Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Multimodal features fusion for gait, gender and shoes recognition

Francisco M. Castro · Manuel J. Marín-Jiménez · Nicolás Guil

Received: date / Accepted: date

Abstract The goal of this paper is to evaluate how the fusion of multimodal features (i.e. audio, RGB and depth) can help in the challenging task of people identification based on their gait (i.e. the way they walk), or gait recognition, and by extension to the tasks of gender and shoes recognition. Most of previous research on gait recognition has focused on designing visual descriptors, mainly on binary silhouettes, or building sophisticated machine learning frameworks. However, little attention has been paid to audio or depth patterns associated to the action of walking. So, we propose and evaluate here a multimodal system for gait recognition. The proposed approach is evaluated on the challenging 'TUM GAID' dataset, which contains audio and depth recordings in addition to image sequences. The experimental results show that using either early or late fusion techniques to combine feature descriptors from three kinds of modalities (i.e. RGB, depth and audio) improves the state-of-the-art results on the standard experiments defined on the dataset for the tasks of gait, gender and shoes recognition. Additional experiments on CASIA-B (where only visual modality is available) support the benefits of feature fusion as well.

1 Introduction

Given a video sequence depicting people walking, the goal of this work is to identify them (i.e. assign an identity) based on the way they walk. This problem is known as *gait recognition*. Typically, gait recognition has been based on features extracted from visual information, as sequences of silhouettes [11] or dense trajectories of points [4]. However, the use of audio and depth information has not been extensively studied. This is in part due to the lack of audio tracks and depth maps associated to video sequences, as happens in the most popular databases for gait recognition (e.g. CASIA gait dataset [35]). Recently, Hofmann et al. [12] released 'TUM Gait from Audio, Image and Depth (GAID) database', or 'TUM GAID'. This dataset of gait sequences contains not only image sequences, but also audio recordings and depth information provided by a Microsoft Kinect sensor. The results they present on the RGB channel are based on binary silhouettes (i.e. Gait Energy Image [11]) combined with a 1-Nearest Neighbor classifier on PCA+LDA-compressed data. They also extract diverse audio features in the time and frequency domain, training a linear SVM on top of them. Finally, they use a simple fusion method on the score domain (i.e. after classification from each modality) to take a decision – giving the sum rule the best results during fusion.

Nevertheless, most of the best results presented in [12] are based on depth information, leaving the image and audio information in a second place. In contrast, we propose in this paper a multimodal approach for gait recognition where non silhouette-based visual features take more relevance, as well as the fusion information methods, showing that we are able to obtain an accuracy of 100% in gait recognition by combining two sets of tracklet-based features [4] (one computed in RGB and the other in depth). On the other hand, in gender recognition and shoes recognition, the combination of audio features with the tracklets computed in RGB and depth outperforms the previous results reported in [12].

The contributions of this paper, which extends a preliminary version published in [3], are: (i) a new depth-based descriptor for gait recognition; (ii) a unified approach for using

F.M. Castro and N. Guil
Department of Computer Architecture.
University of Málaga, 29071, Málaga (Spain). Tel.:(+34)952133388

M.J. Marín-Jiménez
Department of Computing and Numerical Analysis. Maimonides Institute for Biomedical Research (IMIBIC).
University of Córdoba, 14071, Córdoba (Spain). Tel.:(+34)957212255
E-mail: fcastro@uma.es, mjmarin@uco.es, nguil@uma.es

audio, depth and visual information for the problem of gait; and, (iii) a multimodal-based approach for gender and shoes recognition. In addition, we carry out a thorough evaluation of the proposed approach on two challenging datasets: TUM GAID and CASIA-B; and, according to the results, we set new state-of-the-art results in both datasets.

The rest of the paper is organized as follows. After presenting the related works, Sec. 2 describes the multimodal features proposed to describe gait. In Sec. 3, different fusion information strategies are discussed. The experiments and results are presented in Sec. 4. Finally, Sec. 5 contains the conclusions of this work.

1.1 Related work

Many research papers have been published in recent years tackling the problem of human gait recognition. For example, in [14] we can find a survey on this problem summarizing some of the most popular approaches based on video information. Some of them use explicit geometrical models of human bodies, whereas others use only image features. The most popular approach is a silhouette-based descriptor, called Gait Energy Image (GEI) [11], where the main idea is to compute a temporal averaging of the binary silhouettes of the target subject. Based on this approach, many improvements have been published like the approach of Liu et al. [19] whose key idea is to compute Histogram of Oriented Gradients (HOG) descriptors from GEI and Chrono-Gait Image (CGI). Martin-Felez and Xiang [22] propose a new ranking model for gait recognition based on GEI. This new formulation allows to leverage training data from different datasets improving the recognition performance. Zeng et al. [36] propose a new approach that uses deterministic learning for building a RBF-based network. This network is trained with the silhouettes of the subject and is able to obtain a representation that combines spatio-temporal information with shape information. With this new approach, they are able to obtain state-of-the-art results in experiments recorded at different times of the year. To deal with different speeds of the gait, Guan and Li [10] build a robust representation based on random subspaces that are able to extract information invariant to the gait speed. In [37], the authors propose a new gait representation based on computing active regions of the silhouettes along the sequence. Away from these methods, Hu et al. [13] use optical flow to represent the gait movement.

Gait recognition based on audio information is a less studied problem due to its high complexity and lack of databases. The main gait dataset with audio information is the TUM GAID dataset [12]. The authors of this dataset present a baseline method based on building a high level descriptor that is composed of a set of low level descriptors like loudness or Mel-Frequency Cepstral Coefficients (MFCC). Finally, this high level descriptor is used as input for a SVM. Geiger et al. [8] propose an improvement of the previous method by selecting the features that produce best results and removing the rest. Thus, the noise is reduced and the results are improved. Later, the same authors propose a new method based on Hidden Markov Models (HMM) [9] that outperforms the previous methods in terms of precision.

The use of depth information to perform gait recognition has increased in the last years due to the availability of cheap devices that can record depth information like Microsoft Kinect. Traditional works use 2D silhouettes extended to 3D (each pixel represents the depth information instead of the colour or the intensity of the image) as basic descriptors with several modifications to improve on the previous approaches. In [26], Sivapalan et al. propose a new descriptor called Gait Energy Volume (GEV) which is an extension to 3D of the traditional silhouette-based descriptor GEI. Hofmann et al. propose another extension of GEI to 3D based on computing histograms of oriented gradients (HOG) over the 3D silhouettes. A novel representation is proposed by Whytock et al. [32], where instead of using the whole silhouette of the subject, a skeleton on the 3D silhouette of the person is computed. By this way, they are able to smooth the movement of the person and to avoid wrong representations due to poor segmentations. Recently, Lopez-Fernandez et al. [20] proposed a new descriptor for sequences of gait volumes coined Gait Entropy Volume (GEnV). GEnV is robust to viewpoint changes, being suitable for unconstrained curved trajectories, however it requires multiple cameras for obtaining the person reconstructions.

Finally, if multiple information sources are available, the natural idea is to try to combine or to fuse the information from them to build a richer representation. The main strategies of fusion are early fusion and late fusion. Early fusion, or feature level fusion, has been widely used and a large number of approaches have appeared in recent years [1]. The most simple early fusion strategy consists of concatenating descriptors of different sources to build a new and larger descriptor. The next natural step is to learn what features of each modality should be selected in order to obtain a better and more compact representation. This is performed in [33] by using a common dictionary that selects the information from each modality. A more complex strategy is the called Multiple Kernel Learning (MKL) [2] which learns a linear or non linear kernel combination and the associated classifier simultaneously. On the other hand, late fusion aims to combine confidence scores of features of different models. In this case, the easiest strategy is to build the combined confidence scores based on weighted scores of different models according to their global accuracy. A more powerful approach is presented in [34] where the method tries to build new confidence scores by removing the possible classification error following a minimization process. An intermedi-

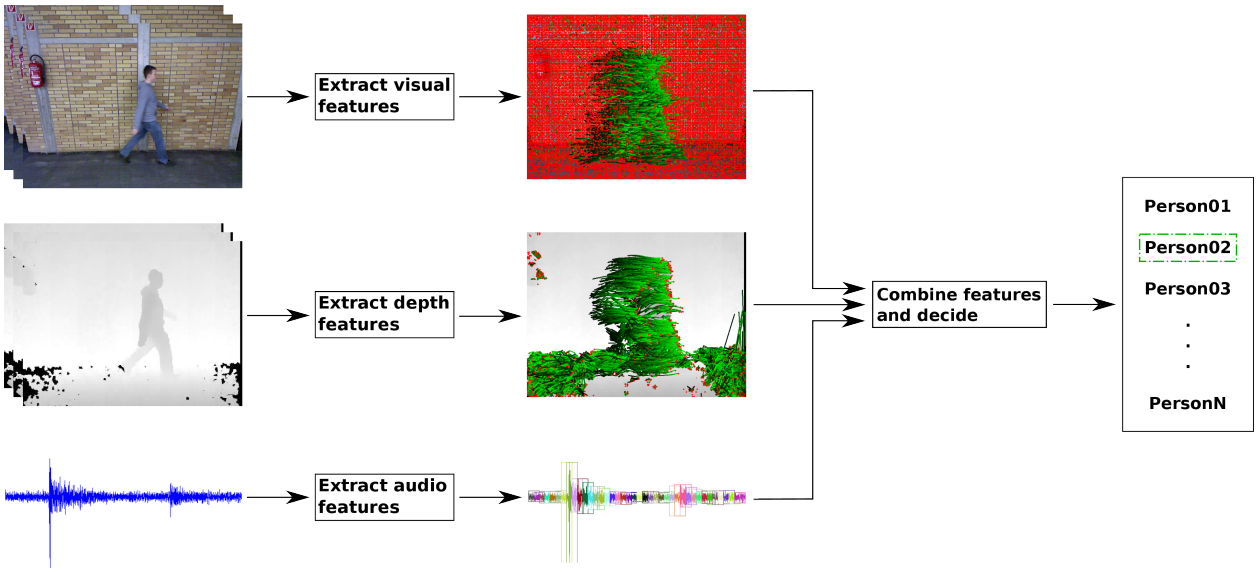

Fig. 1 Proposed pipeline for human gait recognition. Given a video sequence of people walking, audio, depth and visual information are combined in order to assign an identity from a set of predefined ones.

ate approach could be the use of a classifier (e.g. SVM) to learn a combined representation of the scores to maximize the global accuracy as in [21].

2 Multimodal features

In this section we will present how we describe gait sequences from audio, depth and visual information, obtaining a representation from each modality that can be directly fed into a classifier. We adopt the following pipeline to describe a gait sequence: (i) extract low-level features from each modality (i.e. audio, depth and image); (ii) generate a mid-level representation from each modality; (iii) fuse information from the different modalities; and (iv) assign an identity. This pipeline is visually represented in Fig. 1.

2.1 Audio descriptors

When an audio track is available for the target video sequence, we can compute a set of low-level descriptors from it, each representing different properties of the sound. The following descriptors are computed from audio frames (i.e. continuous portion of the audio track) of t seconds (e.g. $t = 0.05$) as in [21]: (*i*) basic statistics: zero-cross, coefficient of skewness, excess kurtosis, flatness and entropy; (*ii*) Mel spectrum; and, (*iii*) Mel-Frequency Cepstral Coefficients (MFCC).

We use 'MIR Toolbox v1.5' [17] for their computation. Then, all those low-level descriptors are concatenated into a single aural descriptor for a given audio frame. The dimensionality of the derived audio descriptor is 58 dimensions (i.e. 5 for basic statistics, 40 for Mel spectrum, 13 for MFCC).

2.2 Visual descriptors

From all the possible visual descriptors existing in the literature for video sequences, we will choose descriptors that represent motion. Therefore, the first step to extract motion-based descriptors is to compute densely sampled trajectories (or tracklets). Those tracklets are computed by following the approach of Wang et al. [31].

The first step is to compute dense optical flow [7] $F = (u_t, v_t)$ on a dense grid (i.e. step size of 5 pixels and over 8 scales). Then, each point $p_t = (x_t, y_t)$ at frame t is tracked to the next frame by median filtering as: $p_{t+1} = (x_t, y_t) + (M * F)|_{(\bar{x}_t, \bar{y}_t)}$, where M is the kernel of median filtering and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of p_t. To minimize drifting effect, the tracking is limited to L frames. We use $L = 15$ as in [15]. As a post-processing step, uninformative and noisy trajectories (e.g. excessively short or showing sudden large displacements) are removed.

As in [4], we are only interested in those tracklets that are generated by people. Therefore, we detect people in image sequences by background subtraction [16], fit a bounding-box to each detection and remove those tracklets that do not share spatio-temporal coordinates with the person bounding-boxes.

2.2.1 DCS features.

Once the local trajectories are computed, one choice is to describe them with the Divergence-Curl-Shear (DCS) descriptor proposed by Jain et al. [15]. As described in [15], the divergence is related to axial motion, expansion and scaling effects, whereas the curl is related to rotation in the image plane. DCS has been already employed by Castro et al. in [4] in the problem of gait recognition with excellent results. Note that a DCS descriptor is typically composed of

four parts: 2D normalized coordinates, a histogram combining Divergence and Curl, a histogram combining Curl and Shear, and a histogram combining Divergence and Shear. Using the default parameters of the software released by [15], this descriptor has 318 dimensions (2D coordinates: 30; Div+Curl: 96; Curl+Shear: 96; Div+Shear: 96).

2.2.2 H2M features.

An alternative descriptor commonly used in conjunction with tracklets is the concatenation of Histograms of Oriented Gradients (HOG) [5], Histograms of Optical Flow (HOF) and Motion Boundary Histograms (MBH) [6], named H2M for compactness.

We use the software of [31] to compute H2M, so the default sizes for HOG, HOF and MBH are 96, 108 and 192, respectively. These, in combination with the 30 dimensions derived from the normalized 2D coordinates of the tracklets, makes a total of 426 dimensions.

2.3 Depth descriptors

Depth information represents the distance between the surface of the element or elements recorded and a reference point, generally the camera that is recording the scene. This can be represented as an image where each pixel represents the distance between that point and the reference point. Therefore, depth information can be treated as another kind of visual stream of data (like RGB, infrared or thermal), and we can compute local trajectories and motion descriptors as in a RGB video (see Fig. 2, bottom). For this aim, we have to pre-process the data provided by the depth sensor to obtain a visual representation of the depth information. This is performed for each frame by scaling the original range of input values for each pixel into a new range compatible with the traditional image representation (e.g. $[0,255]$). Note that, depending on the recording device, the original input range will vary. In our concrete case, TUM GAID dataset (see Sec. 4.1) is recorded with a Microsoft Kinect so the input range is $[0,4000]$ and the selected output range is $[0,255]$.

We propose here a new depth descriptor that uses those depth-normalized images as input. In particular, we compute short-term dense trajectories following the approach previously described in Sec. 2.2. Those depth-based trajectories are then described by using the DCS representation. As is Sec. 2.2, only informative and person-related trajectories are used for further processing.

2.4 Video-level gait representation

In order to build a video-level gait descriptor, we need to summarize the low-level features extracted from each modality. We use here Fisher Vectors (FV) encoding [24], as previously done for gait recognition in [4].

FV encoding, that can be seen as an extension of the Bag of Words (BOW) representation [27], builds on top of a Gaussian Mixture Model (GMM), where each Gaussian corresponds to a visual word. Whereas in BOW, an image is represented by the number of occurrences of each visual word, in FV an image is described by a gradient vector computed from a generative probabilistic model.

The dimensionality of FV is $2ND$, where N is the number of Gaussians in the GMM, and D is the dimensionality of the low-level audio, depth or visual descriptors. In our case, a typical value for the number of Gaussians is $N = 600$. The value of D is given by the dimensionality of the low-level descriptors (e.g., $D = 318$ for DCS, $D = 426$ for H2M) or the number of dimensions obtained after applying PCA (e.g., $D = 150$).

As stated in [24], the capability of description of the FV can be improved by applying it a signed square-root followed by L2 normalization. So, we adopt this finding for our descriptors.

3 Information fusion

When several sources of information are available, a method to fuse those sources is needed. On the one hand, we can combine those sources of information before learning a classification model. This approach is usually known as *early fusion*. A typical example of early fusion is the concatenation of descriptor vectors. On the other hand, we can train independent classifiers from each source of information, and then, define a strategy to fuse the classification or confidence scores. This is known as *late fusion*.

In this section, we start by describing the classification approach we have chosen, and, then, we present six fusion information strategies (three early fusion and three late fusion) that we will evaluate later in the experimental section (Sec. 4).

3.1 Classification

Given a set of video-level descriptors, we train as many binary linear Support Vector Machine (SVM) classifiers [23] as different subject identities have to be learnt. For each binary classifier, the positive class is the target subject, and all the remaining subjects are labelled as negative samples. This setup is usually known as 'one-vs-all'. Therefore, given a test sample, a classification score is assigned by each classifier of the ensemble. Identity is assigned according to the binary classifier that returned the maximum score.

In practice, the Fisher Vector descriptors obtained in Sec. 2.4 for representing the video sequences are compressed by standard Principal Components Analysis (PCA) before being fed into the SVM. We use the implementation available in 'VLFeat library' [29] for classification.

3.2 Early fusion

3.2.1 Feature vector concatenation

The simplest method for information fusion is vector concatenation. Given a set of n row vectors $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$, each computed from a different type of feature, a new feature vector $\hat{\mathbf{f}}$ is defined as the concatenation of the n feature vectors. This approach can be considered as an early fusion method, since the combination of information is carried out before any learning/classification procedure.

3.2.2 Bi-modal codewords

This kind of early fusion [33] builds a unique representation that fuses all the information provided by two different sources. In this manner, instead of concatenating the information, the method learns how to mix the information from different modalities into a unique and richer representation. Given a set of n row vectors $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$, each computed from a different type of modality. The distance between the representation of each modality is computed and a clustering technique is applied to this information in order to select the best features that represent the correlation between modalities. The groups obtained in the clustering process are used to build a new feature vector $\hat{\mathbf{f}}$ that encodes the fused information from the original representation of each modality. Finally, the *average criterion* is applied as it provides the best performance in our experiments. Since this method only can be employed with two modalities, we apply it in pairs (i.e. Visual-Depth, Visual-Audio and Depth-Audio).

3.2.3 Multiple Kernel Learning

Multiple Kernel Learning (MKL) [28] is another approach for early fusion. It is based on the idea of applying a function, or kernel, that maps the input data to a higher dimensionality where the data are linearly separable. From this idea, the authors extend it to use a combination of kernels instead of using a single kernel. Then, the method tries to find an optimal kernel that best maps the information to make it separable. Given a set of k row vectors $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$, each computed from a different type of modality and being $\delta_1, ..., \delta_k$ the $k-$associated distance functions, where $\delta_k = w_k^T f_k$. The algorithm tries to find the optimal kernel $K_{opt} = \sum_k d_k K_k$ where K_k is the k-th kernel matrix (i.e. function of δ_k) and d are the weights. The computation is performed as an SVM optimization framework where the primal problem can be formulated as:

$$\min_{w_k, b, d, \xi \geq 0} \quad \frac{1}{2} \sum_k \frac{w_k^T w_k}{d_k} + C \sum_k \xi_k + \frac{\lambda}{2} ||d||_p^2 \tag{1}$$

$$\text{s.t.} \quad y_i (\sum_k w_k^T f_k + b) \geq 1 - \xi, i = 1, ..., N$$

where $||\cdot||_p$ represents the Euclidean p-norm, ξ_k is the slack parameter, C is the regularization parameter and finally, w_i and and b are the weights and bias of the SVM, respectively. Nevertheless, this formulation is equivalent to concatenate K modalities of each sample. The authors present a richer representation that uses the product of kernels instead of the sum. As in [21], we use a χ^2 distance and a product of exponential kernels of precomputed distance matrices with SVM classifiers as precomputed distance and generalized kernel, respectively.

3.3 Late fusion

3.3.1 Weighted Scores

Weighted scores (WS) is a late fusion method that uses the estimated accuracy of the individual models to assign a weight to each confidence score, obtaining in this way a new combined score. Given a set of n confidence score vectors $\{\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_n}\}$, associated to n models, and their corresponding weighting factors $\{a_1, a_2, ..., a_n\}$, the final score vector $\mathbf{s}_f$ is computed as follows:

$$\mathbf{s}_f = \sum_{i=1}^{n} \mathbf{s}_i * a_i \tag{2}$$

where the sum of the weighting factors is equal 1.

3.3.2 SVM over the scores

SVM over the scores (SVM-OTS) is another late fusion method that builds a SVM classifier using as input the confidence scores of all modalities obtained in a training set. By this way, the classifier is able to automatically learn a relation between modalities that maximizes the classification performance. Given a set of n confidence score vectors $\{\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_n}\}$, obtained from n models, we build a new feature vector $\mathbf{f}_s$ by concatenating those scores. Then, a SVM classifier is trained on those new feature vectors $\{\mathbf{f}_s\}$, to obtain a final fusion score $\mathbf{s}_f$.

3.3.3 Rank Minimization

The method proposed by Ye et al. [34], for the problems of object categorization and video event detection, can be classified into the category of late fusion methods. We will use the term *Rank Minimization* (RM) to denote this method along the paper. After obtaining classification scores from different models, usually trained on different features, we want to combine those scores in order to improve the classification capability of each individual model, obtaining a (hopefully) better score.

Let $\mathbf{s} = [s_1, s_2, ..., s_m]$ be a confidence score vector of a model on m samples. A pairwise relationship matrix T is constructed from $\mathbf{s}$ as:

$$T_{jk} = \text{sign}(s_j - s_k) \tag{3}$$

Given n models, the robust late fusion method of Ye et al. aims at optimizing the following problem:

$$\begin{aligned} &\min_{\hat{T},E_i} \|\hat{T}\|_* + \lambda \sum_{i=1}^{n} \|E_i\|_1, \\ &\text{s.t. } T_i = \hat{T} + E_i, i = 1,...,n, \\ &\quad \hat{T} = -\hat{T}^\top \end{aligned} \tag{4}$$

Where T_i is the pairwise relationship matrix of the i-th model, E_i is a sparse matrix associated to the i-th model, $\hat{T}$ is the estimated rank-2 pairwise relationship matrix consistent among the samples and models, and λ is a positive tradeoff parameter to be cross-validated. Such optimization problem is solved by inexact Augmented Lagrange Multiplier method [18].

As described in [34], given the estimated matrix $\hat{T}$, and assuming that $\hat{T}$ is generated from $\hat{\mathbf{s}}$ as $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$, the new score vector $\hat{\mathbf{s}}$ is computed as

$$(1/m)\hat{T}\mathbf{e} = \arg\min_{\hat{\mathbf{s}}} \|\hat{T}^\top - (\hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top)\|_F^2, \tag{5}$$

treating $(1/m)\hat{T}\mathbf{e}$ as the recovered $\hat{\mathbf{s}}$ after the late fusion of the input scores.

Adaptation of the Rank Minimization method to gait recognition. Let n be the number of sources of information we want to use in our system. In our 'one-vs-all' ensemble of binary SVMs, we have one classifier specialized in a single identity for each source of information. Let us name it c_k^i, where i represents the i-th identity and k represents the k-th source of information.

During test time, from each binary classifier c_k^i, we will obtain a vector $\mathbf{s}_k^i$ of m scores (one per test sample). So, we can compute a pairwise relationship matrix T_k^i from each $\mathbf{s}_k^i$. In other words, if $\mathscr{T}^i = \{T_1^i, T_2^i, ..., T_n^i\}$ is the set of pairwise relationship matrices for a given identity i, we can obtain from $\mathscr{T}^i$ a new vector $\hat{\mathbf{s}}^i$ of identity-specialized scores, that combines the n sources of information. This process is repeated independently for each identity-specialized classifier i, thus, obtaining a new set of scores $\mathscr{S} = \{\hat{\mathbf{s}}^1, \hat{\mathbf{s}}^2, ..., \hat{\mathbf{s}}^N\}$, where N is the number of possible identities. To decide the final identity of each test sample, we seek the maximum over its scores in $\mathscr{S}$, assigning as identity the one of the identity-specialized classifier that generated such combined score.

In the particular case where only a single modality is available, i.e. $n = 1$, the proposed approach is completely valid as well. We will show in the experimental results (Sec. 4.2) that using this strategy to re-score the outputs of the SVM ensemble improves significantly the recognition accuracy. In contrast, one limitation of this method is that it requires having available all the test samples in order to obtain the new scores.

As in [34], we could adopt an *out-of-sample* strategy when we have a new sample at test time not seen during the RM computation. Basically, [34] propose to assign a score based on the feature similarity of the new test sample with regard to the previously seen samples. The reader is referred to [34] for further details on this case.

4 Experiments and results

We present in this section the experiments conducted to validate our proposed pipeline for gait recognition using both visual and audio information. We try to answer the following research questions: (a) *in terms of recognition accuracy, how far can we go by using each modality independently?*; (b) *can we really recognize people by using their gait sound?*; (c) *what fusion information strategy is the most suitable for the proposed features?*; (d) *how much is the improvement, if any, when fusing audio, RGB and depth features in the task of gait recognition?*; and (e) *are tracklet-based and audio descriptors useful for gender and shoes recognition?*

We start by describing the dataset used for our experiments along with the experimental setup, and, then, we present and discuss the results.

4.1 Datasets

TUM Gait from Audio, Image and Depth (GAID) database (TUM GAID) [12]. In TUM GAID 305 subjects perform two walking trajectories in an indoor environment. The first trajectory is performed from left to right and the second one from right to left. Therefore, both sides of the subjects are recorded. Two recording sessions were performed, one in January, where subjects wore heavy jackets and mostly winter boots, and the second in April, where subjects wore different clothes. Some examples can be seen in Fig. 2.

Hereinafter the following nomenclature is used to refer each of the four walking conditions considered: *normal* walk (*N*), carrying a *backpack* of approximately 5 kg (*B*), wearing coating *shoes* (*S*), as used in clean rooms for hygiene conditions, and *elapsed time* (*TN-TB-TS*).

Each subject of the dataset is composed of: six sequences of normal walking (*N1, N2, N3, N4, N5, N6*), two sequences carrying a bag (*B1, B2*) and two sequences wearing coating shoes (*S1, S2*). In addition, 32 subjects were recorded in both sessions (i.e. January and April) so they have 10 additional sequences (*TN1, TN2, TN3, TN4, TN5, TN6, TB1, TB2, TS1, TS2*).

This dataset is very suitable for our experiments since the action is captured by a Microsoft Kinect sensor which provides a video stream, a depth stream and a four-channel audio. Video and depth are recorded at a resolution of 640 × 480 pixels with a frame rate of approximately 30 fps. The four-channel audio is sampled with 24 bits at 16 kHz.

In [12], Hofmann et al. designed the recommended experiments that should be performed in the database. For that purpose, they split the database into three partitions over the

Fig. 2 TUM GAID dataset. People recorded from the same camera viewpoint walking indoors in two seasons. Three situations are included in the dataset: normal walking, walking with a bag and walking with coating shoes. **(top)** RGB source. **(bottom)** Depth source.

Fig. 3 CASIA-B dataset. People recorded from different camera viewpoints walking indoors. Three situations are included in the dataset: normal walking, walking with a bag and walking with coats.

subjects: 100 subjects for training and building models, 50 subjects for validation and 155 subjects for testing. Therefore, once we have selected the parameters of our models, we focus only on the subjects' test partition, training the classifiers on the training sequences of those subjects and testing on the corresponding test sequences. Finally, a set of experiments (Sec. 4.2) are proposed for validating the robustness of the algorithms against different factors.

CASIA-B [35] . In CASIA-B 124 subjects perform walking trajectories in an indoor environment. The action is captured from 11 viewpoints (i.e. from 0^o to 180^o in steps of 18^o) with a video resolution of 320×240 pixels. Three situations are considered: normal walk (*nm*), wearing a coat (*cl*), and carrying a bag (*bg*). Fig. 3 shows some samples of different viewpoints. The authors of the dataset indicate that sequences 1 to 4 of the '*nm*' scenario should be used for training the models. Whereas the remaining sequences should be used for testing: sequences 5 and 6 of '*nm*', 1 and 2 of '*cl*' and 1 and 2 of '*bg*'. Therefore, we follow this protocol in our experiments, unless otherwise stated. This makes a total of 496 video sequences for training, per camera viewpoint.

4.2 Experimental results

For each dataset, we start by describing the experiments performed to give answer to the questions stated at the beginning of Sec. 4. Then, the experimental results are discussed.

4.2.1 Experiments on TUM GAID

Experiment A: baseline. In this experiment we use each modality independently for gait recognition: Visual (DCS and H2M), Depth and Audio features. Those will be used as baseline that we would try to improve with fusion. For Visual (DCS and H2M) and Depth we use the following parameters in PFM: PCAL=150 (PCA applied at descriptor level), PCAH=256 (PCA applied at FV level), K=600 (dictionary size), and a single spatial level where the person bounding-box is vertically split into two non overlapping spatial cells to encode coarse spatial information. For FV-Audio we use PCAL=50, PCAH=75, K=50. These parameters have been chosen through experimentation on a set of possible values and taking those that obtain best results. The results of our experiments along with the accuracy obtained by previous relevant works are summarized in Tab. 1, where each row corresponds to a different method. We have conducted two experiments for visual descriptors (DCS and H2M), one for audio (FV-Audio) and one for depth (DCS-Depth).

Table 1 State-of-the-art on TUM GAID. Percentage of correct recognition on TUM GAID for diverse methods. Each column corresponds to a different scenario. Column 'Avg' is the weighted average computed as the sum of the weighted mean scores of *N, B, S* and *TN, TB, TS*. Best results are marked in bold.

	Method	*N*	*B*	*S*	*TN*	*TB*	*TS*	*Avg*
Visual	GEI [12]	99.4	27.1	52.6	44.0	6.0	9.0	56.0
	SVIM [32]	98.4	64.2	91.6	65.6	31.3	50.0	81.4
	RSM [10]	**100.0**	79.0	97.0	58.0	38.0	**57.0**	88.2
	DCS (this paper)	99.7	99.0	**99.0**	78.1	62.0	54.9	**96.0**
	H2M (this paper)	99.4	**100.0**	98.1	71.9	**63.4**	43.8	95.5
Audio	SVM [12]	44.5	27.4	4.8	3.0	0.0	3.0	23.4
	SVM+feat. sel. [8]	51.9	28.4	4.2	-	-	-	-
	HMM [9]	**65.5**	**36.5**	9.0	-	-	-	-
	FV-Audio (this paper)	62.3	30.3	**9.0**	**12.5**	**18.8**	**12.5**	**32.1**
Depth	depth-GEI [12]	96.8	3.9	88.7	28.0	0.0	22.0	58.8
	GEV [12]	94.2	13.9	87.7	28.0	0.0	22.0	58.8
	DGHEI [12]	99.0	40.3	96.1	50.0	0.0	44.0	74.1
	DCS-Depth (this paper)	90.6	**68.4**	89.4	**68.8**	**71.9**	**59.4**	**81.3**

Focusing on visual descriptors (top part of Tab. 1), it can be observed that DCS outperforms or obtains similar results than previous methods in most cases except in the *TS* case, where it achieves a 2.1% less than the best previous work. Globally, our approach improves the state-of-the-art average from 88.2% to 96.0%. In the case of H2M, it only outperforms DCS in scenarios *B* and *TB*, showing that DCS is a more robust descriptor than H2M.

In the middle of Tab. 1 we can see both our results and the state-of-the-art for audio. Note that only the study

Table 2 Combinations of different modalities on TUM GAID. Combinations performed on each kind of fusion are marked with a ✓.

Combination	Early Fusion	Late Fusion
Visual(DCS)+Audio	✓	✓
Visual(DCS)+Depth	✓	✓
Audio+Depth	✓	✓
Visual(DCS)+Audio+Depth	✓	✓
Visual(DCS)+Visual(H2M)	•	✓
Visual(H2M)+Audio	•	✓
Visual(H2M)+Depth	•	✓
Audio+Depth	•	✓
Visual(DCS)+Visual(H2M)+Audio	•	✓
Visual(DCS)+Visual(H2M)+Depth	•	✓
Visual(H2M)+Audio+Depth	•	✓
Visual(DCS)+Visual(H2M)+Audio+Depth	•	✓

of Hofmann et al. [12] evaluates the temporal cases. The other previous works are exclusively focused on non temporal cases. If we compare our method (row 'FV-Audio') with [12], the results indicate that FV-Audio is better in both temporal and non temporal cases, obtaining an improvement ranging from 9.5% to 18.5% in temporal cases and from a 2.9% to 17.8% in non temporal cases. Analyzing the previous works that focus exclusively on non temporal cases, FV-Audio is close to the results achieved by the best method, HMM, and outperforms the remaining ones.

Focusing on depth (bottom part of Tab. 1), if we compare our method (row 'DCS-Depth') with previous works, it outperforms all of them in temporal cases. The improvement is specially large in *TB*, where the rest of works are not able to classify any sample correctly. Concretely, we obtain an improvement of about 15.4% and a 18.8% in *TN* and *TS* and a 71.9% in *TB* with respect to the best method. In non temporal cases, we are not able to outperform the best method in cases *N* and *S*. However we obtain the best score for case *B*, showing that our method is more robust in scenarios where the subjects suffer substantial changes in their appearance. Finally, our approach improves the state-of-the-art average for depth from 74.1% to 81.3%.

Experiment B: fusion strategies. We use the fusion methods described in Sec. 3 for trying to improve the baseline results. We carry out the experiments with each combination of modalities marked in Tab. 2. Since the baseline results we have obtained with Visual (DCS or H2M) are almost perfect (i.e. $\geq 98.1\%$) for the standard cases, we focus hereinafter on the challenging temporal cases (*TN*, *TB* and *TS*).

We cross-validate the parameters of the evaluated models on the validation set. In the case of Vector Concatenation (Sec. 3.2.1) there are no parameters to tune. For bi-modal codewords (Sec. 3.2.2) we build the codeword using the 85% of the total features (sum of each modality). This percentage has been previously selected by cross-validation. In MKL (Sec. 3.2.3), we use a χ^2 distance and a product of exponential kernels of precomputed distance matrices with SVM classifiers as precomputed distance and generalized kernel, respectively. In Weighted Scores (Sec. 3.3.1), the weighted factors have been obtained as the average of the top five accuracies for each scenario in the training set defined in the dataset. For 'SVM over the scores' (Sec. 3.3.2), we train a generic model that learns how to fuse the information from different modalities. This training process is performed with the unused sequences of the subject test set and the test process is performed with the rest of sequences of that set. Finally, in Rank Minimization (Sec. 3.3.3) we use the following parameters: $\lambda = 1$, $\mu = 0.05$ (normalization factor), $\varepsilon = 10^{-5}$ (max error allowed), $\mu_{\max} = 10^{10}$ (max μ allowed), $\rho = 1.001$ (increment factor of μ).

The results of this experiment are summarized in Fig. 4, where each graph represents a different fusion method. Note that for brevity, only the most relevant combinations are represented. The top row contains early fusion methods and the bottom row contains late fusion methods. We have employed the work of Hofmann et al. [12] for comparison purposes, as it is the only previous work that has applied fusion techniques to TUM GAID dataset. Thus, the label 'Reference' in Fig. 4 indicates the best fusion performance obtained by that work for the three temporal scenarios. It is important to point out that in the previous experiment (Experiment A), FV encoding of visual features ('Visual DCS' and 'Visual H2M') was performed using temporal partitioning in order to increase the number of training samples. However this technique cannot be applied to audio information. As visual and audio information must be fused in the current experiment, we have built a new FV encoding for visual features without temporal partitioning. Thus, the scores achieved by the visual modalities this experiment (Fig. 4) are lower than those in experiment A (Tab. 1).

According to the results, all fusion methods except BiModal outperform the work of Hofmann et al. [12] in all cases. Even with only one modality, we obtain a notable improvement with respect to the 'Reference'. In general, due to the low accuracy of the audio-based models, the combinations that include Audio information do not present important improvements with respect to other combinations that only use Visual (DCS or H2M) or Depth information.

If we focus on the early fusion methods (top row), MKL obtains a 100% of accuracy with the combination of 'Visual(DCS)+Depth' in all scenarios, establishing a new state-of-the-art. With other combinations we obtain lower results than with a single modality, mainly due to the low accuracy of Audio. This behaviour is extensible to the other two early fusion methods in which we obtain worse results when Audio is combined with other modality. With feature vector concatenation of Visual(DCS) and Depth, we obtain an improvement of more than a 9% in all scenarios. Regarding BiModal fusion, the method obtains poor results probably because it was not able to build a good bi-modal dictionary.

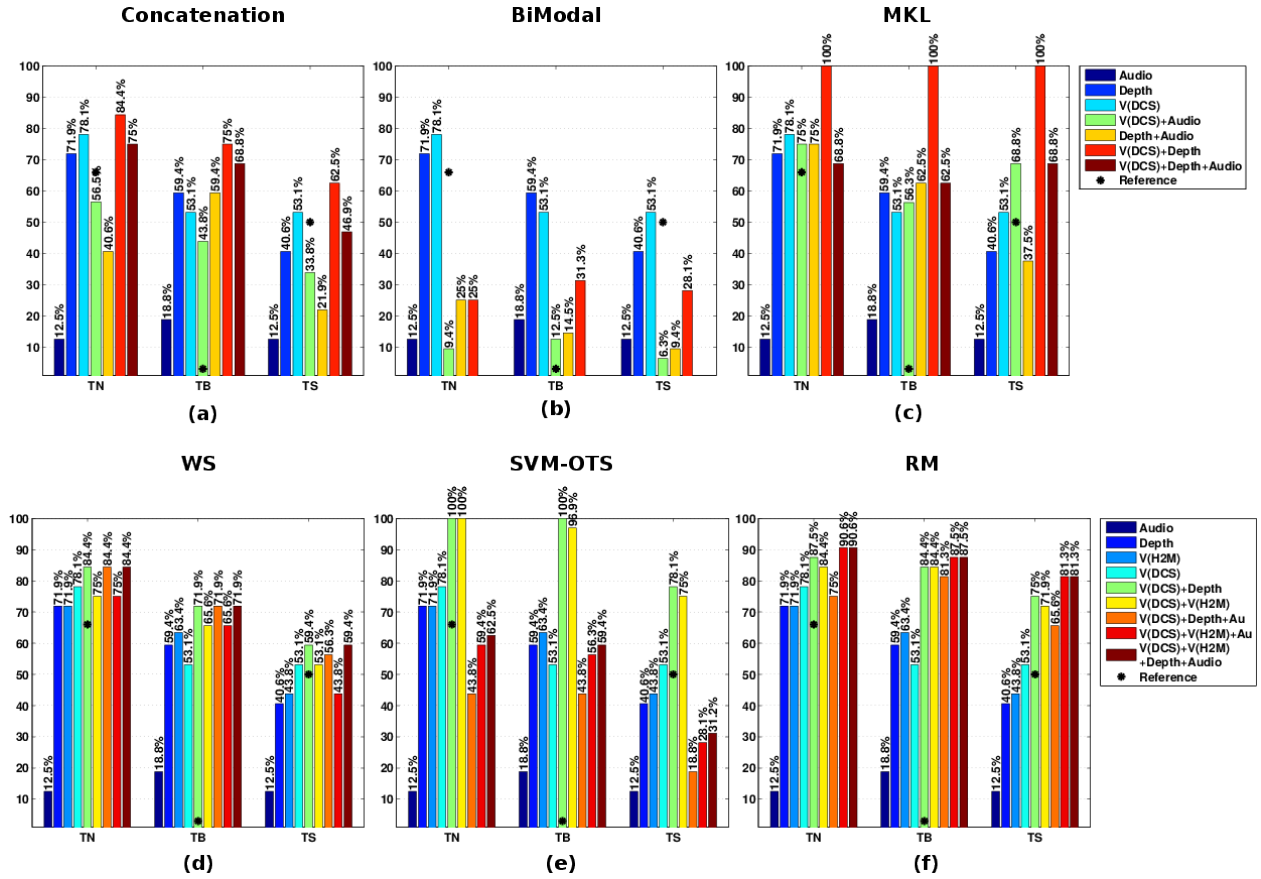

Fig. 4 Gait recognition results in TUM GAID: single modalities vs fusion. Each bar represents either a single modality or a combination of them. In each plot, the results are grouped per scenario. 'Reference' case refers to [12]. **(a)** Early fusion: feature concatenation. **(b)** Early fusion: BiModal codewords. **(c)** Early fusion: Multiple Kernel Learning. **(d)** Late fusion: Weighted Scores. **(e)** Late fusion: SVM over the scores. **(f)** Late fusion: Rank Minimization. (Best viewed in electronic format)

On the other hand, all late fusion methods (bottom row of Fig. 4) improve on single modalities. The best results in scenarios *TN* and *TB* are obtained with SVM-OTS, achieving a perfect accuracy. In contrast, the best result for scenario *TS* is obtained with RM (81.3%). The results obtained by WS are boosted in a 8% by using the simple weighted sum. Regarding RM, we can see that it improves all cases in more than a 20%, even when we fuse Audio data, which helps to boost the results in this case.

In summary, and according to the results, the best early fusion strategy is MKL. Using this strategy, a 100% of accuracy is obtained in all scenarios. On the other hand, the best late fusion method is SVM-OTS, specially in *TN* and *TB* cases where a perfect accuracy is also achieved. In *TS* case we improve the previous results in a 25%. Regarding the best combination of modalities, 'Visual(DCS)+Depth' frequently obtains the best performance, but RM can take advantage of the Audio modality. Finally, if we compare the performance of Visual(H2M) and Depth, we can see that, although in single modality Visual(H2M) has a better accuracy than Depth, when different modalities are fused, Depth helps to achieve a higher accuracy than Visual(H2M), mainly due to the fact that it corresponds to a different source of data.

Table 3 Rank Minimization with single modality in TUM GAID. Percentage of correct recognition on TUM GAID using Rank Minimization with one modality. Each column corresponds to a different scenario. Best results are marked in bold.

Scenario	Audio	Depth	H2M	DCS	Audio-RM	Depth-RM	H2M-RM	DCS-RM
TN	12.5	68.8	71.9	**78.1**	9.4	68.8	75.0	**84.4**
TB	18.8	**71.9**	63.4	53.1	18.8	78.1	75.0	**81.3**
TS	12.5	**59.4**	43.8	53.1	12.5	53.1	62.6	**71.9**

Experiment C: Rank Minimization method on single modalities. Using as input the classification scores obtained during *experiment A*, we use the Rank Minimization method (Sec. 3.3.3) to optimize them, obtaining a better accuracy. We use the same parameters used in previous experiments.

The results of this experiment are summarized in Tab. 3, where each row shows a different scenario. As we can see, when the initial scores are good enough (40% seems enough to obtain a significant improvement), RM is able to improve the accuracy significantly. In this experiment, both visual features 'DCS' and 'H2M' show important improvements in accuracy. However, Depth is only able to outperform the case *TB*, obtaining worse results for the *TS* case. Regarding Audio, RM is not able to improve any case as the initial con-

Table 4 Samples distribution for the gender experiment. Each cell contains the number of samples of the class per scenario.

	N	*B*	*S*
Male	582	194	194
Female	348	116	116

fidence vectors have poor scores for this modality. In summary, we can achieve an improvement between 3% and 28% depending on the tested case by using RM.

Experiment D: gender recognition. Gender recognition can be considered a kind of soft-biometrics extracted from the gait. Therefore, the goal of this experiment is to study how to use gait features to recognize people gender. We use the partitions defined in the dataset [12] for soft-biometric experiments: sequences *N1* to *N6* of the training subject partition are used for train the models, and sequences *N1, N2, N3, N4, N5, N6, B1, B2, S1, S2* of the test subject partition are used for testing, obtaining the accuracy of the methods. Note that in this analysis we only employ the non-temporal information according to the guidelines defined by the authors. Three experiments are defined, one per scenario (i.e. *N*, *B* and *S*), employing the same parameters than in previous section. The number of samples per class is summarized in Tab. 4. The accuracy per scenario (*N, B, S*) is computed as $\frac{\#true\_positives}{\#samples}$.

In Fig. 5 we summarize the results of this experiment. Top and bottom rows display early and late fusion methods, respectively, where each graph represents a different fusion method. The label 'Reference' corresponds to the depth accuracy values achieved by Hofmann et al. [12] in the different scenarios. According to our results, we are able to outperform 'Reference' with a single modality in most cases. Concretely, in Depth modality we outperform the result in case *B* by a 16.8%, obtaining the same accuracy in case *S*. Finally, in case *N* we obtain a 2.8% less than the reference.

Focusing on late fusion methods, the best accuracy for cases *N* and *B* is attained by WS. In case *S* the best score is obtained by SVM-OTS. In early fusion, the best results are obtained by Concatenation, which achieves even better scores than those obtained with late fusion methods. In both kinds of fusion, the best combinations of modalities are 'Visual(DCS) + Depth' and 'Visual(DCS) + Depth + Audio' depending on the case. With these results, we set a new state-of-the-art for all scenarios. In general, the improvement obtained with Audio is low because it obtains worse accuracy than the other modalities and only in some cases it helps to outperform the combinations without Audio. In Tab. 5 we present the confusion matrices for each scenario (*N*, *B*, and *S*) for the Concatenation method with the combination 'Visual(DCS)+Depth' which obtained the best average results. Each cell contains the percentage of samples assigned to the class. The bottom row shows the final accuracy per scenario. As the percentages show, the classification error is lower than 3% for all cases.

Table 5 Confusion matrix for the gender experiment with Concatenation: Visual(DCS)+Depth. Each cell contains the percentage of correct samples assigned to the class. Distribution of samples in Tab.4.

		N		*B*		*S*	
		M	*F*	*M*	*F*	*M*	*F*
CM	*M*	**97.9**	2.1	**97.4**	2.6	**98.6**	1.4
	F	3.2	**96.8**	3.4	**96.6**	5.2	**94.8**
Acc	Concat	97.5		97.1		97.1	

Table 6 Samples distribution for the shoes experiment. Each cell contains the number of samples of the class per scenario.

	N	*B*	*S*
Sneakers	528	176	176
High-Boots	174	58	58
Top-Boots	120	40	40
Loafers	84	28	28
Others	24	8	8

To summarize, the best results for gender recognition are obtained with Concatenation of 'Visual(DCS) + Depth'. In general, Audio features does not help to recognize gender.

Experiment E: shoe type identification. As gender recognition, shoe type identification can be considered a kind of soft-biometrics extracted from the gait. The goal of this experiment is to study how to use gait features to identify types of shoes. Five shoe categories are considered in the dataset: sneakers, high-boots, top-boots, loafers and others (e.g. sandals, ballerina and rubber boots). As in the gender recognition experiment, we use the standard partitions defined in the dataset. The number of samples per class and per scenario is summarized in Tab. 6. Due to the huge differences between the number of samples per class (see Tab. 6), we use Unweighted Average Recall (UAR) to measure the performance of the tested systems: $\frac{1}{N}\sum_{i=1}^{N}\frac{\#true\_positives_i}{\#true\_positives_i+\#false\_negatives_i}$, where N is the number of classes.

For this task, as we observed that the current SVM setup tends to assign a label from the most frequent classes, we have chosen the RUSBoost ensemble method [25] to deal with this issue. In short, RUSBoost is a hybrid sampling/boosting algorithm for learning from skew training data. In our case, we use decision trees as learners due to their good performance in combination with RUSBoost, according to the results presented by its authors. In particular, we build 1000 trees with at least 5 observations per leaf (i.e. each leaf is a class) using a cross-validated learning rate. Then, each leaf of each tree produces a score indicating the likelihood of that class. The final score for each class is computed as the average over the ensemble of trees.

In Fig. 6 we show the results of shoes recognition, where each plot represents a different fusion method. The top row contains early fusion methods and the bottom row contains late fusion methods. 'Reference' corresponds to the results

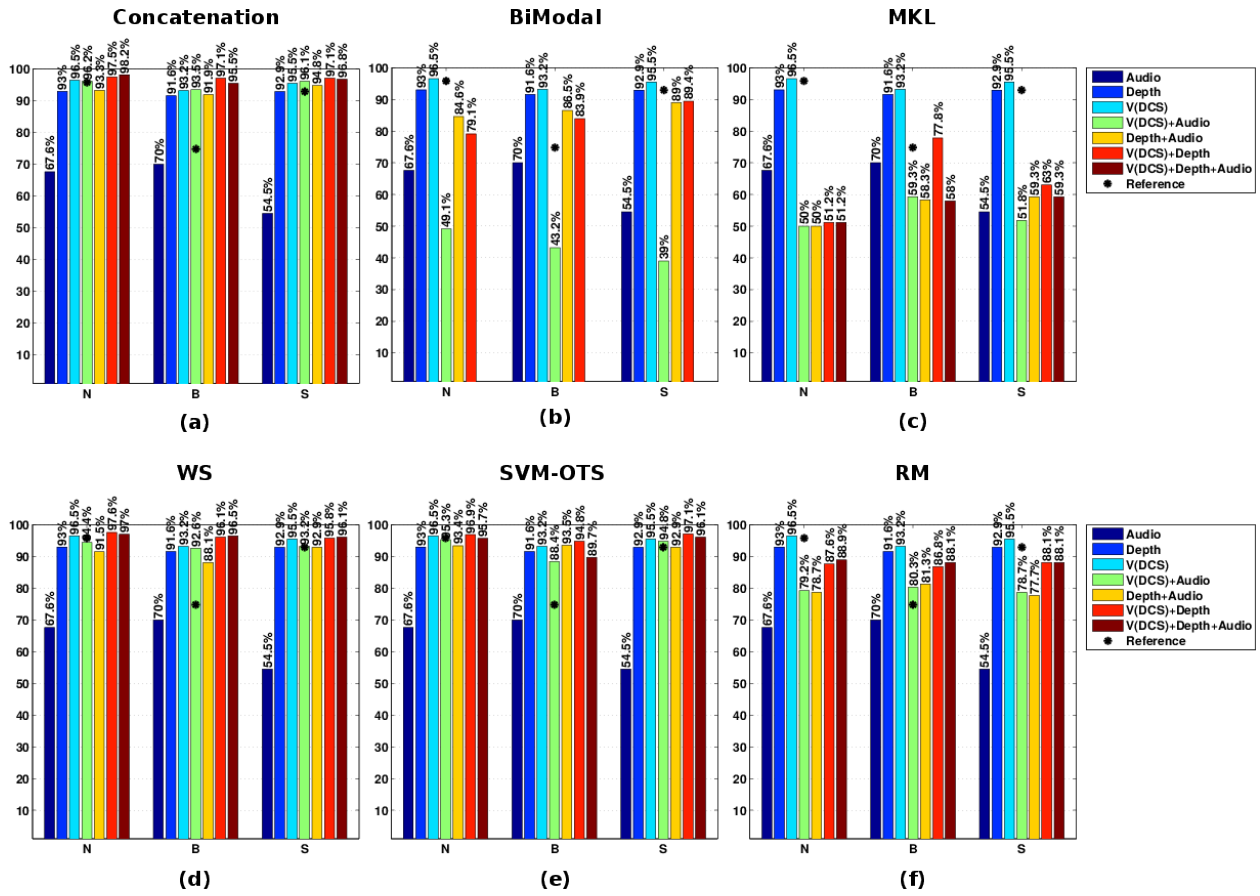

Fig. 5 Gender recognition results in TUM GAID: single modalities vs fusion. Each bar represents either a single modality or a combination of them. In each plot, the results are grouped per scenario. 'Reference' case refers to [12]. **(a)** Early fusion: feature concatenation. **(b)** Early fusion: BiModal codewords. **(c)** Early fusion: Multiple Kernel Learning. **(d)** Late fusion: Weighted Scores. **(e)** Late fusion: SVM over the scores. **(f)** Late fusion: Rank Minimization. (Best viewed in electronic format)

Table 7 Confusion matrix for the shoes experiment with RM: Visual(DCS)+Audio. Each cell contains the percentage of correct samples assigned to the class. Distribution of samples in Tab. 6.

Exp. N	*Sneakers*	*High-Boots*	*Top-Boots*	*Loafers*	*Others*	UAR	[12]
Sneakers	**32.1**	11.4	15.9	17.9	25.0		
High-Boots	1.7	**50.0**	19.0	13.8	15.5		
Top-Boots	10.0	20.0	**52.5**	12.5	5.0	**42.5**	31.4
Loafers	19.9	3.6	25.0	**27.8**	21.4		
Others	12.5	0	25.0	12.5	**50.0**		
Exp. B	*Sneakers*	*High-Boots*	*Top-Boots*	*Loafers*	*Others*	Acc	[12]
Sneakers	**27.8**	11.4	15.9	19.9	25.0		
High-Boots	1.7	**50.0**	19.0	13.8	15.5		
Top-Boots	10.0	20.0	**52.5**	12.5	5.0	**42.5**	30.9
Loafers	17.9	3.6	25.0	**32.1**	21.4		
Others	12.5	0	25.0	12.5	**50.0**		
Exp. S	*Sneakers*	*High-Boots*	*Top-Boots*	*Loafers*	*Others*	Acc	[12]
Sneakers	**22.7**	14.2	16.5	21.0	25.6		
High-Boots	6.9	**48.3**	12.1	8.6	24.1		
Top-Boots	10.0	20.0	**42.5**	15.0	12.5	**39.4**	29.0
Loafers	32.1	0	32.1	**21.4**	14.3		
Others	0	12.5	12.5	12.5	**62.5**		

obtained by Hofmann et al. [12] using fusion. According to the results, we are able to outperform the fusion results of reference with a single modality ('Visual(DCS)') in all cases. Specifically, we improve their results by 8.2%, 4.8% and 1.2% in cases *N*, *B* and *S*, respectively.

Regarding the fusion methods, we can see that, in most cases, the best combination for all methods is 'Visual(DCS)+Audio' due to their high precision. In addition, the impact of Audio in this problem is higher than in the other problems because the sound produced by a shoe is closely related with the type or class of shoe. So, the audio features are more relevant in this problem than in others like gender or gait recognition, where it is really hard to identify those classes by using only sound. In late fusion, the best results are obtained with RM with which we are able to establish a new state-of-the-art in all scenarios. Using this fusion method, we achieve an improvement over 'Reference' of 11.1%, 11.6% and 10.5% in cases *N*, *B* and *S*, respectively. In early fusion, the results obtained are not able to improve single modality 'Visual(DCS)'. Nevertheless, with MKL we are able to improve the best results obtained in 'Reference' by 2.1%, 1.7% and 1.4% in cases *N*, *B* and *S*, respectively.

In Tab. 7 we present the confusion matrices for each scenario (*N*, *B*, and *S*) for the RM method with the combination 'Visual(DCS)+Audio' which obtains the best average results. Each cell contains the percentage of samples

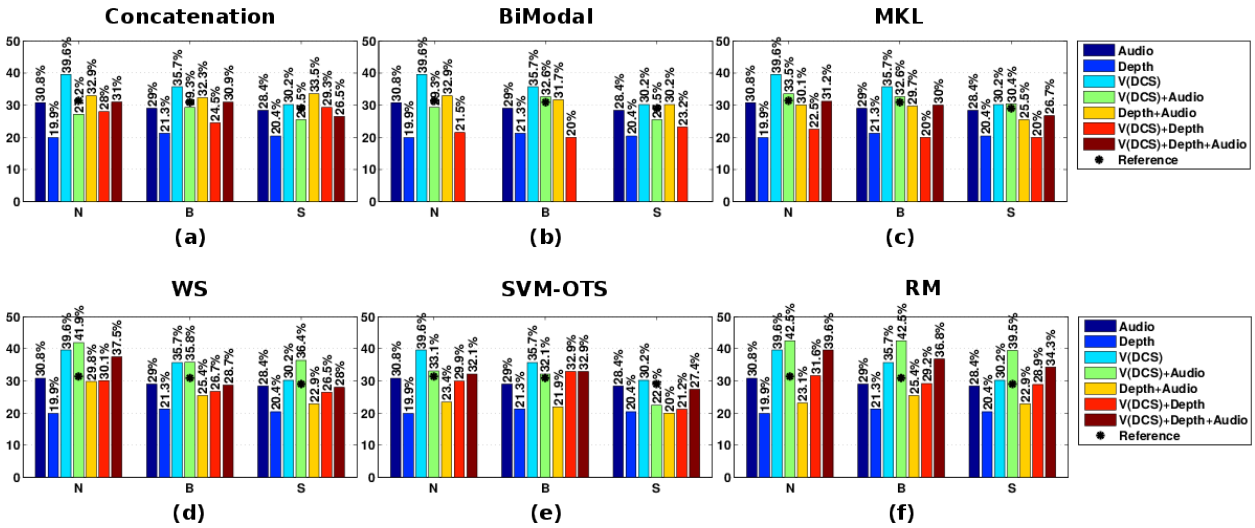

Fig. 6 Shoes recognition results in TUM GAID: single modalities vs fusion. Each bar represents either a single modality or a combination of them. In each plot, the results are grouped per scenario. 'Reference' case refers to [12]. **(a)** Early fusion: feature concatenation. **(b)** Early fusion: BiModal codewords. **(c)** Early fusion: Multiple Kernel Learning. **(d)** Late fusion: Weighted Scores. **(e)** Late fusion: SVM over the scores. **(f)** Late fusion: Rank Minimization. (Best viewed in electronic format)

assigned to the class. The last column shows the final accuracy per scenario (i.e. average of the diagonal values). As the percentages show, recognition values are balanced between classes and, in most cases, the highest percentage corresponds to the correct class.

To summarize, the best results for shoes recognition are obtained with RM and the combination 'Visual(DCS)+Audio'. For this problem, Audio features can help to increase the accuracy in some cases because the sound produced by a shoe is an important characteristic that can help to differentiate different kinds of shoes. Due to the huge difficulties of recognize what kind of shoe wears a person (even for a human), the precision is not so high as in the other problems.

4.2.2 Experiments on CASIA-B

Experiment A: baseline. In this experiment, we use each modality independently to recognize the gait. As the CASIA-B only contains RGB data, we use Visual descriptors (DCS and H2M) to obtain our baseline results. Then, we evaluate different fusion strategies on these two features. We employ the following parameters in PFM: PCAL=150 (PCA applied at descriptor level), PCAH=256 (PCA applied at FV level), K=150 (dictionary size), and a single spatial level where the person bounding-box is vertically split into two non overlapping spatial cells to encode coarse spatial information. These parameters have been chosen through experimentation on a set of possible values and taking those that obtain best results.

Without loss of generality, we report our experimental results on the profile view (i.e. 90^o). The results of our experiments along with the scores obtained by previous relevant works are summarized in Tab. 8, where each row corresponds to a different method. Note that we have trained our model using all cameras to maximize the available training data.

Table 8 State-of-the-art on CASIA-B, camera 90^o. Percentage of correct recognition on CASIA-B for diverse methods on camera 90°. Each column corresponds to a different scenario. Column 'Avg' is the average of scenarios *nm*, *bg* and *cl*. Acronyms: '#subjs' number of subjects used for test; '#train' number of sequences per person used for training; '#test' number of sequences per person used for test. Best results are marked in bold.

	Method	*#subjs*	*#train*	*#test*	*nm*	*bg*	*cl*	*Avg*
Visual	AEI+2DLPP [37]	124	3	3-nm 2-bg-cl	98.4	91.9	72.2	87.5
	GEI [35]	124	4	2	97.6	52.0	32.7	67.8
	iHMM [13]	84	5	1	94.0	45.2	42.9	60.7
	CGI [30]	124	1	1	88.1	43.7	43.0	58.3
	SDL [36]	124	3	3-nm 2-bg-cl	98.4	93.5	**90.3**	**94.1**
	DCS (this paper)	124	4	2	99.6	95.2	73.8	89.5
	H2M (this paper)	124	4	2	91.9	58.5	35.9	62.1
	DCS+H2M (fusion)	124	4	2	**100.0**	**99.2**	72.6	90.6

It can be observed that DCS outperforms previous methods in cases *nm* and *bg*. Only in the *cl* case it achieves results lower than the best previous work, i.e. 16.5% less. Globally, our approach obtain the second best average result of the state-of-the-art. In the case of H2M, it does not outperform DCS in any case, showing that DCS is a more robust descriptor.

Experiment B: fusion strategies. We use the fusion methods described in Sec. 3 for trying to improve the baseline results. We run the experiments with DCS and H2M features using the same fusion parameters of TUM GAID.

Fig. 7 shows the fusion results for each evaluated fusion method and scenario, where each graph represents a different fusion method. Plots *a*, *b* and *c* contain early fu-

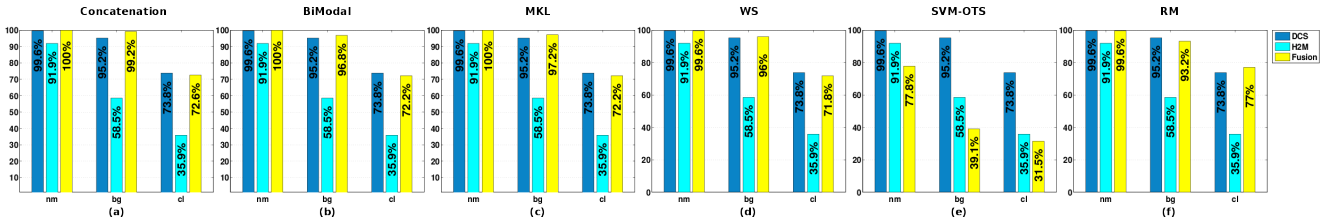

Fig. 7 Gait recognition results in CASIA-B, camera 90°: single modalities vs fusion. Each bar represents either a single modality or a combination of them. In each plot, the results are grouped per scenario. **(a)** Early fusion: feature concatenation. **(b)** Early fusion: BiModal codewords. **(c)** Early fusion: Multiple Kernel Learning. **(d)** Late fusion: Weighted Scores. **(e)** Late fusion: SVM over the scores. **(f)** Late fusion: Rank Minimization. (Best viewed in electronic format)

Table 9 Rank Minimization with single modality in CASIA-B camera 90°. Percentage of correct recognition on CASIA-B using Rank Minimization with one modality. Each column corresponds to a different scenario. Best results are marked in bold.

Scenario	DCS	H2M	DCS-RM	H2M-RM
nm	99.6	91.9	**99.6**	92.3
bg	95.2	58.5	**99.2**	67.7
cl	73.8	35.9	**80.6**	42.7

sion methods, and plots *d*, *e* and *f* contain late fusion methods. For the sake of comparison to the state-of-the-art, fusion results of the best method (Concatenation) are included in bottom row of Tab. 8. We have to remark that for SVM-OTS, as the dataset does not provide a training partition on the subjects, the SVM have been trained with the information of cameras 72° and 108°.

If we focus on early fusion methods, we can see that they improve the results of single modalities for cases *nm* and *bg*. Only in case *cl* they are not able to boost the best scores achieved by single modalities, obtaining values between 1.2% and 1.6% lower. In contrast, late fusion methods do not show significant improvements, only in case *cl* with RM we can outperform the best result of single modality.

In summary, the results show that the best early fusion strategy is Concatenation. With this strategy we establish the best results in cases *nm* and *bg*. On the other hand, the best late fusion method is RM, obtaining the best results in scenario *cl* and almost the best results for the rest of cases.

Experiment C: Rank Minimization method on single modality. Using as input the classification scores obtained during *experiment A*, we use the Rank Minimization method described in Sec. 3.3.3 to obtain a better accuracy. We use the same parameters used in previous experiments.

The results of this experiment are summarized in Tab. 9, where each row shows a different scenario. We can see that the behaviour is similar to the one observed in TUM GAID: RM is able to optimize the scores matrix, improving the accuracy significantly. Both visual features 'DCS' and 'H2M' show important improvements in accuracy. On average, we are able to boost the original results around 4.5%.

4.2.3 Summary

We summarize here the main findings of the experiments carried out in this paper. At feature level, Visual-DCS has resulted to be the most robust descriptor for people identification, gender recognition and shoe type recognition. Combining it with depth has shown to be a good choice for identification and gender recognition, whereas the combination of it with audio is specially convenient for shoe type recognition. In terms of fusion methods, WS is in general a good method for the late fusion case, while either MKL or Concatenation can be used for early fusion. However if the computation cost is a constraint for the target system, our results have shown that late fusion produces similar accuracies to early fusion with less computation.

5 Conclusions

This paper has presented a thorough evaluation of multimodal features for the task of person identification based on gait. The results on 'TUM GAID' dataset show that, although most of the gait information is clearly visual, depth information and, in lesser extent, audio information help to identify people as well. In addition, the combination of visual features ('Visual-DCS') and depth improves the recognition accuracy of the system (compared to just using one of them) in the 'temporal' scenario – people wearing different clothes in different epochs of the year – specially in the cases where people wear a backpack or coating shoes, up to 88% better, achieving a precision of 100% in all scenarios. Another interesting finding is that the Rank Minimization method of Ye et al. [34] can be used on a single modality of gait descriptors combined with an ensemble of 'one-vs-all' SVM classifiers to re-score the individual classification scores, boosting the recognition accuracy. Regarding to the use of gait features for gender and shoes recognition, the results have demonstrated that our descriptors are useful for these tasks. In addition, the fusion of different modalities boosts the accuracy achieving new state-of-the-art results. The results on 'CASIA-B' confirm the observations made on 'TUM GAID': *(i)* the fusion of different features improves the single modality results; and *(ii)*, RM is able to boosts single modality results in all tested cases.

Acknowledgments

This work has been partially funded by project TIC-1692 (Junta de Andalucía), and the Research Project TIN2012-

32952 (Spanish Ministry of Science and Technology). We also thank the reviewers for their helpful comments.

References

1. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6), 345–379 (2010)
2. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality and the SMO algorithm. In: Proceedings of the International Conference on Machine Learning, p. 6 (2004)
3. Castro, F.M., Marín-Jiménez, Guil, N.: Empirical study of audio-visual features fusion for gait recognition. In: Proceedings of the International Conference on Computer Analysis of Images and Patterns, pp. 727–739 (2015)
4. Castro, F.M., Marín-Jiménez, M., Medina-Carnicer, R.: Pyramidal Fisher Motion for multiview gait recognition. In: Proceedings of the International Conference on Pattern Recognition, pp. 1692–1697 (2014)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893. IEEE Computer Society, Washington, DC, USA (2005)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of the European Conference on Computer Vision (ECCV) (2006)
7. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Proc. of Scandinavian Conf. on Image Analysis, vol. 2749, pp. 363–370 (2003)
8. Geiger, J., Hofmann, M., Schuller, B., Rigoll, G.: Gait-based person identification by spectral, cepstral and energy-related audio features. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 458–462 (2013)
9. Geiger, J.T., Kneißl, M., Schuller, B., Rigoll, G.: Acoustic Gait-based Person Identification using Hidden Markov Models. ArXiv e-prints (2014)
10. Guan, Y., Li, C.: A robust speed-invariant gait recognition system for walker and runner identification. In: Int. Conf. on Biometrics (ICB), pp. 1–8 (2013)
11. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 316–322 (2006)
12. Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., 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 **25**(1), 195 – 206 (2014)
13. Hu, M., Wang, Y., Zhang, Z., Zhang, D., Little, J.: Incremental learning for video-based gait recognition with LBP flow. Cybernetics, IEEE Transactions on **43**(1), 77–89 (2013)
14. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **34**(3), 334–352 (2004)
15. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2555–2562 (2013)
16. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Video-Based Surveillance Systems, pp. 135–144. Springer (2002)
17. Lartillot, O., Toiviainen, P.: MIR in Matlab (ii): A toolbox for musical feature extraction from audio. In: ISMIR, pp. 127–130 (2007)
18. Lin, Z., Chen, M., Ma, Y.: The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055 (2010)
19. Liu, Y., Zhang, J., Wang, C., Wang, L.: Multiple HOG templates for gait recognition. In: Proceedings of the International Conference on Pattern Recognition, pp. 2930–2933. IEEE (2012)
20. López-Fernández, D., Madrid-Cuevas, F.J., Carmona-Poyato, A., Muñoz Salinas, R., Medina-Carnicer, R.: Entropy volumes for viewpoint-independent gait recognition. Machine Vision and Applications **26**(7-8), 1079–1094 (2015)
21. Marín-Jiménez, M., Muñoz Salinas, R., Yeguas-Bolivar, E., Pérez de la Blanca, N.: Human interaction categorization by using audio-visual cues. Machine Vision and Applications **25**(1), 71–84 (2014)
22. Martín-Félez, R., Xiang, T.: Uncooperative gait recognition by learning to rank. Pattern Recognition **47**(12), 3793 – 3806 (2014)
23. Osuna, E., Freund, R., Girosi, F.: Support Vector Machines: training and applications. Tech. Rep. AI-Memo 1602, MIT (1997)
24. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 143–156 (2010)
25. Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A.: Rusboost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **40**(1), 185–197 (2010)
26. Sivapalan, S., Chen, D., Denman, S., Sridharan, S., Fookes, C.: Gait energy volumes and frontal gait recognition using depth images. In: Biometrics (IJCB), 2011 International Joint Conference on, pp. 1–6 (2011)
27. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision (ICCV), vol. 2, pp. 1470–1477 (2003)
28. Varma, M., Babu, B.R.: More generality in efficient multiple kernel learning. In: ICML, p. 134 (2009)
29. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)
30. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2164–2176 (2012)
31. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action Recognition by Dense Trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176 (2011)
32. Whytock, T., Belyaev, A., Robertson, N.: Dynamic distance-based shape features for gait recognition. Journal of Mathematical Imaging and Vision **50**(3), 314–326 (2014)
33. Ye, G., Jhuo, I.H., Liu, D., Jiang, Y.G., Lee, D.T., Chang, S.F.: Joint audio-visual bi-modal codewords for video event detection. In: ICMR, p. 39 (2012)
34. Ye, G., Liu, D., Jhuo, I.H., Chang, S.F.: Robust late fusion with rank minimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3021–3028 (2012)
35. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Proceedings of the International Conference on Pattern Recognition, vol. 4, pp. 441–444 (2006)
36. Zeng, W., Wang, C., Yang, F.: Silhouette-based gait recognition via deterministic learning. Pattern Recognition **47**(11), 3568 – 3584 (2014)
37. Zhang, E., Zhao, Y., Xiong, W.: Active energy image plus 2DLPP for gait recognition. Signal Processing **90**(7), 2295 – 2302 (2010)