The HumanID Gait Challenge Problem: Data Sets, Performance, and Analysis

Sudeep Sarkar, *Member*, *IEEE*, P. Jonathon Phillips, *Member*, *IEEE*, Zongyi Liu,
Isidro Robledo Vega, *Member*, *IEEE*, Patrick Grother, and Kevin W. Bowyer, *Fellow*, *IEEE*

Abstract—Identification of people by analysis of gait patterns extracted from video has recently become a popular research problem. However, the conditions under which the problem is "solvable" are not understood or characterized. To provide a means for measuring progress and characterizing the properties of gait recognition, we introduce the HumanID Gait Challenge Problem. The challenge problem consists of a baseline algorithm, a set of 12 experiments, and a large data set. The baseline algorithm estimates silhouettes by background subtraction and performs recognition by temporal correlation of silhouettes. The 12 experiments are of increasing difficulty, as measured by the baseline algorithm, and examine the effects of five covariates on performance. The covariates are: change in viewing angle, change in shoe type, change in walking surface, carrying or not carrying a briefcase, and elapsed time between sequences being compared. Identification rates for the 12 experiments range from 78 percent on the easiest experiment to 3 percent on the hardest. All five covariates had statistically significant effects on performance, with walking surface and time difference having the greatest impact. The data set consists of 1,870 sequences from 122 subjects spanning five covariates (1.2 Gigabytes of data). The gait data, the source code of the baseline algorithm, and scripts to run, score, and analyze the challenge experiments are available at http://www.GaitChallenge.org. This infrastructure supports further development of gait recognition algorithms and additional experiments to understand the strengths and weaknesses of new algorithms. The more detailed the experimental results presented, the more detailed is the possible meta-analysis and greater is the understanding. It is this potential from the adoption of this challenge problem that represents a radical departure from traditional computer vision research methodology.

Index Terms—Gait recognition, human motion analysis, biometrics, human identification, silhouette detection, spatiotemporal correlation.

1 Introduction

Human movement analysis is not new. Biomechanical analysis of gait has been successfully applied in human clinical gait analysis [1]. With regards to gait recognition, a major early result from Psychology is by Johansson [2], who used point light displays to demonstrate the ability of humans to rapidly distinguish human locomotion from other motion patterns. Cutting and Kozlowski [3] showed that this ability also extends to recognition of friends. Since then, there have been various experiments to show that humans can recognize gender, direction of motion, and weight carry conditions. Perhaps the most recent evidence comes from the experiments by Stevenage et al. [4] who show that humans can identify individuals on the basis of their gait signature, without reliance on shape, in the presence of lighting variations and under brief exposures.

Much progress has been made in computer vision-based human motion analysis since the early days of analyzing human motion in terms of groups of rigidly moving points [5], [6]. An excellent snapshot into current work on human movement modeling is available in a recent special issue [7]. Work in computer vision-based human motion modeling can be classified according to the model employed: Articulated versus elastic nonrigid, with and without prior shape modeled [8]; or in terms of whether 2D or 3D models are implicitly or explicitly employed [9]. A more recent, extensive survey [10] looks at more than 130 publications in computer vision-based human motion analysis and classifies them based on the issues addressed: Initialization (eight publications), tracking (48 publications), pose estimation (64 publications), and recognition (16 publications). The review also finds that the three most common assumptions used effectively constrain the scene to be 1) indoors, 2) with static background, and 3) with uniform background color. These assumptions make it difficult to judge the autonomous operation of the developed ideas in real life outdoor situations.

In the specific area of gait recognition, most works have focused on discriminating between different human motion types, such as running, walking, jogging, or climbing stairs [11]. It is only recently that human identification (HumanID) from gait has received attention and become an active area of computer vision [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. The majority of these papers report results that are either limited in the size of the data set (less than 30 people), taken indoors, or have examined performance under a limited

- *S. Sarkar and Z. Liu are with the Department of Computer Science and Engineering, University of South Florida, 4202 E. Fowler Avenue, ENB 118, Tampa, FL 33620. E-mail: {sarkar, zliu4}@csee.usf.edu.*
- *P.J. Phillips and P. Grother are with NIST, 100 Bureau Drive, MS 8940, Gaithersburg, Maryland 20899. E-mail: {jonathon, pgrother}@nist.gov.*
- *I.R. Vega is with the Technological Institute of Chihuahua, Avenida Tecnológico #2909, Chihuahua, Chihuahua, Mexico. E-mail: irobledo@itchihuahua.edu.mx.*
- *K.W. Bowyer is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556. E-mail: kwb@cse.nd.edu.*

Manuscript received 29 Aug. 2003; revised 21 Apr. 2004; accepted 14 June 2004; published online 13 Dec. 2004.
Recommended for acceptance by S. Sclaroff.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0255-0803.

0162-8828/05/$20.00 © 2005 IEEE Published by the IEEE Computer Society

TABLE 1
Summary of Data Sets Used in Current Work on Gait-Based Identification

Data set Used in	Size (subjects, #seqs.)	Scene	Data Covariates
UCSD, 1998 [18], [13]	6, 42	Wall background	Time (minutes)
CMU MoBo, 2001 [35], [25], [27] [21], [22], [19]	25, 600	Indoor, Treadmill	Viewpoint, Walking speeds, Carrying condition, Surface incline
Georgia Tech., 2001 [29], [16], [15], [28]	15, 168	Outdoor	Time (6 months), Viewpoint
	18, 20	Magnetic tracker	Time (6 months)
Maryland, 2001 [25], [27], [21] [22], [19]	25, 100	Outdoor, 30m away	
	55, 222	Outside, Top mounted	Viewpoints (front, side), Time
MIT, 2001 [27], [22], [20]	24, 194	Indoors	Time (13 repeats over 3 month)
Southampton, 2001 [22], [18], [23]	28, 112	Indoors, green background	Time (minutes)
Gait Challenge, 2002 [27], [30], [26], [24] [31], [32], [33], [34]	122, 1870	Outdoor	Viewpoint ($\approx 30°$), Surface (Concrete, Grass), Shoe, Carrying condition, Time (33 repeats, 6 months)

number of conditions; see Table 1. These types of data sets and papers have advanced gait recognition from its very beginnings to the present. To mature gait recognition and to assess its potential requires a larger, more diverse data set. To assist in advancing automatic gait recognition, we introduce the *HumanID Gait Challenge Problem*.

The motivation behind the design of the challenge problem is that, as a research community, we need to answer the following questions:

1. Is progress being made in gait recognition of humans?
2. To what extent does gait offer potential as an identifying biometric?
3. What factors affect gait recognition and to what extent?
4. What are the critical vision components affecting gait recognition from video?
5. What are the strengths and weaknesses of different gait recognition algorithms?

The HumanID Gait Challenge Problem has advanced gait recognition by providing a foundational framework to address these issues. It includes a development data set, a set of 12 experiments, and a baseline algorithm. The baseline algorithm provides a performance benchmark and an initial characterization of automatic gait recognition. The 12 experiments examine the effects of five covariates or factors on gait recognition performance. They provide the foundation to advance automatic gait recognition, to provide an understanding of the critical components in a gait recognition algorithm, and to explain why they are critical. By reporting results on the same experiments, we have quantified improvement in performance of gait recognition algorithms [24], [26], [27], [30], [31], [32], [33]. These results are reported on a previous smaller version of the HumanID Gait Challenge problem [34]. With the full gait challenge problem, it is now possible to quantify the improvement in performance on a large and more detailed set of experiments.

In the short time span of two years, the Gait Challenge problem has already helped guide the evolution of gait algorithms. When the problem was first introduced in 2002, numerous existing algorithms performed perfectly on existing data sets, but performed worse than the baseline algorithm on the Gait Challenge problems. By the end of 2003, researchers had developed algorithms that performed better than the baseline algorithm. We expect this trend to continue. As we shall show, there is adequate room for performance improvement. Apart from spurring the development of better algorithms, it should be possible, as the number of papers reporting performance on the challenge problem increases, to perform analysis on these results. This meta-analysis should make it possible to gain insight and understanding into the critical components for gait recognition and why these components are critical. Such knowledge should help to direct research to further improve gait recognition and processing algorithms.

The key to a successful challenge problem is the data set collected to support the problem. From the data set, a set of experiments are defined. The experiments influence the types of algorithms that will be developed. For the experiments to be effective at influencing the direction of gait research, the design of the experiments needs to solve the *three bears problem*; the experiments must be neither too hard nor too easy, but just right. If performance on the experiments is easily saturated, then the gait recognition community will not be challenged. If experiments are too hard, then it will not be possible to make progress on gait recognition. Ideally, the set of experiments should vary in difficulty, characterize where the gait recognition problem is solvable, and explore the factors that affect performance. A set of experiments cannot meet this ideal unless the appropriate set of data is collected.

The HumanID gait challenge problem data is collected outdoors. The choice of outdoor settings is based on the observations that 1) several indoor data sets are available, 2) nearly perfect gait recognition performances have been reported on indoor data sets, and 3) gait biometrics is most appropriate in outdoor at-a-distance settings, where other biometric sources are harder to acquire. The choice of outdoor setting also forces the development of computer vision algorithms at multiple levels; it does *not* support the divorced tackling of low-level and high-level issues on parallel tracks. Algorithms have to handle complications generated from a person's shadow from sunlight, moving background, and moving shadows due to cloud cover.

For each subject, the challenge gait data set captures gait variations due to five different covariates, which were chosen based on the hypothesis that they either effect gait or effect the extraction of gait features from images. Factors that can affect a persons gait in outdoor settings include surface type, shoe-wear type, and weight carried. Video data of gait is also dependent on the viewpoint. Gait of a person can vary over time. It is important to understand the ability of gait recognition in the presence of these variations. This set of five covariates was selected from a larger list that was arrived at based on discussions with HumanID researchers at CMU, Maryland, MIT, Southampton, and Georgia Tech about potentially important covariates for gait analysis. We, of course, had to choose a subset of the variates from this list based on logistical issues and collection feasibility. There are other possible covariates of interest such as the mood of a person, clothing, speed, and backpack, which were not controlled or exercised in this data set.

Two different conditions were chosen for each of these five covariates: 1) two camera angles, 2) two shoe types, 3) two surfaces (grass and concrete), 4) with and without carrying a briefcase, and 5) two different dates six months apart. We attempted to acquire a person's gait in all possible combinations of these five factors and so there are up to 32 sequences for some persons. The full data set consists of 1,870 sequences from 122 individuals. This data set is significantly larger than those used in present studies and is also unique in the number of covariates exercised. It is the only data set to include walking on a grass surface.

The second part of the challenge problem is the set of 12 challenge experiments of increasing difficulty, as defined by the performance of our baseline algorithm, which is described later. Each experiment consists of definitions of gallery (watch-list) and probe (input data) pairs that differ with respect to one or more covariates. The experiments examine the effect on performance of different camera angles, a change in surface, and the effect of gait sequences acquired months apart. The motivation for the design of the challenge experiments was to focus future developments on the hard aspects of gait recognition from video. Algorithms that can tackle the harder challenge experiments will stand out. It is hoped that future research and papers will provide the gait community with insight into why some factors have a greater effect on performance than others.

The third part of the gait challenge problem is a simple, but effective, baseline algorithm. The baseline algorithm is based on spatial-temporal correlation between silhouettes. Comparisons are made with the silhouettes to reduce the effects of clothing texture artifacts. The baseline algorithm provides performance benchmarks for the experiments. We find that the algorithm, although straightforward, performs quite well on some of the experiments and is quite competitive with the first generation of gait recognition algorithms.

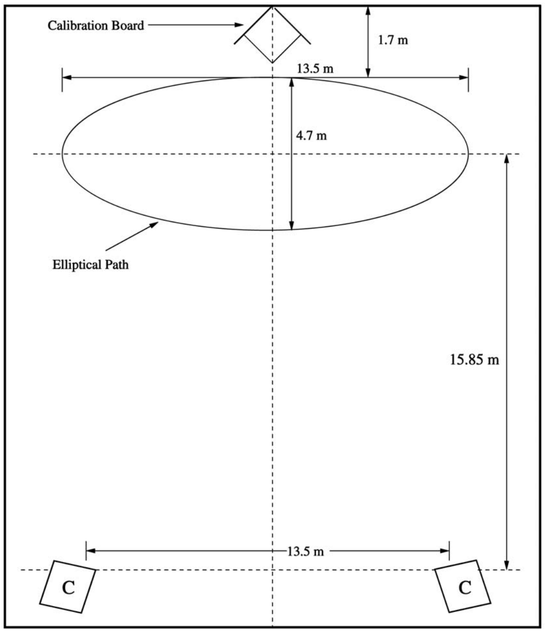


Fig. 1. Camera setup for the gait data acquisition.

The HumanID gait challenge problem touches on the following computer vision problems: matching and comparing temporal signatures, figure and background segmentation, modeling human motion and dynamics, and occlusion. Not all of these aspects are included in the baseline algorithm or will be included in every solution to the problem. However, improvements in performance over the baseline algorithm will touch upon some of these areas. The connection with the challenge problem could serve as the basis for developing and improving algorithms in these areas. In addition, the challenge problem can provide a means for measuring the impact of improvements in algorithms from these areas on a well-defined problem.

In addition to laying out the three above aspects of the challenge problem, this paper explores the questions: 1) Can we quantify the effects of walking surface, elapsed time between sequences, shoe type, viewpoint of camera, and carrying condition on gait recognition? Which condition(s) present the toughest problems? We look at these questions in Section 4. 2) How does the baseline performance change as gallery and probe sets are varied for the different challenge experiments? This is considered in Section 3. 3) What are the error modes of the baseline algorithm? Which subjects are the most difficult to recognize? Better algorithms can probably be designed by concentrating on these subjects and investigating the causes of failure. Section 5 considers these questions.

2 The Data Set

The HumanID gait challenge problem data set was designed to advance the state-of-the-art in automatic gait recognition and to characterize the effects on performance of five conditions. These two goals were achieved by collecting data on a large (122) set of subjects, compared to current standards in gait, spanning up to 32 different conditions, which is the result of all combinations of five covariates with two values each.

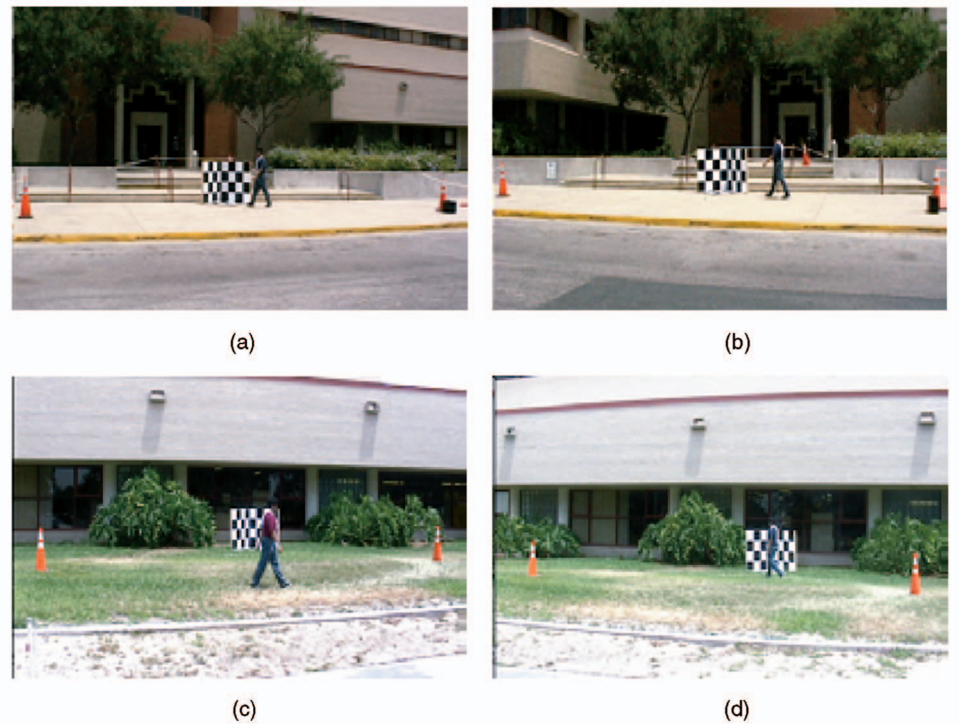


Fig 2. Frames from (a) the left camera for concrete surface, (b) the right camera for concrete surface, (c) the left camera for grass surface, and (d) the right camera for grass surface.

The gait video data was collected at the University of South Florida on May 20-21 and November 15-16, 2001. Participation in the collection process was voluntary. The collection process started with subjects being asked to read, understand, and sign an Institutional Review Board (IRB) approved consent form. The collection protocol had each subject walk multiple times counterclockwise around each of two similar sized and shaped elliptical courses. The basic setup is illustrated in Fig. 1. The elliptical courses were approximately 15 meters on the major axis and five meters on the minor axis. Both courses were outdoors. One course was laid out on a flat concrete walking surface. The other was laid out on a typical grass lawn surface. Each course was viewed by two cameras, whose lines of sight were not parallel, but verged at approximately 30 degrees, so that the whole ellipse was just visible from each of the two cameras. When a person walked along the rear portion of the ellipse, their view was approximately fronto-parallel. Fig. 2 shows one sample frame from each of the four cameras on the two surfaces. The orange traffic cones marked the major axes of the ellipses. The checkered object in the middle is a calibration object that can be used by future algorithms to calibrate the two cameras. We do not use it in this paper. Although data from one full elliptical circuit for each condition is available, we present the challenge experiments on the data from the rear portion of the ellipse. The motivations for the elliptical path are 1) to challenge the development of algorithms that are robust with respect to variations in the fronto-parallel assumption and 2) to provide a data sequence that includes all the views of a person, to help the future development of 3D model-based approaches or 3D visual hull-based approaches. The calibration object and the two views would also help such approaches.

The cameras were consumer-grade Canon Optura (for the concrete surface) and Optura PI (for the grass surface) cameras.[1] These are progressive-scan, single-CCD cameras capturing 30 frames per second with a shutter speed of 1/250 second and with autofocus left on, as all subjects were essentially at infinity. The cameras stream compressed digital video to DV tape at 25 Mbits per second by applying 4:1:1 chrominance subsampling and quantization and lossy intraframe adaptive quantization of DCT coefficients. The 4:1:1 subsampling results in some loss of color resolution, which can affect purely color-based (without luminance) background subtraction schemes. As we shall see later, we do observe some blocking effect in the computed silhouettes, which can be reduced by some smoothing.

The following metadata was collected on each subject: sex (75 percent male), age (19 to 59 years), height (1.47 m to 1.91 m), weight (43.1 kg to 122.6 kg), foot dominance (mostly right), type of shoes (sneakers, sandals, etc.), and heel height. We show the distribution of the number of subjects with respect to age, height, and weight in Fig. 3. Subjects were asked to bring a second pair of shoes so that they could walk the two ellipses a second time in a different pair of shoes. A little over half of the subjects walked in two different shoe types. In addition, subjects were also asked to walk the ellipses carrying a briefcase of known weight (approximately 6 kilograms). Most subjects walked both

[1] Commercial equipment is identified in this work in order to adequately specify or describe the subject matter. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the equipment identified is necessarily the best available for this purpose.

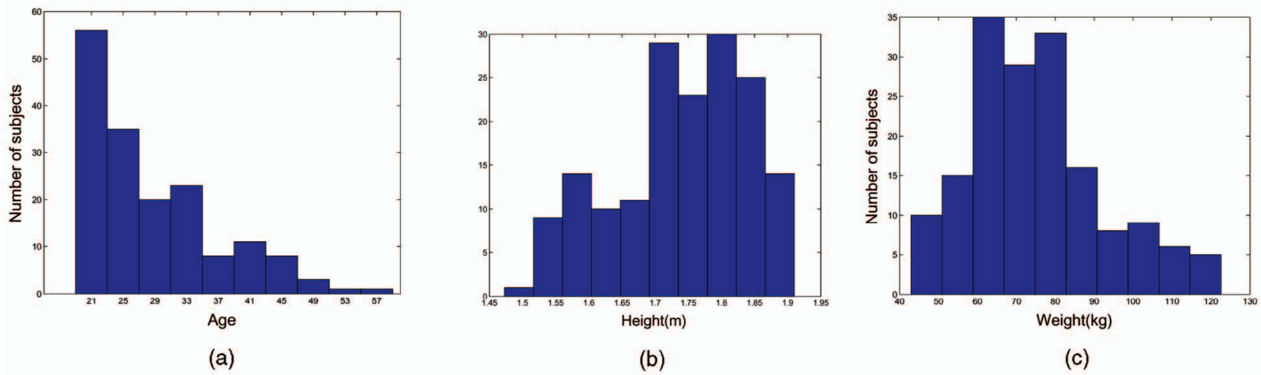


Fig. 3. Metadata statistics in terms of the distribution of the number of subjects with respect to (a) age, (b) height, and (c) weight.

carrying and not carrying the briefcase. In this paper, we denote the values of each of the covariates by the following:

1. Surface type by G for grass and C for concrete,
2. camera by R for right and L for left,
3. shoe type by A or B,
4. NB for not carrying a briefcase and BF for carrying a briefcase, and
5. the acquisition time, May and November, simply by M and N.

There are 33 subjects who were common between the May and November collections, so for them we also have data that exercises the time covariate. Table 2 shows the number of sequences for subjects who participated in the data collection for different covariate combinations.

The imagery was transferred offline from the camera DV tape to files on disc. The camera DV tape was accessed over an IEEE 1394 Firewire interface using Pinnacle's Micro DV 300 PC board. The result was a stand alone video file stored using Sony's (Digital Video) DV-specific "dvsd" codec in a Microsoft AVI wrapper. The transfer from tape to disc was lossless. Finally, the imagery was transcoded from DV to 24-bit RGB using the Sony decoder and the result was written as one 720×480 PPM file per frame. This representation trades off storage efficiency for ease of access. The final sequences contain each subject walking several laps of the course. For the gait data set, we saved frames from the last complete lap, which is from 600 to 700 frames in length. Please note that although the data set contains frames from one whole lap, the results in this paper are on frames from the back portion of the ellipse (see Fig. 1). A subject's size in the back portion of the ellipse is on average 100 pixels in height and 25 to 50 pixels in width.

Because we used two cameras for data acquisition, the data is subsequently synchronized by manually aligning the two sequences by inspection of action in successive frames. Given that the cameras do not accept an external trigger, this human-in-the-loop method gives synchronization to no better than 1/15 second. The data should support some level of stereo analysis, although we do not attempt that in this paper.

TABLE 2
Number of Sequences for Each Possible Combination

Surface	Carry	Shoe	Camera	Time	
				M or N	N
Concrete	NB	A	(L, R)	121	33
	NB	B	(L, R)	60	
	BF	A	(L, R)	121	
	BF	B	(L, R)	60	
Grass	NB	A	(L, R)	122	33
	NB	B	(L, R)	54	
	BF	A	(L, R)	120	
	BF	B	(L, R)	60	

Possible combinations for people who participated in the data collection include surface (G or C), shoe (A or B), camera view (L or R), and carry condition (BF, NB). The last column lists numbers of people who were in both data collections for two cases.

3 The Challenge Experiments

The second aspect of the challenge problem is a set of 12 challenge experiments. The 12 experiments are designed to investigate the effect of five factors on performance. The five factors are studied both individually and in combinations. The results of the baseline algorithm, described later, for the 12 experiments provide an ordering on the difficulty of the experiments.

We structured the challenge tasks in terms of gallery and probe sets, patterned on the FERET evaluations [36]. In biometrics nomenclature, the gallery is the set of people known to an algorithm or system and probes are signatures given to an algorithm to be recognized. In this paper, signatures are video sequences of gait.

To allow for a comparison among a set of experiments and limit the total number of experiments, we fixed one gallery as the control. Then, we created 12 probe sets to examine the effects of different covariates on performance. The gallery consists of sequences with the following covariates: Grass, Shoe Type A, Right Camera, No Briefcase, and collected in May along with those from the *new* subjects in November. This set was selected as the gallery because it was one of the largest for a given set of covariates. The structure of the 12 probe sets is listed in Table 3. The last two experiments study the impact of time. The time covariate implicitly includes a change of shoes and clothes because we did not require subjects to wear the same clothes or shoes in both data collections. We do have record of the shoe types that were used, but since subjects did not necessarily wear the same shoe six months later, the shoes did not match across time for all the subjects; for a subject, a "Shoe A" label in the May data does not necessarily refer to the same shoe as the "Shoe A" label in the November data.

TABLE 3
The Probe Set for Each of the Challenge Experiments

Exp.	Probe	# of	Difference
A	(G, A, L, NB, M/N)	122	View
B	(G, B, R, NB, M/N)	54	Shoe
C	(G, B, L, NB, M/N)	54	Shoe, View
D	(C, A, R, NB, M/N)	121	Surface
E	(C, B, R, NB, M/N)	60	Surface, Shoe
F	(C, A, L, NB, M/N)	121	Surface, View
G	(C, B, L, NB, M/N)	60	Surface, Shoe, View
H	(G, A, R, BF, M/N)	120	Briefcase
I	(G, B, R, BF, M/N)	60	Shoe, Briefcase
J	(G, A, L, BF, M/N)	120	View, Briefcase
K	(G, A/B, R, NB, N)	33	Time, Shoe, Clothing
L	(C, A/B, R, NB, N)	33	Surface, Time, Shoe, Clothing

The gallery set consists of 122 individuals. The probes are specified in terms of the conditions of the covariates: (Surface [C/G], Shoe [A/B], Camera [L/R], Carry [NB/BF], and Time [M/N]). The gallery for all of the experiments is (G, A, R, NB, M/N).

That is why, in Table 3, we use A/B for shoe type in experiments K and L. However, the shoe labels within the May data and within the November data are consistent.

4 The Baseline Algorithm

The third aspect of the challenge problem definition is a baseline algorithm against which future performance improvements can be measured. Ideally, this should be a combination of ″standard″ vision modules that accomplish the task. Drawing from the recent success of template-based recognition strategies in computer vision, we developed a four-part algorithm that relies on silhouette template matching. The first part semiautomatically defines bounding boxes around the moving person in each frame of a sequence. The second part extracts silhouettes from the bounding boxes. The third part computes the gait period from the silhouettes. The gait period is used to partition the sequences for spatial-temporal correlation. The fourth part performs spatial-temporal correlation to compute the similarity between two gait sequences. The baseline algorithm presented in this paper does not the specification of any parameters—it is parameter free.

Locating the bounding boxes in each frame is a semiautomatic procedure. In the manual step, the bounding box is outlined in the starting, middle, and ending frames of a sequence. The bounding boxes for the intermediate frames are linearly interpolated from these manual ones, using the upper-left and bottom-right corners of the boxes. This approximation strategy works well for cases where there is nearly fronto-parallel, constant velocity motion, which is true for the experiments reported here. Fig. 4 shows some examples of the image data inside the bounding box. The bounding boxes are conservatively specified and result in background pixels around the person in each box. These bounding boxes are part of the distributed data set.

4.1 Silhouette Extraction

The second step in the baseline algorithm is to extract the silhouette in the bounding boxes. Following common practice in gait recognition work, we define the silhouette to be the *region* of pixels from a person. Prior to extracting the silhouette, a background model of the scene is built. In the first pass through a sequence, we compute the background statistics of the RGB values at each image location, (x, y), using pixel values *outside* the manually defined bounding boxes in each frame. We compute the mean $\mu_B(x, y)$ and the covariances $\Sigma_B(x, y)$ of the RGB values at each pixel location. For pixels within the bounding box of each frame, we compute the Mahalanobis distance in RGB-space for the pixel value from the estimated mean background value. Based on the Mahalanobis distance, pixels are classified into foreground or background. In our earlier version of the baseline algorithm [34], this decision used a fixed, user-defined threshold. The present version adaptively decides on the foreground and background labels for each frame by estimating the foreground and background likelihood distributions using the iterative expectation maximization (EM) procedure. At each pixel, indexed by k, we have a two-class problem based on a scalar observation—the Mahalanobis distance, d_k. We model the observations as a two-class, $\{\text{Foreground} = \omega_1, \text{Background} = \omega_2\}$, Gaussian Mixture Model (GMM),

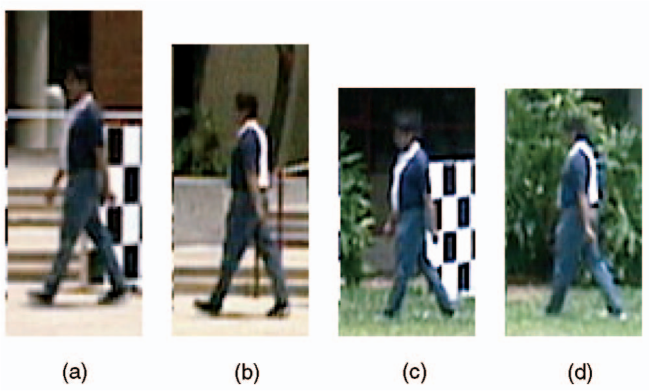


Fig. 4. Sample bounding boxed image data as viewed from (a) left camera on concrete, (b) right camera on concrete, (c) left camera on grass, and (d) right camera on grass.

$$P(d_k) = \sum_{i=1}^{2} P(\omega_i) p(d_k|\omega_i, \mu_i, \sigma_i),$$

where the class likelihood

$$p(d_k|\omega_i, \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(d_k-\mu_i)^2}{2\sigma_i^2}}.$$

For each pixel, we would like to estimate the posterior $P(\omega_1|d_k)$. We iteratively estimate this using the standard EM update equations reproduced below [37]. The estimates from different iterations are distinguished using the superscript:

$$\begin{aligned}
P^{(n+1)}(\omega_i) &= \frac{1}{N}\textstyle\sum_{k=1}^{N} P^{(n)}(\omega_i|d_k) \\
\mu_i^{(n+1)} &= \frac{\left(\sum_{k=1}^{N} P^{(n)}(\omega_i|d_k) d_k\right)}{\left(\sum_{k=1}^{N} P^{(n)}(\omega_i|d_k)\right)} \\
\sigma_i^{(n+1)} &= \frac{\left(\sum_{k=1}^{N} P^{(n)}(\omega_i|d_k)(d_k - \mu_i)^2\right)}{\left(\sum_{k=1}^{N} P^{(n)}(\omega_i|d_k)\right)} \qquad (1) \\
P^{(n+1)}(\omega_i|d_k) &= \frac{\left(p(d_k|\omega_i, \mu_i^{(n)}, \sigma_i^{(n)}) P^{(n)}(\omega_i)\right)}{\left(\sum_{i=1}^{2} p(d_k|\omega_i, \mu_i^{(n)}, \sigma_i^{(n)}) P^{(n)}(\omega_i)\right)}.
\end{aligned}$$

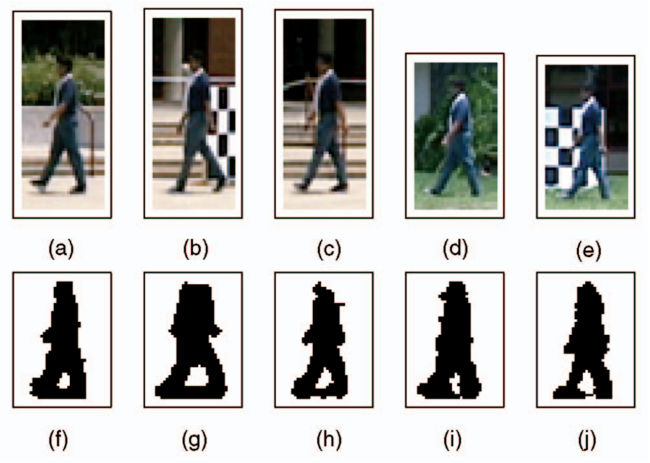


Fig. 5. The bottom row ((f)-(j)) shows sample silhouette frames with a variety of segmentation errors. The raw image corresponding to each silhouette is shown on the top row ((a)-(e)).

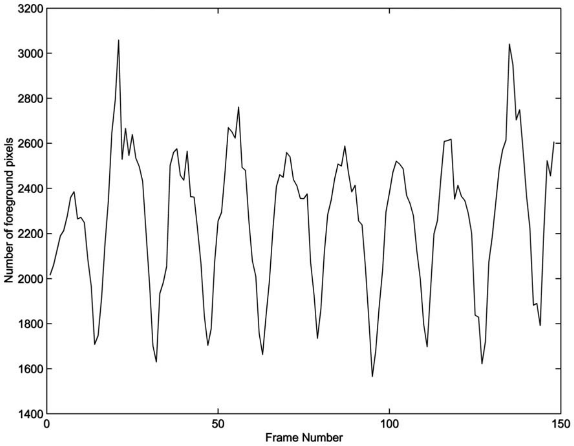


Fig. 6. Cue for gait period—the number of foreground pixels from the bottom half of the silhouettes.

The EM process is initialized by choosing class posterior labels based on the observed distance; the larger the Mahalanobis distance of a pixel, the greater is the initial posterior probability of being from the foreground.

$$\begin{aligned} P^{(0)}(\omega_1|d_k) &= \min(1.0, d_k/255) \\ P^{(0)}(\omega_2|d_k) &= 1 - P^{(0)}(\omega_1|d_k). \end{aligned} \tag{2}$$

We found that, with this initialization strategy, the process stabilizes fairly quickly, within 15 or so iterations.

It is worth mentioning a few words about pre and postprocessing steps that impact overall performance. We have found that if we smooth the computed Mahalanobis distance array (image) using a 9×9 pyramidal-shaped averaging filter or, equivalently, two passes of a 3×3 averaging filter, the visual quality of the silhouette and the recognition performance improves. This smoothing compensates for DV compression artifacts. The convergence of the EM process is faster with these smoothed distances than without, possibly due to a reduction in the noise of the computed Mahalanobis distances. There are two postprocessing steps on the silhouette image computed by EM. First, we eliminate isolated, small, noisy regions by keeping only the foreground region with the largest area. Second, we scale this foreground region so that its height is 128 pixels and occupies the whole height of the 128×88 pixel-sized output silhouette frame. The scaling of the silhouette offers some amount of scale invariance and facilitates the fast computation of a similarity measure. We also center the silhouette along the horizontal direction to compensate for errors in the placement of the bounding boxes. The silhouette is shifted in the horizontal direction so that the center column of the top portion of the silhouette is at column 44.

In most cases, the above strategy results in good quality silhouettes, but there are cases when it has problems. Fig. 5 shows some of these cases. Segmentation errors occur due to:

1. shadows, especially in the concrete sequences,
2. inability to segment parts because they fall just below the threshold and are classified as background,
3. moving objects in the background, such as the fluttering tape in the concrete sequences or moving leaves in the grass sequences or other moving persons in the background, and
4. lingering DV compression artifacts near the boundaries of the person.

There are many other possible scaling and centering options that might reduce the problems that we see in the current silhouettes. One option could be to take into account the entire sequence to decide upon the scaling parameters. However, such strategies would be dependent on the actual path taken by the subject. For instance, in our data set, as the person moves along the elliptical path, the distance of the person from the camera changes, which changes the projected image size. The strategy we use does not use, assume, or estimate the shape of the path taken by the subject. Of course, then our chosen frame by frame method might and does result in erroneous scaling when some part, such as the head, is not detected, but the employed matching strategy, which we shall see later, is resistant to some extent to such errors.

4.2 Gait Period Detection

The next step in the baseline algorithm is gait period detection. Gait periodicity, N_{gait}, is estimated by a simple strategy. We count the number of foreground pixels in the silhouette in each frame over time, $N_f(t)$. This number will reach a maximum when the two legs are farthest apart (full stride stance) and drop to a minimum when the legs overlap (heels together stance). To increase the sensitivity, we consider the number of foreground pixels mostly from the legs, which are selected simply by considering only the bottom half of the silhouette. Fig. 6 shows an instance of the variation of $N_f(t)$. Notice that two consecutive strides constitute a gait cycle. We compute the median of the distances between minima, skipping every other minimum. Using this strategy, we get two estimates of the gait cycle, depending on whether we skipped the first minimum or not. We estimate the gait period by the average of these two medians. Note that this strategy works for near fronto-parallel views, which is the view of choice for gait recognition and would not work for frontal views. However, the failure with respect to viewpoint variation is not drastic. The views in the present data set, on which we show the results, are not strictly fronto-parallel, but include up to 30 degrees variation.

4.3 Similarity Computation

The output from the gait recognition algorithm is a complete set of similarity scores between all gallery and probe gait sequences. Similarity scores are computed by spatial-temporal correlation. Let a probe sequence of M frames be denoted by $\mathbf{S_P} = \{\mathbf{S_P}(1), \cdots, \mathbf{S_P}(M)\}$ and a gallery sequence of N frames be denoted by $\mathbf{S_G} = \{\mathbf{S_G}(1), \cdots, \mathbf{S_G}(N)\}$. The final similarity score is constructed out of matches of disjoint portions of the probe with the gallery sequence. Specifically, we partition the probe sequence into disjoint subsequences of N_{gait} contiguous frames, where N_{gait} is the estimated period of the probe sequence from the previous step. Note that we do not constrain the starting frame of each partition to be from a particular stance. Let the $k\text{th}$ probe subsequence be denoted by $\mathbf{S_{Pk}} = \{\mathbf{S_P}(kN_{gait}), \cdots, \mathbf{S_P}((k+1)N_{gait})\}$. The gallery gait sequence $\mathbf{S_G} = \{\mathbf{S_G}(1), \cdots, \mathbf{S_G}(N)\}$ consists of all silhouettes extracted in the gallery sequence from the back portion of the elliptical path. Note that this gallery sequence is not partitioned. We then correlate each of the subsequences $\mathbf{S_{Pk}}$ with the entire gallery sequence $\mathbf{S_G}$.

There are three ingredients to the correlation computations: frame correlation, correlation between $\mathbf{S_{Pk}}$ and $\mathbf{S_G}$, and similarity between a probe sequence and a gallery sequence, comparing $\mathbf{S_P}$ and $\mathbf{S_G}$.

At the core of the above computation is, of course, the need to compute the similarity between two silhouette frames, $\mathrm{FrameSim}(\mathbf{S_P}(i), \mathbf{S_G}(j))$, which we simply compute to be the ratio of the number of pixels in their intersection to their union. This measure is also called the Tanimoto similarity measure, defined between two binary feature vectors [37]. Thus, if we denote the number of foreground pixels in silhouette $\mathbf{S}$ by $\mathrm{Num}(\mathbf{S})$, then we have,

$$\mathrm{FrameSim}(\mathbf{S_P}(i), \mathbf{S_G}(j)) = \frac{\mathrm{Num}(\mathbf{S_P}(i) \cap \mathbf{S_G}(j))}{\mathrm{Num}(\mathbf{S_P}(i) \cup \mathbf{S_G}(j))}. \tag{3}$$

Note that since the silhouettes have been prescaled and centered, we do not have to consider all possible translations and scales when computing the frame-to-frame similarity. The next step is to use frame similarities to compute the correlation between $\mathbf{S_{Pk}}$ and $\mathbf{S_G}$:

$$\mathrm{Corr}(\mathbf{S_{Pk}}, \mathbf{S_G})(l) = \sum_{j=0}^{N_{gait}-1} \mathrm{FrameSim}(\mathbf{S_P}(k+j), \mathbf{S_G}(l+j)). \tag{4}$$

For robustness, the similarity measure is chosen to be the median value of the maximum correlation of the gallery sequence with each of these probe subsequences. Other choices such as the average, minimum, or maximum did not result in better performance. The strategy for breaking up the probe sequence into subsequences allows us to address the case when we have segmentation errors in some contiguous sets of frames due to some background subtraction artifact or due to localized motion in the background.

$$\mathrm{Sim}(\mathbf{S_P}, \mathbf{S_G}) = \mathrm{Median}_k \left(\max_l \mathrm{Corr}(\mathbf{S_{Pk}}, \mathbf{S_G})(l) \right). \tag{5}$$

TABLE 4
Top Rank Identification Rates (Percentages) for CMU Mobo Data Set Reported by Different Algorithms

Exp ID[1]	1.1	3.4	3.1
Gallery	Slow (25)	Slow (25)	Slow (25)
Probe	Slow (25)	Ball (24)	Fast (25)
CMU [22]	100[1]	92	76
UMD [25], [21]	72		32
UMD [19]	72		12
Georgia Tech.		50	45[2]
MIT [20]	100	50	64[3]
Baseline	92	88	72

[1] As reported in http://www.hid.ri.cmu.edu/HidEval/evaluation.html
[2] As reported in http://www.cc.gatech.edu/cpl/projects/hid/CMUexpt.html
[3] As reported in http://www.ai.mit.edu/people/llee/HID/cmu_data_feat_sel.htm

The number of subjects in the gallery and probes are in parentheses.

5 Performance of Baseline Algorithm

The performance of the baseline algorithm on the challenge experiments establishes a "minimum" performance expected from any vision-based gait recognition algorithm. We show that our baseline algorithm is a reasonable choice by reporting its performance on the CMU Mobo data set [35]. The heart of this section is the baseline performance on all 12 challenge problem experiments. From the results on the 12 esperiments, we are able to rank the difficulty of the experiments. We demonstrate the effectiveness of challenge problems in advancing automatic gait recognition performance by reporting performance of algorithms on the challenge experiemtns. We identify the error modes of the baseline algorithm so that better algorithms can be designed by concentrating on these subjects and investigating the causes of failure.

5.1 Performance of the Baseline Algorithm on Mobo Data Set

Before we establish baseline performance for the challenge experiments, we benchmark the performance of the baseline algorithm on the CMU Mobo data set [35]. The CMU Mobo data is a commonly used data set for which performance has been reported in numerous papers. The CMU Mobo data set consists of sequences from 25 subjects walking on a treadmill positioned in the middle of the room. Each subject is recorded performing three different types of walking: slow walk (2.06 miles/hr), fast walk (2.82 miles/hr), and slow walk holding a ball. Each sequence is 11 seconds long and recorded at 30 frames per second. The data set allows experimenting with speed controlled gait recognition studies. Several papers have published results on this data set, hence, it a good external data set to benchmark the performance of the baseline algorithm. In computing performance scores, we used the silhouettes that were provided with the data set. Table 4 lists the reported identification rates for six algorithms on three commonly reported experiments. The last row lists the performance of the baseline algorithm. For all three experiments, the baseline performance is always the second highest score. Note that given the small data set size, a 4 percent difference in recognition rate represents one subject and, hence, is not statistically different.

TABLE 5
Baseline Performances for the Challenge Experiments in Terms of the Identification Rate P_I at Ranks 1 and 5 and the Verification Rate P_V at a False Alarm Rate of 1 Percent and 10 Percent of Unnormalized (UN), Z-Norm (ZN), and MAD-Norm (MAD)

Exp.	Difference	P_I (%) (at rank)		P_V (%) at P_F = 1%			P_V (%) at P_F = 10%		
		1	5	UN	MAD	ZN	UN	MAD	ZN
A	View	73	88	52	80	82	81	94	94
B	Shoe	78	93	48	80	87	82	94	94
C	Shoe, View	48	78	32	57	65	69	89	94
D	Surface	32	66	24	36	44	61	80	80
E	Surface, Shoe	22	55	16	33	35	52	76	76
F	Surface, View	17	42	10	22	20	45	59	60
G	Surface, Shoe, View	17	38	12	24	28	40	57	55
H	Briefcase	61	85	46	68	72	80	90	91
I	Briefcase, Shoe	57	78	48	60	67	76	85	85
J	Briefcase, View	36	62	22	45	48	64	75	76
K	Time, Shoe, Clothes	3	12	0	3	6	15	27	24
L	Surface, Time, Shoe, Clothes	3	15	0	3	6	18	27	24

All performance scores are in percent.

5.2 Base Results

The performance results for the 12 challenge experiments are reported in the following manner. We match each probe sequence to the gallery sequences, thus obtaining a similarity matrix with a size that is the number of probe sequences by the gallery size. Following the pattern of the FERET evaluations [36], we measure performance for both identification and verification scenarios using cumulative match characteristics (CMCs) and receiver operating characteristics (ROCs), respectively. In the identification scenario, the task is to identify a given probe to be one of the given gallery images. To quantify performance, we sort the gallery images based on computed similarities with the given probe. In terms of the similarity matrix, this would correspond to sorting the rows of the similarity matrix. If the correct gallery image corresponding to the given probe occurs within rank k in this sorted set, then we have a successful identification at rank k. A cumulative match characteristic plots these identification rates (P_I) against the rank k.

In the verification scenario, a system either rejects or accepts if a person is who they claim to be. Operationally, a person presents 1) a new signature, the probe, and 2) an identity claim. The system then compares the probe with the stored gallery sequence that corresponds to the claimed identity. The claim is accepted if the match between the probe and gallery is above an operating threshold, otherwise it is rejected. This decision is made solely on the similarity between a probe signature and the gallery signature that corresponds to the claimed identity, which is the usual practice, and is optimal only if the underlying distributions are not dependent on the probe. However, recent experiments with face recognition methods (FRVT 2002 [38]) showed that similarity score normalization can dramatically increase performance, possibly because it removes the dependencies of the nonmatch scores on the probe. This issue, however, needs a deeper theoretical look in future. Following FRVT 2002, instead of the raw similarity scores, we also report verification performance on gallery normalized similarity scores.

In *normalization*, a similarity score, $\mathrm{Sim}(P_i, G_j)$, between probe, P_i, and gallery signature, G_j, is adjusted by the statistics of the similarity scores between a probe and the full gallery set, $\{G_1, \cdots, G_N\}$. We present results for two normalization functions. The first is z-norm [38], which is

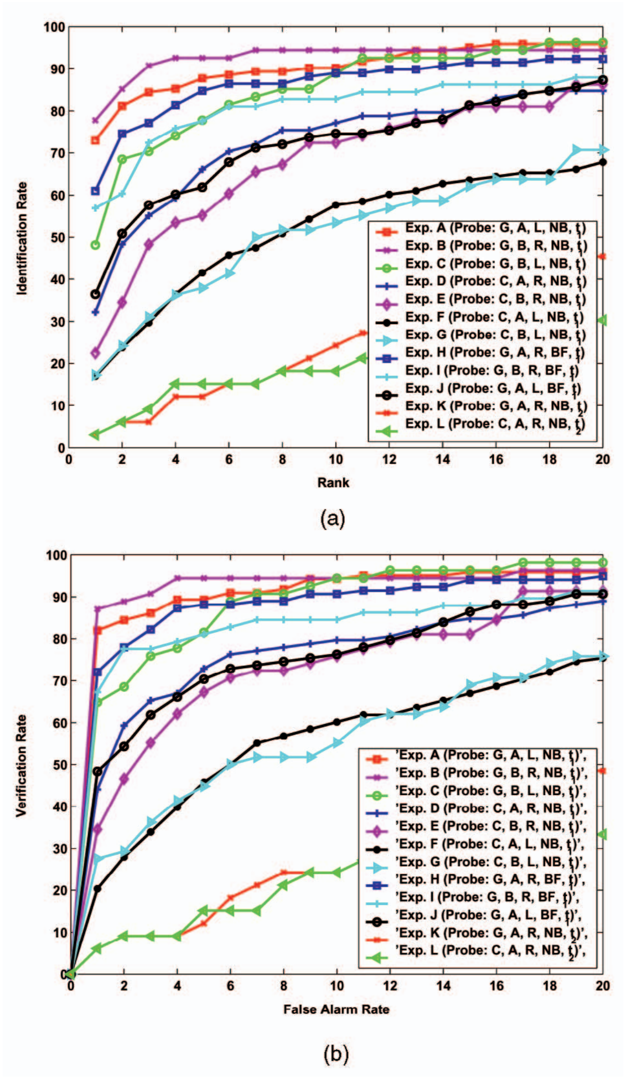
$$\mathrm{Sim_z}(P_i, G_j) = \frac{\mathrm{Sim}(P_i, G_j) - \mathrm{Mean}_j \mathrm{Sim}(P_i, G_j)}{\mathrm{s.d.}_j \mathrm{Sim}(P_i, G_j)}, \qquad (6)$$

where s.d. is standard deviation. For each probe, the normalized scores, most of which are nonmatch scores, except for the one correct match one, will have zero mean and unit standard deviations. The second is MAD-norm, which is

$$\mathrm{Sim_{MAD}}(P_i, G_j) = \frac{\mathrm{Sim}(P_i, G_j) - \mathrm{Median}_j \mathrm{Sim}(P_i, G_j)}{\mathrm{Median}_j |\mathrm{Sim}(P_i, G_j) - \mathrm{Median}_j \mathrm{Sim}(P_i, G_j)|}, \qquad (7)$$

where the denominator is the median of the absolute deviations (MAD) around the median values. The MAD-norm is a robust version of z-norm. For each probe, the MAD normalized scores will have zero first order and unit second order robust statistics. Given these normalized similarity scores, for a given operating threshold, there is a verification rate (or detection rate) and a false accept rate. Changing the operating threshold can change the verification and false accept rates. The complete set of verification and false accept rates is plotted on a receiver operating characteristic (ROC).

Table 5 summarizes the key performance indicators: the identification rate (P_I) at ranks 1 and 5 and the verification rate (P_V) for a false alarm rate of 1 percent and 10 percent. Verification rates are reported for unnormalized, z-normed, and MAD-normed similarity scores. Identification ranges from 3 percent to 78 percent at rank 1 and improves to a range of 12 percent to 93 percent at rank 5. The most striking feature of the verification results is the significant impact that normalization has on performance. At a false accept rate of 1 percent, the z-norm is superior to the MAD-norm and, at a false accept rate of 10 percent, both types of normalization are roughly equivalent. Because of the superiority of the z-norm at a false accept rate of 1 percent, all remaining verification results use the z-normalization procedure. With the z-norm, verification rates at a false accept rate of 1 percent range from 6 percent to 82 percent; at a false accept rate of 10 percent, verification rate ranges from



(a)

(b)

Fig. 7. Baseline performances for the challenge experiments (a) CMC curves and (b) ROCs plotted upto a false alarm rate of 20 percent.

24 percent to 94 percent. These are very encouraging performances, given the straightforward nature of the baseline algorithm. The range of results for the 12 experiments allows for improvement by new algorithms. Fig. 7 plots the CMCs and ROCs of the 12 challenge experiments.

Table 6 lists the identification rates that have been reported by other algorithms on an earlier, smaller (just May data, more than 71 subjects) release of the gait challenge data set. For comparison, we also list the performance of the baseline algorithm on the reduced data set. We see that 1) the ranked order of performance on the different experiments follows that for the baseline algorithm and 2) the performance of the baseline algorithm is competitive with respect to the other algorithms, especially on the hard problems. The performances reported in the table reflect performances published in papers at the end of 2003. The algorithms have evolved since then. Fig. 8 shows the *maximum* identification rates that are being achieved by 2004. Since these scores have not been yet published by the different groups, we report the scores anonymously. As evidence of how the gait challenge problem has already spurred the development of gait recognition algorithms, we also present the corresponding identification rates that were achieved in 2002 by the baseline algorithm and other algorithms. We see that the baseline algorithm has improved; it is now parameter free. We also see that the gait recognition algorithms have improved, however, experiments that compare across surfaces remain challenging.

We can rank the difficulty of the 12 experiments by their identification and verification rates, as reported by the baseline algorithm and corroborated by other algorithms. For instance, Experiment A, where the difference between probe and gallery is just the viewpoint, is easier than Experiment G, where the difference between the gallery and probe is three covariates. The rank of experiments allows for a ranking of the difficulty of the five covariates. From early reported results, this ranking also appears to be somewhat independent of the choice of the gait recognition algorithm, as we see in Table 6. The baseline algorithm-based rankings suggest that shoe type has the least impact, next is about 30 degrees viewpoint, the third is briefcase, then surface type, and time has the most impact, based on the drop in the identification rate due to each of these covariates. We quantify these effects next.

5.3 Impact of Variation in Gallery

The results presented so far are for one gallery set choice. It is well-known that changing the gallery and corresponding probe set changes the recognition rate [36], [38]. In this section, we examine the effect of changing the gallery and corresponding probe set and examine if the

TABLE 6
Reported Top Rank Recognition for Earlier, Smaller, Release of the Gait Challenge Data Set

Exp.	Width Vectors (UMD) [27]	DTW (UMD) [30]	HMM (UMD) [26]	Body Shape (CMU) [24]	HMM (MIT) [31]	Body (CAS) [32]	Baseline
A (view)	52	78	99	87	88	70	87
B (shoe)	40	65	89	81	75	59	81
C (view + shoe)	20	28	78	66	70	51	54
D (surface)	18	10	36	21	25	34	39
E (shoe+surface)	20	10	29	19	15	21	33
F (view+surface)	15	10	24	27	20	27	29
G (view+shoe+surface)	15	10	18	23	10	14	26
# subjects in gallery	71	71	71	71	71	71	71

The numbers for the first two columns are as read from graphs in the cited papers.

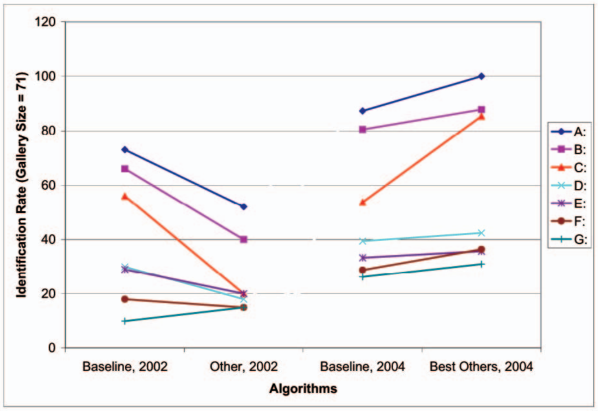


Fig. 8. Improvement in gait recognition algorithms over time with respect to the baseline performance.

order of experiments, based on the baseline recognition rates, changes.

The challenge experiments presented so far use the set (G,A,R,NB,M, or N) as the gallery. To examine the effect of gallery variation, we reran the 12 challenge experiments with different galleries and appropriately modified probe sets. In the challenge experiments, Experiment A examined the effect of change in view. To maintain consistency, the corresponding probe set A for each gallery is a change in view. For example, if the gallery is (C,A,L,NB,M, or N), then the probe set for experiment A should be (C,A,R,NB,M, or N), and so on. We vary the gallery to be one of the following eight cases: (G,A,R), (G,A,L), (G,B,R), (G,B,L), (C,A,R), (C,B,R), (C,A,L), and (C,B,L), with all the remaining two conditions, i.e., Carry and Time, fixed at NB, M, or N. Table 7 summarizes the verification rates at a false alarm rate of 1 percent for the challenge experiments. The first column lists the eight galleries and the remaining columns report recognition rates for changing different covariates. For example, the column labeled Surface + Shoe reports experimental results when the gallery and probe set have different surface and shoe types. The remaining covariates are the same between the gallery and probe set. The performance scores establish bounds on the verification rates for each experiment. The mean and the median score for each experiment provide a proxy for the difficulty level for each experiment. The standard deviation (s.d.) provides a measure of the stability of a covariate. The camera angle or view covariate has the greatest variability in terms of performance.

It is interesting to note that the ordering of the experiments in terms of their difficulty level, as measured by the verification rates, is somewhat invariant to the choice of the gallery set. To quantify the statistical correlation among the ranking of the experiments for the different gallery variations, we use the Friedman test, which is a two-way analysis of performance scores of the n gallery variations for the k experiments. The null hypothesis is that the ratings for the gallery variations are not related. For the data in Table 7, the computed underlying test parameter, which is the Kendall's coefficient of concordance, is found to be 0.96; the maximum correlation being one. The P-value is found to be < 0.0001, which implies that the null hypothesis can be easily rejected. Rejection of the null hypothesis implies that the verification rates for the experiment are different *and* the rates for the different gallery variations are strongly correlated.

The Friedman test does not provide us with a statistical ranking between the experiments, it just tells us if there is one. To rank the experiments, particularly the ones where only one covariate is varied, we use the pairwise Wilcoxon signed rank test [39]. It computes the statistical significance of the null hypothesis that medians of two distributions are equal. Based on this test, along with modified Bonferroni corrections [40] to account for multiple comparisons, for an overall $\alpha = 0.05$ (95 percent significance), we arrive at the following difficulty ranking: (ExpB–Shoe, ExpA–View) $\geq$ (ExpA–View, ExpH–Briefcase) $>$ ExpD–Surface $>$ ExpK–Time.

TABLE 7
Verification Performance Variation at P_F = 1 Percent of Baseline Algorithm Due to Variations in Gallery Type over Eight Possible Combinations

	Experiments											
	View	Shoe	View + Shoe	Surface	Surface + Shoe	Surface + View	Surface + View + Shoe	Briefcase	Briefcase + Shoe	Briefcase + View	Time	Time + Surface
Gallery	A	B	C	D	E	F	G	H	I	J	K	L
(G,A,R)	82	87	65	44	35	20	28	72	67	48	6	6
(G,A,L)	76	82	59	44	35	25	10	75	62	40	3	6
(C,A,R)	54	86	44	32	16	20	14	75	57	24	6	3
(C,A,L)	63	88	49	37	28	17	20	72	59	29	3	6
(G,B,R)	91	82	61	34	24	18	12	69	56	48		
(G,B,L)	89	87	54	34	31	20	20	69	60	46		
(C,B,R)	68	92	41	28	28	26	22	78	67	29		
(C,B,L)	73	83	53	34	35	16	17	67	56	35		
Mean	75	85	53	36	29	20	18	72	61	37	5	5
Median	75	87	54	34	30	20	19	72	60	38	5	6
s.d.	12.8	3.4	8.3	5.6	6.6	3.6	5.9	3.7	4.5	9.5	1.6	1.4

The fixed condition over them being no briefcase and the nonrepeat, i.e., NB, M, or N.

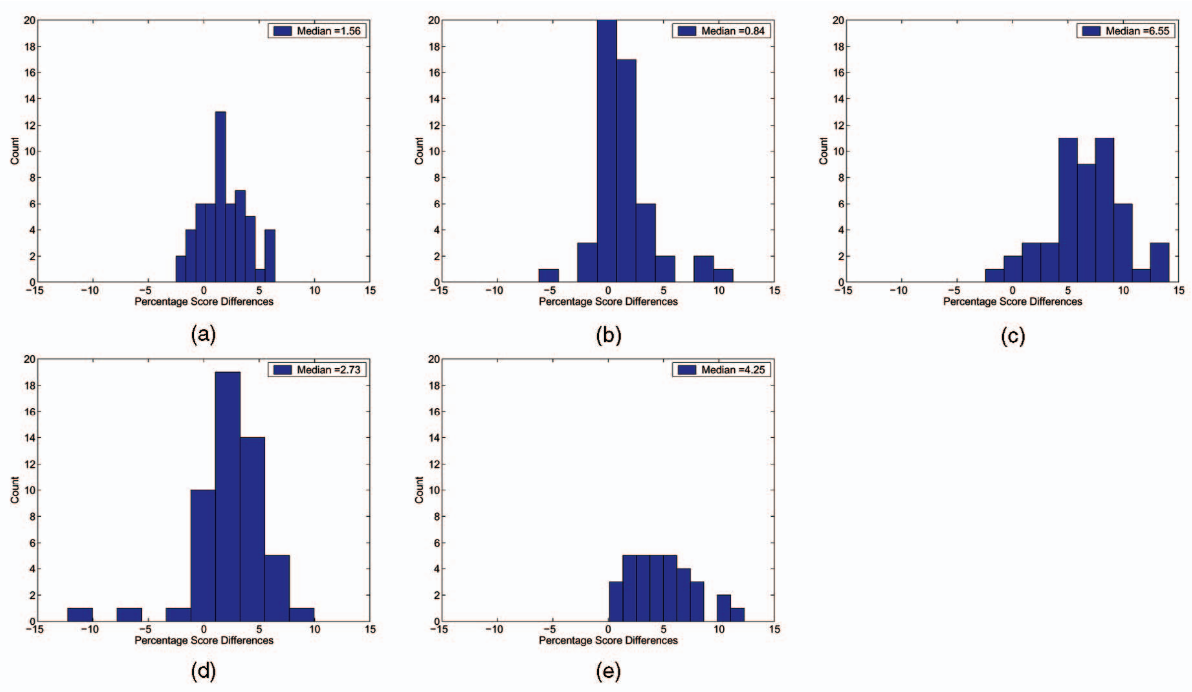


Fig. 9. The distribution of the percentage change in similarity values, $\Delta\mathrm{Sim}_{12}(i)$, between two probes differing with respect to (a) view point, (b) shoe type, (c) surface type, (d) briefcase, and (e) time.

5.4 Covariate Effects

Which covariate has the most impact on recognition? From the baseline recognition results, it appears that time has the most impact as the recognition rates, for Experiments K and L are the lowest. However, using recognition rates as indicators of covariate impact has problems and is at best a gross measure of impact. The recognition rate is a function of both the match and the nonmatch score distributions. This rate can change due to changes in either the match scores, nonmatch scores, or both. This is problematic since the nonmatch scores are a function of identity differences *and* any covariate difference that is present between the gallery and probes. The effect of a covariate is more cleanly captured by its impact on just the match scores.

We quantify the effect of a covariate on recognition by comparing the match scores for two probe sets, over the same set of individuals, that differ with respect to a specific covariate, but are similar in all other aspects. Therefore, if we want to study the effect of viewpoint on performance, for instance, we can consider the probes in Experiments B and C, which differ with respect to just viewpoint. For shoe type, we use the probes for Experiments A and C; for surface we use the probes for Experiments B and E; for briefcase we use the probes in Experiments B and I; and for time, we use the probe in Experiment A and the probe specified by (G, A/B, L, NB, N).

Let a similarity score for the ith subject in two choices of the probe sets, Probe 1 and Probe 2, be $\mathrm{Sim}_1(\mathbf{S}_{\mathbf{P}_i}, \mathbf{S}_{\mathbf{G}_i})$ and $\mathrm{Sim}_2(\mathbf{S}_{\mathbf{P}_i}, \mathbf{S}_{\mathbf{G}_i})$, respectively. The change in similarity for subject i, given by

$$\Delta\mathrm{Sim}_{12}(i) = \frac{\mathrm{Sim}_1(\mathbf{S}_{\mathbf{P}_i}, \mathbf{S}_{\mathbf{G}_i}) - \mathrm{Sim}_2(\mathbf{S}_{\mathbf{P}_i}, \mathbf{S}_{\mathbf{G}_i})}{\mathrm{Sim}_2(\mathbf{S}_{\mathbf{P}_i}, \mathbf{S}_{\mathbf{G}_i})},$$

quantifies the effect of a covariate on subject i. The distribution of these $\Delta\mathrm{Sim}_{12}(i)$ for all the subjects that are common between the probes and the gallery would provide an idea of the net effect of the covariate. If the distribution is centered around zero, this would signify no impact. If the drop is large, then we can infer that the distribution of the match scores, upon changing that covariate, would overlap more with the nonmatch scores, with consequent drop in recognition performance.

Fig. 9 shows the distribution of the score changes between probes differing with respect to view point, shoe type, surface type, briefcase, and time. Notice how the distribution shifts as we go from shoe type to viewpoint to briefcase to time to surface type differences. The median percentage increases in similarity scores for shoe, viewpoint, briefcase, time, and surface are 0.84, 1.56, 2.73, 4.25, and 6.55, respectively. The Wilcoxon signed rank test [39] can be used to compute statistical significance of the null hypothesis that the population median of the score changes is 0. It is a nonparametric test that takes into account the magnitude as well as the rank and is more sensitive than the Sign-Test or the Student t-test, especially for small numbers. Using this test, we find that we can easily reject the null hypothesis that the population median of the score changes for each covariate is 0 (with $\mathrm{P\text{-}values} < 0.001$), i.e., the score changes for all the covariates are significantly different from zero.

We can also compute the statistical significance for the ordering of the covariate impact ranking by performing pairwise Wilcoxon signed rank test. However, we have to be careful to take into account the multiple comparisons; in general, the individual pairwise comparisons must be performed at a tighter significance level than the desired overall significance level. We use the modified Bonferroni significance level-based testing of the individual pairwise testing [40]. The individual comparisons, of which we had 10, were rank ordered from most to least significant. So as to

TABLE 8
Modified Bonferroni Test for 10 Pairwise Tests of the Impact of the Covariates to Achieve an Overall Significance of 0.05

Factor Pairs	Surface Brief	Surface Shoe	Surface View	Shoe Brief	View Time	View Brief	Shoe Time	Brief Time	View Shoe	Surface Time
Wilcoxon P-value	0	0	0	0.0038	0.0068	0.0582	0.0674	0.0674	0.0992	0.2783
Modified-α	0.0055	0.0055	0.0062	0.0071	0.0083	0.0100	0.0125	0.0166	0.0250	0.0500
Reject Null Hypo	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No

achieve an overall significance level of 0.05, for the kth rank we use a cutoff of $\alpha/(10 - k + 1)$. Table 8 lists which of the pairwise null hypotheses we can reject. Based on the results, statistically speaking, the score changes due to shoe, view, briefcase, and time are similar, whereas the scores changes due to time and surface are similar. Thus, (view, shoe, briefcase, time) $\leq$ (time, surface).

The pairwise statistical tests in Table 8 clearly suggest that the impacts due to change in surface type and time are different from the impact of the other covariates. They seem to impact gait at a more fundamental level than other covariates. For example, we have found that the surface and time covariates impact the gait period more than other covariates. Fig. 10 plots the histogram of the differences in gait period for the same subject across views, surface, shoe-type, time, and carrying conditions. If a covariate does not impact the gait period, then the histogram should be peaked around zero. However, we notice that for surface-type and time, the histogram spreads to large values, which points to significant differences in gait period. The histogram for the carrying condition (briefcase and no briefcase) has a peak to the left of that for the surface-type.

5.5 Study of Failures

Is there a pattern to the failure in identification? Are there subjects who are difficult to recognize across all conditions? Is there an "easy to recognize" subset of subjects? Answers to these questions will help identify the hard sequences to work on in future. To answer such questions, we look at the pattern of failures in identification for each subject across different experiments. We partition the data set into subsets of subjects who are easy, moderate, and hard to identify based on the percentage of experiments in which a subject was correctly identified. Note that we considered percentages instead of absolute numbers since all subjects did not participate in all experiments. We consider a subject easy to identify if the subject was identified in more than 80 percent of the experiments that he/she participated in; in our data set there are 12 such subjects. We consider a subject hard to identify if the subject is correctly identified in less than 40 percent of the experiments; there are 56 subjects in this category. The rest of the subjects are considered moderately difficult to recognize; there are 54 subjects in this category. Fig. 11 shows some samples from each class. It is not obvious to us from visually observing the images or the associated silhouettes the reason why some subjects are hard to recognize. There are bad quality silhouettes, e.g., with missing head regions or missing leg regions, in all of the classes of subjects. Clothing or shadows also do not seem to play a role. However, to rule out any of these on a firm basis, future in-depth statistical correlation studies will have to be conducted.

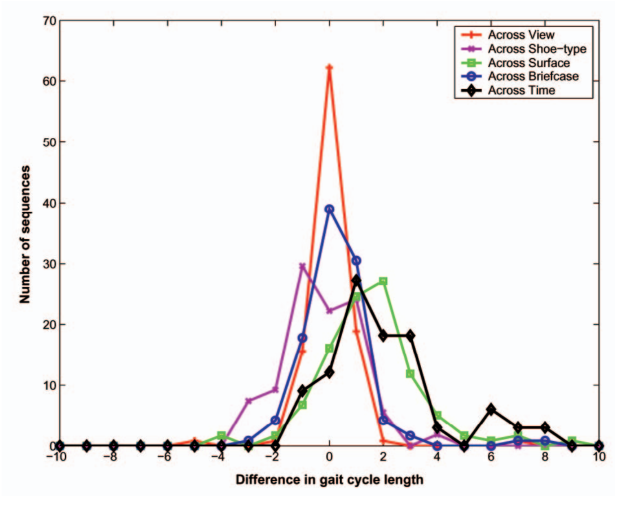


Fig. 10. Distribution of period differences across conditions.

6 Conclusions and Discussion

The HumanID gait challenge problem provides a set of 12 experiments of increasing difficulty. The 12 experiments examine the impact of five covariates on performance. The five covariates are camera angle, shoe type, grass or concrete surface, carrying or not carrying a briefcase, and time. The identification performance varies from 78 percent on the easiest experiment to as low as 3 percent on the hardest experiment. For verification, performance varies from 87 percent to 6 percent at a false alarm rate of 1 percent.

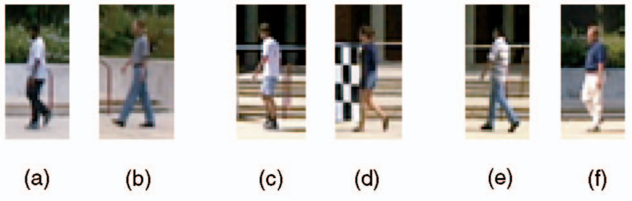


Fig. 11. Samples of subjects: (a) and (b) are easy to identify, (c) and (d) have moderate levels of identification difficulty, and (e) and (f) are hard to identify.

The results from the 12 experiments in Table 5 and Table 7 establish a baseline performance.

6.1 Significant Findings

We investigated two methods for normalizing similarity scores for verification performance. Overall, we found that performing normalization significantly increased performance, with the z-norm method being better than the MAD method. For performance on sequences taken on different days, the unnormalized verification rate at a false accept rate of 1 percent was zero and 6 percent after performing z-normalization (experiments K and L). For experiment B, change in shoe type, performance increased from 48 percent for unnormalized to 87 percent z-normalized similarity scores.

Focused analysis of the study of the impact of a covariate on match-score distribution suggests that shoe type has the least effect on performance, but the effect is nevertheless statistically significant. This is followed by either a change in camera view or carrying a brief case. Carrying a brief case does not affect performance as much as one might expect (Section 4). This effect is marginally larger than changing shoe type, but is substantially smaller than a change in surface type. In future experiments, it may be interesting to investigate the effect of carrying a backpack, rather than a briefcase, or to vary the object that is carried.

One of the factors that has a large impact is time, resulting in lower recognition rates for changes when matching sequences over time. This dependence on time has been reported by others too, but for indoor sequences and for less than six month differences. When the difference in time between gallery (the prestored template) and probe (the input data) is in the order of minutes, the identification performance ranges from 91 percent to 95 percent [18], [22], [23], whereas the performances drop to 30 percent to 45 percent when the differences are in the order of months and days [20], [22], [27] for similar sized data sets. Our speculation is that other changes that naturally occur between video acquisition sessions are very important. These include change in clothing worn by the subject, change in the outdoor lighting conditions, and inherent variation in gait over time. For applications that would require matching across days or months, these would most likely be the important variables. However, there are many applications, such as short term tracking across many surveillance cameras, for which these long term related variations would not be important.

The other factor with large impact on gait recognition is walking surface. With the subject walking on grass in the gallery sequence and on concrete in the probe sequence, rank 1 recognition is only 32 percent. Performance degradation might be even larger if we considered other surface types, such as sand or gravel, that might reasonably be encountered in some applications. The large effect of surface type on performance suggests that an important future research topic might be to investigate whether the change in gait with surface type is predictable. For example, given a description of gait from walking on concrete, is it possible to predict the gait description that would be obtained from walking on grass or sand? Alternatively, is there some other description of gait that is not as sensitive to change in surface type?

6.2 Gait versus Face

One of the open questions is the potential for gait to perform identification. We address this question by comparing our gait results with face recognition. Our analysis provides a rough guide to the current state of gait recognition. Face recognition performance has been well characterized by a number of evaluations, the most recent being the Face Recognition Vendor Test (FRVT) 2002 [38]. Because gallery size is different in the gait challenge problem and FRVT 2002, comparison is made for verification performance at a false accept rate of 1 percent. Unlike identification, verification performance is not a function of gallery size. Since the gait challenge problem performs recognition from outdoor video, we need to look at face recognition results from outdoor images. In FRVT 2002, there are two results on outdoor facial images. In both cases, the gallery is of indoor full frontal images. In the first result, the probe set consists of outdoor images taken on the same day as the gallery images. Verification performance varied for different systems ranging from 54 percent to 5 percent, with a median of 34 percent. From Table 5, gait performance varied from 87 percent to 20 percent on the 10 experiments where the gallery and probe set sequences were taken on the same day. The median performance score was 57 percent. In the second set of outdoor face recognition results, the probe set consists of outdoor images taken on a different day than the gallery image of a person; the median difference in time is about five months. Verification performance varied from 47 percent to 0 percent for different systems, with a median of 22 percent. Experiments K and L in the gait recognition problem, which have probes from six months later, are comparable to this scenario. The recognition rate for both experiments is 6 percent. A number of caveats need to be mentioned in this analysis. The FRVT 2002 performance numbers are from a blind evaluation on sequestered data. This is not the case for our gait results. On the other hand, the results in this paper are for a baseline algorithm at the beginning of intense research of automatic gait recognition. This compares to a decade of intensive development in automatic face recognition. Using the respective performances only as a rough guide, we see that video-based gait as an outdoor at-a-distance biometric has the potential to be 1) competitive with faces and 2) as a biometric to be fused with face.

6.3 The Greater Context

Human identification through analysis of gait information extracted from video is an important problem for computer vision. On the practical side, there are valuable potential applications in the area of video surveillance and security. Progress on gait recognition will aid progress on related problems such as characterizing human activity in video. General solutions to the gait problem will address fundamental computer vision problems that include segmentation and handling of occlusion. The process of solving this problem will identify which fundamental problems in computer vision and pattern recognition need further research. In turn, this problem will provide a method for measuring progress on the fundamental computer vision and pattern recognition problems.

The HumanID gait challenge problem provides for a scientific basis for advancing and understanding automatic gait recognition and processing. One aspect of this is that researchers wishing to work on a new algorithm will not have to invest the substantial start-up costs of acquiring a data set large enough to lend credibility to their results. Advancements in gait can be quantified by performance on the challenge experiments. The baseline algorithm makes it possible for researchers to focus on developing new techniques for one component of the baseline algorithm. The new component can be substituted for the baseline component and performance can be computed for the new component. This provides a measure of the effectiveness of the new component to the gait algorithm. As the number of researchers reporting performance results on the challenge problem increases, the potential to understand what are the critical components of gait algorithms work increases. The understanding increases because meta-analysis is possible on the different papers reporting challenge problem results. The more detailed the experimental results presented, the more detailed the possible meta-analysis is and the greater the understanding is. For example, if multiple research groups report results on different silhouettes, the greater the understanding of how silhouettes effect performance. It is this potential from the adoption of this challenge problem that represents a possible revolution in computer vision research methodology.

Acknowledgments

This research was supported by funds from the Defense Advanced Research Projects Agency (DARPA) HumanID program under contracts F49620-00-1-00388 (USF) and ONR N000140210410 (UND) and by the US National Science Foundation under Grant No. 0130768. The authors would like to thank Stan Janet and Karen Marshall from NIST for very meticulously assembing the data for distribution and creating the bounding boxes for the sequences. They would also like to thank the HumanID researchers at CMU, Maryland, MIT, Southampton, and Georgia Tech for discussion about potentially important covariates for gait analysis. They also thank Dr. Pat Flynn from the University of Notre Dame for testing the baseline algorithm code and scripts.

References

[1] M.W. Whittle, "Clinical Gait Analysis: A Review," *Human Movement Science,* vol. 15, pp. 369-387, June 1996.

[2] G. Johansson, "Visual Motion Perception," *Scientific Am.,* vol. 232, pp. 75-88, June 1976.

[3] J.E. Cutting and L.T. Kozlowski, "Recognition of Friends by Their Walk," *Bull. of the Psychonomic Soc.,* vol. 9, pp. 353-356, 1977.

[4] S.V. Stevenage, M.S. Nixon, and K. Vince, "Visual Analysis of Gait as a Cue to Identity," *Applied Cognitive Psychology,* vol. 13, pp. 513-526, Dec. 1999.

[5] B. Flinchbaugh and B. Chandrasekaran, "A Theory of Spatio-Temporal Aggregation for Vision," *Artificial Intelligence,* vol. 17, pp. 387-407, 1981.

[6] J.A. Webb and J.K. Aggarwal, "Structure from Motion of Rigid and Jointed Objects," *Artificial Intelligence,* vol. 19, pp. 107-130, 1982.

[7] A. Hilton and P. Fua, "Modeling People toward Vision-Based Understanding of a Person's Shape, Appearance, and Movement," *Computer Vision and Image Understanding,* vol. 81, pp. 227-230, Mar. 2001.

[8] J. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding,* vol. 73, pp. 428-440, Mar. 1999.

[9] D. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding,* vol. 73, pp. 82-98, Jan. 1999.

[10] T. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Computer Vision and Image Understanding,* vol. 81, pp. 231-268, Mar. 2001.

[11] *Motion-Based Recognition,* M. Shah and R. Jain, eds. Kluwer, 1997.

[12] S. Niyogi and E. Adelson, "Analyzing Gait with Spatiotemporal Surfaces," *Computer Vision and Pattern Recognition,* 1994.

[13] J. Little and J. Boyd, "Recognizing People by Their Gait: The Shape of Motion," *Videre,* vol. 1, no. 2, pp. 1-33, 1998.

[14] J. Shutler, M. Nixon, and C. Carter, "Statistical Gait Description via Temporal Moments," *Proc. Fourth IEEE Southwest Symp. Image Analysis and Interpretation,* pp. 291-295, 2000.

[15] A. Bobick and A. Johnson, "Gait Recognition Using Static, Activity-Specific Parameters," *Computer Vision and Pattern Recognition,* pp. I:423-430, 2001.

[16] R. Tanawongsuwan and A. Bobick, "Gait Recognition from Time-Normalized Joint-Angle Trajectories in the Walking Plane," *Computer Vision and Pattern Recognition,* pp. II:726-731, 2001.

[17] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated Face and Gait Recognition from Multiple Views," *Computer Vision and Pattern Recognition,* pp. I:439-446, 2001.

[18] J. Hayfron-Acquah, M. Nixon, and J. Carter, "Automatic Gait Recognition by Symmetry Analysis," *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication,* pp. 272-277, 2001.

[19] C. BenAbdelkader, R. Cutler, and L. Davis, "Motion-Based Recognition of People in Eigengait Space," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 267-272, 2002.

[20] L. Lee and W. Grimson, "Gait Analysis for Recognition and Classification," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 155-162, 2002.

[21] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-Based Recognition of Humans Using Continuous HMMs," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 336-341, 2002.

[22] R. Collins, R. Gross, and J. Shi, "Silhouette-Based Human Identification from Body Shape and Gait," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 366-371, 2002.

[23] L. Wang, W. Hu, and T. Tan, "A New Attempt to Gait-Based Human Identification," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 115-118, 2002.

[24] D. Tolliver and R. Collins, "Gait Shape Estimation for Identification," *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication,* 2003.

[25] A. Kale, N. Cuntoor, and R. Chellappa, "A Framework for Activity Specific Human Identification," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* 2002.

[26] A. Sunderesan, A.K.R. Chowdhury, and R. Chellappa, "A Hidden Markov Model Based Framework for Recognition of Humans from Gait Sequences," *Proc. IEEE Int'l Conf. Image Processing,* 2003.

[27] N. Cuntoor, A. Kale, and R. Chellappa, "Combining Multiple Evidences for Gait Recognition," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* 2003.

[28] R. Tanawongsuwan and A. Bobick, "Performance Analysis of Time-Distance Gait Parameters under Different Speeds," *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication,* 2003.

[29] A. Johnson and A. Bobick, "A Multi-View Method for Gait Recognition Using Static Body Parameters," *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication,* pp. 301-311, 2001.

[30] A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa, "Gait Analysis for Human Identification," *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication,* 2003.

[31] L. Lee, G. Dalley, and K. Tieu, "Learning Pedestrian Models for Silhouette Refinement," *Proc. Int'l Conf. Computer Vision,* 2003.

[32] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette Analysis-Based Gait Recognition for Human Identification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 12, pp. 1505-1518, Dec. 2003.

[33] I.R. Vega and S. Sarkar, "Representation of the Evolution of Feature Relationship Statistics: Human Gait-Based Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, pp. 1323-1328, Oct. 2003.

[34] P.J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm," *Proc. Int'l Conf. Pattern Recognition,* pp. 385-388, 2002.

[35] R. Gross and J. Shi, "The CMU Motion of Body MOBO Database," technical report, Carnegie Mellon Univ., 2001.

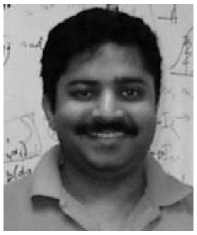
[36] P.J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

[37] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* Wiley, 2001.

[38] P.J. Phillips, D. Blackburn, M. Bone, P. Grother, R. Micheals, and E. Tabassi, "Face Recogntion Vendor Test," http://www.frvt.org, 2002.

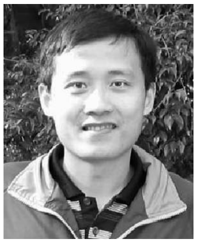
[39] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics,* vol. 1, pp. 80-83, 1945.

[40] J. Jaccard and C.K. Wan, *LISREL Approach to Interaction Effects in Multiple Regression.* Sage Publications, 1996.



Sudeep Sarkar received the BTech degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1988. He received the MS and PhD degrees in electrical engineering, on a University Presidential Fellowship, from The Ohio State University, Columbus, in 1990 and 1993, respectively. Since 1993, he has been with the Computer Science and Engineering Department at the University of South Florida, Tampa, where he is currently a professor. His research interests include perceptual organization in single images and multiple image sequences, biometrics, gait recognition, color-texture analysis, and performance evaluation of vision systems. He has coauthored one book and coedited another book on perceptual organization. He was the recipient of the US National Science Foundation CAREER award in 1994, the USF Teaching Incentive Program Award for undergraduate teaching excellence in 1997, the Outstanding Undergraduate Teaching Award in 1998, and the Theodore and Venette Askounes-Ashford Distinguished Scholar Award in 2004. He has served on the editorial boards for the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1999-2003) and the *Pattern Analysis & Applications Journal* (2000-2001). He is currently serving on the editorial board of the *Pattern Recognition Journal* and the *IEEE Transactions on Systems, Man, and Cybernetics, Part-B*. He is a member of the IEEE and the IEEE Computer Society.

P. Jonathon Phillips received the BS degree in mathematics and the MS degree in electronic and computer engineering from George Mason University, Virginia, and the PhD degree in operations research from Rutgers University, New Jersey. He is a leading technologist in the fields of computer vision, biometrics, face recognition, and human identification. Currently, he is at the National Institute of Standards and Technology (NIST) and has been assigned to the Defense Advanced Projects Agency (DARPA). Prior to joining NIST, he developed and designed the FERET database collection and FERET evaluations at the US Army Research Laboratory. He was awarded the Department of Commerce Gold Medal for his work as the test director for the Face Recognition Vendor Test (FRVT) in 2002. His work has been reported in print media of record including the New York Times and the Economist. He has organized two conferences and workshops on face recognition and three on empirical evaluation. He has coedited three books on face recognition and empirical evaluation. He has been guest editor of special issues or sections of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *Computer Vision and Image Understanding*. His current research interests include computer vision, face recognition, biometrics, digital video processing, developing methods for evaluating biometric algorithms, and computational psychophysics. He is a member of the IEEE and the IEEE Computer Society.



Zongyi Liu received the BS degree in business from Shenzhen University, Shenzhen, China, in 1997 and the MS degree in computer science and application from the University of Electronic Science and Technology of China in 2000. He is currently a PhD candidate at the University of South Florida. His research interests are computer vision-based gait biometrics, pattern recognition, motion, and image segmentation.

Isidro Robledo Vega received the BSc degree in industrial engineering in electronics in 1989 and the MSc degree in electronics engineering with a computer science option in 1996 from the Instituto Tecnológico de Chihuahua, Mexico, and the PhD degree in computer science and engineering from the University of South Florida, Tampa, in 2002. He is currently a professor of the division of posgraduate studies and research at the Instituto Tecnológico de Chihuahua, Mexico. His research interests include human motion analysis, computer vision, digital image processing, and artificial intelligence. He is a member of the IEEE and IEEE Computer Society.

Patrick Grother received the BSc degree in physics in 1988 and the MSc degree in computer engineering in 1990, both from Imperial College, London. As a staff member at the National Institute of Standards in Technology, he is responsible for the evaluation of biometric systems, for which he received the Department of Commerce Gold Medal in 2003. He is interested in pattern recognition, fusion, data mining, and image processing.

Kevin W. Bowyer received the PhD degree in computer science from Duke University, North Carolina. He is currently the Schubmehl-Prein Department Chair of the Department of Computer Science and Engineering at the University of Notre Dame. He was previously a member of the faculty at the Department of Computer Science and Engineering at the University of South Florida, the Institute for Informatics at the Swiss Federal Technical Institute (Zurich), and the Department of Computer Science at Duke University. He has served as editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, as North American Editor of the *Image and Vision Computing Journal*, and as chair of the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence. He received an Outstanding Undergraduate Teaching Award from the USF College of Engineering in 1991 and Teaching Incentive Program Awards in 1994 and 1997. He is author of the textbook *Ethics and Computing-Living Responsibly in a Computerized World* (IEEE press/ Wiley Press, 2001, second edition) and has conducted several NSF-sponsored faculty workshops on the theme of teaching ethics and computing. He is a fellow of the IEEE and the IEEE Computer Society. His photograph is an infrared image, chosen in accordance with this paper's biometrics subject matter.