Gait-based Person Identification using Motion Interchange Patterns

Gil Freidlin, Noga Levy, Lior Wolf

Tel-Aviv University, Israel

Abstract. Understanding human motion in unconstrained 2D videos has been a central theme in Computer Vision research for decades, and over the years many attempts have been made to design effective representations of video content. In this paper, we apply to gait recognition the Motion Interchange Patterns (MIP) framework, a 3D extension of the LBP descriptors to videos that was successfully employed in action recognition. This effective framework encodes motion by capturing local changes in motion directions. Our scheme does not rely on silhouettes commonly used in gait recognition, and benefits from the capability of MIP encoding to model real world videos. We empirically demonstrate the effectiveness of this modeling of human motion on several challenging gait recognition datasets.

Keywords: MIP, LBP, Gait Recognition, CASIA, TUMGAID

1 Introduction

Human gait is a valuable biometric characteristic describing the coordinated, cyclic movements of a walking person. Gait analysis is available where other biometrics cannot be measured, as gait can be recognized from a distance, does not require cooperation or even awareness of the subject, and works well on low resolution videos as recorded by standard surveillance cameras. The main challenge of gait recognition is the inherent large variability due to physical factors such as injuries or fatigue, carrying a load or wearing motion restrictive clothes.

Over the years many attempts have been made to design effective representations of video content. These range from high-level shape representations, to methods which consider low-level appearance and motion cues. In the task of Action recognition, the video representation aims to distinguish among human actions regardless of their performer. Interestingly, motion representations developed for action recognition and applied for gait recognition [13, 15, 5, 31, 9] demonstrate good perception within the same action (walking).

In this work, we adopt the Motion Interchange Patterns (MIP) [20] representation that was developed for action recognition applications. MIP encodes motion directly from video frames, and does not require preprocessing such as extracting the silhouette from the background or finding the cycles of the motion as other methods do. This rich local representation of human motion produces a discriminative signature of human cyclic gait motion. We suggest adaptations of the original MIP scheme to gait based identification.

2 Gait Recognition

Gait recognition approaches can be roughly divided into model-based and model-free categories. The model-based family of methods use knowledge about the body shape for the gait analysis. Model matching is performed in each frame in order to measure the physical gait parameters such as trajectories, limb length and angular speed.

Model-free techniques capture gait characteristics by analyzing the feature distribution over the space and time extent of the motion. These techniques often rely on extracting the human silhouette in every frame under the assumption that the interesting information about gait pattern lies in the body shape and contour. Popular methods such as the GEI variants [11] estimate the gait period and average the silhouettes over each cycle. Motion features are then computed either directly on the silhouette characteristics or by modeling the silhouette sequence using, for example, optical flow [24] or dynamic texture descriptors [22].

The human silhouette represents human body motions in a compact and efficient way but requires background subtraction, a challenging task for realistic backgrounds. Identification performance is sensitive to the silhouettes quality, hence silhouette-based methods are not well adjusted to unconstrained environment. Additionally, relying merely on silhouettes might miss out details containing significant motion information.

In a recent line of work, descriptors extracted directly from video frames, that were originally developed for action recognition, are applied to gait recognition. A few examples are LBP descriptors [16], HOG variants [13, 15, 5], histogram of 3D binary patterns [31] and dense trajectories [9] .

3 Action Recognition Descriptors

A central family of action recognition approaches uses low-level representation schemes of the information in a video. These approaches can be further categorized as local descriptors [25], optical flow based methods [1] and dynamic-texture representations [35].

Local descriptors [21, 33, 27] capture the locality of the human motion in time and space. As a first stage, pixels that are potentially significant to understand the scenario are detected and the region around them is represented by a local descriptor. To represent the entire video, these descriptors are processed and combined using, for example, a bag-of-words representation [26]. A major drawback of this approach is the sensitivity to the number of interest points detected. When a small number of interest points is detected, there is insufficient information for recognition. Videos with too much motion (such as waves or leaves in the wind) may provide a lot of information irrelevant for recognition.

The optical flow between successive frames [1, 30], sub-volumes of the video [18], or surrounding the central motion [7, 8] is highly valuable for Action Recognition. A drawback of optical flow methods is committing too soon to a particular motion estimate at each pixel. When these estimates are mistaken, they affect subsequent processing by providing incorrect information.

Dynamic-texture representations extend existing techniques for recognizing textures in 2D images to time-varying "dynamic textures" [19, 12]. One such technique

is Local Binary Patterns (LBP) [28], that extracts texture using local comparisons between a pixel and the pixels surrounding it, and encodes these relations as a short binary string. The frequencies of these binary strings are combined to represent the entire image region.

The Local Trinary Patterns (LTP) descriptor of [35] is an LBP extension to videos. An LTP code of a pixel is a trinary string that is computed by considering the relations among patches centered around the pixel in consecutive frames. A video is partitioned into a regular grid of non-overlapping cells and the histograms of the LTP codes in each cell are then concatenated to represent the entire video.

In this work, we adopt a dynamic-texture based representation, the Motion Interchange Patterns (MIP) [20], a leading video representation that was developed and evaluated on action recognition applications. This representation reflects the range of possible changes in motion and their likelihoods of occurring at each pixel in the video. Static edges are indicated by identifiable combinations of the MIP values, and may be ignored by subsequent processing. MIP codes also allow effective camera motion compensation, required in unconstrained videos.

4 Motion Interchange Patterns

Given an input video, the MIP encoding [20] assigns eight trinary strings consisting of eight digits each, to every pixel in every frame. A single digit compares the compatibility of one motion in a specific direction from the previous frame to the current frame, and one motion in another direction from the current frame to the next one. Figure 1 illustrates the motion structure extracted from comparing different patches.

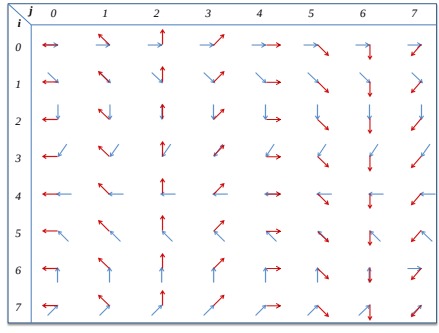


Fig. 1. Representation of motion comparisons of patches from three successive frames. For a given pixel and frame, blue arrows show the motion from a patch in the preceding frame and red arrows show the motion to a patch in the succeeding frame.

The code of a given pixel p in the current frame, denoted $S(p)$, is constructed by considering eight possible 3×3 patches around p in both preceding and successive

frames. Each digit in $S(p)$ refers to a pair of patches, one from the preceding frame and another from the following frame, out of 64 such pairs.

The sum of squared differences (SSD) patch-comparison operator is used to set the matching bit. Denote by SSD1 (SSD2) the sum of squared differences between the patch in the previous (next) frame and the patch in the current frame, as depicted in Figure 2. Each trit, $S_{i,j}(p)$, is computed as follows, for some threshold parameter θ:

$$S_{i,j}(p) = \begin{cases} 1 & if \quad SSD1 - \theta > SSD2 \\ 0 & if \quad |SSD2 - SSD1| \leq \theta \\ -1 & if \quad SSD1 < SSD2 - \theta \end{cases} \tag{1}$$

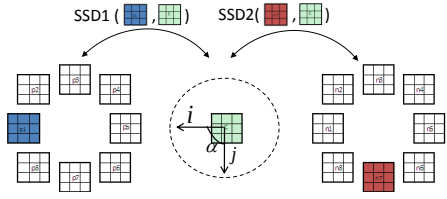


Fig. 2. Each trinary digit in the MIP encoding represents a comparison of two SSD scores, both referring to the same central patch (in green). SSD1 is computed between the central patch and a patch in the previous frame (in blue), and SSD2 is computed between the central patch and a patch in the next frame (in red).

A value of -1 indicates that the former motion is more likely and 1 indicates that the latter is more likely. The 0 value indicates that both are compatible in approximately the same degree or that there is no motion is this location. MIP compares all eight motions to the eight subsequent motions, obtaining a comprehensive characterization of the change in motion at each video pixel.

MIP Global Descriptor Denote by i and j the patch locations taken from the previous and following frames respectively, and let α be the angle between direction i and direction j out of the eight possible angle values. There are eight (i, j) pairs for each α, and the concatenation of their $S_{i,j}(p)$ values creates a trinary string. Each 8-trit string is separated into two binary strings, a positive string indicating the ones and a negative string indicating the minus ones, and translated into an integer in the range 0-255. Each pixel obtains 16 integer values, two values per α, that represent the complete motion interchange pattern for that pixel.

For each angle α, two histograms of size 256 are pooled (for the values taken from the positive and negative binary strings ,separately) from a 16×16 patch around each image pixel and concatenated, thus creating 512-dimensional MIP features. A dictionary containing 5000 code words is constructed using k-means on a random subset of MIP features (50000 in our experiments), taken from the encoded gallery set videos. Then, each local string is assigned to the closest word in the dictionary. Denote by u^{α} the histogram of the dictionary code words in the entire movie, normalized to the sum

of one and containing the square root of each element. The global descriptor of a video clip is a concatenation of the eight u^{α} histograms of all channels.

5 MIP-based Gait Recognition

Our baseline method employs MIP encoding on videos to find a motion signature of a walking person. We compute the MIP encoding for each video, and then use the local features to create a global descriptor for the whole video as described in section 4.

The MIP encoding is well adapted to gait recognition. The MIP descriptor is a normalized histogram of a bag-of-words of the patterns, hence contains pattern frequencies and does not require finding the gait cycles explicitly. We assume that each video contains at least one gait cycle. Moreover, significant motion patterns tend to repeat in each cycle while noise is random, and are therefore better represented in the histogram.

Another advantage is that MIP does not require silhouette extraction but rather works directly on the video frames. When MIP encoding is applied to moving silhouettes, the boundaries of the body motion are well encoded but other relevant details in the raw video are lost (e.g. the hand swing when passing over the body).

Designed for the action recognition task, MIP implicitly decodes all moving objects in the scene. Therefore, in a video clip containing a single walking person, MIP implicitly decodes the moving person without prior knowledge of the body location, while other methods require external human detection [15] or bounding box assignment. However, when the scene contains other consistently moving objects, their motion is encoded as well, hence narrowing down the area of interest might be needed.

We suggest two modifications of MIP adjusted for gait recognition - confounding details removal and temporal MIP.

Confounding Details Removal. MIP is an appearance-based method, hence, along with the action of interest, it encodes other details that can be misleading in the background or outfit. The standard MIP partly overcomes confusing information by downscaling the input images into a fixed size (100×134 in our experiments) before applying MIP. However, the degraded image quality affects the expressiveness of pose description that might be valuable for analyzing the motion, for example in the elbows region. Hence, after downscaling we upscale the frames to their original size by interpolation and compute MIP on the original size frames. We acquire a precise MIP encoding of moving body parts represented by significantly more features compared to MIP on the downscaled images, without being distracted by misleading details. This form of filtering is more suitable compared to conventional direct low-pass filtering on the original image, as it tends to remove textures while keeping depth boundaries without distorting the moving shape. By removing confounding patterns, the weight of the motion patterns relevant for gait identification is increased, thus improving the representation of the motion in the learned dictionaries.

As shown in Figure 3, the resulting encoding follows the moving body parts accurately.

Temporal MIP. The local motion pattern used in the standard MIP compares local motion in a three-sequential-frame scope, symmetrical in both preceding and successive

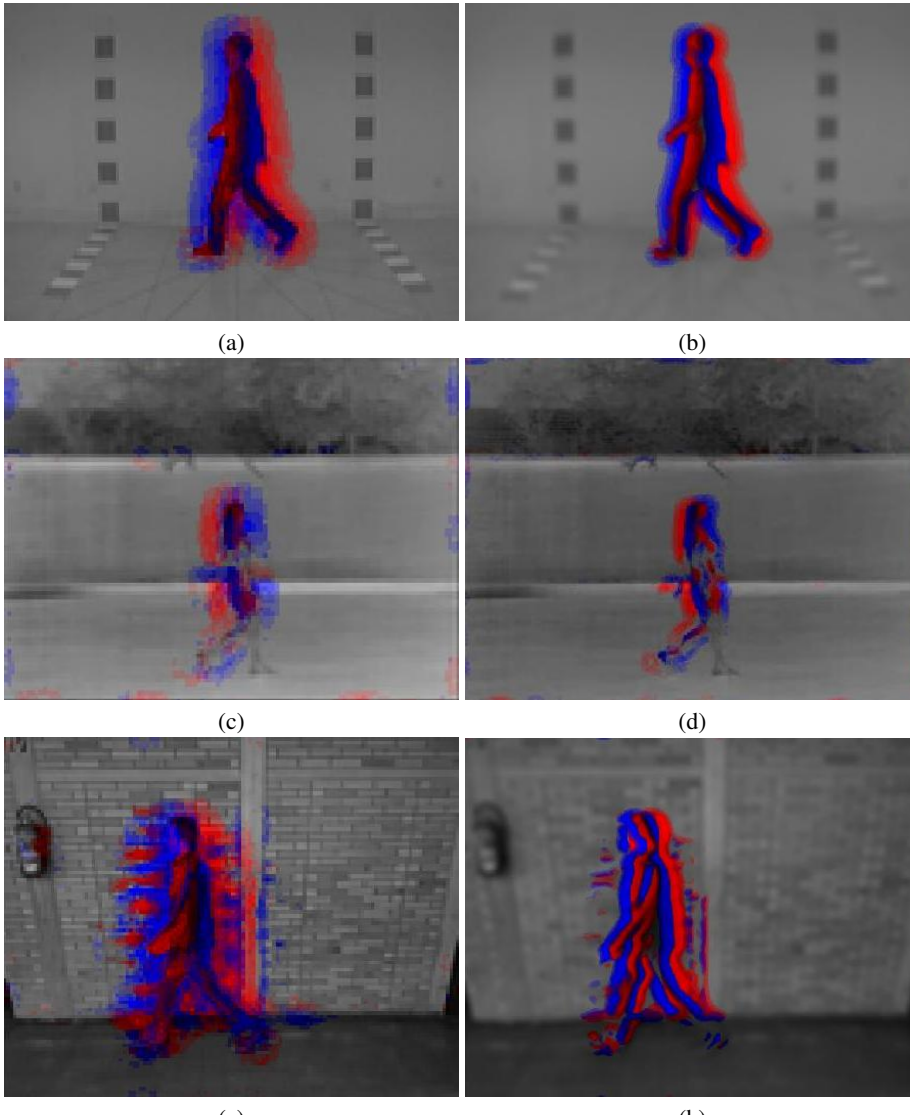


Fig. 3. MIP encoding. The first row contains images from CASIA-B, the second row contains images from CASIA-C, and the bottom row contains images from TUMGAID. In each row, the left image shows the standard MIP encoding and the right image shows MIP with confounding details removal. The encoding after details removal is sharpened and represent the moving human body in greater accuracy. The coded motions are illustrated by color coding pixels by their 8-trit strings content, for a specific α between the compared directions. Blue - motion from the previous frame to the current frame, red - motion from the current frame to the next frame. In image (e), the bricks shape within the shade is encoded, contributing misleading motion patterns. In image (h), details removal is applied and the shade is not encoded as a part of the moving object.

directions. The temporal MIP suggested here enlarges the temporal scope by considering temporal a-symmetric scopes of motion.

The MIP encoding described in section 4 is computed for a given frame t on frames $t-1$, t and $t+1$. The temporal MIP further encodes MIP on frames $t-2$, t and $t+1$ and on frames $t-1$, t and $t+2$, and illustrated in Figure 4. A normalized histogram is constructed separately for every α in each of these encodings. Finally, the global descriptor is a concatenation of all 24 histograms. According to our experiments, extending the temporal scope to the symmetric five frames encoding does not improve performance either by its own or when concatenated with the suggested encoding.

Fig. 4. Visualization of the Temporal MIP extension. Standard MIP encodes three successive frames, $t-1$, t and $t+1$ (solid arrows). Temporal MIP additionally encodes frames $t-1$, t and $t+2$ (dotted arrows), and frames $t-1$, t, and $t+2$ (dashed arrows). Frame t is emphasized in red.

Figure 5 describes the features extracted by the three MIP components of the temporal MIP on examples from CASIA-B and CASIA-C datasets, both on the downscaled frames and on the original size frames after details removal. The details removal variant is computed on the frames enlarged to their original size, thus produces significantly more features to describe the same action compared to standard MIP.

6 Classification

Given a gallery set, each image is represented by a global descriptor. These descriptors are used to train a multiclass linear SVM classifier. For N different subjects (class labels), N binary classifiers are obtained in the One-vs-All scheme. Prediction of a new example is performed by extracting its global descriptor, applying all binary classifiers and choosing the subject whose matching classifier gains the highest confidence score.

7 Experiments

We demonstrate our method on the CASIA-B and CASIA-C datasets and on the recently published TUM-GAID dataset. These datasets are challenging, containing various walking styles such as walking in different paces, walking while wearing a coat and carrying a bag or wearing restrictive shoes. Variation in the time of recording given in the TUM-GAID dataset are not tested here.

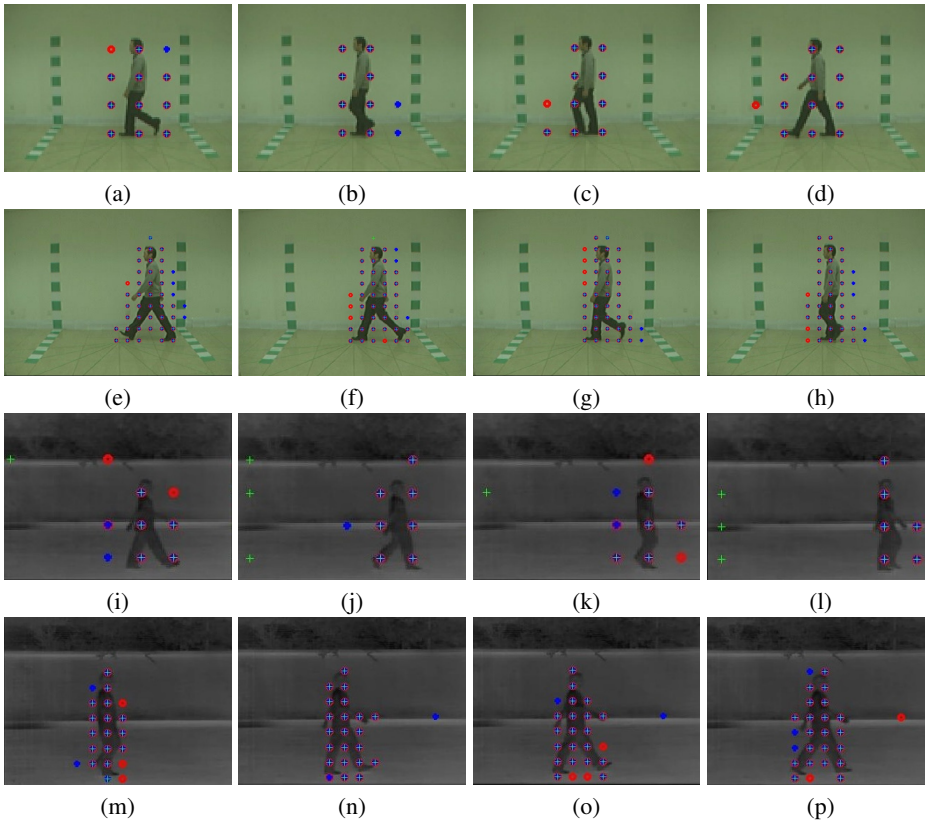


Fig. 5. Representation of the Temporal MIP local features on walking people from the CASIA datasets. Images (a)-(d) show temporal MIP features on a video taken from CASIA-B , (e)-(h) show the details removal variant on the same video. Images (i)-(l) show temporal MIP features on a video taken from CASIA-C, and (m)-(p) show the details removal variant. In the details removal variant, MIP is applied on the full sized frames and hence contains more features. Legend: green pluses - standard MIP features, blue stars - MIP features on frames $(t-2, t, t+1)$, red circles - MIP features on frames $(t-1, t, t+2)$. The features of all three encodings participating in the temporal MIP tend to occur in similar locations.

We test the performance of our method for standard MIP and temporal MIP representations, both with and without confounding details removal, and compare to the results reported by other methods on these datasets.

Performance is evaluated by the classification accuracy – the rate of correct identification by the first match. Experimentally, in most cases our method is comparable or superior to the other approaches, and the temporal MIP and confounding details removal adjustments usually outperform the vanilla MIP classification.

Table 1. The evaluation protocols for the CASIA-B dataset. Gallery and probe size represents the number of examples taken for each of the 124 subjects participating in the evaluation test. (a) first set of experiments, the protocol is defined in [38], (b) second set of experiments, the protocol is defined in [16]

Gallery	Probe
NN - first 4	NN - last 2
NN - first 4	BG - 2
NN - first 4	CL - 2

(a)

Gallery	Probe
NN - 5	NN -1
NN - 6	CL - 2
NN - 6	BG - 2
CL - 1	CL - 1
CL - 2	NN - 6
CL - 2	BG - 2
BG - 1	BG - 1
BG - 2	NN - 6
BG - 2	CL - 2

(b)

7.1 CASIA-B

The CASIA-B dataset [38] is a large multi-view gait database, containing 124 subjects captured from 11 views. For each subject, three walking styles are recorded - six video clips of normal walk (NN), two of carrying a bag (BG), and two of wearing a coat (CL). CASIA-B was recorded in a controlled indoor environment, with no textured outfits. Therefore, the performance of the details removal MIP in this case is equivalent to a direct encoding of the frames in their original resolution with no filtering applied.

In this work, only recordings captured from a lateral viewpoint are considered. The protocols used for testing are described in Table 1. The first set of experiments follows the evaluation protocol suggested in [38]. It uses as gallery the first four normal walk (NN) sequences per subject and three probe sets, one per each walking style. The second set of experiments follows the evaluation protocol in [16] and contains all gallery-probe combinations of walking styles.

Table 2 compares the performance on the first set of experiments. The results on the left refer to probe NN, and the results on the rigth refer to probes BG and CL. All compared methods except LBP-Flow [16] rely on silhouette extraction. Our method achieves good performance on the NN probe, and the details removal variants generalize well to the other walking styles, outperforming the other methods on the BG probe by $\sim 5\%$, and achieving the second best result on the CL probe.

Table 3 compares performance of standard MIP against LBP-Flow [16] for all combinations of walking styles per gallery and probe, following the evaluation protocol given in [16]. When the gallery and probe contain different walking styles, all existing sequences are used in both gallery and probe. When the gallery and probe share the same walking style, cross-validation is performed with one example per subject as the probe and the other examples in the gallery, and the average performance is reported. In all combinations, MIP variants outperform LBP-FLOW by a large gap.

Table 2. Comparison on CASIA-B dataset from a lateral viewpoint. The model is trained on normal walking and tested separately on each of the walking styles. Left - comparison of the performance on the normal (NN) style probe, right - comparison of the performance on carrying a bag (BG) and wearing a coat (CL). (*) The *Robust* method [17] is trained on three examples per subject and tested on the remaining examples, differently from the protocol defined in Table 1

Method	**NN**
MIP	95.96
Temporal MIP	96.37
MIP + Detail removal	98.79
Temporal MIP + Detail removal	99.19
LBP-FLOW [16]	94
HWLD [31]	**100**
GEI+ nn [38]	97.6
GEI + LDA [11] (results from [4])	83.1
PSC [23]	97.7
FDEI - Wavelet [4]	90.3
FDEI - Frieze [4]	91.1
IDTW [37]	83.5

Method	**BG**	**CL**
MIP	87.9	55.64
Temporal MIP	88.3	57.66
MIP + Details removal	**98.38**	83.87
Temporal MIP + Details Removal	97.98	77.82
LBP-FLOW [16]	45.2	42.9
HWLD [31]	92.2	**96.5**
GEI+ nn [38]	32.7	52.0
GFI Fusion [3]	83.6	48.8
Cross-view [2]	78.3	44.0
Robust(*) [17]	91.9	78.0
PRWGEI [36]	93.1	44.4

Table 3. Comparison on CASIA-B dataset of all combinations of gallery and probe against LBP-FLOW, following the protocol specified in Table 1(b). The number of examples in the gallery and probe indicated the number of examples for each subject out of 124 subjects

Gallery	**NN**			**BG**			**CL**		
Probe	**NN**	**BG**	**CL**	**NN**	**BG**	**CL**	**NN**	**BG**	**CL**
MIP	95.96	89.11	66.12	75	87.5	50.8	51.34	54.43	87.9
LBP-FLOW [16]	94	45.2	42.9	45.2	64.2	25	36.9	22.6	57.1

7.2 CASIA-C

The CASIA-C dataset [32] contains video of lateral view captured at night and recorded by a fixed low resolution infra-red camera. There are 153 subjects walking in four walking styles with 10 movies per subject: four movies for normal walking (fn), and two movies per each of the other walking styles – slow pace (fs), quick pace (fq) and carrying a bag (fb).

Table 4 summarizes the evaluation protocol used for CASIA-C dataset. In the experiments referring to gallery and probe that share the same walking style (within), the probe contains one example per subject and the other examples serve as the gallery. Each experiment is repeated with different probe examples for k times, where k is the number of examples per subject in the relevant walking style. We report the average accuracy on the k repetitions. In the experiments training on one walking style and evaluating on a different walking style (cross), all available sequences are used.

Table 5 shows the classification accuracy when training on normal walking and evaluating on all walking styles. The MIP variants outperform all compared methods, and the confounding details removal boosts performance on the bag carrying test set. Table 6 summaries the results when learning on the slow pace, quick pace and carrying

Table 4. The evaluation protocol for CASIA-C dataset: (a) the gallery and probe are from the same walking style, (b) cross style experiments. The number of examples per subject taken as gallery and as probe is specified, for each of the 153 subjects participating. CV stands for cross-validation

Gallery	Probe	Remarks
fn - 3	fn - 1	4-fold CV
fs - 1	fs - 1	2-fold CV
fq - 1	fq - 1	2-fold CV
fb - 1	fb - 1	2-fold CV

(a)

Gallery	Probe			
fn - 4		fs - 2	fq - 2	fb - 2
fs - 2	fn - 4		fq - 2	fb - 2
fq - 2	fn - 4	fs - 2		fb - 2
fb - 2	fn - 4	fs - 2	fq - 2	

(b)

Table 5. Results on CASIA-C dataset for a gallery containing normal walking style and evaluated on all probe sets. The first column refers to the normal walking probe. (*) The PSA results in [23] refer to a random subset of 50 subjects (out of 153 subjects)

Method	Within	Cross		
		fs	fq	fb
MIP	99.34	**95.09**	**98.69**	96.73
Temporal MIP	99.34	93.79	**98.69**	97.05
MIP+Details removal	99.34	92.15	98.36	99.02
Temporal MIP + Details Removal	99.34	92.16	**98.69**	**99.34**
WBP [22]	99.02	86.3	89.5	80.7
PSA(*) [23]	98	92	92	93
Gait curves [6]	91	65.4	69	25
Bag Of Gait [29]	**99.84**	91.23	95.78	89.82
Pseudo Shape [32]	98	82.4	91.8	24.4
GEI [38]	96	74	83	60
HTI [32]	94	85	88	51

a bag train sets, evaluated within the same walking style and on the other styles. MIP variants outperform the compared methods on most combinations.

7.3 TUM-GAID

The TUM-GAID [14] is a recently published dataset with 305 subjects, captured indoor from a lateral viewpoint. The movies were taken by a 3D-depth camera and provide matching audio. In this work we only use the 2D RGB images of the recorded subjects. For each subject, three walking styles are recorded - normal walking (N), carrying a backpack (B) and wearing coating shoes (S). A subset of 32 people is recorded again after a three months period in all walking styles (TN, TB, TS).

The evaluation protocol designed in [14] defines a test set containing 155 subjects. For recognition, the gallery consists of four normal walk recordings per each of the 155 subjects and the probe is divided into six test sets, for each walking style and recording phase. The experiments conducted here use the N, B and S probe sets. Table 7 shows the evaluation protocol used for those probe sets.

Table 6. Results on the CASIA-C dataset. The top two rows refer to the gallery and probe walking styles respectively. (*) The PSA results in [23] refer to a random subset of 50 subjects (out of 153 subjects).

	Within			**Cross**								
Gallery	**fs**	**fq**	**fb**	**fs**			**fq**			**fb**		
Probe	**fs**	**fq**	**fb**	**fn**	**fq**	**fb**	**fn**	**fs**	**fb**	**fn**	**fs**	**fq**
MIP	99	**99.34**	99	**93.13**	89.54	**88.23**	95.75	**84.31**	92.48	92.97	83.98	90.52
Temporal MIP	**99.34**	**99.34**	**99.34**	91.17	87.25	85.94	96.95	83.98	**94.44**	93.95	**86.93**	91.5
MIP + Details Removal	99	**99.34**	**99.34**	87.41	66.33	80.07	97.05	62.41	88.88	96.73	84.31	91.83
Temporal MIP + Details Removal	**99.34**	**99.34**	**99.34**	85.78	66.33	78.43	**97.22**	62.41	91.83	**97.22**	85.29	**93.46**
WBP [22]	95	96	96	88	61	71	84	61	71	81	70	80
PSA(*) [23]	98	96	96		**93**							
Gait curves [6]	85	79.1	81									

Table 7. Evaluation protocol for the N, B and S probe sets from the TUMGAID dataset as defined in [14]. The number of examples per subject taken as gallery and as probe is specified for each of the 155 subjects

Gallery	**Probe**
N - first 4	N - last 2
N - first 4	B - 2
N - first 4	S - 2

Table 8 compares our results to other methods. This comparison is challenging, as all methods apart from MIP and GEI [13] employ the depth information provided by the dataset.

MIP and MIP variants cope well with all walking styles. When normal walk is used for both training and testing, all presented methods show very good performance. The RSM method [10] achieves the best performance, utilizing the depth information to extract high quality silhouettes. When training on normal walk and testing on either (B) or the coating shoes probe (S), Mip and temporal MIP outperform all other methods. Temporal MIP gains the highest accuracy on the backpack carrying probe, while MIP wins temporal MIP by a small margin on the coating shoes probe (S).

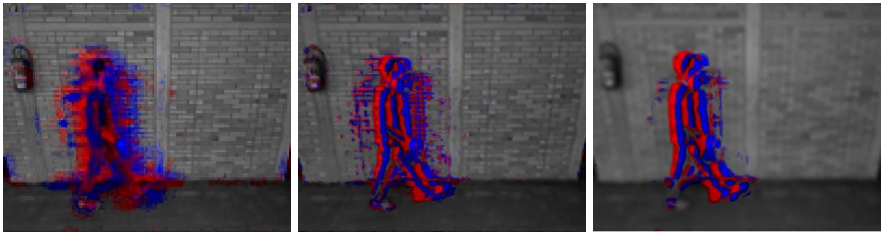
Although the TUM-GAID dataset is captured indoor, it contains a challenging background of a brick wall nearby the subjects. Due to the lighting conditions, the subjects cast shadows on the wall, which follow them and vary in shape and direction.

When applying MIP, the shadow is encoded along with the movement, as shown in Figure 6(a) and Figure 3(e). Hence, the shaded area contributes motion patterns to the MIP encoding. Since the background contains repetitive strong edges and colored bricks, the filtering in the details removal pre-process does not eliminate these undesirable patterns that clearly reflects the brick edges, as shown in Figure 6(b).

Elimination of these edges is done by applying a Gaussian filter (3×3, $\sigma = 1$) on each frame after downsampling, and then upscaling the frame to the original size.

Table 8. Results on the TUM-GAID dataset trained on normal walking and evaluated on three walking styles. N - normal walking, B - carrying a backpack and S - wearing coating shoes. All compared methods except for our method and GEI utilize depth information

Method	**N**	**B**	**S**
MIP	98.06	95.8	**97.42**
Temporal MIP	98.38	**97.42**	96.77
MIP + Details Removal	97.41	90.96	89.35
Temporal MIP + Details Removal	97.74	94.19	91.61
GEI (results from [13])	94.2	13.9	87.7
Depth-GHEI [13]	96.8	3.9	88.7
Depth-GEI [13]	99	40.3	96.1
GEV [13]	99.4	27.1	52.6
Unimodal RSM [10]	**100**	79	97
SVIM [34]	98.4	64.2	91.6



(a) (b) (c)

Fig. 6. Detail removal preprocessing for TUMGAID dataset. (a) Low resolution MIP encoding shows the shaded area is encoded, creating motion patterns caused by the shade and the patterned wall. (b) After applying detail removal preprocessing (downsampling then upsampling again to the original frame size) misleading motion patterns that reflects the bricks pattern are still exists in the current shaded area. (c) the result of the new preprocessing flow using a gaussian filtering to suppress the strong edges, now following mostly the moving body.

Figure 6(c) demonstrates the new encoding, which focuses on the moving body while avoiding the misleading wall and shadow patterns.

The standard MIP encoding performs better on this dataset over the details removal MIP encoding. The reason might be the information found in the shadow, that is coded when no details removal is applied. Since all scenes in this dataset were recorded in the same location, in similar conditions and from the same viewpoint, the information encoded in the shaded area might contribute to identification.

8 Summary and Conclusions

Most methods applied to gait recognition involve a preprocessing step of silhouette extraction, making them sensitive to the silhouettes quality and unstable in unconstrained environments.

In this work, we examine the the Motion Interchange Patterns, designed to directly represent motion in unconstrained 2D videos, on gait recognition datasets. Following our observations, we suggest two adaptations of MIP to the task of gait recognition – a temporal extension of the encoded motion, and confounding details removal that enables the analysis of the frames in their original size without getting lost in confounding details.

Employing MIP is a step towards motion analysis that is perceptive enough to identify people from a distance, in real world sequences and under various appearances.

Acknowledgments. Portions of the research in this paper use the CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. TPAMI 32(2), 288–303 (2010)
2. Bashir, K., Xiang, T., Gong, S.: Cross view gait recognition using correlation strength. In: BMVC. pp. 1–11 (2010)
3. Bashir, K., Xiang, T., Gong, S., Mary, Q.: Gait representation using flow fields. In: BMVC. pp. 1–11 (2009)
4. Chen, C., Liang, J., Zhao, H., Hu, H., Tian, J.: Frame difference energy image for gait recognition with incomplete silhouettes. Pattern Recognition Letters 30(11), 977–984 (2009)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
6. DeCann, B., Ross, A.: Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In: SPIE Defense, Security, and Sensing. pp. 76670Q–76670Q. International Society for Optics and Photonics (2010)
7. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 726–733 (2003)
8. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8 (2008)
9. Gong, W., Sapienza, M., Cuzzolin, F.: Fisher tensor decomposition for unconstrained gait recognition. Training 2, 3 (2013)
10. Guan, Y., Wei, X., Li, C.T., Marcialis, G.L., Roli, F., Tistarelli, M.: Combining gait and face for tackling the elapsed time challenges. In: Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on. pp. 1–8. IEEE (2013)
11. Han, J., Bhanu, B.: Individual recognition using gait energy image. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(2), 316–322 (2006)
12. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: Real-time detection of violent crowd behavior. In: CVPRW, 2012 IEEE Computer Society Conference on. pp. 1–6. IEEE (2012)
13. Hofmann, M., Bachmann, S., Rigoll, G.: 2.5 d gait biometrics using the depth gradient histogram energy image. In: Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on. pp. 399–403. IEEE (2012)
14. Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., Rigoll, G.: The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. Journal of Visual Communication and Image Representation 25(1), 195–206 (2014)

15. Hofmann, M., Rigoll, G.: Improved gait recognition using gradient histogram energy image. In: Image Processing (ICIP), 2012 19th IEEE International Conference on. pp. 1389–1392. IEEE (2012)
16. Hu, M., Wang, Y., Zhang, Z., Zhang, D., Little, J.J.: Incremental learning for video-based gait recognition with lbp flow. Cybernetics, IEEE Transactions on 43(1), 77–89 (2013)
17. Iwashita, Y., Uchino, K., Kurazume, R.: Gait-based person identification robust to changes in appearance. Sensors 13(6), 7884–7901 (2013)
18. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. vol. 1, pp. 166–173. IEEE (2005)
19. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. In: BMVC. vol. 1, p. 2 (2008)
20. Kliper-Gross, O., Gurovich, Y., Hassner, T., Wolf, L.: Motion interchange patterns for action recognition in unconstrained videos. In: Computer Vision–ECCV 2012, pp. 256–269 (2012)
21. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 2046–2053 (2010)
22. Kusakunniran, W., Wu, Q., Li, H., Zhang, J.: Automatic gait recognition using weighted binary pattern on video. In: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on. pp. 49–54 (2009)
23. Kusakunniran, W., Wu, Q., Zhang, J., Li, H.: Pairwise shape configuration-based psa for gait recognition under small viewing angle change. In: Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on. pp. 17–22. IEEE (2011)
24. Lam, T.H., Cheung, K.H., Liu, J.N.: Gait flow image: A silhouette-based gait representation for human identification. Pattern recognition 44(4), 973–987 (2011)
25. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2-3), 107–123 (2005)
26. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 2, pp. 2169–2178. IEEE (2006)
27. Liu, J., Yang, Y., Saleemi, I., Shah, M.: Learning semantic features for action recognition via diffusion maps. Computer Vision and Image Understanding 116(3), 361–377 (2012)
28. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24(7), 971–987 (2002)
29. Qin, J., Luo, T., Shao, W., Chung, R., Chow, K.: A bag-of-gait model for gait recognition
30. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
31. Sivapalan, S., Chen, D., Denman, S., Sridharan, S., Fookes, C.: Histogram of weighted local directions for gait recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on. pp. 125–130. IEEE (2013)
32. Tan, D., Huang, K., Yu, S., Tan, T.: Efficient night gait recognition based on template matching. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. vol. 3, pp. 1000–1003. IEEE (2006)
33. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 3169–3176. IEEE (2011)
34. Whytock, T., Belyaev, A., Robertson, N.M.: Dynamic distance-based shape features for gait recognition. Journal of Mathematical Imaging and Vision pp. 1–13 (2014)

35. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 492–497. IEEE (2009)
36. Yogarajah, P., Condell, J.V., Prasad, G.: P rw gei: Poisson random walk based gait recognition. In: Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on. pp. 662–667. IEEE (2011)
37. Yu, S., Tan, D., Huang, K., Tan, T.: Reducing the effect of noise on human contour in gait recognition. In: Advances in Biometrics, pp. 338–346 (2007)
38. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. vol. 4, pp. 441–444. IEEE (2006)