Human Gait Identification from Extremely Low Quality Videos: an Enhanced Classifier Ensemble Method

Yu Guan[1], Yunlian Sun[2], Chang-Tsun Li[1],
and Massimo Tistarelli[2]
Department of Computer Science, University of Warwick,
Coventry, CV4 7AL, UK[1]
Department of Sciences and Information Technology,
University of Sassari, 07100, Sassari, Italy[2]
{g.yu, c-t.li}@warwick.ac.uk[1]
{elaine.sun717, mtista}@gmail.com[2]

Abstract

Nowadays, surveillance cameras are widely installed in public places for security and law enforcement, but the video quality may be low due to limited transmission bandwidth and storage capacity. In this paper, we propose a gait recognition method for extremely low quality videos, which have a frame-rate at one frame per second (1 fps) and resolution of 32×22 pixels. Different from popular temporal reconstruction-based methods, the proposed method uses the Average Gait Image (AGI) over the whole sequence as the appearance-based feature description. Based on the AGI description, we employ a large number of weak classifiers to reduce generalization errors. The performance can be further improved by incorporating the model-based information into the classifier ensemble. We find that the performance improvement is directly proportional to the average disagreement level of weak classifiers (i.e., diversity), which can be increased by using the model-based information. We evaluate our method on both indoor and outdoor databases (i.e., the low quality versions of OU-ISIR-D and USF databases), and the results suggest that our method is more general and effective than other state-of-the-art algorithms.

1 Introduction

Gait is a behavioural biometric characteristic which has been successfully used for remote human identification. Compared with recognition techniques based on other biometric characteristics like fingerprint or iris, gait recognition can be applied at a larger distance and without the cooperation from subjects. However, several factors may affect the performance of automatic gait recognition systems. These factors can be roughly divided into three categories: 1) environmental, e.g., walking surface, camera viewpoint, occlusion, etc., 2) subject-related, e.g., shoe type, carrying condition, clothing, speed, etc., and 3) video-quality-related, e.g., low frame-rate or low resolution. Nowadays, surveillance cameras are widely installed in public places such as airports, government buildings, streets and shopping malls to prevent criminal activities. However, the video quality may be low due to limited transmission bandwidth and storage capacity. The corresponding gait sequences in such low quality videos may have much lower resolution if subjects are far away from the surveillance cameras, and much fewer "clean" gait frames available if subjects are significantly suffered from the occlusions in the crowded areas. The lack of frames may have a similar effect as extremely low frame-rate (e.g., 1 fps). As such, it is desirable to propose gait recognition algorithms that are robust to low quality gait videos. In [1], phase synchronisation was used, and gait probe videos with low frame-rate were matched with high frame-rate gallery videos. This approach fails whenever both the probe and gallery samples are from low frame-rate videos. Several temporal reconstruction-based methods were proposed to deal with such dual low frame-rate problem. In [2], Al-Huseiny et al. proposed a level-set morphing approach for temporal interpolation. In [3], a temporal Super Resolution (SR) method was employed to build a high frame-rate gait sequence period based on multiple periods of low frame-rate gait sequences. Based on [3], Akae et al. [4] applied an exemplar of high frame-rate image sequences to improve the temporal SR quality. A unified framework of example-based and reconstruction-based periodic temporal SR was proposed in [5]. It works well for gait recognition from (dual) extremely low frame-rate videos (with small gait fluctuations). Most of these works attempt to recover the high frame-rate gait sequences in the first place. However, in the applications of human identification given extremely low frame-rate videos (e.g., 1 fps), they either have low performance due to the generated high levels of reconstruction artifacts [2, 4], or have to assume that the amount of motion among the gait periods is the same, which is not feasible when there are large gait fluctuations [5].

Instead of using the popular temporal reconstruction-based concept, which may have the afore-mentioned problems, in this work we use the average gait image over the whole sequence to compute features both for biometric reference and biometric probe. A large number of weak classifiers are generated and combined to reduce generalization errors. The main differences between this work and our previous work [6] can be summarised as follows: 1) We fuse model-based information to further enhance the ensemble accuracy, and study the performance gain from the perspective of individual classifier accuracy and the diversity among them. 2) More experiments have been performed and more results are reported.

The remainder of this paper is organised as follows: Section 2 briefly reviews the related work and highlights the main contributions. The proposed method is demonstrated in Section 3. In Section 4, the method is experimentally evaluated, and a comparison with other methods is also provided. Conclusions are drawn in Section 5.

2 Related Work and Proposed Advances

Gait recognition methods can be roughly divided into two categories: model-based and appearance-based approaches. Model-based methods(e.g.,[7, 8]) use human body parameters as features for recognition, while appearance-based methods are more general, and most of them perform classification based on pixel intensities(e.g.,[5, 6, 9–16]). Compared with model-based methods, appearance-based approaches can also work well with relatively low resolution data when body parameters are difficult to estimate. Gait Energy Image (GEI) [13] is one of the most popular appearance-based feature description due to its simplicity and effectiveness. GEI encodes a gait cycle of binarised aligned silhouettes into a grayscale image through an averaging operation which can smooth the segmentation errors, and also significantly reduce the computational cost. Average Gait Image (AGI) [16] has similar properties to GEI. It is the average image over the whole gait sequence, instead of over a single gait cycle. Compared with GEI, there is no requirement for gait cycle detection, which is difficult to perform given the low frame-rate [6]. Several silhouettes and their AGI samples (derived from extremely low quality videos) from the OU-ISIR-D [14] and USF [15] databases are shown in Fig. 1.

However, direct AGI comparison makes the classification process prone to

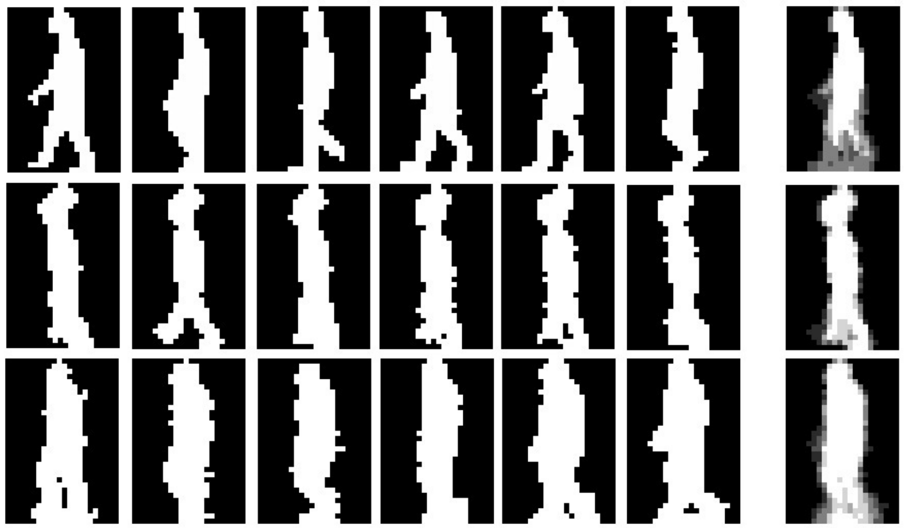


Figure 1: Extremely low quality gait sequences (derived from videos with frame-rate at 1 fps and resolution of 32×22 pixels). Top/middle row: from the indoor OU-ISIR-D [14] database with low/large gait fluctuations; bottom row: from the outdoor USF database [15]. In each row, the rightmost image is the AGI corresponding to the whole gait sequence.

errors when covariate factors exist (e.g., shoe, camera viewpoint, carrying condition, clothing, speed, etc.). Various gait recognition algorithms have been proposed, which may be robust to one or more covariate factors (e.g., [5, 6, 9–13]). Out of them, Random Subspace Method (RSM) is one of the most effective. RSM was initially proposed by Ho [17] to build random decision forests and it was applied to face recognition by Wang and Tang [18]. Guan et al. applied this concept to gait recognition [10]. Since the effect of unknown covariate factors on a query gait is unpredictable, the pre-collected gallery data (capture in a certain walking condition) used for training can be less representative. In this case, overfitting to the less representative training data can be the main reason that hamper the performance of the learning-based algorithms [9]. By combining a large number of weak classifiers, the generalization errors can be significantly reduced [6, 9–11]. RSM is a general framework which is robust to a large number of covariates such as shoe, carrying condition, (small changes in) camera viewpoint [10], clothing [11], speed [9], etc. However, RSM's performance is limited when intra-class variations are extremely large (e.g., due to elapsed time). Given that every biometric modality may have its upper bound in terms of distinctiveness, multimodal fusion can be a viable solution[19]. In light of this, multimodal-RSM was proposed in [12], by using face information to reinforce the gait-based weak classifiers, thus reducing the effect of the elapsed time covariate. Although the results suggest that face information [12, 20] from surveillance cameras is useful to some extent for identifying subjects, when subjects are too far away from the camera with extremely low resolution, face information may become unreliable.

For classifier ensembles (such as RSM), the diversity of the predictions (i.e., predicted labels by the base classifiers) is important and can be measured by means of different metrics [21]. One of the most popular metrics is the *pair-wise disagreement measure*, which is directly proportional to the number of *different* outputs (correct or wrong) between any pair of base classifiers in the whole multiple classifier system [21]. This measure was initially proposed by Skalak [22]. Ho [17] applied it to random decision forests. In this work, we use this metric to explore the relationship between diversity and ensemble accuracy. Our contributions can be summarised as follows:

1. The difference between the gallery and probe (due to extremely low quality) is analysed as a similar effect caused by normal covariate (such as carrying condition), which can be solved using the RSM concept. We further improve the ensemble performance by using the model-based information to strengthen the RSM-based weak classifiers. The experimental results suggest that our method is more general and effective than other methods.

2. This work demonstrates an effective way of fusing model-based methods and appearance-based methods. Since both information can be derived from the same gait video, this is especially useful when the footage quality is extremely low with other modalities/information unavailable.
3. We study three metrics in the classifier ensemble, i.e., individual classifier accuracy, ensemble accuracy, and diversity. From the perspective of individual classifier accuracy and diversity, we explain the performance gain (i.e., enhanced ensemble accuracy) through incorporating model-based information into RSM.

3 Proposed Method

In this section, we first introduce how to use RSM [6] for feature extraction, then we propose two simple model-based methods for extremely low quality videos. Model-based information can be used to enhance the weak classifiers without sacrificing the diversity of the whole multiple classifier system. Finally, the diversity measurement used to evaluate our system is described.

3.1 RSM

In this work, we use AGI [16] as the appearance-based feature description. Given a sequence with T gait frames $F_t(t = 1, 2, ..., T)$, AGI is defined as:

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

In extremely low frame-rate environments (e.g., 1 fps gait videos), the major benefit of using AGI is that the gait period, which is difficult to estimate, is not required. However, when the video recording time is also short, the major problem is the lack of frames (e.g., $T = 5$ frames for a sequence), which makes the averaging operation less effective. For example, when the walking starting stances of the probe and reference sequence are different, through averaging operation, although the static parts (e.g., head) can be relatively stable [6], the dynamic parts (e.g., legs or arms) can be rather diverse. Assuming such effect caused by extremely low frame-rate are intra-class variations that the gallery data fails to capture and in this case, a general model based on RSM concept [10] can be used to reduce such generalization errors.

Given c classes/gait sequences in the gallery, there are c AGIs. Let m be the pixel number of an AGI, after concatenating each two-dimensional AGI, the gallery can be represented as $A = [a_1, a_2, ..., a_c] \in \mathbb{R}^{m \times c}$. Then the scatter matrix S can be estimated:

$$S = \frac{1}{c}\sum_{i=1}^{c}(a_i - \mu)(a_i - \mu)^T, \tag{2}$$

where $\mu = \frac{1}{c}\sum_{i=1}^{c} a_i$. The eigenvectors of S can be computed and the leading v eigenvectors are retained as candidates to span the random subspaces.

L random subspaces can be generated and each projection matrix (e.g., feature extractor) is formed by randomly selecting s eigenvectors (from the v candidates). Let the projection matrix be $R^l \in \mathbb{R}^{s \times m}, l \in [1, L]$, and each gallery sample $a_i \in \mathbb{R}^m, i \in [1, c]$ can be projected as L sets of coefficients $\alpha_i^l \in \mathbb{R}^s, l = 1, 2, ..., L$ as the new gait representations:

$$\alpha_i^l = R^l a_i, \quad l = 1, 2, ..., L, \quad i \in [1, c]. \tag{3}$$

During a comparison trial, for the l^{th} subspace first R^l is used for feature extraction, and then Euclidean distance is adopted to measure the dissimilarity. For example, the distance between a query AGI $x \in \mathbb{R}^m$ and a certain class

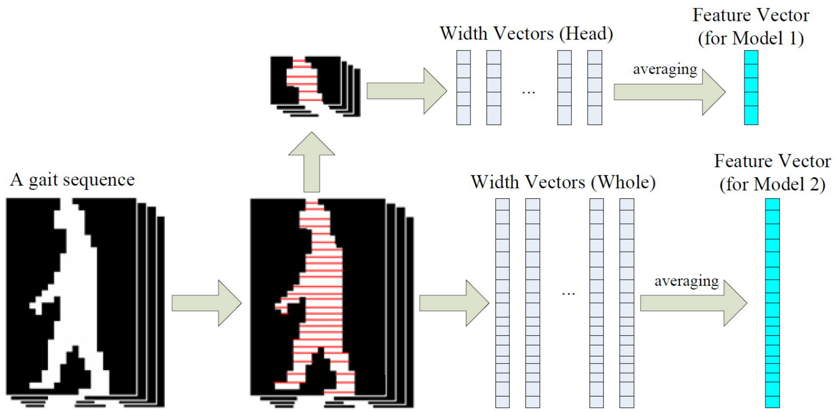


Figure 2: The process of generating feature vectors for model-based methods.

$\alpha_i^l, i \in [1, c]$ is:

$$d(x, \alpha_i^l) = \|R^l x - \alpha_i^l\|, \quad i \in [1, c], \quad l \in [1, L], \tag{4}$$

which can be updated using the model-based information for further processing.

3.2 Model-based Method for Low Quality Videos

In low resolution environments (e.g., 32×22 pixels per frame, see Fig. 1), it is difficult to estimate the body parameters that are used as classical model-based features, e.g., angles of hip and thigh [8], or stride and leg length [7], etc. For gait images with poor segmentation quality, in [23] Kale et al. used the silhouette widths (of each row) as model-based features, which are easy to estimate.

Motivated by [23], we use widths from the binarised aligned silhouette as model-based features, under the assumption that low resolution and poor segmentation quality may have similar effect on gait images. We also employ the widths (of each row) from the static silhouette area (i.e., the head area) as model-based features, which are less affected by the low frame-rate [6]. Since each gait sequence may have several frames, the corresponding average width vectors for *the whole silhouette* and *the head area* are used as input feature vectors for model-based classification. In this work, the classification models based on head width vector and silhouette width vector are referred to as Model 1 and Model 2, respectively. The process of generating the model-based feature vectors for Model 1 and Model 2 is illustrated in Fig. 2.

During a comparison trial, the distance between two sequences can be measured based on the Euclidean distance. Let the dimensionality of the feature vector (for Model 1 or Model 2) be r, given the gallery consisting c classes/sequences $B = [b_1, b_2, ..., b_c] \in \mathbb{R}^{r \times c}$, and distance between a query gait $y \in \mathbb{R}^r$ and a certain class $b_i, i \in [1, c]$ is:

$$d(y, b_i) = \|y - b_i\|, \quad i \in [1, c], \tag{5}$$

which can be used as model-based information to update RSM-based classifiers.

3.3 Fusion with Model-based Information

Multimodal fusion is a popular way to enhance the performance by combining two or more biometric modalities, and it has demonstrated its effectiveness in a large number of biometric applications [19]. However, it assumes that multiple modalities are available, which may not hold in less controlled environments. For single modality, one popular form of fusion is to combine the classification scores derived from different feature spaces, and this is especially useful for biometric data with limited information. In [13], real GEI and synthetic GEI were generated from the same gait silhouettes, before their classification scores were fused at a score level. Although these two sources maybe highly correlated, the enhanced performance suggests that there may exist some complementary power between the two different feature representations [13]. Due to the fact that correlated information may be combined to boost the performance, we fuse model-based and appearance-based methods in this work. We aim to enhance the performance of RSM, an appearance-based method, by incorporating the model-based information, which is used to update the classifiers of the L subspaces. Given a query gait and the gallery, through (4) we can get the corresponding distance vector $d^l \in \mathbb{R}^c$ for the l^{th} subspace. By using the min-max rule [24], it can be normalized as:

$$d^l_{rsm} = \frac{d^l - \min(d^l)}{\max(d^l) - \min(d^l)}, \quad l \in [1, L]. \tag{6}$$

Similarly, through (5) we can get the distance vector for the model-based methods. Let the distance vector be d_{model} after min-max normalization, then the fused distance vector (for the l^{th} subspace) d^l_{fusion} can be updated using the weighted sum rule:

$$d^l_{fusion} = \omega d_{model} + (1 - \omega) d^l_{rsm}, \quad l \in [1, L], \tag{7}$$

where $\omega \in [0, 1]$ is a weight factor for the model-based information. Based on the fused distance vector, the Nearest Neighbour (NN) classifier is then used for classification for the l^{th} subspace. The final classification decision is achieved through majority voting among all the L classifiers. Given a query gait and c classes in the gallery with labels $[W_1, W_2, ..., W_c]$, the optimal class label $\hat{W}_i$ is:

$$\hat{W}_i = \underset{W_i}{\operatorname{argmax}} \sum_{l=1}^{L} \triangle^l_{W_i}, \quad i \in [1, c], \tag{8}$$

where

$$\triangle^{l}_{W_i} = \begin{cases} 1, & \text{if } d^{l}_{fusion}(W_i) = \min(d^{l}_{fusion}), \\ 0, & \text{otherwise}, \end{cases} \quad i \in [1, c]. \tag{9}$$

The weight factor ω is important for the score-level fusion for each subspace. It is not appropriate to set ω too high or too low. According to (7), we can see that our method becomes the conventional RSM system in our previous work [6] when $\omega = 0$. On the other hand, when $\omega = 1$, it becomes the model-based method. Intuitively, ω should be a small value, given the fact that model-based features in such low quality videos are less reliable. A detailed evaluation of ω is provided in Section 4.3.

3.4 Diversity Measurement

Diversity among the classifiers is deemed to be a key issue in classifier ensemble [21]. Yet the relationship between ensemble accuracy and diversity is still unclear, which may depend on the specific applications and the metrics used for diversity measurement [21]. In this work, we use the *pair-wise disagreement measure* [17, 22] to explore the relationship of diversity and the ensemble accuracy of of our proposed gait recognition system.

Let $Z = [z_1, z_2, ..., z_N]$ be the test set, and each sample z_j includes both AGI vector $x_j \in \mathbb{R}^m$ and model-based feature vector $y_j \in \mathbb{R}^r$, i.e., $z_j = \{x_j, y_j\} \in \mathbb{R}^{m+r}, j = 1, 2, ..., N$. After fusing model-based information into the RSM system through (7), for simplicity we can represent the l^{th} classifier as $D_l : \mathbb{R}^{m+r} \longrightarrow \{0, 1\}, l \in [1, L]$ such that for $z_j \in Z$, $D_l(z_j) = 1$ if the classification is correct, and $D_l(z_j) = 0$, otherwise. Based on the classification results from two classifiers D_l and D_k, we can count the numbers with respect to four different output scenarios as follows:

$$N^{11} = \forall_{j\in[1,N]} \quad \{\#(D_l(z_j) = 1 \wedge D_k(z_j) = 1)\}, \tag{10}$$

$$N^{10} = \forall_{j\in[1,N]} \quad \{\#(D_l(z_j) = 1 \wedge D_k(z_j) = 0)\}, \tag{11}$$

$$N^{01} = \forall_{j\in[1,N]} \quad \{\#(D_l(z_j) = 0 \wedge D_k(z_j) = 1)\}, \tag{12}$$

$$N^{00} = \forall_{j\in[1,N]} \quad \{\#(D_l(z_j) = 0 \wedge D_k(z_j) = 0)\}. \tag{13}$$

The disagreement of two classifiers D_k and D_l is equal to the ratio between the number of cases on which D_k and D_l make different predictions (i.e., case N^{10} and case N^{01}) to the total number of test samples N [21], i.e.,

$$Dis(D_k, D_l) = \frac{N^{10} + N^{01}}{N}, \qquad k, l \in [1, L], \tag{14}$$

where $N = N^{11} + N^{10} + N^{01} + N^{00}$. We can also write (14) as:

$$Dis(D_k, D_l) = 1 - \frac{(N^{00} + N^{11})}{N}, \qquad k, l \in [1, L]. \tag{15}$$

For a multiple classifier system D consisting of L base classifiers, the diversity $Div(D)$ is defined as the average disagreement level of all the $L(L-1)/2$ classifier pairs, i.e.,

$$Div(D) = \frac{2}{L(L-1)} \sum_{l=1, l<k}^{L} Dis(D_k, D_l). \tag{16}$$

$Div(D)$ tends to be low when the average base classifiers are either *too weak* or *too strong*, since on test set Z the outputs of most classifier pairs will be more likely to be either *Both Wrong* (i.e., case N^{00}) or *Both Correct* (i.e., case N^{11}), which are inversely proportional to *Disagreement*, according to (15).

Our aim is to enhance the ensemble accuracy by strengthening the weak classifiers (through fusing the model-based information) without sacrificing the diversity. The experimental evaluation of the relationship among diversity, individual classifier accuracy and ensemble accuracy is provided in Section 4.3.

4 Experimental Evaluation

In this section, we first introduce the datasets and experimental settings used. Then we discuss the performance sensitivity with respect to the random feature number s used for each classifier. By using the model-based information, we explain the enhanced ensemble accuracy in terms of the individual classifier accuracy and diversity. Finally, we compare our system with other state-of-the-art methods for gait recognition from videos with extremely low quality.

4.1 Dataset and Configuration

The proposed method is evaluated on the *extremely low quality versions* of the indoor OU-ISIR-D database [14] and the outdoor USF database [15]. Both databases provide the binarised aligned silhouettes. The original resolution and frame-rate in OU-ISIR-D database are 128×88 pixels and 60 fps [14], while they are 128×88 pixels and 30 fps in the USF database [15]. The intention of this work is to propose a system that is capable of dealing with extremely low video quality. Therefore, we down-sample the afore-mentioned databases to create *extremely low quality versions* with lower resolution (i.e., 32×22 pixels) and frame-rate (i.e., 1 fps) in a manner similar to [1, 5].

The OU-ISIR-D database consists of two datasets, namely, DB-high (i.e., with small gait fluctuations) and DB-low (i.e., with large gait fluctuations). For DB-high/DB-low, there are 100 subjects (1 subject per sequence) for both the gallery and probe. For the outdoor USF database, 12 experiments were initially designed by Sarkar et al. for algorithm evaluations against covariate

Table 1: Datasets configuration. DB-high/DB-low has low/high gait fluctuations; DB-outdoor has high levels of segmentation errors, and camera viewpoint covariate.

Dataset	DB-high	DB-low	DB-outdoor
#Subject	100	100	122
#Seq. per Subject	1	1	1
Resolution (pixels)	32×22	32×22	32×22
Frame-rate (fps)	1	1	1
#Frames per Seq.	6	6	$4 \sim 7$

factors such as camera viewpoint, shoe, carrying condition, walking surface, and elapsed time, etc. Since in this work we focus on human gait recognition from extremely low quality videos, instead of evaluating our method on the 12 experiments (with default quality), only the extremely low quality version of USF dataset A is used, and we refer to it as DB-outdoor in this work. DB-outdoor includes 122 subjects (1 subject per sequence) for both the gallery and probe, which are captured in different camera viewpoints (about 30° difference). A summary of the datasets configuration is shown in Table 1.

The proposed model-based methods (Model 1 or Model 2) use the average width vector of a certain area (head or the whole silhouette) as feature template, as shown in Fig. 2. For Model 1, we simply define the head area is roughly the topmost 1/3 of the whole silhouette. Specifically, for videos with resolution 32×22 pixels, the dimensionality for feature vector corresponding to Model 1 (resp. Model 2) is $r = 10$ (resp. $r = 32$).

The recognition accuracy is used as the performance measurement. Due to the random nature, the results of different runs may vary to some extent. We repeat all the experiments 10 times and report the mean values. In Table 3, we also report results in terms of the mean and standard deviation.

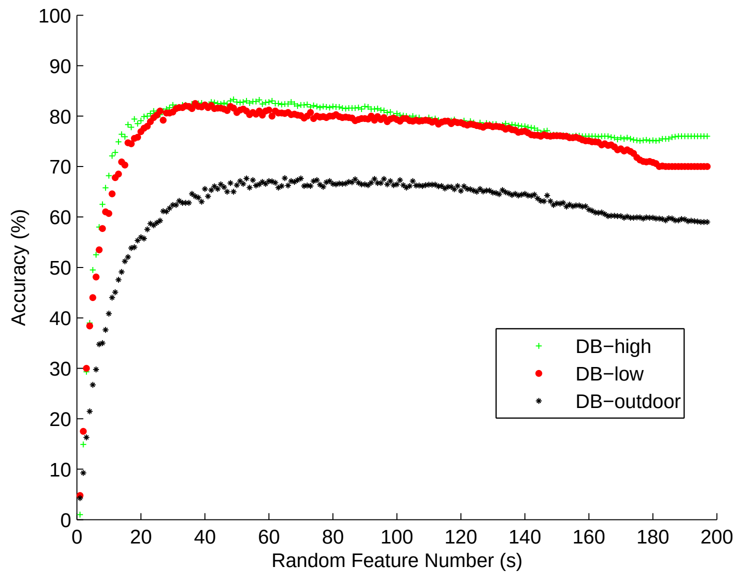


Figure 3: The accuracy distribution (%) with respect to random feature number (s), given both the probe and gallery videos with frame-rate at 1 fps and resolution of 32×22 pixels.

4.2 The Effect of Random Feature Number

For the initial eigenspace construction, we choose eigenvectors corresponding to the leading 200 eigenvalues (i.e., $v = 200$), which preserves 100% of the variance. For each random subspace, the corresponding base classifier has some generalization power for the unselected features (i.e., unselected subspaces) [17]. But the underfitting problem may arise if random feature number s is too small. On the other hand, the base classifier may be overfitted if random feature number s is too large.

By setting the classifier number $L = 1000$, we check the sensitivity of the random feature number s within the range $[2, v - 2]$ on the three datasets (i.e., DB-high, DB-low, and DB-outdoor). The accuracy distribution with respect to the random feature number s shown in Fig. 3 clearly indicates the effect of

Table 2: Accuracy (%) of Model 1/Model 2 on the three datasets, given both the probe and gallery videos with frame-rate at 1 fps and resolution of 32×22 pixels.

-	DB-high	DB-low	DB-outdoor
Model 1	57	51	6.56
Model 2	59	62	35.25

fusion based on underfitted (e.g., with $s \leq 20$) or overfitted (e.g., with $s \geq 160$) classifiers. They tend to have lower accuracies, and the reasons may be: 1) for underfitted/weak classifiers, there is not enough information for them to make the correct classification; 2) for overfitted/strong classifiers, due to the highly overlapped feature set, they tend to make the same prediction (i.e., lack of diversity), which makes the fusion less effective.

4.3 Performance Gain Analysis

To enhance the overall fusion performance, we aim to use ancillary information (from model-based methods) to enhance the underfitted classifiers. Although Model 1 and Model 2 may be less reliable in terms of discriminative power (see Table 2), they may provide information from a different perspective.

To measure the effect through fusing model-based information, three metrics are used, i.e., ensemble accuracy ($Acc_{ensemble}$), individual classifier accuracy (Acc_{indiv}), and diversity (Div). $Acc_{ensemble}$ is the performance through majority voting (see (8)). Acc_{indiv} is the average performance of the L classifiers. Given a test set $Z = [z_1, z_2, ..., z_N]$ and classifier $D_l : \mathbb{R}^{m+r} \longrightarrow \{0, 1\}$, we have

$$Acc_{indiv} = \frac{1}{L} \sum_{l=1}^{L} \sum_{j=1}^{N} \frac{D_l(z_j)}{N}. \tag{17}$$

Div is the average of the disagreement levels between any two distinct classifiers in a multiple classifier system (see (16)).

Based on different random feature numbers $s = \{20, 40, 80, 160\}$, we conduct experiments on various values of the weight factor ω within the range $[0, 1]$ with an interval of 0.1. It is worth noting that when $\omega = 0$, the proposed system is the same as the conventional RSM system proposed in our previous work [6] without fusing model-based information. On the three datasets (both the probe and gallery videos with frame-rate at 1 fps and resolution of 32×22 pixels), the performance distributions with respect to ω are shown in Fig. 4-6. Note we do not report the results corresponding to Model 1 on DB-outdoor, which provides extremely unreliable information in the outdoor environment (with 6.56% accuracy, see Table 2). From Fig. 4-6, we can observe:

1. Generally, $Acc_{ensemble} \propto Div$.
2. Without fusing model-based information (i.e., $\omega = 0$), Div is relatively low (e.g., $Div \approx 10\%$) when the base classifiers are either too strong (e.g., see Fig. 4(d), 5(d) when $Acc_{indiv} \approx 70\%$) or too weak (e.g.,see Fig. 6(a) when $Acc_{indiv} \approx 6\%$).
3. $Div \propto Acc_{indiv}$ holds only for weak classifier ensemble (e.g.,with $Acc_{indiv} \leq 30\%$). When $Acc_{indiv} \geq 30\%$, however, Div starts to decrease with respect to the strengthening base classifiers.
4. Model-based information may enhance Acc_{indiv} when low weight is assigned (e.g., $\omega \leq 0.5$). The enhancements tend to be more significant for underfitted classifiers than overfitted classifiers.

Diversity is important for classifier ensemble, and our experimental results suggest that $Acc_{ensemble} \propto Div$ in our gait recognition scenarios. Diversity can be increased by enhancing the weak classifiers (with lower Acc_{indiv}), which can be constructed based on a small number of random features, e.g., $s = 20$. Model-based information can then be used to increase the diversity, and thus enhance $Acc_{ensemble}$. Given the fact it is not beneficial to fuse stronger base classifiers, it is preferable to assign the model-based information lower weight in order to preserve the diversity.

For model-based information, although Model 1 has lower accuracies in the indoor datasets (see Table 2), it can provide some ancillary information to the RSM system (with higher $Acc_{ensemble}$). One explanation is that compared with Model 2 (based on the whole silhouette), Model 1 (based on the head area) is less correlated with the RSM-based weak classifiers, which are derived from the whole body. DB-outdoor is more challenging due to higher levels of segmentation errors and camera viewpoint covariate, and in the case the model-based information can be extremely unreliable (see Table 2). Nevertheless, compared with the best results from conventional RSM without fusing model-based information (see Fig. 3), incorporating Model 2 (with 35.25% accuracy) into underfitted classifiers can still reduce the error rate by up to 5%.

Gait identification from extremely low quality videos is a challenging task due to the lack of information. In the low resolution condition, it is also difficult to capture other modalities (e.g., profile face [12]) to enhance the performance of the gait recognition system. Experimental results suggest the effectiveness of our method in this limited condition, since both model-based information and RSM-based classifiers can be derived from the low quality silhouettes. Although model-based information may be less reliable, they may reveal the data structure from a different perspective. Using such information may enhance the RSM-based weak classifiers without sacrificing the diversity.

4.4 Algorithms Comparisons

On DB-high and DB-low, we compare our method with other algorithms, i.e., morphing-based reconstruction (Morph) [2], Periodic Temporal SR (PTSR) [4], and Example-based and Reconstruction-based Temporal SR (ERTSR) [5]. We

Table 3: Algorithms comparisons in terms of accuracy (%) on DB-high/DB-low, given both the probe and gallery videos with frame-rate at 1 fps and resolution of 32×22 pixels.

-	DB-high	DB-low
Morph [2]	52	N/A
PTSR [4]	44	N/A
ERTSR [5]	87	N/A
RSM ($s = 20$) [6]	79.50±1.90	75.20±2.49
RSM ($s = 40$) [6]	82.10±0.57	81.80±1.75
RSM+Model 1($s = 20$)	**90.80±1.48**	**88.40±1.43**
RSM+Model 1($s = 40$)	88.50±1.35	87.50±1.18

directly quote the results of Morph, PTSR, and ERTSR from [5], which are based on the same experimental settings as ours.

As stated in Section 4.3, higher performance can be achieved by combining low-weighted model-based information (Model 1) with weaker classifier ensemble. In Table 3, we report our results (mean and standard deviation of 10 runs) based on $s = \{20, 40\}$, $\omega = 0.2$. It is also worth noting that our results are less sensitive to s and ω within a certain range, as shown Fig. 4-5.

From Table 3, we can see that reconstruction-based methods ([2, 4]) tend to have low accuracies. This is because significant artifacts can be generated due to the extremely low frame-rate and low resolution, and reconstruction-based methods [2, 4] are not able to cope with those artifacts effectively. EPTSR [5] can greatly improve the accuracy by assuming that the degree of motion is the same among gait cycles. However, this assumption does not hold when there are large gait fluctuations (e.g., on DB-low) [5]. Compared with the three methods, the RSM-based method in our previous work [6] is more adaptive and can be applied in both DB-high an DB-low with reasonable accuracies. In this work, fusing model-based information into the RSM system can further reduce the error rate. This effect is more significant for weaker classifier ensembles (e.g., with $s = 20$).

5 Conclusions

In this paper, we propose an enhanced classifier ensemble method for gait identification from extremely low quality videos. By incorporating the model-based information into the RSM-based weak classifiers, the diversity of the classifiers can be enhanced, which is positively correlated to the ensemble accuracy. We also find that it is less beneficial to combine stronger base classifiers, since in this case they tend to have the same prediction, which contributes negatively to the diversity of the whole multiple classifier system. Compared with other state-of-the-art algorithms, our method delivers significant improvements in terms of identification accuracy and generalization capability. In the future, we will 1) investigate other ancillary information which can be derived from the low quality videos; 2) design an adaptive mechanism on deciding the weight ratio between RSM-based gait classifiers and the given ancillary information.

6 Acknowledgement

This work was partially supported by the EU COST Action IC1106.

References

[1] Mori, A., Makihara, Y., and Yagi, Y., "Gait recognition using period-based phase synchronization for low frame-rate videos," in *Proc. Int'l Conf. Pattern Recognition*, 2010, pp. 2194–2197.

[2] Al-Huseiny, M., Mahmoodi, S., and Nixon, M., "Gait learning-based regenerative model: A level set approach," in *Proc. Int'l Conf. Pattern Recognition*, 2010, pp. 2644–2647.

[3] Makihara, Y., Mori, A., and Yagi, Y., "Temporal super resolution from a

single quasi-periodic image sequence based on phase registration," in *Proc. Asian Conf. Computer Vision*, 2011, pp. 107–120.

[4] Akae, N., Makihara, Y., and Yagi, Y., "Gait recognition using periodic temporal super resolution for low frame-rate videos," in *Proc. Int'l Joint Conf. Biometrics*, 2011, pp. 1–7.

[5] Akae, N., Mansur, A., Makihara, Y., and et al.,, "Video from nearly still: An application to low frame-rate gait recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1537–1543.

[6] Guan, Y., Li, C.-T., and Choudhury, S., "Robust gait recognition from extremely low frame-rate videos," in *Proc. Int'l Workshop on Biometrics and Forensics*, 2013, pp. 1–4.

[7] Bobick, A. and Johnson, A., "Gait recognition using static, activity-specific parameters," in *Proc. of Computer Vision and Pattern Recognition*, 2001, pp. 423–430.

[8] Cunado, D., Nixon, M., and Carter, J., "Using gait as a biometric, via phase-weighted magnitude spectra," in *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication*, 1997, pp. 95–102.

[9] Guan, Y. and Li, C.-T., "A robust speed-invariant gait recognition system for walker and runner identification," in *Proc. Int'l Conf. Biometrics*, 2013, pp. 1–8.

[10] Guan, Y., Li, C.-T., and Hu, Y., "Random subspace method for gait recognition," in *Proc. Int'l Conf. Multimedia and Expo Workshops*, 2012, p-p. 284–289.

[11] Guan, Y., Li, C.-T., and Hu, Y., "Robust clothing-invariant gait recogni-

tion," in *Proc. Int'l Conf. Intelligent Information Hiding and Multimedia Signal Processing*, 2012, pp. 321–324.

[12] Guan, Y., Wei, X., Li, C.-T., and et al.,, "Combining gait and face for tackling the elapsed time challenges," in *Proc. Int'l Conf. Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[13] Han, J. and Bhanu, B., "Individual recognition using gait energy image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006, 28, pp. 316–322.

[14] Makihara, Y., Mannami, H., Tsuji, A., and et al.,, "The ou-isir gait database comprising the treadmill dataset," *IPSJ Trans. on Computer Vision and Applications*, 2012, 4, pp. 53–62.

[15] Sarkar, S., Phillips, P., Liu, Z., and et al.,, "The humanid gait challenge problem: data sets, performance, and analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005, 27, pp. 162–177.

[16] Veres, G., Gordon, L., Carter, J., and et al.,, "What image information is important in silhouette-based gait recognition?," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2004, pp. 776–782.

[17] Ho, T. K., "The random subspace method for constructing decision forests," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1998, 20, pp. 832–844.

[18] Wang, X. and Tang, X., "Random sampling lda for face recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2004, pp. 259–265.

[19] Jain, A., Ross, A., and Prabhakar, S., "An introduction to biometric recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, 2004, 14, pp. 4–20.

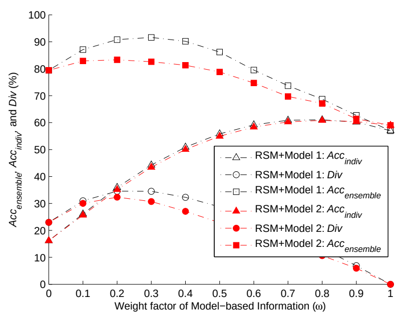
[20] Lovell, B., Bigdeli, A., and Mau, S., "Invited paper: Embedded face and biometric technologies for national and border security," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2011, pp. 117–122.

[21] Kuncheva, L. I. and Whitaker, C. J., "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, 2003, 51, pp. 181–207.

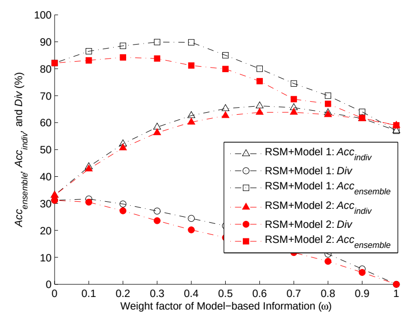
[22] Skalak, D. B., "The sources of increased accuracy for two proposed boosting algorithms," in *Proc. American Association for Arti Intelligence Workshop*, 1996, pp. 120–125.

[23] Kale, A., Sundaresan, A., Rajagopalan, A. N., and et al.,, "Identification of humans using gait," *IEEE Trans. on Image Processing*, 2004, 13, pp. 1163–1173.

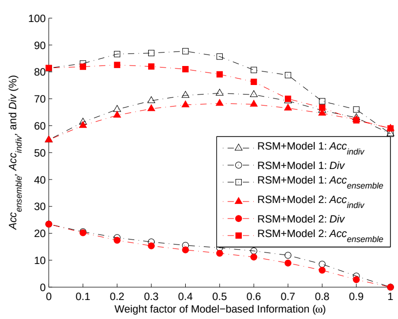
[24] Jain, A., Nandakumar, K., and Ross, A., "Score normalization in multimodal biometric systems," *Pattern Recognition*, 2005, 38, pp. 2270–2285.



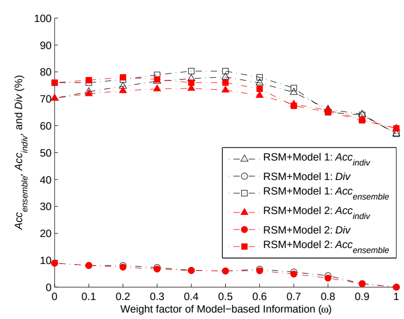
(a) $s = 20$



(b) $s = 40$

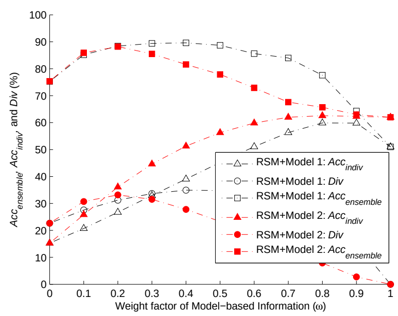


(c) $s = 80$

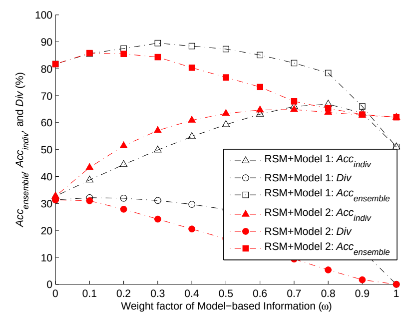


(d) $s = 160$

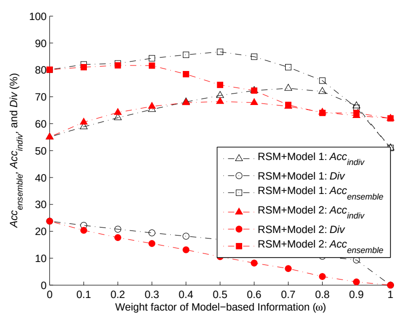
Figure 4: On DB-high, performance distributions with respect to the weight of model-based information (ω). (a)-(d) are based on classifier ensemble with random feature number $s = \{20, 40, 80, 160\}$, respectively.



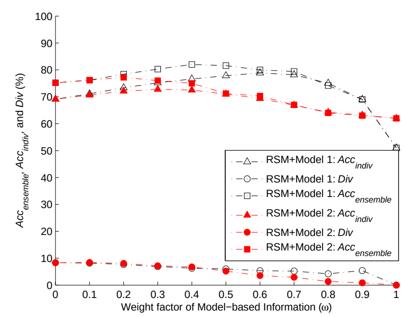
(a) $s = 20$



(b) $s = 40$

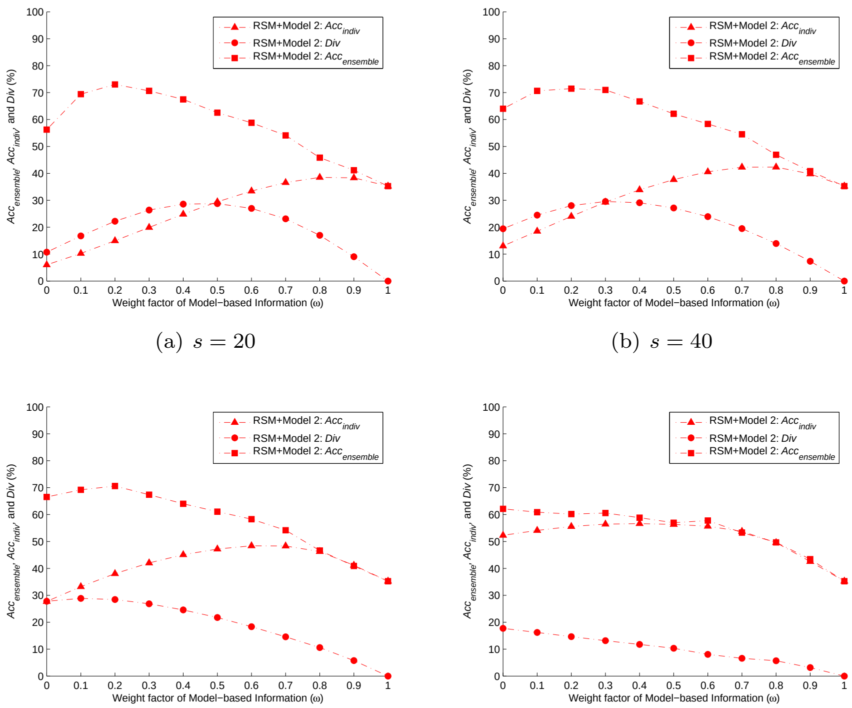


(c) $s = 80$



(d) $s = 160$

Figure 5: On DB-low, performance distributions with respect to the weight of model-based information (ω). (a)-(d) are based on classifier ensemble with random feature number $s = \{20, 40, 80, 160\}$, respectively.



(a) $s = 20$ (b) $s = 40$

(c) $s = 80$ (d) $s = 160$

Figure 6: On DB-outdoor, performance distributions with respect to the weight of model-based information (ω). (a)-(d) are based on classifier ensemble with random feature number $s = \{20, 40, 80, 160\}$, respectively.