Spatio-Temporal Alignment and Hyperspherical Radon Transform for 3D Gait Recognition in Multi-View Environments

C. Canton-Ferrer, J.R. Casas, M. Pardàs
Image Processing Group – Technical University of Catalonia
Barcelona, Spain

Abstract

This paper presents a view-invariant approach to gait recognition in multi-camera scenarios exploiting a joint spatio-temporal data representation and analysis. First, multi-view information is employed to generate a 3D voxel reconstruction of the scene under study. The analyzed subject is tracked and its centroid and orientation allow recentering and aligning the volume associated to it, thus obtaining a representation invariant to translation, rotation and scaling. Temporal periodicity of the walking cycle is extracted to align the input data in the time domain. Finally, Hyperspherical Radon Transform is presented as an efficient tool to obtain features from spatio-temporal gait templates for classification purposes. Experimental results prove the validity and robustness of the proposed method for gait recognition tasks with several covariates.

1. Introduction

Gait analysis is a promising research direction towards contact-free biometrics for person recognition. Automatic gait recognition is attractive because it enables the identification of a potentially uncooperative subjects from a distance, with a variety of possible applications. Moreover, these algorithms must be robust to pose variations, perspective changes, low resolution and noisy input images.

Biometric identification techniques based on vision-based gait recognition have to deal with two important issues: the appearance variations during the walking cycle due to the relative position between the camera and the subject, and the generation of informative features including spatio-temporal information towards maximizing the distinguishability among subjects. The first problem is usually found in monocular systems and is tackled by requiring the user to walk following a determinate path to ensure that a correctly aligned lateral view of the subject is obtained [1, 6]. In some works, multi-camera systems are employed towards being robust to appearance changes us-

ing calibration information to infer 3D information in the form of aligned synthetic views [12, 17], homographically normalized body part trajectories [9], the canonical sagital plane of the subject [13, 19]. In the field of feature generation for gait recognition we can find the gait energy image proposed by [6] that encodes space and time information in a template image and has being widely employed [1, 19]. Another useful representation is the Radon transform of the gait energy image producing a sparse set of features [1]. Other techniques employ normalized sequences of limbs positions [9] or 3D reconstructions [16].

Due to the high dimension of feature spaces, classification techniques employed in this field aim at a dimension reduction and/or feature selection. Linear techniques such as PCA, MDA and LDA have been thoroughly used [6, 1]. In some cases, fusion of classifiers allowed integrating information from multiple views at feature level [8] or combining information from multiple modalities such as face and gait [17].

The current article presents a robust solution to person recognition using information provided by multiple views. In order to overcome appearance variations due to perspective, a 3D voxel reconstruction of the scene is obtained. Information provided by a tracking system allows estimating the centroid and orientation of the subject under study and translating and rotating its associated volume to a common spatial reference frame. The obtained set is scale, translation and rotation invariant. Time alignment is achieved by estimating the walking cycle out of the obtained invariant volume data. The Hyperspherical Radon Transform is introduced as a robust technique to analyze spatio-temporal data by integrating the aligned set through a set of hyperplanes that integrate information from both space and time. A first dimension reduction is performed through a variance analysis for feature selection and the LDA algorithm is applied afterwards. Finally, effectiveness of the proposed algorithm is assessed by mean on quantitative metrics over an annotated multi-camera dataset. Real-time performance of the proposed algorithm is also achieved proving its validity for real systems.
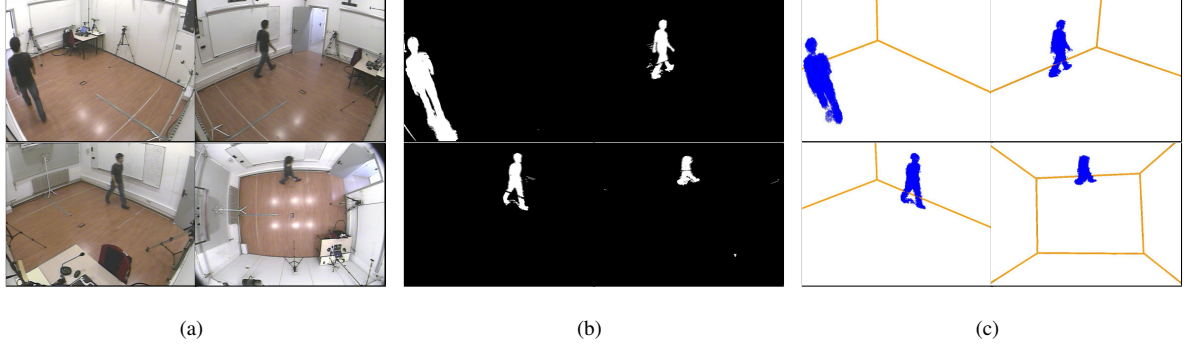
1

(a)　　　　　　　　　　(b)　　　　　　　　　　(c)

Figure 1. Multi-camera input data sample. In (a), a sample of the original images. In (b), the foreground segmentation of the input images employed by the Shape-from-Silhouette algorithm and, in (c), projection of the binary 3D voxel reconstruction.

2. Data generation

For a given frame in the video sequence, a set of N_C images are obtained from the N_C cameras (see a sample in Fig.1(a)). Each camera is modeled using a pinhole camera model based on perspective projection with camera calibration information available [7]. Then, foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and subtraction technique [18] as shown in Fig.1(b).

Redundancy among cameras is exploited by means of a Shape-from-Silhouette technique [3]. This process generates a discrete binary occupancy representation of the 3D space (voxels) denoted as $\mathcal{V}_t$, $1 \leq t \leq N_T$, and shown in Fig.1(c). A voxel is labeled as foreground or background by checking the spatial consistency of its projection on the N_C segmented silhouettes. Usually, this data is noisy and may present holes and spurious blobs due to shadows and reflections. However, the presented technique can cope with such inaccuracies and still provide satisfactory results.

2.1. Spatio-temporal alignment

Prior to compute any transformation, it is required to preprocess the original input data $\mathcal{V}_{1:N_T}$ in order to obtain a representation invariant to spatial scale changes, rotations and translations. Some vision approaches to gait recognition achieve this spatial invariance by constraining the user to follow a determinate trajectory [1, 6] to get frontoparallel images. Other techniques using information provided by multiple cameras generate invariant representations as synthetic views [17, 19] or body part trajectories [9].

Time alignment is usually a required feature of the input data towards a coherent temporal analysis. This process is usually achieved by detecting the start and end of the walking cycles in a sequence and analyzing their statistics either in the temporal domain [5, 9] or using motion templates such as the gait energy image [1, 6].

Trajectory analysis for spatial invariance

The elements of the computed 3D reconstructions, the voxels, are directly related with the physical scale of the objects in the scene, hence rendering the set $\mathcal{V}_{1:N_T}$ invariant to perspective and scaling issues. Without loss of generality, we presume that the XY of our coordinate system is the ground plane and the Z axis is normal to it. Translation and rotation invariance is achieved by first estimating the centroid of the subject, $\mathbf{c}_t$, and its orientation on the XY plane, α_t, and then applying the following linear transformation to every element $\mathbf{x} \in \mathcal{V}_t$:

$$\mathbf{x}' = \left(\begin{array}{cc} \mathbf{R}\left(\alpha_t\right) & \mathbf{0} \\ \mathbf{0} & 1 \end{array} \right) (\mathbf{x} - \mathbf{c}_t)^\top, \qquad (1)$$

where $\mathbf{R}(\alpha_t)$ stands for the rotation matrix on the XY plane. Due to this transformation, the voxels of the resulting set are localized in a specific region of the space thus allowing to crop out the empty zones and finally obtain the new aligned voxel set denoted as $\mathcal{S}_t$ of dimension $s_x \times s_y \times s_z$. It is required for the centroid estimation $\mathbf{c}_t$ to be accurate and robust to noisy 3D measurements and the sparse sampling approach presented in [2] provided satisfactory results. Angle α_t is obtained from the velocity vector associated to sequence $\mathbf{c}_t$ using a Kalman filter. An example of this process is depicted in Fig.2.

Walking cycle detection for time alignment

Let us define the walking cycle of each subject as the category set $\mathcal{G} = \{C^L_{N_L} \cdots \, C^N \cdots \, C^R_{N_R}\}$, where C^N represents the neutral pose and the subsequences $\{C^L_1 \cdots \, C^L_{N_L}\}$ and $\{C^R_1 \cdots \, C^R_{N_R}\}$ are the left and right forward leg movement respectively (see Fig.4). Note that left and right subcycles may not have the same duration, therefore $N_L = N_R$ cannot be granted.

Once the aligned sets $\mathcal{S}_{1:N_T}$ have been computed, we will label some of its element to one of the categories contained in $\mathcal{G}$. Hence, every $C \in \mathcal{G}$ will have associated a list
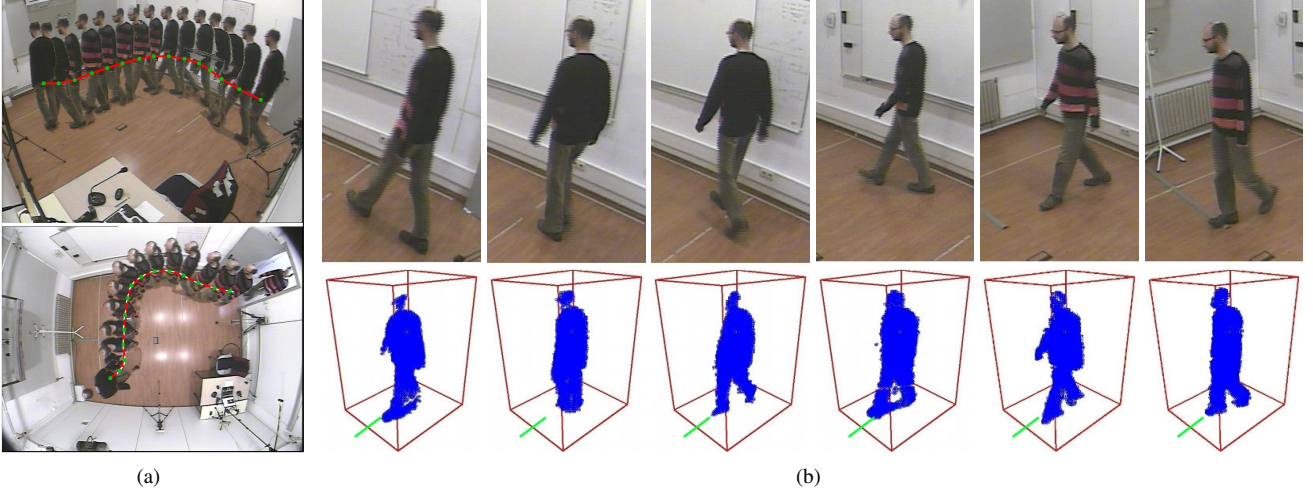
Figure 2. Volume invariance to scale, translation and rotation. In (a), an example of the tracking of the centroid $\mathbf{c}_t$ for a walking random path from two perspectives. In (b), the resulting aligned volume after applying transformation of Eq.1 and the original images.

of the time instants that represent an instance of that specific category within the walking cycle. First, the start and end of every walking cycle within the input sequence are estimated by analyzing the step width of $\boldsymbol{\mathcal{S}}_{1:N_T}$, that is the maximum volume span in the X axis, shown in Fig.3(a). Extrema of this function are employed to estimate the period of every walk subcycle, N_L and N_R, and allow recognizing the start and end of walking cycles. The analysis of the volume associated to the legs allows defining a likelihood function to assess whether the left or right leg is stepping forward. Essentially, the lower section of set $\boldsymbol{\mathcal{S}}_t$ up to $z = 1$ m is selected and four regions are defined: front-left ($\boldsymbol{\mathcal{S}}_t^{\text{FL}}$), front-right ($\boldsymbol{\mathcal{S}}_t^{\text{FR}}$), rear-left ($\boldsymbol{\mathcal{S}}_t^{\text{RL}}$) and rear-right ($\boldsymbol{\mathcal{S}}_t^{\text{RR}}$). The likelihood of having the left or right leg forward is expressed as:

$$p(\text{left}) \propto \left|\boldsymbol{\mathcal{S}}_t^{\text{FL}}\right| + \left|\boldsymbol{\mathcal{S}}_t^{\text{RR}}\right|, \quad p(\text{right}) \propto \left|\boldsymbol{\mathcal{S}}_t^{\text{FR}}\right| + \left|\boldsymbol{\mathcal{S}}_t^{\text{RL}}\right|, \quad (2)$$

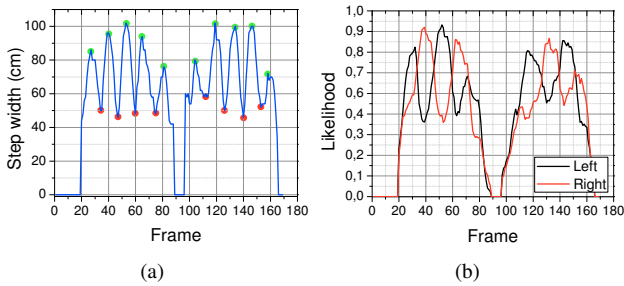where operator $|\cdot|$ stands for the number of foreground voxels of the enclosed volume. See Fig.3(b) for an example.

Figure 3. Gait cycle analysis. In (a), the step width computed over the aligned input data $\boldsymbol{\mathcal{S}}_t$ and the detected extrema, depicted as red and green dots. In (b), the likelihood of having either the left or right leg forward.

Finally, once every element of $\boldsymbol{\mathcal{S}}_{1:N_T}$ has been assigned to a class $C \in \boldsymbol{\mathcal{G}}$ or has been disregarded, the mean of all volumes associated to every class is computed, and will be denoted as $\boldsymbol{\mathcal{W}}_C$. This collection of volumes will define the spatio-temporal set $\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{W}}} = \{\boldsymbol{\mathcal{W}}_{C_{N_L}^L} \cdots \boldsymbol{\mathcal{W}}_{C^N} \cdots \boldsymbol{\mathcal{W}}_{C_{N_R}^R}\}$, also regarded as a gait template for a determinate subject. This set is invariant to scaling, rotation, translation and has been properly aligned and timely averaged within the walking cycle thus being suitable as a person's gait template for the forthcoming classification algorithm. An example of this set is shown in Fig.4. For the sake of notation simplicity in the next sections, ordinal time instants are assigned to volumes within $\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{W}}}$, centered at the neutral pose C^N, thus becoming $\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{W}}} = \{\boldsymbol{\mathcal{W}}_{-N_L} \cdots \boldsymbol{\mathcal{W}}_0 \cdots \boldsymbol{\mathcal{W}}_{N_R}\} \equiv \{\boldsymbol{\mathcal{W}}(x, y, z, t)\}$.

3. Hyperspherical Radon Transform

The Radon transform [4] and its variants [10] have been found useful for dimension reduction and to obtain informative features in image classification problems. Within the scope of this paper, this technique has been used for monocular gait recognition using template images [1] and to process 3D data for search and retrieval tasks [20]. The usual approach to define the transformation variables of the Radon transform is through circular or spherical coordinate systems. In this paper, an extension of the Radon transform in hyperspherical coordinates is presented to deal with the aligned spatio-temporal gait template set $\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{W}}}$.

Let $\mathbb{R}^4$ be the space-time framework where $\mathbf{x} \in \mathbb{R}^4$ stands for coordinates (x, y, z, t). Let S_ρ be the hypersphere with radius $\rho \in \mathbb{R}$ centered at $\mathbf{p}_0 = (s_x/2, s_y/2, s_z/2, 0)$ and $\boldsymbol{\eta} \in \mathbb{R}^4$ a unit vector posed in the coordinate system of
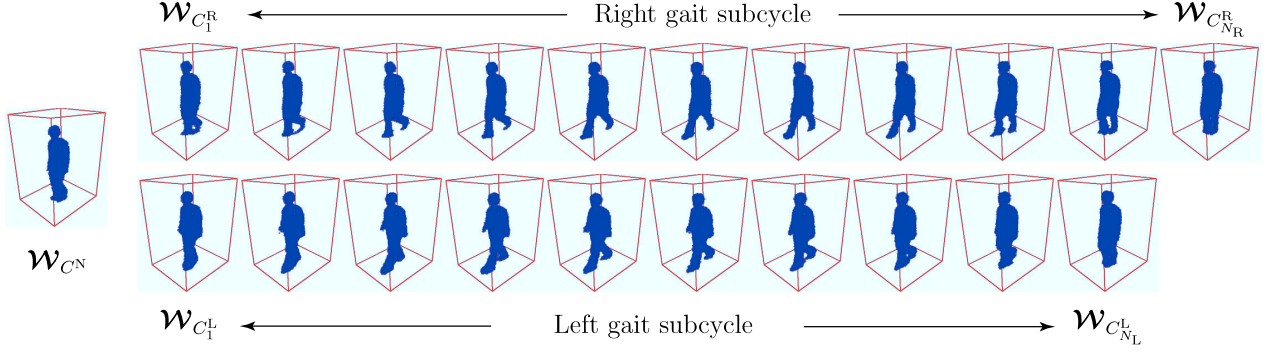
Figure 4. Example of the spatio-temporally aligned set $\mathcal{G}_{\mathcal{W}}$ for a given individual whose left walk subcycle is shorter than the right one.

this hypersphere, described by angles ϕ, θ and ψ as

$$\eta_x = \sin\phi\sin\theta\cos\psi, \qquad \eta_y = \sin\phi\sin\theta\sin\psi,$$
$$\eta_z = \sin\phi\cos\theta, \qquad \eta_t = \cos\phi, \qquad (3)$$

with $\{\phi,\theta\} \in [0,\pi]$, $\psi \in [0,2\pi]$. Let us define the hyperplane $\pi(\boldsymbol{\eta},\rho) = \{\mathbf{x}|\mathbf{x}^\top\boldsymbol{\eta} = \rho\}$ normal to S_ρ at point $\mathbf{p} = \mathbf{p}_0 + \boldsymbol{\eta}\rho$. The hyperspherical Radon Transform, $\mathcal{R}(\boldsymbol{\eta},\rho) : \mathbb{R}^4 \to \mathbb{R}^4$ of the spatio-temporal volume $\mathcal{W}(\mathbf{x})$ over this hyperplane is defined as [4]:

$$\mathcal{R}(\boldsymbol{\eta},\rho) = \int_{\mathbf{x}\in\pi(\boldsymbol{\eta},\rho)} \mathcal{W}(\mathbf{x})d\mathbf{x}. \qquad (4)$$

A more descriptive representation of this transform is obtained when using hyperspherical coordinates together with Dirac's delta function in a discrete domain as

$$\mathcal{R}(\phi,\theta,\psi,\rho) = \sum_{k=1}^{K} \mathcal{W}(\mathbf{x}_k)\delta(\mathbf{x}_k\boldsymbol{\eta} - \rho). \qquad (5)$$

It must be noted that hyperplane $\pi(\boldsymbol{\eta},\rho)$ spans over time and space hence the transformed coefficients will encode information from both domains and is robust to variations due to spurious noisy voxels. This integrating hyperplane can be understood as a 3D plane that shifts along time (see Fig.5) and intersects with $\mathcal{W}(\mathbf{x})$ to produce $\mathcal{R}(\boldsymbol{\eta},\rho)$.

The Radon transform is very suitable for gait representation and recognition. During the walking cycle, there are two noticeable variations: the appearance changes among different time instants produced by the limbs movement and the variations among subjects when performing a walking cycle. This means that the Radon transform over a properly space and time aligned set guarantees some specific coefficient will vary considerably through time and among subjects. Therefore, the study of these coefficients will allow distinguishing among different gait templates.

A discrete set of values for the parameters (ϕ,θ,ψ,ρ) is defined with a step $\Delta(\phi,\theta,\psi) = 10^\text{o}$ and $\Delta\rho = 5$ cm with $\rho \in [0, \max(s_x, s_y, s_z, N_\text{R}, N_\text{L})]$. With this transformation, the dimension of the analyzed data is reduced up

to a 25% of the original size of $\mathcal{G}_{\mathcal{W}}$. It has been tested empirically that employing smaller steps did not yield to a performance gain. Since the employed hyperplanes are defined in a discrete domain, their elements (shown in Fig.5) are precomputed using a rasterization technique [11] and stored towards a faster computation of $\mathcal{R}(\boldsymbol{\eta},\rho)$.

The hyperspherical Radon transform is not invariant to scaling, translation and rotation thus the preprocessing of input data $\mathcal{V}_{1:N_\text{T}}$ ensures that transformed coefficients from different subjects will be comparable.

4. Feature Extraction and Classification

A direct comparison of the obtained Radon coefficients of several subjects, $\mathcal{R}^i$, $1 \leq i \leq N_\text{S}$, can be performed but the obtained recognition rate will low ($\sim 60\%$). Moreover, the dimension of a transformed gait template $\mathcal{R}^i$ is still high, turning out the generation of this set a computationally hard task. The study of the coefficients in $\mathcal{R}^i$ showed that this set is sparse, hence a first step to reduce
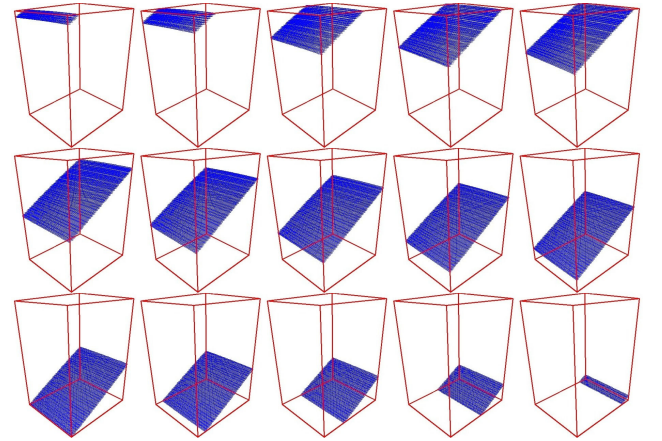

Figure 5. Example of a hyperplane $\pi(\boldsymbol{\eta},\rho)$, for $\phi = \pi/8$, $\theta = \pi/6$, $\psi = 0$ and $\rho = 0$. Snapshots of $\pi(\boldsymbol{\eta},\rho)$ at consecutive time instants are displayed in order to shown this 4D hyperplane.

the dimensionality of the problem is to discard evaluating those hyperplanes yielding to a null transform. A variance analysis [14] is conducted over $\mathcal{R}^i$ showing that, for the data employed in this paper, only a 2% of the coefficients of $\mathcal{R}^i$ have a significant variance and might be usable as features to distinguish among different subjects. Let us denote $\widetilde{\mathcal{R}}^i$ as the set of the selected Radon coefficients exhibiting the maximum variance.

Linear dimension reduction techniques have been applied over $\widetilde{\mathcal{R}}^i$, $\forall i$, in order to further reduce its dimension and class separability. A linear discriminant analysis has been conducted to obtain a final 30 feature feature set assigned to each subject, F^i. Classification of an unknown subject, F^U, has been performed using a minimum distance criterion:

$$\min_i \| F^U - F^i \|. \tag{6}$$

5. Experiments and Results

There is a number of datasets intended for vision-based gait but, although some of them datasets contain video data from multiple cameras, few provide both synchronization among cameras or calibration information. In order to test the validity of our algorithm we have recorded a dataset containing 28 people, 22 men and 6 women, walking in a 4x5 meters room surveyed by 5 calibrated and sync cameras at 25 fps with 768x576 pixels. The training part of this dataset presents the different subjects walking straight in the scenario while the testing part includes several covariates of the walking cycle (carrying a bag, wearing slippers or no shoes, wearing a coat). Invariance of the proposed algorithm to scaling, rotation and translation has been tested by three particular covariates that involved walking in diagonals, zig-zag or randomly.

Obtained recognition results for this scenario are shown in Fig.6. It can be seen that for standard cases such as straight walking the recognition rate is the highest and this performance decreases gradually as the complexity of the path grows. The abrupt changes in the trajectory of the subject yields to noisy angle estimations therefore producing a misalignment of the input data fed to the hyperspherical Radon transform.

Once the right Radon coefficients are located, their computation and the LDA analysis has been processed in real time on a 3GHz desktop computer.

6. Conclusions and Future Work

This paper introduces a novel approach to gait recognition and two main contributions are presented. First, a unified space-time representation of input data in the form of a time evolving volume set associated to a walking cycle. This representation is invariant to scale, rotation and
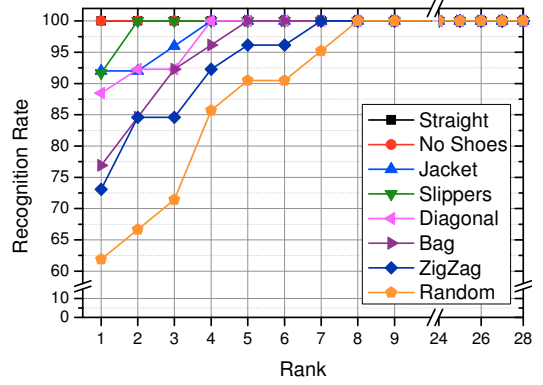

Figure 6. Cumulative match scores for the proposed algorithm when analyzing input data with several covariates.

translation changes. In order to analyze this input, the hyperspherical Radon transform is introduced as an effective algorithm to produce a sparse set of features through the integration of the spatio-temporal volume over a set of hyperplanes. Further dimension reduction using LDA yield a high class separability. Results over an annotated dataset containing 28 subjects with a number of covariates proved that the proposed method is effective for gait recognition tasks.

The proposed method is not straighforward applicable to the sequences of the widely known Gait Challenge problem [15] due to issues with the calibration of the cameras and the adequateness of the data. The combination of the 3D representation and the ability to analyze spatio-temporal data through the proposed Radon transform might produce high recognition rates on large datasets. Future steps aim at proving this afirmation.

Future research lines involve applying the presented scheme on sequences with moderate occlusions yielding to noisy 3D reconstructions. The extension of other integral transforms to the spatio-temporal domain is also under study. As a contribution for further comparison, the authors will release the employed multi-camera dataset.

References

[1] N. Boulgouris and Z. Chi. Gait recognition using radon transform and linear discriminant analysis. *IEEE Trans. on Image Processing*, 16(3):731–740, 2007. 1, 2, 3

[2] C. Canton-Ferrer, R. Sblendido, J. R. Casas, and M. Pardàs. Particle filtering and sparse sampling for multi-person 3D tracking. In *Proc. of IEEE Int. Conf. on Image Processing*, pages 2644–2647, 2008. 2

[3] G. Cheung, T. Kanade, J. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 714–720, 2000. 2

[4] S. Deans. *The Radon Transform and some of its applications*. Wiley, 1983. 3, 4

[5] S. Dockstader, M. Berg, and A. Tekalp. Stochastic kinematic modeling and feature extraction for gait analysis. *IEEE Trans. on Image Processing*, 12(8):962–976, 2003. 2

[6] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006. 1, 2

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2

[8] X. Huang and N. Boulgouris. Human gait recognition based on multiview gait sequences. *EURASIP Journal on Advances in Signal Processing*, 2008:1–8, 2008. 1

[9] F. Jean, A. Branzan, and R. Bergevin. Towards view-invariant gait modelling: computing view-normalized body part trajectories. *Pattern Recognition*, 42:2936–2949, 2009. 1, 2

[10] A. Kadyrov and M. Petrou. The trace transform and its applications. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 23(8):811–828, 2001. 3

[11] C. Lincke, C. Wüthrich, and P. Guitton. An exact weaving rasterization algortithm for digital planes. In *Proc. Int. Conf. on Computer Graphics and Visualization*, volume 2, pages 395–402, 1999. 4

[12] A. López, C. Canton-Ferrer, and J. R. Casas. Virtual view appearance representation for human motion analysis in multi-view environments. In *Proc. European Signal Processing Conf.*, 2010. 1

[13] A. Roy-Chowdhury, A. Kale, and R. Chellappa. Towards a view invariant gait recognition algorithm. In *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 143–150, 2003. 1

[14] D. B. Rubin. Matching to remove bias in observational studies. *Biometrica*, 29(1):159–183, 1973. 5

[15] S. Sarkar, P. Phillips, Z. Liu, I. Robledo Vega, P. Grother, and K. Bowyer. The HumanID gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 27(2):162–177, 2005. 5

[16] R. D. Seely, S. Samangooei, L. Middleton, J. Carter, and M. Nixon. The university of southampton multi-biometric tunnel and introducing a novel 3D gait dataset. In *Proc. IEEE Workshop on Biometrics: Theory, Applications and Systems*, 2008. 1

[17] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 439–446, 2001. 1, 2

[18] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 252–259, 1999. 2

[19] A. Tyagi, J. Davis, and M. Keck. Multiview fusion for canonical view generation based on homography constraints. In *Proc. ACM Int. Workshop on Video Surveillance and Sensor Networks*, pages 61–70, 2006. 1, 2

[20] D. Zarpalas, P. Daras, A. Axenopoulos, D. Tzovaras, and M. Strintzis. 3D model search and retrieval using the spherical trace transform. *EURASIP Journal on Advances in Signal Processing*, 2007:1–14, 2007. 3