Combining Gait and Face for Tackling the Elapsed Time Challenges

Yu Guan, Xingjie Wei, Chang-Tsun Li
Department of Computer Science,
University of Warwick, Coventry, CV4 7AL, UK
g.yu,x.wei,c-t.li@warwick.ac.uk

Gian Luca Marcialis, Fabio Roli
Department of Electrical and Electronic Engineering,
University of Cagliari, 09123, Cagliari, Italy
marcialis, roli@diee.unica.it

Massimo Tistarelli
Department of Sciences and Information Technology,
Univeristy of Sassari, 07100, Sassari, Italy
tista@uniss.it

Abstract

*Random Subspace Method (RSM) has been demonstrated as an effective framework for gait recognition. Through combining a large number of weak classifiers, the generalization errors can be greatly reduced. Although RSM-based gait recognition system is robust to a large number of covariate factors, it is, in essence an unimodal biometric system and has the limitations when facing extremely large intra-class variations. One of the major challenges is the elapsed time covariate, which may affect the human walking style in an unpredictable manner. To tackle this challenge, in this paper we propose a multimodal-RSM framework, and side face is used to strengthen the weak classifiers without compromising the generalization power of the whole system. We evaluate our method on the TUM-GAID dataset, and it significantly outperforms other multimodal methods. Specifically, our method achieves very competitive results for tackling the most challenging elapsed time covariate, which potentially also includes the changes in shoe, carrying status, clothing, lighting condition,*etc.

1. Introduction

Compared with other biometric traits like fingerprint or iris, gait recognition can be applied at a distance without requiring the cooperations from subjects, and it has gained considerable attentions in the past decade. However, for automatic gait recognition systems, covariate factors (*e.g.*, camera viewpoint, carrying condition, clothing, shoe, speed, video frame-rate, *etc.*) may affect the performance. These factors have been extensively studied in the literatures [3–7, 9, 10, 14, 16–21, 23, 24], and great performance improvements are made. Existing gait recognition methods can be roughly divided into two categories: model-based and appearance-based approaches. Model-based methods (*e.g.*, [1]) aim to model the human body structure for recognition, while appearance-based approaches (*e.g.*, [3–7,9,10,14,16–21,23,24]) are more general and they can perform classification regardless of the underlying body structure. Compared with model-based methods, appearance-based approaches can also work well in the low resolution environments, when body structure is difficult to construct.

Gait Energy Image (GEI) [7] is a popular feature template and it is widely used in recent appearance-based algorithms due to its simplicity and effectiveness [3, 5–7, 9, 10, 14, 16, 18, 20, 21, 23, 24]. GEI is the average silhouette over one gait cycle, which encodes a number of binarized silhouettes into a grayscale image [7, 16]. The averaging operation can, not only smooth the segmentation errors but also significantly reduce the computational cost [7]. Several GEI samples from the newly released TUM Gait from Audio, Image and Depth (TUM-GAID) database [9] are illustrated in Fig. 1. However, when the walking condition of the query gait is different from the gallery, direct GEI matching makes the classification process prone to errors [16]. To reduce such effect, various feature extraction algorithms have been proposed in the previous works (*e.g.*, [5, 7, 20, 21]). Among these algorithms, Random Subspace Method (RSM) [5] is the most effective one. To overcome the problem caused by overfitting the less representative training data (*i.e.*, gallery), a large number of weak classifiers are combined. The RSM framework is robust to a large number of covariate factors such as shoe, (small changes in) view, carrying condition [5], clothing [6], speed [3], frame-rate [4], *etc*. However, the effectiveness of RSM has not been fully validated against the most challenging elapsed time covariate, which potentially also includes the changes of clothing, carrying condition, weight, fatigue, *etc*. Elapsed time may have an unpredictable effect on gait, making the gait recognition system less reliable.

It is an open question to handle extremely large intra-class variations (*e.g.*, elapsed time) for unimodal biometric systems, and in this case building multimodal system could be an effective way to enhance the performance [13]. In the context of gait recognition, it is natural to fuse face [10,17,23,24]. In [17], improved performance was achieved when fusing lateral gait and frontal face. It is more practical to fuse lateral gait and side face (referred to as gait and face in this paper), since both modalities can be acquired using the same camera. In [23], Zhou and Bhanu performed a score-level fusion of gait and the Enhanced Side Face Image(ESFI), they found that improving face image quality can further enhance the fusion performance. They also used a feature-level fusion strategy by concatenating the ESFI and gait into a new feature template for classification [24]. Alpha matte was used by Hofmann *et al.* to segment gait and face images with better qualities, before a score-level fusion. Gait can also be combined with gait from different feature spaces [7] or cameras/sources [9, 18]. In [7], GEI and synthetic GEI templates were generated from the same silhouettes to different feature spaces, and they were fused in the score level. In [9], GEI, depth, and audio information were fused to tackle the elapsed time challenges and encouraging performance was achieved. In [18], by concatenating gait information from 3 cameras from different views into a new template, promising results were achieved on a small temporal dataset.

Performance can be improved when fusing gait and other modalities. However, the role of feature extraction is largely neglected by the afore-mentioned methods [7, 9, 10, 18, 23, 24]. In this work, we aim to extend the state-of-the-art RSM to a multimodal-RSM framework. Multi-class Kernel Fisher Analysis (KFA) [15] is used for face feature extraction. Then we use the corresponding face score to strengthen the gait-based weak classifiers, before the majority vote. By assigning lower weight to the face score, the diversity of the weak classifiers is preserved. The experimental results suggest that although face recognition at a distance is less reliable, it can effectively assist gait recognition. By fusing gait and face under the proposed multimodal-RSM framework, the performance is dramatically improved for tackling the hard problems such as elapsed time.

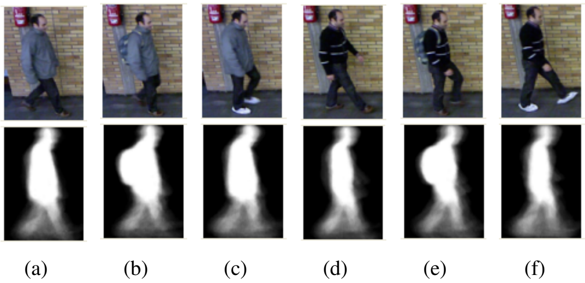


Figure 1. Gait images from the TUM-GAID dataset for a subject in 6 different conditions. (a): normal; (b): backpack; (c): coating shoes; (d): elapsed time + normal; (e): elapsed time + backpack; (f): elapsed time + coating shoes. Top row includes the gait RGB images while the bottom row includes the corresponding GEIs.

2. Motivation

Compared with other covariate factors, only a few works systematically study the effect of elapsed time. In [18], Matovski *et al.* investigated the effect of elapsed time (up to 9 months) based on 25 subjects by fusing the gait information from 3 cameras in different views, and their results suggest that: 1) irrespective of other covariates, short term elapsed time does not affect the recognition significantly; 2) the accuracies may drop rapidly when the potential covariates (*e.g.*, clothing) are included. Since in real-world scenarios it is unrealistic to have other covariates perfectly controlled, our objective in this work is to tackle the elapsed time challenge in a less constrained environment. Besides, different from the work in [18] by using 3 different cameras, the modalities we are fusing (*i.e.*, gait and face) can be acquired using a single camera.

In RSM systems, weak classifiers with lower dimensionality tend to have better generalization power [8]. However, they may face an underfitting problem if the dimensionality is too low. It is desirable to use an independent source to strengthen the weak classifiers. Although face at a distance may be less reliable, it may provide some complementary information for gait. In [11], Jain *et al.* demonstrated that the error rate of fingerprint system can be further reduced by integrating soft biometric information. Similarly, in this work we treat the less reliable face as a soft biometric trait by assigning lower weight to its score. After summing up the weighted face score and each gait score (corresponding to each weak classifier), the final result is achieved by majority vote among the updated classifiers. It is also worth mentioning that by assigning lower weight to face score, the face information is less likely to smooth the diversity (among the updated weak classifiers), which is crucial for multiple classifier systems [2].

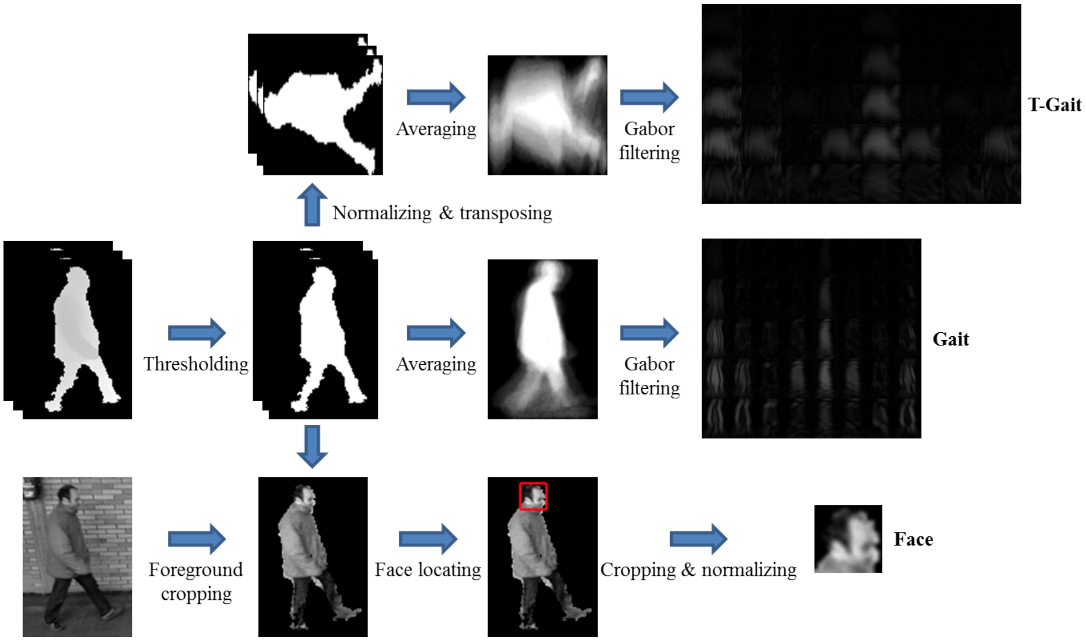


Figure 2. The process of generating the 3 feature templates, *i.e.*, *T-Gait, Gait, Face*

3. Gait Recognition

3.1. Gait Feature Templates

Gabor-filtered GEI (referred to as *Gait*) has been demonstrated to be an effective feature template for gait recognition [3, 20, 21]. Given a GEI sample, Gabor functions from 5 scales and 8 orientations are employed to generate the *Gait* template (see Fig. 2). Another gait feature template derived from the same silhouettes is also employed in this work. After transposing the optimized-GEI defined in [14], we use the corresponding Gabor-filtered features (referred to as *T-Gait*) as the second gait template (see Fig. 2). For a gait sequence, by using *Gait* (resp. *T-Gait*), RSM [5] can extract the random features in the column direction (resp. row direction) in the first place. For computational efficiency, similar to [21], we use the subsampled version of both templates.

3.2. RSM for Gait Recognition

In this section, we introduce the feature extraction process using the RSM framework [5]. First 2DPCA [22] is used to decorrelate the feature space (in column direction). Given n *Gait/T-Gait* samples $I_i (i = 1, 2, \dots, n)$ in the gallery, the scatter matrix S can be estimated using:

$$S = \frac{1}{n} \sum_{i=1}^{n} (I_i - \mu)^T (I_i - \mu) \qquad (1)$$

where $\mu = \frac{1}{n} \sum_{i=1}^{n} I_i$. The eigenvectors (of S) associated with zero eigenvalues are removed, while the rest are preserved as candidates for the random subspace construction. L random spaces are generated, with the projection matrices $R^1, R^2, ..., R^L$ formed by randomly selecting N eigenvectors from non-zero candidates. Each sample can be projected into L subspaces, and we use the corresponding coefficients as the new feature descriptors for a certain subspace.

Given the l^{th} subspace, to achieve the optimal class separability, 2DLDA is further employed (in row direction) to project the coefficients corresponding to the gallery samples into the canonical space. There is a projection matrix W^l maximizing the ratio of the between-class scatter matrix S_b^l to the within-class scatter matrix S_w^l, *i.e.*,

$$\underset{W^l}{\operatorname{argmax}} \operatorname{trace}(((W^l)^T S_w^l W^l)^{-1}((W^l)^T S_b^l W^l)) \qquad (2)$$

For the l^{th} subspace, let W^l(resp. R^l) be the canonical s-pace (resp. eigenspace) projection matrix, then the feature extraction can be performed. Given a gait sequence with n_p *Gait/T-Gait* samples $I_t (t = 1, 2, \dots, n_p)$, the corresponding extracted feature templates Q_t^l are:

$$Q_t^l = W^l (I_t R^l) \qquad (t = 1, 2, \dots, n_p) \qquad (3)$$

After feature extraction for the the l^{th} subspace using (3), Nearest Mean (NM) classifier can be used to get the gait score for the l^{th} subspace. For RSM-based gait recognition system [5], among the L base classifiers, a decision-level fusion (*e.g.*, majority vote) is normally used for the final classification decision.

4. Face Recognition

4.1. Face Cropping

The process of face cropping is demonstrated in Fig. 2: first we use the binarized depth mask on the corresponding grayscale image to get the whole human body from the background. Then, a line-by-line scanning is performed to locate two landmarks (*i.e.*, the top-most pixel and the right-most pixel of the upper body area). A pre-defined face area is then cropped based on the two landmarks. Finally the cropped faces are aligned by the landmarks and normalized to a size of 18 × 18 pixels, and they are referred to as *Face* templates in this work.

4.2. KFA for Face Recognition

In this work, multi-class KFA [15] is used for face feature extraction. KFA first performs nonlinear mapping from the input space to a high dimensional feature space by using the kernel trick. Then LDA can be employed in the new feature space.

Let $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m \times n}$ be the data matrix of n training samples (*i.e.*, gallery) in the input space, and x_i denotes the concatenated vector from the i^{th} *Face* template. Assume there are c classes and $n_1, n_2, ..., n_c$ are the number of training samples for each class where $\sum_{i=1}^{c} n_i = n$. Let $f : \mathbb{R}^m \to F$ be a nonlinear mapping from the input space to the feature space. Then the data matrix in the feature space can be represented as: $Y = [f(x_1), f(x_2), ..., f(x_n)]$. Generally, we expect different classes to be well separated while samples within the same class to be tightly related. Similar to (2), this leads to optimizing $J_1 = \mathrm{trace}(S_m^{-1} S_b)$ where S_b is the between-class scatter matrix while S_m is the mixture scatter matrix in the feature space.

However, it is difficult to evaluate S_m and S_b in the high dimensional feature space. In KFA, a kernel matrix K is defined as: $K = Y^T Y$ where $K_{ij} = (f(x_i) \cdot f(x_j)), i, j = 1, 2, ..., n$. So optimizing J_1 leads to solving the following eigenvalue problem by replacing S_m and S_b with the kernel matrix K:

$$KZK\alpha = \lambda KK\alpha, \tag{4}$$

where α (resp. λ) denotes the eigenvector (resp. eigenvalue). Here $Z \in \mathbb{R}^{n \times n}$ is a block diagonal matrix: $Z = diag\{Z_1, Z_2, ..., Z_c\}$ where Z_j is a $n_j \times n_j$ matrix with elements all equal to $\frac{1}{n_j}, j = 1, 2, ..., c$. Let $A = [\alpha_1, \alpha_2, ..., \alpha_r] \in \mathbb{R}^{n \times r}$ be the eigenvectors corresponding to the $r (r \leqslant c - 1)$ largest eigenvalues, and for a face vector x, the KFA features can be extracted by using:

$$F = A^T B \tag{5}$$

where $B = [f(x_1) \cdot f(x) \quad f(x_2) \cdot f(x) \quad ... \quad f(x_n) \cdot f(x)]^T$. Actually, the kernel matrix K can be computed by a kernel function instead of explicitly performing the nonlinear mapping. Here we use the fractional power polynomial kernel function [15] as:

$$k(a, b) = (f(a) \cdot f(b)) = sign(a \cdot b)(abs(a \cdot b))^d \tag{6}$$

where $sign(\cdot)$ is the sign function and $abs(\cdot)$ is the absolute value operator. We empirically set $d = 0.8$. By using KFA, the linear model is able to capture the nonlinear patterns in the original data. After feature extraction, we can get the face score by a NM classifier.

5. Fusion Strategy

In the context of the RSM framework, we update the voters/base classifiers by fusing face score and each gait score out of the L subspaces, before the majority vote. For the l^{th} subspace, we first use the min-max normalization [12] on the face score and the l^{th} gait score, respectively. Then the voters are updated using weighted sum rule. Specifically, for the l^{th} subspace, given the normalized gait score S^l_{gait} and the normalized face score S_{face}, the updated score $S^l_{face+gait}$ corresponding to the l^{th} classifier is defined as:

$$S^l_{face+gait} = \omega S_{face} + S^l_{gait}. \tag{7}$$

Due to the fact that the same face score S_{face} is used to update a total number of L gait-based classifiers, the face score weight ω is negatively correlated with the diversity of this multiple classifier system. For RSM-based multiple classifier systems, previous studies [5, 8] empirically demonstrated that the performance does not decrease with the increasing number of classifiers (*i.e.*, non-decreasing diversity). From the perspective of diversity, intuitively it is preferable to assign ω a small value.

In [11], when the less reliable soft biometric traits (gender, ethnicity, and height) with low weights assigned were fused in the fingerprint system, significant performance gain was achieved. In the context of human identification at a distance, although face is less reliable due to low resolution or the presence of other covariates, it may provide some complementary information for gait when using it as the ancillary information with low weight assigned.

Due to the afore-mentioned issues, we use face as a soft biometric trait with its score low weight assigned, and the corresponding evaluations are provided in Section 6.2. After the score is updated for each voter, majority vote is applied for the final classification decision. We also fuse two multimodal-RSM systems (*i.e.*, based on source *T-Gait* + *Face* and source *Gait* + *Face*) in the decision level. In this work, we simply combine equal number voters from each source, before the majority vote. For example, the decision-level fusion with 1000 voters denotes 500 voters from source *T-Gait* + *Face* and 500 voters from source *Gait* + *Face*, respectively.

-	Gallery	Probe					
Walking Condition	N	N	B	S	TN	TB	TS
#Seq.	155×4	155×2	155×2	155×2	16×2	16×2	16×2
#*Gait/T-Gait* per Seq.	1	1	1	1	1	1	1
#*Face* per Seq.	4	1	1	1	1	1	1

Table 1. Experimental settings on the TUM-GAID dataset. Abbreviation note : N - Normal, B - Backpack, S - Shoe, TN - Time+Normal, TB - Time+Backpack, TS - Time+Shoe.

6. Experiments

We conduct experiments on the newly released TUM-GAID dataset [9]. This dataset simultaneously contains RGB video, depth and audio with 305 subjects in total. In [9], Hofmann *et al.* designed an experimental protocol (based on 155 subjects) to evaluate the robustness of algorithms against covariate factors like shoe, carrying condition (5kg backpack), elapsed time (January/April) which also potentially includes changes in clothing, lighting condition, *etc*.

In the TUM-GAID dataset, the depth images in the tracked bounding box are provided and we can get the corresponding binarized silhouettes/masks by thresholding and aligning. But note that our method can also be applied on RGB videos, and in this case the silhouettes can be segmented using some background substraction methods (*e.g*., [10, 19]). Then we can acquire the *Face* template using the method introduced in Section 4.1. The process of getting the 3 feature templates used in this paper is illustrated in Fig. 2. There is 1 gait template (*i.e*.,*Gait/T-Gait*) corresponding to a gait sequence. For face, since there are small view changes in a sequence, in the gallery we select 4 *Face* templates from each sequence to capture more intra-class variations. For probe, only 1 *Face* template is selected from each sequence. The experimental settings of the gallery and probe sets are shown in Table 1.

There are two main parameters in the RSM framework [5], namely, the random subspace dimension N, and classifier number L. It was verified that in the RSM framework the performance does not decrease with the increasing number of classifiers [5, 8]. In this work, following [5], we set $L = 1000$. In the RSM framework when the subspace dimension N is too large it faces an overfitting problem [5] and its performance converges to the one of traditional 2D-PCA+2DLDA. Classifiers with small value of N can usually generalize well [8], but underfitting may occur when N is too small. Our aim is to strengthen the weak classifiers by fusing the additional face information (to avoid underfitting). In this case, it does not sacrifice the generalization power of the whole system by using a small value of N, and we set $N = 2$ in this work.

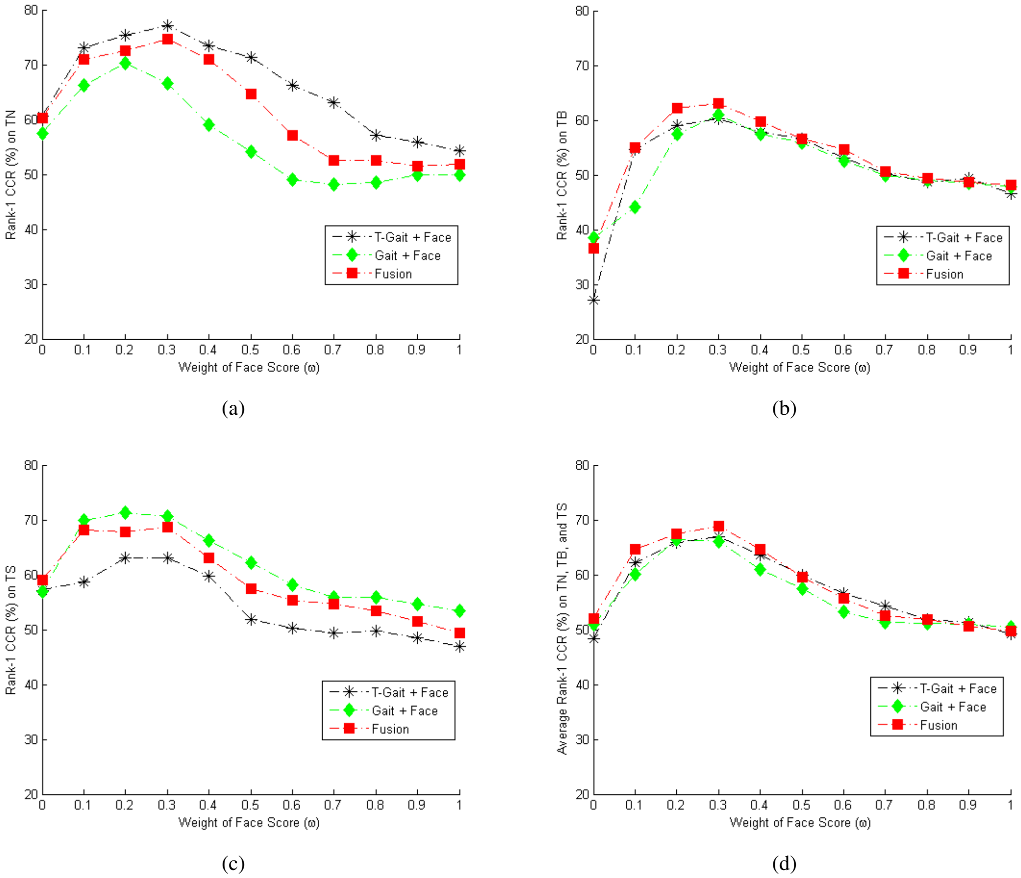
To evaluate the performance of the algorithms, we use the rank-1/rank-5 Correct Classification Rate (CCR). Rank-1 (resp. rank-5) CCR shows the correct subject is ranked as the top 1 candidate (resp. top 5 candidates). Due to the random nature, the results of different runs may vary to some extent. We repeat all the experiments 10 times and the overall rank-1 performance statistics (mean, standard deviation, maxima and minima) of the proposed multimodal-RSM system are reported in Table 4, which indicates the stability of our method. For the rest of the paper, we only report the mean values.

Experiment	N	B	S	TN	TB	TS
#Seq.	310	310	310	32	32	32
Rank-1 CCRs						
Face + KFA	89	72	71	44	**38**	44
Gait + RSM	**100**	**79**	**97**	58	**38**	**57**
T-Gait + RSM	99	62	92	**61**	27	**57**

Table 2. The rank-1 CCRs (%) by using single modality on the 6 probes from the TUM-GAID dataset.

6.1. Identification using Single Modality

Experiments based on 3 single modalities (*i.e*., *Face* only, *Gait* only, and *T-Gait* only) are conducted, respectively. The rank-1 CCRs corresponding to the 6 probe sets are illustrated in Table 2. Generally, based on a certain modality, reasonable results can be achieved on probe sets N, B, and S. However, when elapsed time is taken into account (*i.e*., on probe sets TN, TB, and TS), the CCRs decrease significantly. It was experimentally verified that carrying condition has little impact on RSM-based gait recognition algorithms [3, 5]. However, when the object carried is heavy (5kg backpack in TUM-GAID dataset [9]), it may change the whole walking style to some extent and thus affects the performance. Similarly, for the elapsed time covariate, subjects may also change their walking styles in an unpredictable manner due to potential changes in carrying status, shoe, clothing, emotion, fatigue, *etc*. The coupled effect may have significant impact on the recognition accuracies when only gait trait is used. Although face may be affected by lighting condition and low resolution, intuitively it is less sensitive to the heavy object carried, shoe, clothing, *etc*. It may provide some additional information to enhance the performance of the gait recognition system.



(a)

(b)

(c)

(d)

Figure 3. Using multimodal-RSM to tackle the elapsed time challenges on probes TN, TB, and TS. Fusion denotes source *T-Gait + Face* and source *Gait + Face* are fused in the decision level (with 500 classifiers from each source).

6.2. Tackling the Elapsed Time Covariate using Multimodal-RSM

In this section, by using face as ancillary information, we apply the proposed multimodal-RSM to tackle the challenging elapsed time covariate. Over probes TN, TB, and TS, the rank-1 CCR distributions with respect to face score weight are reported in Fig.3(a)-3(c). Fig. 3(d) summarizes the average performance distribution of the 3 probes with respect to face score weight. From these figures, we can observe:

1. Higher performance can be achieved when the weight of the face score is relatively low (*e.g.*, $0.1 \leq \omega \leq 0.3$). For example when $\omega = 0.3$, the performance gains over gait-based RSM (when $\omega = 0$) are upto $15\%, 25\%$, and 10% for probes TN, TB, and TS, respectively.
2. Based on low weight of face score, *T-Gait + Face* has higher performance on probe TN (Fig.3(a)), while *Gait + Face* yields higher accuracies on probe TS (Fig. 3(c)). When fusing these two sources (*i.e.*, *Gait + Face* and *T-Gait + Face*) in the decision level, higher rank-1 CCRs can be achieved on the most challenging probe TB (Fig.3(b)). On the 3 probe sets, decision-level fusion of these two sources can always deliver stable performance (Fig. 3(a)-3(d)).

Generally, the performance is very competitive when the weight is within a certain range of small values (*e.g.*, $0.1 \leq \omega \leq 0.3$). The performance decreases when setting ω too high or too low. There are two extreme cases according to (7): when $\omega \to \infty$, the performance converges to the one of face recognition system (*i.e.*, KFA + *Face* in this paper), and when $\omega = 0$, it becomes a gait-based RSM system. For a general multimodal-RSM system, given the fact that it is difficult to collect representative validation data (which covers all the possible covariates) for parameter tuning, it remains an open question to find the optimal ω. Nevertheless, experimental results suggest that very significant performance gains can be achieved by assigning face score a relatively low weight. In this case, the weak classifiers are strengthened without sacrificing the diversity of the whole multiple classifier system.

Due to the unpredictable nature, elapsed time is often deemed as the most challenging covariate factor in the con-

Experiment	N	B	S	TN	TB	TS	Overall
#Seq.	310	310	310	32	32	32	-
Rank-1 CCRs							
GEI + Eigenface [10]	97	63	65	47	50	28	71.9
Audio + Depth + GEI [9]	99	59	95	66	3	50	80.2
Proposed method	**100**	**96**	**99**	**75**	**63**	**69**	**95.6**
Rank-5 CCRs							
GEI + Eigenface [10]	100	84	79	63	63	50	85.0
Audio + Depth + GEI [9]	100	85	99	**81**	28	72	91.5
Proposed method	**100**	**98**	**100**	80	**74**	**78**	**97.3**

Table 3. Algorithms comparison in terms of rank-1/rank-5 CCRs (%). Overall denotes the weighted average.

Mean	Std	Max	Min
95.55	0.22	95.81	95.22

Table 4. The rank-1 CCR statistics (%) over 10 runs of our multimodal-RSM system

text of human identification at a distance. Based on our multimodal-RSM framework, performance is significantly improved by fusing these two independent yet complementary modalities. Specifically, by assigning relatively low weight, face is used as an ancillary information to strengthen the gait-based weak classifiers without sacrificing the diversity of the whole multiple classifier system. The general performance can be further boosted by fusing classifiers from the two sources (*i.e.*, *Gait* + *Face* and *T-Gait* + *Face*) in the decision level. For the rest of this paper, we only report the results of this decision-level fusion (with $\omega = 0.3$).

6.3. Algorithms Comparison

Over the 6 probe sets, we compare our multimodal-RSM with the 3 unimodal-based methods (from Section 6.1), as shown in Fig.4. It clearly indicates the effectiveness of our fusion method for tackling the most challenging elapsed time covariate. It also has superb performance when the subject carries 5kg backpack (probe B). Since our results are based on 10 runs, the corresponding maxima, minima, mean and standard deviation are listed in Table 4, from which we can see the performance of our system is highly stable.

We also compare our method with two recently proposed multimodal methods [9, 10]. We implement the GEI + Eigenface [10] and quote the results of Audio + Depth + GEI from [9]. Table 3 illustrates the performance of the three methods in terms of rank-1/rank-5 CCRs, and the performance of our method is generally much higher. Specifically, for tackling the elapsed time, the method in [10] has lower performance in probe TS while the method in [9] (when face information is not fused) only has 3% rank-1 CCR in probe TB. Compared with them, our method consistently has much higher performance in these cases. However, the coupled effect of elapsed time and heavy backpack (probe TB) has larger impact on our system, with only 63% rank-1 CCR achieved, and the accuracy is much lower than the ones on probe TN (by 12%) and probe TS (by 6%). Nevertheless, compared with the gait-based RSM systems, fusing face as additional information can dramatically reduce the error rates. Motivated by this observation, in the future, we will fuse more soft biometric traits (*e.g.*, height, gender, age, *etc.*) into the proposed multimodal-RSM system to further boost the performance.

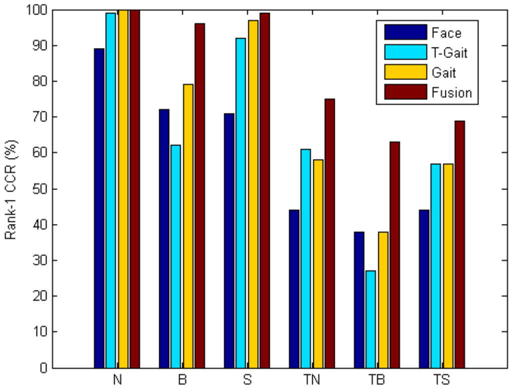


Figure 4. Unimodal vs. Multimodal. Fusion denotes source *T-Gait* + *Face* and source *Gait* + *Face* are fused in the decision level.

7. Contributions, Limitations, and Future Work

In this work, we extend the existing RSM framework [5] by allowing other modalities (*e.g.*, face) to be fused. Face is deemed as a soft biometric trait to provide additional information to strengthen the weak classifiers in terms of discrimination capability, without compromising the generalization power of the whole system. Based on the updated voters (by fusing face information), combining the *T-Gait*-based and *Gait*-based RSM systems in the decision level can further reduce the error rate. The proposed multimodal-

RSM system has much higher performance than the unimodal systems and other multimodal methods.

Although additional experiments and theoretical findings are necessary to draw the final conclusions on the benefit of fusing face information, this work empirically demonstrates an effective way on combining multi-modalities information to tackle the most challenging elapsed time covariate, which may also potentially include the changes of clothing, shoe, carrying status, *etc*. However, our experiments are based on indoor environments with well segmented silhouette and face, while in the outdoor environments, segmented data based on background substraction methods [10, 19] can be relatively noisy. In the future, we will evaluate our method on the more challenging outdoor databases (*e.g.*, the USF dataset [19]). Besides, we will also explore how to effectively fuse other soft biometric traits like age, gender, height, *etc*. into this RSM-based system to further improve the performance on challenging problems.

References

[1] D. Cunado, M. Nixon, and J. Carter. Using gait as a biometric, via phase-weighted magnitude spectra. In *Proceedings of 1st International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 95–102, March 1997.

[2] L. Didaci, G. Fumera, and F. Roli. Diversity in classifier ensembles: Fertile concept or a dead end? In *Proceedings of International Workshop on Multiple Classifier Systems (MC-S)*, May 2013 (In Press).

[3] Y. Guan and C.-T. Li. A robust speed-invariant gait recognition system for walker and runner identification. In *Proceedings of IAPR International Conference on Biometrics (ICB)*, June 2013, (In Press).

[4] Y. Guan, C.-T. Li, and S. D. Choudhury. Robust gait recognition from extremely low frame-rate videos. In *Proceedings of International Workshop on Biometrics and Forensics (I-WBF)*, pages 1–4, April 2013.

[5] Y. Guan, C.-T. Li, and Y. Hu. Random subspace method for gait recognition. In *Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 284–289, July 2012.

[6] Y. Guan, C.-T. Li, and Y. Hu. Robust clothing-invariant gait recognition. In *Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pages 321–324, July 2012.

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

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

[9] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 2013, (In Press).

[10] M. Hofmann, S. Schmidt, A. N. Rajagopalan, and G. Rigoll. Combined face and gait recognition using alpha matte preprocessing. In *Proceedings of IAPR International Conference on Biometrics (ICB)*, pages 390–395, 2012.

[11] A. Jain, S. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proceedings of the International Conference on Biometric Authentication (ICBA)*, pages 731–738, 2004.

[12] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[13] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, Jan. 2004.

[14] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *Proceedings IEEE 12th International Conference on of Computer Vision Workshops (ICCVW)*, pages 1058–1064, 2009.

[15] C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):725–737, 2006.

[16] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: averaged silhouette. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 211–214, 2004.

[17] Z. Liu and S. Sarkar. Outdoor recognition at a distance by fusing gait and face. *Image Vision Computing*, 25(6):817–832, 2007.

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

[19] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother, and E. Ortiz. The humanid gait challenge problem: data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

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

[21] D. Xu, Y. Huang, Z. Zeng, and X. Xu. Human gait recognition using patch distribution feature and locality-constrained group sparse representation. *IEEE Transactions on Image Processing*, 21(1):316–326, Jan. 2012.

[22] J. Yang, D. Zhang, A. Frangi, and J. Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137, Jan. 2004.

[23] X. Zhou and B. Bhanu. Integrating face and gait for human recognition at a distance in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1119–1137, 2007.

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