the B setup leads to a substantial drop in performance. Using audio features, all setups lead to a UAR slightly above 60%, which is always significantly (at least at the 0.05 level) worse then the results obtained with video and depth features. Note that feature-level fusion did not lead to an improvement in gender recognition accuracy owing to the large discrepancy between the performance of the modalities and therefore, the results are not displayed here.

Test set results: Soft biometrics						
a) Gender	N		B		S	
%	UAR	WAR	UAR	WAR	UAR	WAR
DGHEI	**95.6**	**95.8**	**66.4**	**74.8**	**90.9**	**92.9**
Audio	61.0	61.3	60.3	62.9	63.2	65.2
b) Shoe type	N		B		S	
%	UAR	WAR	UAR	WAR	UAR	WAR
DGHEI	**33.7**	60.4	30.6	31.6	26.8	48.7
Audio	30.9	53.9	27.9	52.3	**31.3**	41.3
Fusion (feature level)	31.7	60.4	29.9	42.3	26.6	**52.3**
Fusion (score level)	31.4	**63.4**	**30.9**	**56.8**	29.0	**52.3**
c) Age	N		B		S	
–/years	CC	MAE	CC	MAE	CC	MAE
DGHEI	**.41**	**3.27**	**.33**	4.73	**.43**	**3.35**
Audio	.09	3.52	-.01	**3.58**	-.05	4.32
d) Height	N		B		S	
–/cm	CC	MAE	CC	MAE	CC	MAE
Silhouette height	.73	**4.66**	**.74**	**4.56**	.72	**4.70**
DGHEI	**.77**	5.30	**.74**	6.11	**.74**	5.55
Audio	.36	7.83	.31	7.87	.16	8.37

Table 7: Results for soft biometrics experiments on the test set (155 subjects) for three different recording variations: normal (N) with six recordings per person, backpack (B) with two recordings and coating shoes (S) with two recordings. For gender and shoe type classification, unweighted and weighted average recall (UAR/WAR) in percent are reported. For age and height estimation, cross correlation (CC) and mean absolute error (MAE) in years and cm, respectively, are given. To assess the MAE with respect to the database distribution, note that, in the test set the mean absolute deviation is 3.3 years for age and 8.1 cm for height. Thus, regression results are only slightly better than guessing.

Shoe type recognition rates are around 30 % UAR and likewise significantly above chance level, both with video or audio features. Audio features are better for the B setup. Here, WAR is significantly better (at

the 0.001 level) for audio. Video features are in general better for the N and S setup. Since both modalities achieve similar recognition rates, fusion is applied. For UAR, no significant improvement can be achieved, while for WAR, score-level fusion improves the results of all three setups. Note that UAR and WAR results diverge strongly for shoe type recognition. This is because of the unbalanced distribution of different shoe types in the database. A balancing strategy was implemented in order to rectify this imbalance. Thereby, the training data of shoe types with smaller amounts of data are upsampled in order to have a roughly uniform distribution over different shoe types. However, no significant improvement in UAR could be achieved. One reason could be that there are just not enough training data of the smaller classes to create a well generalized classifier. The results shown here are based on experiments without this balancing strategy.

Using audio data, no significant results are achieved for age estimation. While at first sight, the resulting MAE for age estimation appears very promising (e. g. compared to [49]), it should be kept in mind that most of the subjects in the database are between 20 and 30 years old. The mean absolute age deviation in the test set is only 3.3 years. Therefore, the resulting MAE needs to be evaluated in this context.

Both for young and old subjects, the algorithm mostly predicts ages between 23 and 27 years. Therefore, the MAE is close to the mean absolute age deviation in the test set. This leads to the conclusion that the extracted audio features are uncorrelated with the age of the persons.

Intuitively, older subjects should produce significantly different sounds of footsteps as compared to younger subjects, due to different pace, for example. However, the features used in this study are not customized for this task and are thus not robust enough to estimate the age correctly. Additionally, the bias in age in the database (only a small portion of older subjects) makes it difficult for the learning algorithm to learn the sounds of these subjects.

With depth-based features, correlation coefficients around 0.4 are achieved in the age estimation task. The predicted ages cover a broader range and ages of older subjects are estimated better. Therefore, a better correlation coefficient is achieved as compared to the audio features. However, for some younger persons, a higher age is predicted, which especially in the B experiment leads to a high MAE. Due to the age bias in the database, the MAE is not better than the mean absolute deviation. Using features which are more suited for the task of age estimation should lead to better results.

For height estimation, better results are obtained: With depth-based features (DGHEI), correlation coefficients above 0.7 are achieved with the best MAE being as low as 5.30 cm for the N setup. As an additional baseline, height estimation from silhouette height is presented. Support Vector Regression is used on the average pixel height over a sequence to estimate the height. This simple approach outperforms regression from the height-normalized DGHEI. On the test set, the mean absolute height deviation is 8.1 cm, which reveals the potential of height regression. For audio data, reasonable results are achieved at least in the N and B setup.

10. Outlook and Conclusion

In this article we have contributed to the field of identification of humans by the way they walk (i. e. gait recognition) in three ways:

(1) We have presented the TUM GAID database, which comprises data of 305 subjects, in various covariates. Most of all, this database is the first to simultaneously contain RGB data, depth data, as well as audio data. The database is publicly and freely available[3] and is meant to foster research in multimodal gait recognition.

(2) Our second contribution is a precise definition of experimental setups. We define classical human identification experiments with gallery and probe sets. As opposed to previous experiment definitions, we split the whole database roughly into two halves leaving the first half for development and tuning approaches. This is contrary to previous databases, where tuning and development always had to be carried out directly on the test data. In this way, overfitting and optimization on the test set is prevented. Besides the human identification experiments, we also precisely defined experiments for soft biometrics. Here, the first half of the data is further split into a training and validation set, which can be used to tune soft biometrics algorithms. Metadata for gender, shoe type, age and height is provided which makes gait-based recognition and regression on these characteristics possible. While the metadata is roughly balanced and thus allows for good analysis, further recordings (e. g. of older subjects) could make the data even better suited for soft-biometrics in a wider range.

(3) Finally, we presented several baseline algorithms and according results for reference both for the human identification task as well as for the soft biometrics tasks. All modalities, that is RGB data, depth data and audio data, could successfully be applied. In general, vision-based and depth-based approaches outperformed the audio-based approaches. Using only a single modality, it can be seen that depth – and especially depth gradient histograms – lead to superior results. This shows the value of depth information for person identification. However, the results also show that recognizing subjects as well as soft biometrics recognition is very well possible with audio features. Fusion of visual and audio features partially resulted in additional performance gains, especially in setups, where vision features show low performance (e. g. backpack variation). However, also fusion of vision-based and depth-based features led to improvement. The overall best results were obtained in case of fusing all three modalities, i. e. vision, depth and audio. This shows that fusing data captured with different sensors can be highly beneficial. In future work, more sophisticated fusion schemes may possess the potential to even further leverage to benefits of multimodal fusion.

Overall, the aim of the presented article is to encourage further research in multimodal gait recognition, as well as in fusion of those modalities. The presented experimental setups are meant to be used in the same

[3] `www.mmk.ei.tum.de/tumgaid`

way by all researchers using the database to allow for competitive performance analysis. Even though the baseline algorithms shown in this work show creditable results, there is much room for improvement. For the identification experiments, model building was done employing only the test set while the develeopment set was used solely for parameter tuning. Future work can make extensive use of the provided development data for model building or learning of background models. This also includes the possibility of specifically learning the time covariate, which was not possible with the setups used with earlier databases. Due to the fact that faces are captured at a relatively high resolution, the database can potentially be used for combined analysis of face and gait recognition, which is a highly relevant direction of research that has only received limited attention in the past. Our current baseline approaches use separate methods for video and audio. Future work will investigate combined audio-visual gait models which will tie the modalities closer together. In our experiments, using depth data led to the best results. With the provided database, future work can further analyse depth as a modality and the relatively young field of depth-based gait recognition can be brought forward.

11. Acknowledgements

This research was supported by the ALIAS project (AAL-2009-2-049) co-funded by the EC, the French ANR and the German BMBF. The authors would like to thank Adriana Anguera for her valuable input.

References

[1] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2) (2006) 316–322.

[2] Y. Huang, D. Xu, T.-J. Cham, Face and human gait recognition using image-to-class distance, IEEE Transactions on Circuits and Systems for Video Technology 20 (3) (2010) 431–438.

[3] A. Kale, A. Sundaresan, A. Rajagopalan, N. Cuntoor, A. RoyChowdhury, V. Krueger, Identification of humans using gait, IEEE Transactions on Image Processing 13 (9) (2004) 1163–1173.

[4] Z. Liu, S. Sarkar, Improved gait recognition by gait dynamics normalization, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (6) (2006) 863–876.

[5] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, K. W. Bowyer, The humanID gait challenge problem: Data sets, performance, and analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 162–177.

[6] M. Hofmann, G. Rigoll, Improved gait recognition using gradient histogram energy image, in: IEEE International Conference on Image Processing, Orlando, FL, USA, 2012, pp. 1389–1392.

[7] M. Hofmann, S. Schmidt, A. Rajagopalan, 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.

[8] R. D. Seely, S. Samangooei, L. Middleton, J. N. Carter, M. S. Nixon, The university of southampton multi-biometric tunnel and introducing a novel 3D gait dataset, in: Proc. of IEEE International Conference on Biometrics: Theory, Applications and Systems, Washington, DC, USA, 2008, pp. 1–6.

[9] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, C. Fookes, Gait energy volumes and frontal gait recognition using depth images, in: Proc. of International Joint Conference on Biometrics, Washington, DC, USA, 2011, pp. 1–6.

[10] M. Hofmann, S. Bachmann, G. Rigoll, 3d 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.

[11] Y. Shoji, T. Takasuka, 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.

[12] K. Mäkelä, J. Hakulinen, 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.

[13] C. BenAbdelkader, R. Cutler, L. Davis, Stride and cadence as a biometric in automatic person identification and verification, in: Proc. of IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 2002, pp. 372–377.

[14] C. Yam, M. Nixon, J. Carter, Automated person recognition by walking and running via model-based approaches, Pattern Recognition 37 (5) (2004) 1057–1072.

[15] D. Tao, X. Li, X. Wu, S. Maybank, General tensor discriminant analysis and gabor features for gait recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1700–1715.

[16] A. Sundaresan, A. Chowdhury, R. Chellappa, A hidden markov model based framework for recognition of humans from gait sequences, Proc. of IEEE International Conference on Image Processing (2003) 90–93.

[17] A. Kale, A. Roy-Chowdhury, R. Chellappa, Fusion of gait and face for human identification, in: Proc. Acoustics, Speech, and Signal Processing 2004, Vol. 5, 2004, pp. 901–904.

[18] X. Zhou, B. Bhanu, Feature fusion of side face and gait for video-based human identification, Pattern Recognition 41 (3) (2008) 778–795.

[19] T. Zhang, X. Li, D. Tao, J. Yang, Multimodal biometrics using geometry preserving projections, Pattern Recognition 41 (3) (2008) 805–813.

[20] T. K. M. Lee, S. Ranganath, S. Sanei, Fusion of chaotic measure into a new hybrid face-gait system for human recognition, in: Proceedings of the 18th International Conference on Pattern Recognition - Volume 04, ICPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 541–544.

[21] G. Shakhnarovich, L. Lee, T. Darrell, Integrated face and gait recognition from multiple views, in: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Vol. 1, 2001, pp. 439–446.

[22] X. Zhou, B. Bhanu, Feature fusion of face and gait for human recognition at a distance in video, in: Proc. of the 18th Int. Conf. on Pattern Recognition, Vol. 4, Hong Kong, China, 2006, pp. 529–532.

[23] B. She, Framework of footstep detection in in-door environment, Proc. of International Congress on Acoustics (2004) 715–718.

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

[25] A. Itai, 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.

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

[27] K. Kalgaonkar, 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.

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

[29] J. Little, J. Boyd, Recognizing people by their gait: The shape of motion, Videre: Journal of Computer Vision Research 1 (2) (1998) 1–32.

[30] R. Gross, J. Shi, The CMU motion of body (mobo) database, Tech. rep., CMU (2001).

[31] A. Y. Johnson, A. F. Bobick, A multi-view method for gait recognition using static body parameters, in: Proc. of International Conference on Audio- and Video-Based Biometric Person Authentication, Halmstad, Sweden, 2001, pp. 301–311.

[32] A. Kale, N. Cuntoor, R. Chellappa, A framework for activity-specific human identification, in: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Orlando, FL, USA, 2002.

[33] N. Cuntoor, A. Kale, R. Chellappa, Combining multiple evidences for gait recognition, in: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, China, 2003, pp. 6–10.

[34] R. T. Collins, R. Gross, J. Shi, Silhouette-based human identification from body shape and gait, in: Proc. of IEEE Conference on Face and Gesture Recognition, Washington, DC, USA, 2002, pp. 351–356.

[35] J. S. M. Nixon, J. Carter, M. Grant, Experimental plan for automatic gait recognition., Tech. rep., University of Southampton (2001).

[36] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: Proc. of IEEE International Conference on Pattern Recognition, Vol. 4, Hong Kong, China, 2006, pp. 441–444.

[37] Y. Makihara, H. Mannami, A. Tsuji, M. Hossain, K. Sugiura, A. Mori, Y. Yagi, The OU-ISIR gait database comprising the treadmill dataset, IPSJ Transactions on Computer Vision and Applications 4 (2012) 53–62.

[38] D. Matovski, M. Nixon, S. Mahmoodi, J. Carter, The effect of time on gait recognition performance, Information Forensics and Security, IEEE Transactions on 7 (2) (2012) 543 –552.

[39] M. Hofmann, S. Sural, G. Rigoll, Gait recognition in the presence of occlusion: A new dataset and baseline algorithms, in: Proc. of International Conference on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, 2011, pp. 99–104.

[40] www.xbox.com/kinect/, accessed 13.07.2012.

[41] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, 1999, pp. 2246–2252.

[42] A. Criminisi, P. Perez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, Image Processing, IEEE Transactions on 13 (9) (2004) 1200–1212.

[43] M. Camplani, L. Salgado, Efficient spatio-temporal hole filling strategy for kinect depth maps, in: IS&T/SPIE International Conference on 3D Image Processing (3DIP) and Applications, Vol. 8290, Burlingame, CA, USA, 2012, pp. 82900E 1–10.

[44] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. of International Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 2005, pp. 886–893.

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

[46] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, 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.

[47] T. Ho, The random subspace method for constructing decision forests, Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (8) (1998) 832–844.

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The weka data mining software: an update, SIGKDD Explorations 11 (1) (2009) 10–18.

[49] Y. Makihara, M. Okumura, H. Iwama, Y. Yagi, Gait-based age estimation using a whole-generation gait database, in: Proc. of the International Joint Conference on Biometrics (IJCB2011), Washington DC, USA, 2011, pp. 1–6.