Performance analysis for automated gait extraction and recognition in multi-camera surveillance

**Michela Goffredo · Imed Bouchrika ·
John N. Carter · Mark S. Nixon**

Abstract Many studies have confirmed that gait analysis can be used as a new bio-metrics. In this research, gait analysis is deployed for people identification in multi-camera surveillance scenarios. We present a new method for viewpoint independent markerless gait analysis that does not require camera calibration and works with a wide range of walking directions. These properties make the proposed method particularly suitable for gait identification in real surveillance scenarios where people and their behaviour need to be tracked across a set of cameras. Tests on 300 synthetic and real video sequences, with subjects walking freely along different walking directions, have been performed. Since the choice of the cameras' characteristics is a key-point for the development of a smart surveillance system, the performance of the proposed approach is measured with respect to different video properties: spatial resolution, frame-rate, data compression and image quality. The obtained results show that markerless gait analysis can be achieved without any knowledge of camera's position and subject's pose. The extracted gait parameters allow recognition of people walking from different views with a mean recognition rate of 92.2% and confirm that gait can be effectively used for subjects' identification in a multi-camera surveillance scenario.

Keywords Gait analysis · Gait recognition · Biometrics · Multi-view ·
Surveillance · Object handover

M. Goffredo · I. Bouchrika (✉) · J. N. Carter · M. S. Nixon
School of Electronics and Computer Science, University of Southampton,
Southampton SO17 1BJ, UK
e-mail: ib04r@ecs.soton.ac.uk

M. Goffredo
e-mail: mg2@ecs.soton.ac.uk

J. N. Carter
e-mail: jnc@ecs.soton.ac.uk

M. S. Nixon
e-mail: msn@ecs.soton.ac.uk

🙎 Springer

1 Introduction

Activity monitoring for security purposes and work flow control in critical infrastructures can be successfully achieved with a distributed network of video sensors, instead of having one high resolution camera with a limited field of view. Multiple-cameras, in fact, provide a solution to wide and complex areas where overlapped and non-overlapped views allow more accurate surveillance of human action. The aim of a smart surveillance system is to automatically track people across a network of cooperating active cameras and detect abnormal behaviours. In this context, "object handover" is defined as the subject's identification and tracking across a distributed network of cameras, that can have both overlapped and non-overlapped views.

In this research, we propose gait analysis as a solution for subjects' identification across a network of cameras. Recognizing individuals by the way they walk is a particularly challenging research area, where the potential for personal identification is supported by a rich literature, including medical and psychological studies [11, 30]. The suitability of gait analysis and recognition for surveillance systems emerges from the fact that gait can be perceived from a distance as well as its non-invasive nature. Although gait recognition is not sufficiently mature to be used in real world applications such as visual surveillance, it overcomes most of the limitations that other biometrics suffer from such as face, fingerprints, and iris recognition which can be obscured in most situations where serious crimes are involved. Furthermore, gait analysis can be potentially deployed in numerous applications for event detection and activity recognition including walking, running, jumping and load carriage. Recently, Bouchrika et al. [4] shown the use of gait periodicity for detection of walking pedestrians.

The completely unobtrusiveness without any subject cooperation or contact for data acquisition make gait particularly attractive for identification purposes in camera handover. In real surveillance scenarios, however, we need a system that operates in an unconstrained environment where maybe there is no information regarding the relative position between the camera and the walking subject [38, 45]. 2D view independent gait identification is a recent research area and the approaches found in literature extract some gait parameters which are independent of the human pose [1, 43] by analysing the silhouette shape over time or aim at synthesising the sagittal view of the human body from any other arbitrary views [16, 24]. The latter choice is justified by the fact that the lateral view has proven recognition capability in a great number of works [20, 33, 46]. However, they need information on the camera parameters [24] or rely on the use of reflective markers on the body [16].

In this study, a novel 2D markerless view independent gait analysis algorithm is presented and tested for assessing it applicability for object handover in multi-camera scenarios. The method does not need camera calibration or pre-knowledge of subject pose. Since the choice of the cameras' characteristics is a key-point for the development of a smart surveillance system, the performance of the proposed approach is measured with respect to different video properties.

1.1 Related work

Gait recognition techniques at the state of the art can be divided into 3D [2, 29] and 2D approaches [20, 33, 46]. In the first group, identification relies on parameters

extracted from the 3D limb movement. These methods use a large number of digital cameras and the 3D reconstruction is achieved after a camera calibration process. On the other hand, the 2D gait biometric approaches extract explicit features describing gait by means of human body models [5, 34] or silhouette shape [10, 42, 44]. A rich variety of data has been collected for evaluation of 2D gait biometrics. The widely used and compared databases on gait recognition include: the University of South Florida [36]; Carnegie Mellon University [17]; and the University of Southampton [39] data. The majority of methods and databases found in the literature use a single camera positioned with specific orientations with respect to the subject's walking direction (generally capturing the walk from the lateral view) and a large number of papers for gait recognition have been published [19, 33].

Several strategies have been employed for coordinating subject tracking between multiple overlapping camera views [3, 26]. They have assumed that 3D ground calibration information was available for each camera or estimate it by analysis of training data. Recently there has been some interest in tracking objects between blind regions [21, 23, 27]. They make use of Kalman filtering assuming the ground plane is known [9] or assume that the transition models are known or hand-labelled [13, 37]. Some approaches recover the relative cameras' position and the subject's movement across them with a statistical approach [28, 41]: two cameras are considered connected if objects seen departing in one camera are seen arriving in the other and the statistical dependence is characterized by the distribution of observation transformations between cameras, such as departure to arrival transition times.

While there is a large number of papers regarding the tracking between blind regions, a few works address the issue of subject's identification across a network of video sensors [6, 14, 25, 31, 40]. Appearance information is indeed one of the most popular solutions for solving the object handover issue since it is relatively robust towards size and orientation changes [6, 14, 25, 31]. Possible appearance features are colour templates, histograms, moments, signatures as dominant colours, and partitive colour layouts. Since illumination, camera distortion and object resolution change in a multi-camera setting, these approaches include an inter-camera distortions and illumination compensation [6, 14, 22]. However, in some environments like the military or industrial scenarios, where everyone dresses in identical clothes, appearance information cannot be applied. Furthermore, the colour-based approaches work in a short time period and are not suitable for a long-term workflow control. To this extent, Stillman et al. used face identification in a multimodal approach where acoustic and visual data are combined [40]. However, in typical surveillance applications cameras are usually located far away from the subjects, which result in low-resolution face images. Another concern is the face orientation: most face-based methods work only for frontal images and the accuracy of identification quickly decreases even with the slight orientation differences.

2 Theory

The proposed method for the view independent gait analysis is based on two consecutive stages: markerless joints' estimation from the image sequence and viewpoint rectification. Gait parameters are then used for identification.

2.1 Markerless joints' position estimation

Let $\mathbf{S}(\mathbf{x}, \mathbf{y}, t_\varphi)$ be the RGB frame (of size $R \times C$ pixels) at time $\mathbf{T} = [t_1, t_2, \ldots, t_\varphi, \ldots, t_F]$ where (x_1, y_1) is the top-left corner of the image. By applying a background subtraction method based on the threshold of the three components of the color space YUV, the binary image $\overline{\mathbf{S}}(\mathbf{x}, \mathbf{y}, t_\varphi)$ has been extracted. The pixels $(\mathbf{x}_{sil}, \mathbf{y}_{sil})$ containing the human silhouette have been selected as the object with maximum area [18]. The lower limbs pose estimation algorithm is based on the proportions of the human body segments, following the medical results of anatomical studies [12]:

$$y'_{hip} = \min\left(\mathbf{y}_{sil}\right) + 0.5 \cdot H$$

$$y'_{knee} = \min\left(\mathbf{y}_{sil}\right) + 0.75 \cdot H$$

$$y'_{ankle} = \min\left(\mathbf{y}_{sil}\right) + 0.90 \cdot H \tag{1}$$

where H is the silhouette's height.

Subsequently, the shins and thighs have been analysed separately. Let $\hat{\mathbf{X}}$ be the set of horizontal coordinates where $\overline{\mathbf{S}}(\hat{\mathbf{X}}, y'_{knee}, \hat{t}_\delta) = 1$, then the subset of frames

$$\hat{\mathbf{T}} = \left[\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_\delta, \ldots, \hat{t}_{D \leq F}\right] \tag{2}$$

where the legs do not overlap are those where $\hat{\mathbf{X}}$ forms 2 single connected components larger than $0.02H$ pixels.

Consequently, for each frame $\hat{t}_\delta$, the shin extraction algorithm, based on the linear approximation of the skeleton of the portion of image that includes the lower leg, is applied on $\overline{\mathbf{S}}(\mathbf{x}, \mathbf{y}, \hat{t}_\delta)$. The sub-images containing the shins (labeled as $\mathbf{S}_{shin1}$ and $\mathbf{S}_{shin2}$) are extracted from

$$\mathbf{S}_{lower\ leg}(\mathbf{x}, \mathbf{y}, \hat{t}_\delta) = \begin{cases} \overline{\mathbf{S}}\left(\mathbf{x}, y_s, \hat{t}_\delta\right) & \text{if } y_s \in \mathbf{y}_{shin} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where the vertical coordinates belonging to the two shins $\mathbf{y}_{shin} = [y_1, y_2, \ldots, y_s, \ldots, y_S]$ are defined in the following way:

$$\mathbf{y}_{shin} = \left[y'_{knee}, y'_{knee} + 1, \cdots, \min\left(y'_{ankle}, y_{overlap} - 1\right)\right] \tag{4}$$

which includes the possibility of feet overlapping with $y_{overlap}$: vertical coordinate where the set of correspondent horizontal coordinates $\mathbf{x}_{overlap}$ belong to a single object larger than $0.02H$ pixels.

Therefore, the left and right shins are defined by the pixels with coordinates

$$\begin{cases} \mathbf{x}^{skel}_{shin\ell}\left(\hat{t}_\delta\right) = [x_{1\ell}, x_{2\ell}, \ldots, x_{s\ell}, \ldots, x_{S\ell}] \\ \mathbf{y}^{skel}_{shin\ell}\left(\hat{t}_\delta\right) = \mathbf{y}_{shin} \end{cases} \quad \ell = \{1, 2\} \tag{5}$$

where

$$x_{s\ell} = \sum_{j=1}^{C} x_j \cdot \mathbf{S}_{shin\ell}\left(x_j, y_s, \hat{t}_\delta\right) \bigg/ \sum_{j=1}^{C} \mathbf{S}_{shin\ell}\left(x_j, y_s, \hat{t}_\delta\right) \tag{6}$$

Then, the shins can be linearly approximated by the first order polynomial with coefficients

$$\mathbf{p}_\ell\left(\mathbf{x}^{skel}_{shin\ell}, \hat{t}_\delta\right) = p_{\ell 0}\left(\hat{t}_\delta\right) + p_{\ell 1}\left(\hat{t}_\delta\right) \cdot \mathbf{x}^{skel}_{shin\ell}\left(\hat{t}_\delta\right) \tag{7}$$

Therefore, the angle between the shins and the vertical axis (called knee angle) at frame $\hat{t}_\delta$ is $\alpha_\ell(\hat{t}_\delta) = \pi - \arctan\left[p_{\ell 1}(\hat{t}_\delta)\right]$. The procedure is applied for each frame in the subset $\hat{\mathbf{T}} = [\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_\delta, \ldots, \hat{t}_D]$.

In order to also be able to estimate α_1 and α_2 when the legs overlap, the extracted shins are labeled as 'left' and 'right' according to the following rationale without losing generalities: the subject is walking for at least two gait cycles; in t_1 the shin labeled as *shin*1 belongs to the right leg; the labels swap after the legs overlap.

Subsequently, a 3rd order polynomial interpolation of α_1 and α_2 has been applied and the knees angles are achieved during the whole video sequence. This choice has been determined experimentally.

The upper legs orientation is extracted at every frame $\mathbf{T} = \left[t_1, t_2, \ldots, t_\varphi, \ldots, t_F\right]$ with a coarse to fine hips estimation procedure where at first, the hips position is achieved with

$$
\begin{cases}
x'_{hip\ell} = \dfrac{1}{P} \cdot \sum_{j=1}^{P} \widetilde{x}_j + (2\ell - 3) \cdot H \cdot \mu \cdot 10^{-3} \\[2ex]
y'_{hip\ell} = y'_{hip} \cdot (2\ell - 3) \cdot \left(\dfrac{\widetilde{x}_P - \widetilde{x}_1}{2}\right) \cdot \sin\left(0.3 \cdot \mu\right)
\end{cases}
\tag{8}
$$

where $\widetilde{\mathbf{X}} = \left[\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_j, \ldots, \widetilde{x}_P\right]$ is the subset of P ($P \le C$) horizontal coordinates such as $\overline{\overline{\mathbf{S}}}(\widetilde{\mathbf{X}}, y'_{hip}, t_i) = 1$.

Equation 8 puts in relationship the horizontal hips position and the walking direction μ, calculated with respect to the horizontal axes of the image reference system. These relationships have been obtained with a regressive analysis of the 3D Georgia Tech motion capture data (www.cc.gatech.edu/cpl/projects/hid) by considering different camera positions.

The angle μ is estimated as the angle of inclination of the straight line which approximates the heel strikes points with coordinates:

$$
\left(\mathbf{x}_{str}, \mathbf{y}_{str}\right) = \left(\mathbf{x}_{feet}, \mathbf{y}_{feet}\right) \Bigg| \sum_{i=1}^{P} \boldsymbol{SK}\left(\mathbf{x}_{feet}, \mathbf{y}_{feet}, t_i\right) > \tau
\tag{9}
$$

where $\boldsymbol{SK}$ is the skeleton [32] of the image, $\mathbf{x}_{feet}$ and $\mathbf{y}_{feet}$ are the coordinates corresponding to the portions of the silhouette belonging to the feet and the threshold τ is automatically set so that at least 4 heels strikes are extracted (in accordance with the hypothesis listed above).

Subsequently, an improved hips pose estimation is obtained with a linear approximation of the thighs by the first order polynomial with coefficients

$$
q_{\ell 0}(t_\varphi) = \frac{y^m_{shin\ell} \cdot x'_{hip\ell} - y'_{hip\ell} \cdot x^m_{shin\ell}}{x'_{hip\ell} + x^m_{shin\ell}}
$$

$$
q_{\ell 1}(t_\varphi) = \frac{1}{2} \cdot \left(a'_\ell + \frac{y'_{hip\ell} - q_{\ell 0}}{x'_{hip\ell}}\right)
\tag{10}
$$

where $y^m_{shin\ell} = \min\left(\mathbf{y}_{shin}\right)$ and $x^m_{shin\ell}$ is the mean value of the correspondent horizontal coordinates. a'_1 and a'_2 are the slope of the straight lines that approximate respectively the left and right edges of the positions of silhouette belonging to the thighs.
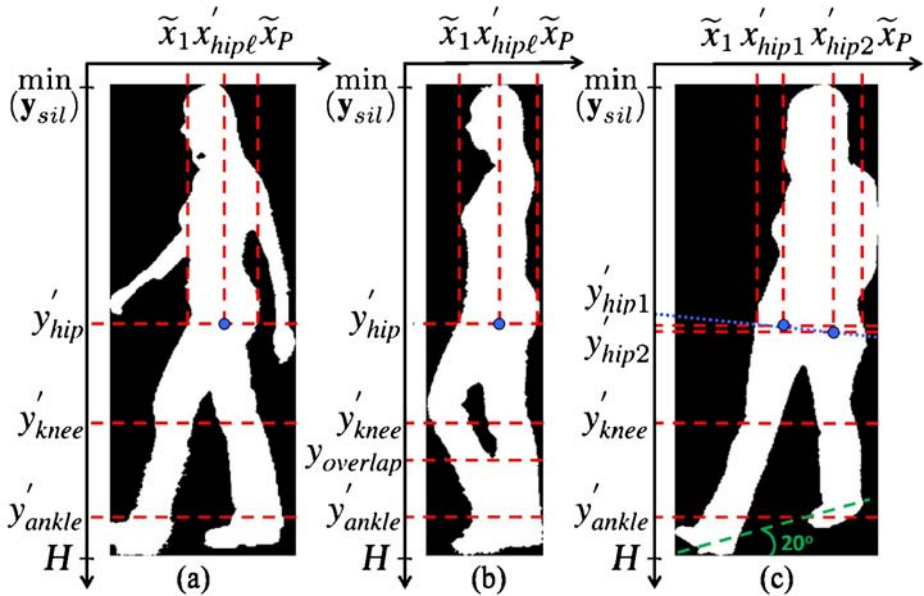
Fig. 1 Markerless gait analysis: $\mu = 0^o$ view with non-overlapped (**a**) and overlapped (**b**) feet; $\mu = 20^o$ view (**c**)

Therefore, the angle between the thighs and the vertical axis at every frame t_φ is $\gamma_\ell(t_\varphi) = \pi - \arctan[q_{\ell 1}(t_\varphi)]$.

Ultimately, the thighs are labeled as 'left' or 'right' in accordance with the corresponding shin labeling. Figure 1 shows the shins and thighs extraction procedure. The angles $\alpha_1, \alpha_2, \gamma_1$ and γ_2 are then corrected by projecting the limbs trajectories on the lateral plane, as explained next.

2.2 View-point independent rectification

The method proposed by the authors in [16] is based on four main assumptions: the nature of human gait is cyclic; subjects walk along a straight line; the distances between the bone joints are constant; and the articulated leg motion is approximately planar. Therefore, the multiple periods of linear gait motion appear analogous to a single period viewed from many cameras related by linear translation. Following this rationale, the positions of the points of interest, i.e. the leg joints, lie in an auto-epipolar configuration consistent with the imaged motion direction. The epipole is thus estimated by computing the intersection of the set of lines formed by linking the correspondent points of interest in each phase of the gait cycle. In order to find these correspondences, the gait periodicity is calculated by applying the stereo transformation that maps the epipole vector e to the ideal point $[1, 0, 0]^T$ and then by computing the cost based on dot product between matching limb segment vectors.

Let $\mathbf{j}_i^\ell$ be the set of joints positions for each leg $\ell = \{1, 2\}$ at the ith frame in the image reference system. After estimating the periodicity of gait, assuming linear velocity between consecutive frames, the set of points of interest $\mathbf{j}_i^\ell$ are recomputed in order to lie on straight lines starting from the epipole. At first the set of points

and the epipole are mapped to the unit square and re-normalized to the unit norm $\|\mathbf{e_0}\| = 1$ respectively. Subsequently, the optimal points are found by estimating the positions $\mathbf{j}_i^\ell$ that lie on the epipolar line and that satisfies the condition

$$\mathbf{j}_i^{\ell^T} [\mathbf{e_0}] \times \mathbf{j}_i^\ell = 0 \tag{11}$$

Therefore the back projected rays, formed from a set of optimal points, intersect in a single worldspace point: the epipole. The back projection of all sets of points generates the cluster of 3D points for an assumed single period of reconstructed gait motion. The Direct Linear Transform, DLT, is then used in order to triangulate each worldspace point $\mathbf{J}^\ell$

$$\mathbf{j}_i^\ell \times \mathbf{P}_i \cdot \mathbf{J}^\ell = 0 \tag{12}$$

with the set of camera projection matrices

$$\mathbf{P}_i = \left[\mathbf{R}_\mathbf{e}^T, -i\mathbf{e}_0 \right] \tag{13}$$

where $\mathbf{j}_i^\ell$ is the image of the worldspace point $\mathbf{J}^\ell$ in the ith period image, $\mathbf{R}_\mathbf{e}^T$ is the 3 by 3 rotation matrix that aligns the epipolar vector with the X (horizontal) axis, and i is an integer describing the periodicity of the subject's translation as the number of frames in a sequence.

Considering the assumption that the articulated leg motion is approximately planar, the 3D limb points can be reasonably fitted to two planes. Since the epipolar vector is aligned with the X axis, the ideal point $[1, 0, 0, 0]^T$ do lie on each of the worldspace planes. Therefore, the pencil of planes that intersect this ideal point have the form $p = [0, v2, v3, v4]^T$. Consequently the problem is reduced to finding two lines within the YZ plane cross section data.

After computing the mean $[y, z]^T$ of the point distribution, the translation $\mathbf{H}_t$ that maps this point to the origin is applied. The two cross section plane lines $l_1 = [v_2, v_3, v_4]^T$ and $l_2 = [v_2', v_3', v_4']^T$ are then achieved by orthogonal regression and then aligned parallel with the Y (vertical) axis by applying a rotation $\mathbf{H}_r$. The intersection point of the two lines is then called u and is given by the cross product between the two lines.

Consequently, the pair of transformed lines are mapped to $l_\ell' = \mathbf{H}_r \cdot l_\ell$ and the rotation matrix $\mathbf{H}_r$ and the perspective transformation $\mathbf{H}_\alpha$

$$\mathbf{H}_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \alpha & 0 & 1 \end{pmatrix} \tag{14}$$

are applied to the point $\mathbf{u}'$ in order to transform it to the ideal point $[1, 0, 0]^T$.

Since $\mathbf{u}'$ lies on the Y axis and has the form $[y, 0, w]^T$, the transformation $\mathbf{H}_\alpha \mathbf{u}'$ gives $\alpha = -w/y$ and the corresponding line mapping $\mathbf{H}_\alpha l_\ell'$ effectively zeros the first component of the two normal lines. Since the lines are parallels, they are normalized

$$l_1' = [0, 1, -c_1]^T \quad l_2' = [0, 1, -c_2]^T \tag{15}$$

so that is it possible to find the point (c_1, c_2) of intersection with Z (depth) axis. A further similarity transform $\mathbf{H}_s$ that translates the midpoint $(c_1, c_2)/2$ to the origin

and scales in the Z direction to rectify the lines to the form $l = [0, 1, \pm 1]^T$ is then applied.

The translation by ± 1 mapping the selected set of points onto the $Z = 0$ plane is then computed with the matrix $\mathbf{H}_b$. The combined set of transformations thus forms the limb plane transformation $\mathbf{H} = \mathbf{H}_\beta \mathbf{H}_s \mathbf{H}_\alpha \mathbf{H}_r \mathbf{H}_t$.

Therefore, the projection transform mapping the back projected points into the image can be decomposed as:

$$\mathbf{j}_i^\ell = \mathbf{P}_i \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{H}^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{H} \end{pmatrix} \mathbf{J}^\ell \tag{16}$$

where

$$\mathbf{H} = \begin{pmatrix} 1 & m_2 & m_3 & m_4 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -\beta \\ 0 & -\alpha & 0 & 1 \end{pmatrix} \tag{17}$$

The corresponding transformation of worldspace points $[u, v, 0, w]^T$ into the image is given by

$$\mathbf{j}_i^\ell = \mathbf{H}_p \, [u, v, w]^T \tag{18}$$

where

$$\mathbf{H}_p = \begin{bmatrix} \mathbf{e}' & m_2' - \alpha \cdot (m_4' - i \cdot \mathbf{e}') & (m_4' - i \cdot \mathbf{e}') - \beta \cdot m_3' \end{bmatrix} \tag{19}$$

with

$$m_i' = \mathbf{R}_e^T \qquad \mathbf{e}' = \mathbf{R}_e^T \, [1, 0, 0]^T \tag{20}$$

Finally the sets of optimal $Z=0$ plane points is found by solution of the

$$\mathbf{x}_{i,\beta}' \times \mathbf{H}_p \, (i, \beta) \, \overline{\mathbf{w}} = 0 \tag{21}$$

for each point $\mathbf{w}$ in order to minimize the reprojection error.

Structure on the $Z=0$ plane has been recovered up to an affine ambiguity $\mathbf{H}_\mu$ that maps the imaged circular points $[1, \mu, \pm j\lambda, o]^T$ back to their canonical positions $[1, \pm j, 0]^T$:

$$\mathbf{H}_\mu = \begin{pmatrix} 1 & 0 & 0 \\ -\dfrac{\mu}{\lambda} & \dfrac{1}{\lambda} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{22}$$

For estimating the metric structure, the lengths of the articulated limbs is assumed to be known and constant over all the frames. Thus the squared distance between two points be written

$$\mathbf{D}_\ell^2 = \Delta \mathbf{j}_i^{\ell \mathbf{T}} \Delta \mathbf{j}_i^\ell \tag{23}$$

where

$$\Delta \mathbf{j}_i^\ell = [u_1 - u_0, v_1 - v_0]^T \tag{24}$$

if we consider the distance between points 0 and 1. If $\Delta\mathbf{j}_i^\ell$ and $\Delta\mathbf{j}_{i+1}^\ell$ are the pose difference vectors for a limb segment at two consecutive frames, then the equal limb length constraint can be written

$$\Delta\mathbf{j}_i^{\ell\mathbf{T}} \cdot \mathbf{H^T} \cdot \mathbf{H} \cdot \Delta\mathbf{j}_i^\ell = \Delta\mathbf{j}_{i+1}^{\ell\mathbf{T}} \cdot \mathbf{H^T} \cdot \mathbf{H} \cdot \Delta\mathbf{j}_{i+1}^\ell \tag{25}$$

Therefore, writing $\Delta\mathbf{j}_i^{\ell\mathbf{T}}$ and the element of the matrix $\mathbf{M} = \mathbf{H}^T\mathbf{H}$ as $m = [\mathbf{M}_{11}, \mathbf{M}_{12}, \mathbf{M}_{22}]$, the equation is

$$\left| \delta x_1^2 - \delta x_2^2 \quad 2\left(\delta x_1^2 \delta y_1^2 - \delta x_2^2 \delta y_2^2\right) \quad \delta y_1^2 - \delta y_2^2 \right| m = 0 \tag{26}$$

Since m is defined up to scale then a minimum of two corresponding pose constraints are required. All constrains formed from all sets of combinations of same limb frame poses are stacked on each swing plane.

The rectification matrix $\mathbf{H}_\mu$ is formed from the extracted parameters of $\mathbf{H}^T\mathbf{H}$, where $\mu = -m_2/m_3$ and

$$\lambda = \sqrt{m_1/m_3 - \mu^2} \tag{27}$$

The ideal epipole $[1, 0, 0]^T$ is then mapped by $\mathbf{H}_\mu$ to $\left[1, -\mu/\lambda, 0\right]^T$ so a rotation $\mathbf{H}_r$ is necessary in order to align the epipole back along the X axis such that $\mathbf{H}_a = \mathbf{H}_r\mathbf{H}_\mu$ is the affine transform that recovers metric angles and length ratios on both planes. Points on the metric plane $\mathbf{w}$ are then mapped into the image as:

$$\mathbf{j}_i^\ell = \mathbf{H}_p \mathbf{H}_a^{-1} \left(\mathbf{H}_a \overline{\mathbf{u}}\right) = \overline{\mathbf{H}}\overline{\mathbf{w}} \tag{28}$$

Scaling is then applied to both planes in order to transform each first limb segment to unit length. The mean set of limb lengths for both planes is estimated as $\mathbf{d}$, $\mathbf{d}'$. These lengths are related by the inter-plane scaling: $\mathbf{d}_i = \tau\mathbf{d}_i'$. A minimal solution to this trivial set of linear equations requires at least one valid length correspondence within the set of limb segments. With $\mathbf{H}_\tau$ now known the optimal first limb segment length $\mathbf{D}_1$ on the first plane can be evaluated. The scaling transform $\mathbf{H}_s$ that maps $\mathbf{D}_1$ to the unit length and update both sets of points and projection homographies is then calculated.

After recovering the fronto-parallel structure of the subject's gait, the representation of the leg joints function $\left[\mathbf{J}_x^\ell(t), \mathbf{J}_y^\ell(t)\right]$ is found by fitting a modified Fourier series to the data with fixed fundamental frequency f_0 and period T:

$$\mathbf{J}_x^\ell(t) = v_x t + \sum_{k=1}^n A_k \cos\left(2\pi k f_0 \left(t + \frac{(\ell - 1)T}{2}\right) + \phi_k\right) + \mathbf{J}_{x0}^\ell \tag{29}$$

analogously for $\mathbf{J}_y^\ell(t)$. Thus, the projection of the leg joints on the lateral plane is obtained with an optimized procedure in the following way

$$\breve{\mathbf{J}}^\ell(t) = \left[h_1 \; h_2 \; h_3\right] g\left(t + \frac{(\ell - 1)T}{2} : f_0, \mathbf{D}_\ell, v_x, v_y, F\right) \tag{30}$$

where $g(t)$ is the bilateral Fourier series function with coefficients F and h are the values of the inverse normalization transform matrix.

Therefore, starting from a video sequence from a single camera and without any calibration, the proposed markerless system, in junction with [16], estimates the gait parameters projected on the lateral plane.

2.3 View-point invariant gait identification

The processing and derivation of good gait features from this trajectory-based data, is still an unsolved and challenging problem due to the complexity of the human visual perception system as well as the compound nature of gait motion inherent in the numerous variables associated with it including kinematics, kinetics and anthropometrics [8, 42]. An important issue in gait recognition is the derivation of appropriate features that can capture the discriminative individuality from the subject's gait. Such features should respond to crucial criteria such as robustness and invariance to weather conditions, clothing and operating conditions.

In order to identify a subject by their gait, we derive the angular measurements, anthropometric measurements as well as the trunk spatial displacement which best describe the gait kinematics. The use of angular motion is very common in gait analysis and recognition. The angles of the joints including the hip and the knee; are considered the most important kinematics of the lower limbs. The anthropometric measurements include the subject height and lengths of the lower limbs. Feature selection is employed to derive as many discriminative cues as possible whilst removing the redundant and irrelevant gait features which may degrade the recognition rate. It is practically infeasible to run an exhaustive search for all the possible combinations of features in order to obtain the optimal subset for recognition due to the high dimensionality of the feature space. For this reason, we employed the Adaptive Sequential Forward Floating Selection (ASFFS) search algorithm. The algorithm uses a validation-based evaluation criterion which is proposed to find the subset of features that minimises the classification errors as well as ensure good separability between the different classes. In contrast to the voting scheme used in the KNN, the evaluation function uses different weights w to signify the importance of the most nearest neighbours. The probability score for a sample s_c to belong to class c is expressed in the following equation (31):

$$f(s_c) = \frac{\sum_{i=1}^{N_c-1} z_i w_i}{\sum_{i=1}^{N_c-1} w_i} \tag{31}$$

where N_c is the number of instances in class c, and the weight w_i for the ith nearest instance is related to proximity as:

$$w_i = (N_c - i)^2 \tag{32}$$

The value of z_i is defined as:

$$z_i = \begin{cases} 1 & \text{if } nearest(s_c, i) \in c \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

such that the $nearest(s_c, i)$ function returns the ith nearest instance to the sample s_c. The Euclidean distance metric is employed to find the nearest neighbours.

3 Experimental tests

The proposed approach for subject identification from different view-points has been tested both on synthetic and on real data. Different video properties have been

Table 1 Angle's errors (in terms of mean and standard deviation) on synthetic gait with different spatial resolution

Image resolution/ mean sil's height	m_e (deg)	σ_e (deg)	Proc. time (s/frame)
500×490/200	2.04	2.12	1.44
250×245/100	2.80	2.15	0.97
163×163/50	3.06	3.56	0.37

considered in order to compare the performance of the method with respect to the cameras' properties.

3.1 Performance analysis of gait feature extraction

A quantitative validation of gait analysis has been obtained in a virtual environment: synthetic video sequences (500×490 pixels, 30fps) of a humanoid walking for one gait cycle have been created with Poser 7® (by e-frontier) in order to provide the ground truth for validation purposes. Different spatial resolutions, image qualities and compressions have been considered. A range of spatial resolutions (250×245, 163×163 pixels) have been considered by sub-sampling the original images with factors 2 and 3. Table 1 shows the comparative results in terms of mean (m_e) and standard deviation (σ_e) of joint's angles compared to the ground truth over the gait cycle.

The results, with a mean value of 2.6 ± 32.61, are particularly encouraging and present the same magnitude to the ones obtained with 3D markerless systems or 2D complex model based ones [24, 46]. Reducing the image resolution and silhouette's height clearly affect the algorithm accuracy, as predictable. On the other hand, a 50-pixels-height silhouette still maintains an acceptable result. Furthermore, Table 1 shows the mean time for processing one frame (Matlab7® code processed with 2GHz Intel CPU): the model-free proposed approach allows to extract the kinematic information with computational cost per silhouette's pixel lower than $2 \cdot 10^{-4}$s.

In addition, to simulate background subtraction imprecision, zero-mean Gaussian noise has been added to the synthetic images. The standard deviation (σ_n) of the added noise varies from 0 (original data) to 150 corresponding to PSNR from ∞ to 7.01dB.

Results in Table 2 show m_e and σ_e with respect to the different PSNR. Obviously the error increases with the level of noise added to the frames but it is lower than 5 degrees even at high noise levels. The noise robustness allows to understand how the proposed method depends on an accurate silhouette extraction and thus to extend the approach in a real context.

Table 2 Angle's errors (in terms of mean and standard deviation) on synthetic gait with different zero-mean Gaussian noises

σ_n / PSNR(dB)	m_e (deg)	σ_e (deg)
0/∞	2.04	2.12
5/28.12	2.30	2.13
25/24.25	3.01	3.38
50/21.89	3.30	3.72
100/15.23	3.25	3.68
150/7.01	4.25	3.49

Fig. 2 Mean angle's errors on synthetic gait with different image compressions
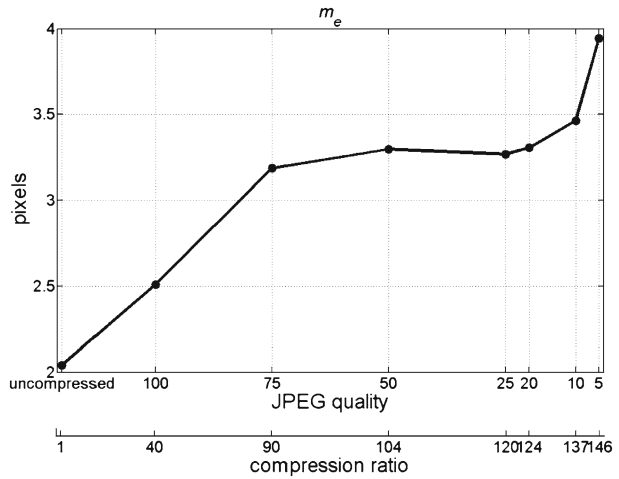
Since images from a network of cameras can be compressed for data storage problems, different JPEG compressions have been compared. Seven quality factors between 100 to 5 have been used for compressing the synthetic images, allowing a compression ratio $CR = S_u/S_c$ between 40 to 146, where S_u and S_c are the sizes of the uncompressed and compressed images respectively.

Figure 2 shows m_e with respect to the image compression. The value obtained with uncompressed images is also reported. The mean error increases rapidly with the compression, then remaining constant at about 3.3 pixels for JPEG qualities between 75 and 25.

The proposed method for gait analysis has been tested on real data from the CASIA-B database where 12 young healthy subjects walk along straight lines with 6 different camera orientations (36°, 54°, 72°, 90°, 108°, 126°). The 90° corresponds to the side view walking direction as shown in Fig. 3. The video sequences have a spatial resolution and frame rate of $320{\times}240$ pixels and 25fps respectively. The method has been applied to the video sequences, the limbs pose has been estimated frame by frame and the hip and knee angles have been extracted for each camera position and for each subject. Figure 3 shows an example of the limbs pose estimation in the 6 directions. The frames have been compressed with different JPEG qualities and the joints' position has been compared with the one extracted from the uncompressed images. The mean joints' distance is then measured:

$$m_d = \frac{\sum_{i=1}^{I} \sum_{v=1}^{V} \sum_{j=1}^{J} \sqrt{\left(x_{ivj}^{c} - x_{ivj}^{u}\right)^2 + \left(y_{ivj}^{c} - y_{ivj}^{u}\right)^2}}{I \cdot V \cdot J} \tag{34}$$

where the superscripts $'u'$ and $'c'$ indicate the joints' position from the uncompressed and compressed images respectively. I is the number of individuals, V the number of video sequence per person and J the number of tracked joints.

Figure 4 depicts m_d with respect to the image compressions. The results are in accordance with the one obtained on synthetic images and show an increase of vertex's de-location especially for qualities higher than 25. The results obtained with different compressions confirm the applicability of gait analysis in real
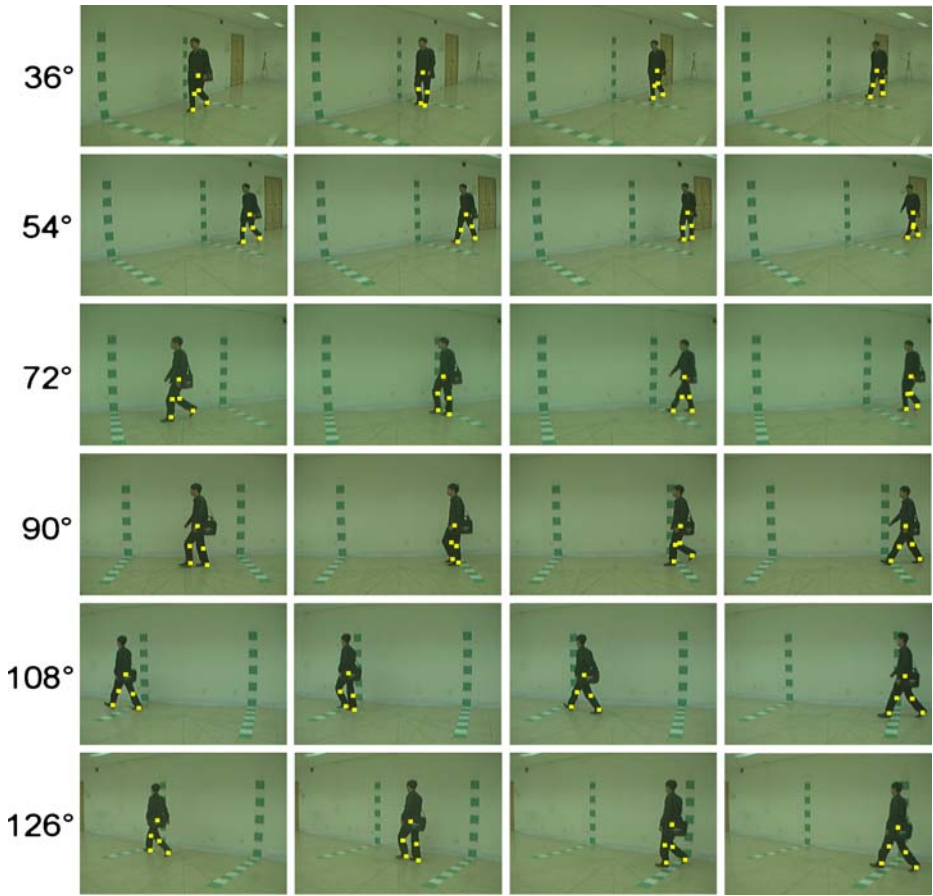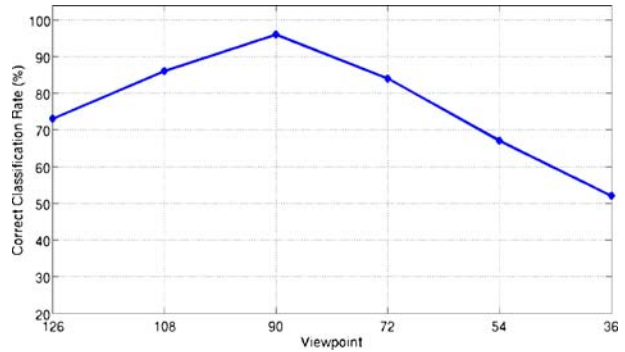
Fig. 3 Markerless joints extraction in different view points

Fig. 4 Mean joints' distance on real gait with different image compressions
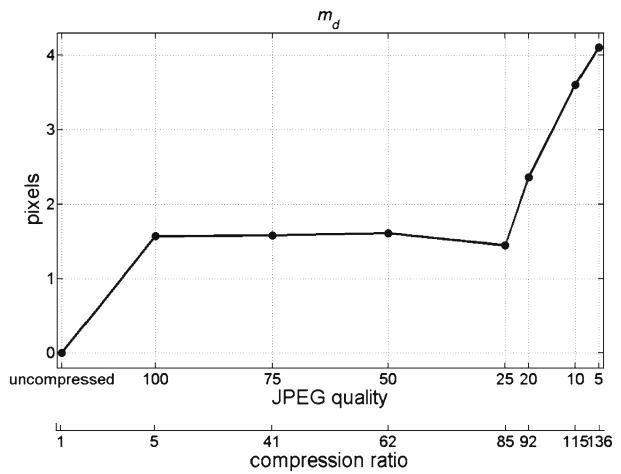
Fig. 5 Gait recognition from different viewpoints using non-rectified data


surveillance scenarios where cameras usually presents JPEG compression qualities higher than 50.

3.2 Viewpoint analysis of gait recognition

In order to assess the performance of gait recognition from different viewpoints using a single uncalibared camera, a set of 275 video sequences with 12 different subjects recorded at 6 viewpoints are taken from the Casia gait database [7]. The Correct Classification Rate (CCR) is computed using the K-nearest neighbour (KNN) classifier with the Leave-one-out cross-validation rule. The KNN rule is applied at the classification phase due to its low complexity and hence fast computation besides the ease of comparison to other methods. In the leave-one-out validation, every instance from the original sample is used for testing and is validated against the remaining observations. This is repeated for all the observations in the dataset. The recognition rate is computed as the average of all validations.

To investigate the effects of the viewpoint, an initial experimental is carried out to measure the recognition rate using the non-rectified gait data. The CCR is computed separately for every viewpoint. Figure 5 shows the variation of the CCRs with respect to the different viewpoints. The highest recognition rate is 96% for the side view whilst the recognition rate decreases when the walking direction changes from the lateral view. This is because of the variations in the extracted angular features from

Fig. 6 Mean angle's errors on synthetic gait with different image compressions
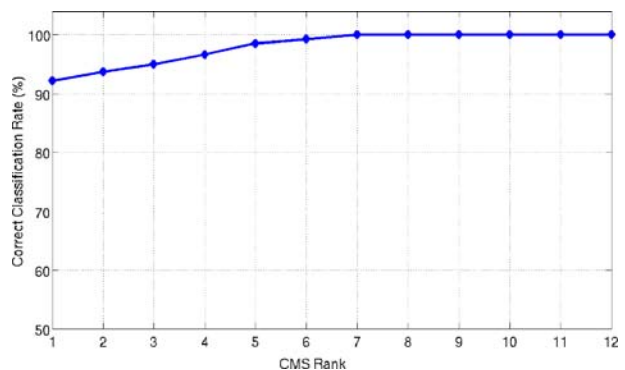
Table 3 Video frame rate effects on the performance of gait recognition

Video frame rate	Recognition rate
25 *fps*	92.2%
15 *fps*	81%

different views. Therefore, non-rectified gait angular data cannot be used directly for biometric applications. This was confirmed by the recent research study by Goffredo [15].

In the second experiment, the viewpoint rectification method described in Section 2.2 is applied to reconstruct the gait angular features into the lateral plane (i.e. sideview) Based on a set of 275 video sequences of walking subjects, a high recognition rate of 92.2% is achieved for all the viewpoints combined together. It is worth to note that the database has other covariate factors including clothing and load carriage. This is to signify the importance of the gait reconstruction into translating gait as a biometric method into real world surveillance and monitoring applications.

Another useful evaluation measure is the Cumulative Match Score (CMS) which was introduced by Phillips et al. in the FERET protocol [35] for the evaluation of face recognition algorithm. The measure assesses the ranking capabilities of the recognition system by producing a list of scores that indicates the probabilities that the correct classification for a given test sample is within the top n matched class labels. A classification score of 92.2% and 100% are achieved at the 1st and 7th rank respectively. The CMS score at 1st rank is the correct classification rate. Figure 6 shows the CMS curve of the gait identification for the rectified data.

Furthermore, a number of experiments are carried out using the same video set to investigate the algorithm potentials for data recorded at low frame rate. The performance error is simulated by dropping a number of frames from every 25 frames (25 is the original frame rate) of the video sequences which is equivalent to changing the frame rates. Table 3 shows the recognition rates for data recorded at different frame rates. It is observed that the algorithm performance is not much affected even when dropping 40% of the frames as the algorithm predicts the joint positions for the missing frames through the use of data fitting.

4 Conclusions

A smart surveillance system aims to automatically track people across a network of cameras and detect abnormal behaviours. In this context, the subject's identification and tracking between the cameras is a critical point. Possible solutions found in the literature are based on colour or appearance information. However, most approaches to a date limit their application to a short time period and environments where subjects do not wear uniforms.

We have proposed gait analysis as a solution for subjects' identification across a network of cameras. The completely unobtrusiveness without any subject cooperation or contact for data acquisition make gait particularly attractive for identification purposes in camera handover.

A novel 2D markerless view-independent gait analysis algorithm has been presented: the method does not need camera calibration or prior knowledge of subject pose. Since the choice of the cameras' characteristics is a key-point for the development of a smart surveillance system, the performance of the proposed approach has been measured with respect to different video properties.

Tests on synthetic and on real video sequences allowed performance evaluation of the proposed approach with respect to different spatial resolution, frame-rate, data compression and image quality. The obtained results show that gait analysis can be efficiently used for view-independent subjects' identification with commercially available video cameras.

References

1. BenAbdelkader C, Davis LS, Cutler R (2002) Motion-based recognition of people in eigengait space. In: Proc IEEE conf AFG, pp 267–274. http://csdl.computer.org/comp/proceedings/fgr/2002/1602/00/16020267abs.htm
2. Bhanu B, Han J (2003) Human recognition on combining kinematic and stationary features. In: Proc int conf AVBPA, pp 600–608
3. Black J, Ellis T, Rosin P (2002) Multi view image surveillance and tracking. In: Workshop on motion and video computing. Proc, pp 169–174. doi:10.1109/MOTION.2002.1182230
4. Bouchrika I, Nixon MS (2007) Gait-based pedestrian detection for automated surveillance. In: International conference on computer vision systems
5. Bouchrika I, Nixon MS (2007) Model-based feature extraction for gait analysis and recognition. In: LNCS, vol 4418. Springer, New York
6. Cai Y, Chen W, Huang K, Tan T (2007) Continuously tracking objects across multiple widely separated cameras. In: ACCV (1), pp 843–852
7. Casia gait database (2006). http://www.sinobiometrics.com
8. Chau T (2001) A review of analytical techniques for gait data. Part 1: fuzzy, statistical and fractal methods. Gait Posture 13(1):49–66
9. Chilgunde A, Kumar P, Ranganath S, WeiMin H (2004) Multi-camera target tracking in blind regions of cameras with non-overlapping fields of view. In: British machine vision conference
10. Collins R, Gross R, Shi J (2002) Silhouette-based human identification from body shape and gait. In: Proc. IEEE conf. on automatic face and gesture recognition, pp 351–356
11. Cutting J, Kozlowski L (1977) Recognizing friends by their walk: gait perception without familiarity cues. Bull Psychon Soc 9:353–356
12. Dempster WT, Gaughran GRL (1965) Properties of body segments based on size and weight. Am J Anat 120:33–54
13. Fisher R (2002) Self-organization of randomly placed sensors, p IV: 146 ff
14. Gilbert A, Bowden R (2006) Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity, pp 125–136
15. Goffredo M, Seely RD, Carter JN, Nixon MS (2008) Markerless view independent gait analysis with self-camera calibration. In: Proc. IEEE conf. on automatic face and gesture recognition
16. Goffredo M, Spencer N, Pearce D, Carter JN, Nixon MS (2007) Human perambulation as a self calibrating biometric. In: LNCS, vol 4778. Springer, New York
17. Gross R, Shi J (2001) The cmu motion of body (mobo) database. Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh
18. Haralick RM, Shapiro LG (1992) Computer and robot vision, vol 1. Addison-Wesley, Reading
19. Hu W, Tan T, Wang L, Maybank S (2004) A survey on visual surveillance of object motion and behaviors. IEEE Trans Syst Man Cybern 34(3):334–352. doi:10.1109/TSMCC.2004.829274

20. Huang P, Harris C, Nixon M (1999) Recognising humans by gait via parametric canonical space. J Artif Intell Eng 13(4):359–366
21. Huang T, Russell S (1997) Object identification in a Bayesian context. In: IJCAI97, pp 1276–1282
22. Ilie A, Welch G (2005) Ensuring color consistency across multiple cameras. In: Tenth IEEE conf. on computer vision, vol 2, pp 1268–1275. doi:10.1109/ICCV.2005.88
23. Javed O, Rasheed Z, Shafique K, Shah M (2003) Tracking across multiple cameras with disjoint views. In: Proc. ninth IEEE conf. on computer vision, vol 2, pp 952–957. doi:10.1109/ICCV.2003.1238451
24. Kale A, Chowdhury A, Chellappa R (2003) Towards a view invariant gait recognition algorithm. In: Proc IEEE conf AVSS, pp 143–150
25. Kang S, Min J, Paik JK (2001) Multiple-camera tracking system with seamless object handover. SPIE, pp 27–34. doi:10.1117/12.424949. http://link.aip.org/link/?PSI/4303/27/1
26. Khan S, Shah M (2003) Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Trans Pattern Anal Mach Intell 25(10):1355–1360. doi:10.1109/TPAMI.2003.1233912
27. Kettnaker V, Zabih R (1999) Bayesian multi-camera surveillance. IEEE Conf Comput Vision Pattern Recogn 2:259. doi:10.1109/CVPR.1999.784638
28. Makris D, Ellis T, Black J (2004) Bridging the gaps between cameras. In: Proc IEEE Comput Vis Pattern Recognit 2:II–205–II–210. doi:10.1109/CVPR.2004.1315165
29. Middleton L, Wagg DK, Bazin AI, Carter JN, Nixon MS (2006) Developing a non-intrusive biometric environment. In: Proc IEEE int conf IROS, pp 723–728. doi:10.1109/IROS.2006.282619
30. Murray MP, Drought AB, Kory RC (1964) Walking patterns of normal men. J Bone Joint Surg 46:335
31. Niu C, Grimson E (2006) Recovering non-overlapping network topology using far-field vehicle tracking data. In: ICPR (4), pp 944–949
32. Nixon M, Aguado A (2007) Feature extraction & image processing, 2nd edn. Academic, London
33. Nixon MS, Carter JN (2006) Automatic recognition by gait. Proc IEEE 94(11):2013–2024
34. Orrite-Urunuela C, del Rincon J, Herrero-Jaraba J, Rogez G (2004) 2d silhouette and 3d skeletal models for human detection and tracking. Proc IEEE Conf ICPR 4:244–247
35. Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The FERET evaluation methodology for face recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104
36. Phillips P, Sarkar S, Robledo I, Grother P, Bowyer K (2002) The gait identification challenge problem: data sets and baseline algorithm. Proc Int Conf Pattern Recogn 1:385–388. doi:10.1109/ICPR.2002.1044731
37. Rahimi A, Dunagan B, Darrell T (2004) Simultaneous calibration and tracking with a network of non-overlapping sensors. Proc Comput Vision Pattern Recogn 1:I–187–I–194. doi:10.1109/CVPR.2004.1315031
38. Sardis M, Anagnostopoulos V, Doulamis N (2009) Applied surveillance using biometrics on agents infrastructures. In: 5th Conference on artificial intelligence applications and innovations, Greek, April 2009
39. Shutler J, Grant M, Nixon MS, Carter JN (2002) On a large sequence-based human gait database. In: Proc int conf recent advances in soft computing, pp 66–72
40. Stillman S, Essa I (2001) Towards reliable multimodal sensing in aware environments. In: PUI '01: proc. workshop on perceptive user interfaces. ACM, New York, pp 1–6 doi:10.1145/971478.971499
41. Tieu K, Dalley G, Grimson W (2005) Inference of non-overlapping camera network topology by measuring statistical dependence. In: Tenth IEEE conf. on computer vision, vol 2, pp 1842–1849. doi:10.1109/ICCV.2005.122
42. Veres G, Gordon L, Carter JN, Nixon MS (2004) What image information is important in silhouette-based gait recognition? Proc IEEE Conf CVPR 2:II–776–II–782
43. Wang L, Tan T, Hu W, Ning H (2003) Automatic gait recognition based on statistical shape analysis. IEEE Trans IP 12(9):1120–1131
44. Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identification. IEEE Trans PAMI 25(12):1505–1518
45. Yamada Y (1999) Advanced method for improvement of obscure video image. In: Proc. IEEE 33rd annual carnahan conference on security technology, pp 440–445
46. Zhang R, Vogler C, Metaxas D (2007) Human gait recognition at sagittal plane. J Image Vis Comput 25(3):321–330

Michela Goffredo was born in Rome, Italy, on May the 12nd 1978. She received the Electronic Engineering degree from the University of Roma Tre in 2003. Her thesis dealt with markerless human posture analysis.

She got the PhD in Biomedical Electronics at the Department of Applied Electronics, University of Roma Tre, where she was involved in research on Image Processing for Human Movement Analysis.

She is presently working on Gait Biometrics at the School of Electronics and Computer Science of the University of Southampton and on Arm Motor Control at the Department of Applied Electronics, University of Roma Tre. Her research interests are in image and video processing and human motion analysis.

Imed Bouchrika received the Bsc and PhD degrees in Electronics and Computer Science from the University of Southampton (United Kingdom) in 2004 and 2008 respectively. Since 2008, he is working as a research fellow at the Information: Signals, Images, Systems (ISIS) Research Group of the University of Southampton. His research areas are image processing, gait biometrics and visual surveillance.

John N. Carter graduated with a BA in Experimental Physics from Trinity College, Dublin, Ireland followed by a PhD in Astrophysics from the University of Southampton England. In 1985 he changed discipline and joined the Department of Electronic Engineering also at Southampton as a lecturer researching in Signal and Image Processing. He is currently a Senior Lecturer in the Information: Signals, Images and Systems Research Group. In the past he has worked on programs as diverse as Diesel Engine Diagnostics and Vocal Tract imaging. A recent success in this field has been the development of a new dynamic form of Magnetic Resonance Imaging, which makes it possible to reconstruct high time resolution multi-planar views of the human vocal tract while a subject is repeating a short phrase. His current major interests are in the general area of Four-Dimensional Image Processing. That is, analysing sequences of images to extract both two and threedimensional features, exploiting coherence over the whole sequence, i.e. imposing simple smoothness and continuity constraints. This has applications in object tracking and feature detection where it is tolerant of high levels of noise and missing data. This has found application in biometrics, in particular Automatic Gait Analysis. Dr. Carter is a member of the IEEE.

Mark S. Nixon is the Professor in Computer Vision at the University of Southampton UK. His research interests are in image processing and computer vision. His team develops new techniques for static and moving shape extraction which have found application in automatic face and automatic gait recognition and in medical image analysis. His team were early workers in face recognition, later came to pioneer gait recognition and more recently joined the pioneers of ear biometrics. Amongst research contracts, he was Principal Investigator with John Carter on the DARPA supported project Automatic Gait Recognition for Human ID at a Distance.

He chaired BMVC 98 and with Josef Kittler he chaired the Audio Visual Biometric Person Authentication (AVBPA 2003), was Publications Chair for the International Conference on Pattern

Recognition (ICPR 2004) and co-chaired the IEEE 7th International Conference on Face and Gesture Recognition FG2006. He has recently been program chair on BTAS '08 and ICB '09.

His vision book, co-written with Alberto Aguado, Feature Extraction and Image Processing reached its Second Edition in 2007 and with Tieniu Tan and Rama Chellappa, his book Human ID based on Gait which is part of the new Springer Series on Biometrics, was published in 2005. Dr. Nixon is a member of the IEEE.