Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Image and Vision Computing xx (2006) 1–14
image AND vision COMPUTING
www.elsevier.com/locate/imavis

Zernike velocity moments for sequence-based description of moving features

J.D. Shutler [a,*], M.S. Nixon [b]

[a] Remote Sensing Group, Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, UK
[b] Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

Received 20 September 2004; received in revised form 14 December 2005; accepted 15 December 2005

Abstract

The increasing interest in processing sequences of images motivates development of techniques for sequence-based object analysis and description. Accordingly, new velocity moments have been developed to allow a statistical description of both shape and associated motion through an image sequence. Through a generic framework motion information is determined using the established centralised moments, enabling statistical moments to be applied to motion based time series analysis. The translation invariant Cartesian velocity moments suffer from highly correlated descriptions due to their non-orthogonality. The new Zernike velocity moments overcome this by using orthogonal spatial descriptions through the proven orthogonal Zernike basis. Further, they are translation and scale invariant. To illustrate their benefits and application the Zernike velocity moments have been applied to gait recognition—an emergent biometric. Good recognition results have been achieved on multiple datasets using relatively few spatial and/or motion features and basic feature selection and classification techniques. The prime aim of this new technique is to allow the generation of statistical features which encode shape and motion information, with generic application capability. Applied performance analyses illustrate the properties of the Zernike velocity moments which exploit temporal correlation to improve a shape's description. It is demonstrated how the temporal correlation improves the performance of the descriptor under more generalised application scenarios, including reduced resolution imagery and occlusion.

Keywords: Statistical moments; Zernike moments; Motion; Velocity moments; Gait

1. Introduction

Moving object description is a growing area of computer vision research, traditionally an arena dominated by tracking algorithms. The developments in this area were previously limited not least by the storage requirements of image sequences. With the advance of digital video (DV), and the explosion of storage capacities, the analysis and storage of image sequences has become viable, enabling increased interest. Tracking algorithms [1] generally locate the region or feature of interest in the first frame and then track it throughout the remainder of the sequence. This requires good initialisation in the first image and assumes that in later images tracked objects are not overcome by noise or occlusion. This kind of approach enables real-time performance, a major benefit of these algorithms. With the ever increasing available computing power, alternative approaches that process the complete image-sequence are appearing. For example, the velocity Hough transform for conic sections [2] and its extension for arbitrary shapes [3] process a complete image sequence, overcoming the problems of image noise and occlusion by exploiting temporal correlation, treating the image sequence as a single entity rather than individual images. These approaches locate the perimeter of a moving shape by searching for a particular motion. However, a great deal of information can be held within a shape's perimeter—motivating techniques enabling holistic moving shape description.

Statistical moments, e.g. [4] describe a shape with respect to its axes, producing holistic descriptions encoding information including mass, centroid and variation across axes. Mukundan [5] provides descriptions of most of the current moment techniques, along with background information and applications. In general the different types of moments fall into two categories, orthogonal and non-orthogonal. Orthogonal moments produce features that are less correlated than their non-orthogonal counterparts. Further, the orthogonality property enables simple, accurate signal reconstruction from the generated moments. Moments that are non-orthogonal tend

* Corresponding author. Address: Remote Sensing Group, Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, UK.
Tel.: +44 1752 633417.
E-mail addresses: jams@pml.ac.uk, jamie@zepler.org (J.D. Shutler).

0262-8856/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.imavis.2005.12.001

to be simpler to implement, computationally less expensive and include descriptors that have a range of useful properties, i.e. scale, translation and rotation invariance. Their highly correlated features (as a result of their non-orthogonal nature) make reconstruction more difficult. This correlation requires the need for high accuracy in the calculations when interested in the high frequency components of the image and/or when analysing large datasets.

There have been many studies using two-dimensional moments for image recognition purposes. However, to date, most applications use single images. Hoey [6] used Zernike polynomials to study facial motion by generating flow fields which provided input to hidden Markov models. Little [7] used moments to characterise optical flows between images for gait recognition. These techniques still only link adjacent images, and do not consider the complete sequence. Rosales [8] described motion by producing one image that contained information from a complete sequence, building on the work done by Davis [9]. Rosales's system was based on Hu [10] invariant moments and was used to recognise types of motion, e.g. sitting down or kicking; due to several images being compressed into one, subtle differences between subjects are lost due to self occlusion and overlapping of data.

For this work, we began by looking at a traditional statistical method of moments to describe the motion of a person through multiple images. Unfortunately, this does not provide a very detailed description of the motion, as there is no information linking the images of the sequence, since they are treated as separate entities. By using the general theory of moments a method has been developed that not only contains information about the pixel structure of the moving object, but also how its movement flows between images. Through analysing image sequences the temporal information can be exploited and the possibility of describing deforming shapes becomes apparent. Accordingly, we describe a new technique called velocity moments, enabling the holistic statistical description of temporal image sequences. We present this new technique to enable the application of statistical moments to image sequences. To aid its characterisation while demonstrating its beneficial attributes, we apply it to human gait recognition, an emergent biometric.

This paper is structured as follows. Firstly, Section 2 briefly reviews non-orthogonal and orthogonal statistical moments. Velocity moments are then introduced in Section 3. Section 4 uses human gait classification to illustrate their application. Section 5 details the performance attributes of the Zernike velocity moments analysing the effects of reduced resolution imagery and occlusion. Conclusions are then drawn.

2. Background theory

Statistical moments are applicable to many different aspects of image processing, ranging from invariant pattern recognition and image encoding to pose estimation. Moments of an image [10], describe the image content (or distribution) with respect to its axes. They are designed to capture both global and detailed geometric information about the image. In continuous form an image can be considered as a two-dimensional Cartesian density distribution function $f(x,y)$. With this assumption, the general form of a moment of order $(p+q)$, evaluated over the complete image plane ξ is:

$$M_{pq} = \int\int_{\xi} \psi_{pq}(x,y) f(x,y) \mathrm{d}x\, \mathrm{d}y; \quad p,q = 0,1,2,\ldots,\infty \tag{1}$$

The *weighting kernel* or *basis* function is ψ_{pq}. This produces a weighted description of $f(x,y)$ over the entire plane ξ. The basis functions can have a range of useful properties that may be passed onto the moments, producing descriptions which can be invariant under rotation, scale, translation and orientation. For image analysis a discrete version is required, for this conversion we assume that ξ is divided into square pixels of dimensions $\Delta A = 1 \times 1$, with constant intensity I over each pixel so $P_{xy} = I(x,y)\Delta A$.

2.1. Non-orthogonal moments

Early work by Hu [10] applied statistical moments to image analysis defining the Cartesian moments which in discrete form are:

$$m_{pq} = \sum_{x=1}^{M}\sum_{y=1}^{N} x^p y^q P_{xy} \tag{2}$$

Extending them to include translation invariance Hu defined the Centralised moments

$$\mu_{pq} = \sum_{x=1}^{M}\sum_{y=1}^{N} (x-\bar{x})^p (y-\bar{y})^q P_{xy} \tag{3}$$

where M and N are the image dimensions, $p+q$ is the order and P_{xy} is the pixel value at position (x,y). $\bar{x}$ and $\bar{y}$ are the x and y centres of mass (COMs)

$$\bar{x} = \frac{m_{10}}{m_{00}} \qquad \bar{y} = \frac{m_{01}}{m_{00}} \tag{4}$$

which describe a unique position within the field of view.

Cartesian moments, Eq. (2) are formed using a monomial basis set $x^p y^q$. This basis set is non-orthogonal and this property is passed onto the Cartesian moments. These monomials increase rapidly in range as the order increases, producing highly correlated descriptions. This can result in important descriptive information being contained within small differences between moments, which can lead to the need for high computational precision.

2.2. Orthogonal moments

Moments produced using orthogonal basis sets also exist. These orthogonal moments have the advantage of needing lower precision to represent differences to the same accuracy as the monomials. The orthogonality condition also simplifies the reconstruction of the original function from the generated moments as each descriptor (or moment) is independent (uncorrelated). Many orthogonal sets exist (Legendre, Zernike,

Pseudo Zernike, etc.). However, the Zernike moment formulation appears to be one of the most popular, outperforming the alternatives [11] (in terms of noise resilience, information redundancy and reconstruction capability). The pseudo-Zernike formulation proposed by Bhatia and Wolf [12] further improved these characteristics. However, here we study the original formulation of these orthogonal invariant moments.

Complex Zernike moments [4] are constructed using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc $(x^2+y^2)\leq 1$. For a discrete image with current pixel P_{xy} the Complex Zernike moments are defined as

$$A_{mn} = \frac{m+1}{\pi} \sum_x \sum_y P_{xy}[V_{mn}(x,y)]^* \tag{5}$$

where $x^2 + y^2 \leq 1$

where $m=0,1,2,\ldots,\infty$ defines the order and (* denotes the complex conjugate. While n is an integer (that can be positive or negative) depicting the angular dependence, or rotation, subject to the conditions

$$m-|n| = \text{even}, \quad |n| \leq m \tag{6}$$

and $A^*_{mn} = A_{m,-n}$ is true. The Zernike polynomials $V_{mn}(x,y)$ expressed in polar coordinates are

$$V_{mn}(r,\theta) = R_{mn}(r)\exp(jn\theta) \tag{7}$$

where (r,θ) are defined over the unit disc, $j=\sqrt{-1}$ and $R_{mn}(r)$ is the orthogonal radial polynomial, defined as

$$R_{mn}(r) = \sum_{s=0}^{(m-|n|)/2} (-1)^s F(m,n,s,r) \tag{8}$$

and

$$F(m,n,s,r) = \frac{(m-s)!}{s!\left(\frac{m+|n|}{2}-s\right)!\left(\frac{m-|n|}{2}-s\right)!} r^{m-2s} \tag{9}$$

where $R_{mn}(r)=R_{m,-n}(r)$ and if the conditions in Eq. (6) are not met, then $R_{mn}(r)=0$. To calculate the Zernike moments, the image (or region of interest) is first mapped to the unit disc using polar coordinates (r,θ), where the centre of the image is the origin of the unit disc. Those pixels falling outside the unit disc are not used in the calculation. Translation and scale invariance can be achieved by normalising the image using the Cartesian moments prior to calculation of the Zernike moments [13] using

$$h(x,y) = f\left(\frac{x}{a}+\bar{x}, \frac{y}{a}+\bar{y}\right) \quad \text{where} \quad a = \sqrt{\frac{\beta}{m_{00}}} \tag{10}$$

β is the new predetermined mass, m_{00} is the shape's original mass, $\bar{x}$ and $\bar{y}$ are the shape's COMs and $h(x,y)$ is the new translated and scaled function. Due to the structure of the Zernike moments, the pixel descriptions are weighted in favour of their distance from the origin of the unit disc. Those pixels lying closer to the perimeter of the unit disc will have more weight than those lying closer to the origin. As r approaches unity the radial polynomials display steeper gradients and converge. The higher order polynomials (and their corresponding moments) will have improved capability to describe image detail due to their increased oscillations, especially in the region before convergence where their frequency increases. Thus, image detail which is encoded around the region of convergence will be more correlated. Finally, the absolute value of a Zernike moment is rotation invariant as reflected in the mapping of the image to the unit disc. Relationships between Cartesian and Zernike moments can be exploited to aid understanding and/or possible computation speed increases, e.g. [14].

3. Velocity moments

One method of developing a statistical moment technique to analyse image sequences is to stack the images into a three-dimensional XYT (x,y plus time) block, and then apply a 3D descriptor to these data. Data in this form could be described using conventional 3D moments [15], treating time as the z-axis. However, this method confounds the separation of the time and space information, as they are embedded in the data and not specific to the descriptor. Time is fundamentally different from space, thus, we intend to acknowledge this by treating it separately. To analyse image sequences we reformulate the moment descriptor to incorporate time, enabling the separation and/or combination of the time and spatial descriptions. To achieve this, a method of motion description within the moment basis is required. The COM describes a unique global position within the field of view. The COM is guaranteed to exist, independent of the distribution, allowing the use of this low order moment as the basis of a generic framework, as previously established with the centralised moments [10]. The difference between consecutive COM descriptions in an image sequence enables a description of motion in either axis.

Our new velocity moments are based around the COM description and are primarily designed to describe a moving and/or changing shape in an image sequence. The method enables the structure of a moving shape to be described, together with associated motion information. The velocity moments are calculated from a sequence of images. Their generalised structure is:

$$A_{mn\alpha\gamma} = \sum_{i=2}^{I} \sum_x \sum_y \text{USP}_{i_{xy}} \tag{11}$$

The shape's structure (in each image i) contributes through each pixel $P_{i_{xy}}$ and the weighting function S. Here, S is either a centralised Cartesian polynomial [10], or a Zernike polynomial [4]. Motion, or velocity, is introduced through U as the differences between consecutive COMs in the image sequence. The Cartesian monomials were first studied due to their simplicity and ease of computation. The orthogonal Zernike moments are well-established in pattern recognition, providing an ideal platform to enable the analysis of the new framework on an orthogonal basis.

3.1. Cartesian velocity moments

The Cartesian velocity moments [16] are computed from a sequence of I ($M\times N$) images as

$$\mathrm{vm}_{pq\alpha\gamma} = \sum_{i=2}^{I}\sum_{x=1}^{M}\sum_{y=1}^{N} U(i,\alpha,\gamma)S(i,p,q)P_{i_{xy}} \tag{12}$$

where $S(i,p,q)$ arises from the centralised moments

$$S(i,p,q) = (x-\bar{x}_i)^p(y-\bar{y}_i)^q \tag{13}$$

and $U(i,\alpha,\gamma)$ introduces velocity as

$$U(i,\alpha,\gamma) = (\bar{x}_i-\bar{x}_{i-1})^{\alpha}(\bar{y}_i-\bar{y}_{i-1})^{\gamma} \tag{14}$$

$\overline{x_i}$ is the current COM in the x-direction, while $\bar{x}_{i-1}$ is the previous COM in the x-direction, $\bar{y}_i$ and $\bar{y}_{i-1}$ are the equivalent values for the y-direction. The image sequence begins at image index $i=1$; summation commences at $i=2$ to ensure that the first velocity calculation $U(2,\alpha,\gamma)$ is defined, achieving invariance to the start position. It can be seen that the equation can easily be decomposed into averaged centralised moments (vm_{1100}), and then further into an averaged Cartesian moment (vm_{1100} with $\bar{x}_i=\bar{y}_i=0$). The velocity moments for which $\alpha=\gamma=0$ are

$$\mathrm{vm}_{pq00} = \sum_{i=2}^{I}\sum_{x=1}^{M}\sum_{y=1}^{N} (x-\bar{x}_i)^p(y-\bar{y}_i)^q P_{i_{xy}} \tag{15}$$

which are the averaged centralised moments. Setting $p=q=0$ produces

$$\mathrm{vm}_{00\alpha\gamma} = \sum_{i=2}^{I}\sum_{x=1}^{M}\sum_{y=1}^{N} (\bar{x}_i-\bar{x}_{i-1})^{\alpha}(\bar{y}_i-\bar{y}_{i-1})^{\gamma} P_{i_{xy}} \tag{16}$$

which is a summation of the difference between COMs of successive images (i.e. the distance travelled). The structure of Eq. (12) allows the image structure to be described together with velocity information from both the x- and y-directions. To produce velocity values in pixels per image, Eq. (12) is normalised according to

$$\overline{\mathrm{vm}}_{pq\alpha\gamma} = \frac{\mathrm{vm}_{pq\alpha\gamma}}{A(I-1)} \tag{17}$$

where A is the average area (number of pixels) of the moving object.

3.2. Zernike velocity moments

The Zernike velocity moments [17] are expressed as:

$$A_{mn\alpha\gamma} = \frac{m+1}{\pi}\sum_{i=2}^{I}\sum_{x}\sum_{y} U(i,\alpha,\gamma)S(m,n)P_{i_{xy}} \tag{18}$$

They are bounded so that $(x^2+y^2)\leq 1$, while the shape's structure contributes through the orthogonal polynomials:

$$S(m,n) = [V_{mn}(r,\theta)]^* \tag{19}$$

Velocity is introduced as before (Eq. (14)), and normalisation is produced by substituting $\mathrm{vm}_{pq\alpha\gamma}$ with $A_{mn\alpha\gamma}$ in Eq. (17). The coordinate values for $U(i,\alpha,\gamma)$ are calculated using the Cartesian moments and then translated to polar coordinates. If we consider first the horizontal motion case only, the angle of the vector θ for a difference in x position is either 0 or π radians. The value used is dependent on the direction of movement. If the movement is left to right then $x=r\cos(0)=r$ where r is the length of the vector from the previous COM to the current COM, i.e. the velocity in pixels/image. Alternatively, if the movement is right to left then $x=-r$. The mapping to polar coordinates results in a sign change that can be used to detect the direction of motion. Similarly for the y-direction velocity, the values of θ are either $\pi/2$ or $3\pi/2$ radians producing motion of r and $-r$ pixels, respectively.

It is possible (through a minor modification to Eq. (18)) to produce rotation invariant Zernike velocity moments. This modification increases the correlation of the descriptions, a characteristic that may not be desirable for classification problems. The total number of Zernike velocity moments M_{T} for order m, rotations $n>0$, velocities α and γ, subject to the conditions of Eq. (6) are:

$$M_{\mathrm{T}} = mn\alpha\gamma \tag{20}$$

Zernike moments are orthogonal producing less correlated descriptors than Cartesian moments. One advantage of this is the ability to easily reconstruct the original signal from the moment values. The orthogonality property is passed onto the spatial descriptions of the velocity moments. Intuitively for the Zernike velocity moments it can be seen that each individual image's spatial descriptions remain orthogonal, just weighted by velocity $U(i,\alpha,\gamma)$. However, the overall description of the sequence becomes correlated, due to the high similarity between consecutive images.

If we consider Zernike velocity moments describing just spatial information (no motion) of a moving rigid shape, then the correlation between images is exploited, and is advantageous, refining the description of the rigid shape as the sequence increases in length. The final Zernike velocity moments of this sequence can be considered as refined (or averaged) Zernike moments of a single image, the descriptions of which are orthogonal. Alternatively, if the shape is moving and deforming (non-rigid), such as a person walking, then the spatial correlation between consecutive image descriptions is reduced. The Zernike velocity moments are a weighted sum of the Zernike moments over multiple consecutive images. The weighting (velocity) is real-valued and scalar, and the spatial description of each consecutive image in the sequence are orthogonal. The final descriptors of the moving and deforming shape are temporally correlated due to the use of the image sequence.

In the simple case of a rigid shape the motion information may not be of interest. However, if there are larger changes in shape between consecutive images, and if the motion in the sequence is non-linear, then this information becomes potentially more interesting. In this case, the intra-sequence motion is linked to each image's spatial description and can be

exploited. Section 4 on human gait analysis aims to exploit these properties, using the velocity moments to describe a temporal image sequence of a shape, which (as the sequence progresses) alters in both composition and motion.

3.3. Cartesian versus Zernike

Due to the monomials around which the Cartesian velocity moments are based, they will produce descriptors which are highly correlated. Therefore, when wanting to distinguish between features generated from a large database, the need for high precision in the calculation becomes increasingly more important (refer to Section 2.1). The point at which this becomes an issue will be application dependent, governed by the size of the database and the features of interest. Alternatively, the orthogonality of the Zernike velocity moments provides less correlated descriptions, even when analysing large databases. These less correlated descriptions provide improved performance in the presence of noise in comparison to the Cartesian moments. Furthermore, due to the less correlated descriptions a lower accuracy in the calculation can be used to achieve the equivalent descriptive power to that of a set of Cartesian velocity moments. Rotational invariance also becomes a possibility.

3.4. Interpreting the velocity moments

The exact descriptions captured by statistical moments can be difficult to explain intuitively, especially as their order increases. Due to their simplistic nature the Cartesian moments allow some interpretation, for example Cartesian and centralised moments have been shown to capture descriptions of spatial symmetry and asymmetry [18]. However, the complicated nature of the Zernike moment descriptions is more difficult to explain intuitively. Furthermore, higher order moments (spatial orders >4) become increasingly difficult to interpret, as these will describe the higher frequency components of the image. Here, follows some descriptions of the velocity moments towards aiding the reader's interpretation.

The velocity components (differences between COMs, Eq. (14)) are determined using the centralised moments and thus allow some intuitive interpretation. If we consider the velocity moments A_{0010} and vm_{0010} containing purely motion information, these both describe the mean between-image x-direction velocity. A_{0020} and vm_{0020} describe the averaged magnitude of between-image x-direction velocity (or a measure of absolute range of motion or variance). A_{0030} and vm_{0030} will both give a description of the kurtosis (or asymmetry) of the between-image x-direction velocity. Alternatively, the Cartesian velocity moment vm_{2010} provides a description of the mean between-image x-direction motion coupled with the corresponding x-direction spatial variance of each image in the sequence (measured about the centre of mass). For example, if analysing a sequence of images of a walking person (as viewed from the side), vm_{2010} will describe the mean x-direction motion between consecutive images coupled with the range in pixel spread of each image. This produces a description of the forward motion coupled with the spread (or swing) of the subjects' limbs. In the same way, vm_{3020} provides a description of the mean spatial kurtosis the sequence's image's x-axis distribution coupled with the range in x-direction motion between each image pair. It is important to note that $\mathrm{vm}_{3000}+\mathrm{vm}_{0020}\neq\mathrm{vm}_{3020}$ (and likewise for A_{3020}), as the motion and spatial description for each image pair are linked within the calculation, rather than the average motion and average spatial description within the complete sequence being combined, e.g. in this sense the velocity moments are not linear.

4. Human gait

Gait is defined as the 'manner of walking or forward motion' [19]. It is primarily determined by muscular and skeletal structure. One of the earliest documented examples of recognition by gait was Shakespeare who wrote in The Tempest [Act 4 Scene 1]

> "High'st Queen of state, Great Juno comes; I know her by her gait"

An early documented example of psychological gait observations was by Johansson [20] who attached point light displays onto specific points on a subject. Johansson then showed that people could distinguish human motion from the movement of the lights alone. Early computer vision work studied the mechanics of a subject's hip and leg motion for recognition, e.g. [21,22]. Interest has since increased in this emergent biometric, generating a plethora of different algorithms and approaches [7,23–30]. Naturally to allow its application as an operational biometric, interest has also turned to the study of the covariates of gait, looking at the effects of different viewing angles, footwear and the carrying of objects. The largest study to date being the HumanID gait challenge [31]. The HumanID study, (culminating in the gait challenge) enabled the development of improved databases, e.g. [31–33], two of which are analysed here.

In general, all of the computer vision gait recognition approaches thus far can be categorised into two groups: model-based and holistic approaches. Model-based approaches tend to be computationally expensive and model just the lower body motion, e.g. [7,30]. Holistic approaches use the whole, or complete body and/or motion as a cue to identity. For example, BenAbdelkader [23] applied eigen analysis to self-similarity maps and [29] applied principal component analysis and supervised learning techniques to spatial–temporal silhouettes. Also, temporal-symmetry [26] and area measures [25] have both demonstrated good classification results.

We are primarily interested in the recognition capability of the velocity moments when applied to sequences of moving features. Therefore, we demonstrate the application of the velocity moments using human gait classification, producing a holistic description of temporal motion.

4.1. Methodology

As a person walks, variations in both horizontal and vertical motions exist. This can be seen in Fig. 1 which shows the x and y COM variations for two sequences of the same subject, walking for one complete gait cycle (heel strike to heel strike). It can be seen that intra-subject variations exist for these two sequences. These differences are due to variations in the subject's walking, sampling issues and noise in the extraction process. Consequently, each variation in motion is linked to the spatial descriptions of each image. Therefore, by using the velocity moments we can produce descriptions that link both the person's motion and their corresponding shape, in each stage of their gait cycle.

Due to the nature of the encoding of information in the Zernike polynomial (Section 2.2), the Zernike moments will efficiently describe the extremities of the subject as they move. Details including the head, arms and legs will appear closer to the perimeter of the unit disc mapping than the torso. This means that the characteristics that are most likely to vary between subjects (i.e. leg, arm and head shape/movement) are described efficiently, whereas details including explicit torso shape will not be as efficiently encoded.

From a human vision approach, it is suggested that both shape and motion information are important when observing a person's gait. Motion can be split into two types, gross motion (e.g. the subject's overall forward movement) and intimate motion (e.g. the particular way in which a subject swings their arms as they move for ward). Thus, in our classification approach we apply the velocity moments to two different image sets. Firstly, a set of binary silhouettes, or spatial templates (STs) for each subject sequence is obtained by background removal. These will provide shape and gross motion information. Optical flow images or temporal templates (TTs) are then computed to provide intimate motion information. These dense optical flow fields [34] describe the motion in a local region around each pixel. The algorithm searches amongst a limited set of displacements for the displacement that minimises the absolute difference between the image patch in one image and the corresponding patch in the next image. This technique has previously been used in gait recognition [7,35] and produces results which are consistent with human psychophysics. We use the magnitude of the dense optical flow fields as Huang [36] demonstrated that for human gait the magnitude appeared most important.

The velocity moments are then calculated for these two image sets, STs and TTs. Due to the periodic nature of gait, analysis is performed on one complete gait cycle. Those velocity moments suitable for classification are then selected using the single-factor ANOVA technique and the Scheffe post-hoc test [37,38]. Due to the small number of gait sequences available per subject, the ANOVA method is only used as a guide as the resulting variance estimates will not be precise. The single-factor ANOVA selects features that singularly separate portions of the dataset, thus separation of the complete dataset can be achieved by combining features which complement each other. This analysis is to demonstrate the application of the velocity moments and we are not primarily interested in optimal feature set selection, nor classification. Therefore, final selection is achieved by ANOVA guided manual selection, resulting in possible non-optimal results. The selected moments are used to produce a multidimensional feature space for classification, rather than combining them prior to classification. For our purposes combining moment features prior to classification is avoided as this can introduce further problems, such as amplifying noise [39].

Finally, classification of these selected features is possible through a number of different methods. Here, we have chosen to use a simple classifier so as to avoid getting trapped in the intricacies of classifier theory. Thus, classification of the moment features is achieved using the *k-nearest neighbour* technique ($k=1$ and 3) using the leave one out rule with cross validation. Doubtless the overall classification results could be improved by using a more powerful technique. A description of each database is included and the results are presented in terms of classification analyses.

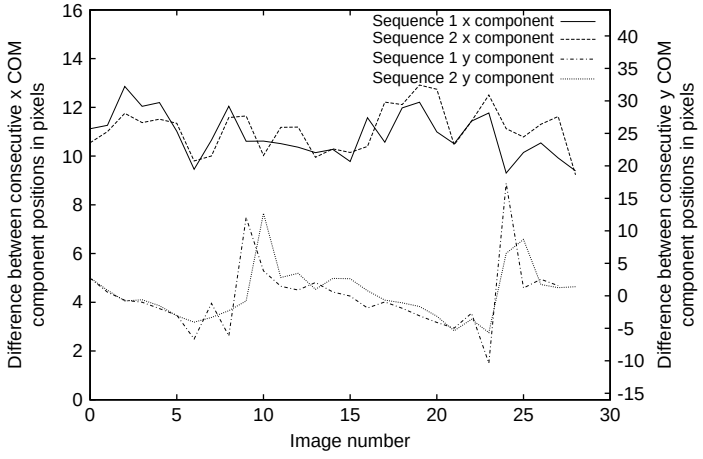

Fig. 1. Two sequences of the x and y COM variations for one subject's complete gait cycle (heel strike to heel strike of the same foot).

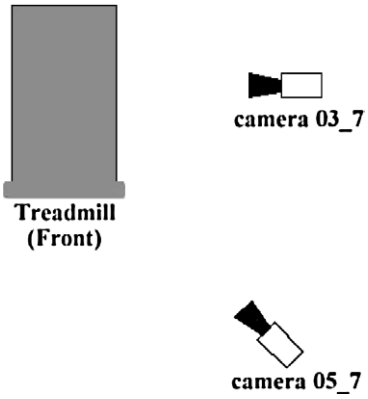

Fig. 2. A plan view of the treadmill and cameras for the CMU database.

4.2. Processing issues

When the subject is mapped onto the unit disc prior to the Zernike moment calculation for each image, care must be taken to ensure that no part of the subject's shape falls on the perimeter of, or outside the unit disc. The value of β for Eq. (10) is set so that the mapped pixels' coordinates are within 90% of the unit disc's radius. This also reduces the effect of the converging polynomials as r approaches unity (avoiding increased correlation in the descriptions).

4.3. CMU database

The Carnegie Mellon University Robotics Laboratory (CMU) database [32] (developed within the HumanID research program), consists of STs of 25 subjects walking on a treadmill under two different viewing geometries, with subjects walking at two different speeds. The computation of the TTs was not possible at the time of these analyses as the full database was only available as STs. Thus, we concentrate on the classification properties of the STs. For analysis we have partitioned the database into four different subsets. Two of the subsets view the subject from the side (normal to the subject's walking direction), the remaining two view from an oblique angle ($\simeq 45°$ from normal), as shown in Fig. 2 and summarised in Table 1.

The STs within the database were generated from the original colour data using a simple background subtraction technique. This result was median filtered to remove the effects of noise caused by variations in lighting, etc. producing the final STs as supplied within the database. Fig. 3 shows an example ST from the database and its original colour data (courtesy of CMU). All subjects have four sequences (per database subset) of them walking for one complete gait cycle,

Table 1
The partitioned CMU database

CMU subset	Camera	Walking speed	Viewing angle
CMU_03_7_s	03_7	Slow	Normal
CMU_03_7_f	03_7	Fast	Normal
CMU_05_7_s	05_7	Slow	Oblique
CMU_05_7_f	05_7	Fast	Oblique

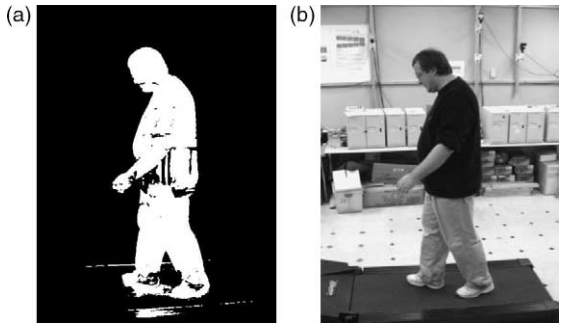

Fig. 3. Example (a) ST image from the CMU_03_7_s subset and (b) the original colour image.

heel strike to heel strike of the same foot, producing STs with no forward velocity. However, fluctuations about the mean x position may exist.

For the *CMU_03_7_s* subset $M_T = 784$ Zernike velocity moments ($n = m = 0 \ldots 12$, $\mu = \gamma = 0 \ldots 3$) were calculated on the 100 sequences. Due to the large amount of data this took multiple days to process on a cluster of eight 1 GHz machines. Consequently, this list was reduced to those moments whose Fisher statistic (F) satisfied $F > 30$, along with those moments which proved useful in previous analyses [39]. This reduced the moment list for the remaining three database subsets to 90, 43 of which included velocity information (both x and/or y). Table 2 shows the k-nn classification results for each database subset using all 90 velocity moments, showing results $\geq 90\%$ ($k = 1, 3$) for all four cases. A further reduced feature set was achieved using the one-way ANOVA technique with the Scheffe post-hoc tests. The F statistic results for the manually selected moments were all $F > 28$ and the majority were $F \gg 28$ where the F_{crit} values are 1.66(5%) and 2.05(1%). These results enable the rejection of the ANOVA null hypothesis as differences between within- and between-class means exist. Table 3 shows the classification results for this manually refined list of six velocity moments, all of which are over $\geq 87\%$ ($k = 1, 3$) (parentheses are used for indices exceeding nine). The moments used (for both camera views) to classify the fast and slow walks are identical, while between camera views they differ.

4.4. SOTON database

The Southampton (SOTON) database (also developed within the HumanID research program) consists of 50 subjects, with four sequences of each subject, a total of 200 sequences [33]; a far larger dataset than previous analyses. The subjects walked around a continuous bone-shaped track, the main shank of which

Table 2
The classification results for the four CMU database subsets using 90 velocity moments

Camera	Classification $k=1$		Classification $k=3$	
	Slow (%)	Fast (%)	Slow (%)	Fast (%)
CMU_03_7	100.00	100.00	100.00	100.00
CMU_05_7	100.00	99.00	99.00	95.00

Table 3
The classification results for the CMU database subsets using six velocity moments

Camera	Zernike velocity moments	Classification $k=1$		Classification $k=3$	
		Slow (%)	Fast (%)	Slow (%)	Fast (%)
CMU_03_7	A_{8202}, $A_{(12)400}$, A_{2000} A_{2200}, $A_{(11)(11)01}$, A_{4002}	91.00	91.00	90.00	87.00
CMU_05_7	A_{8000}, $A_{(12)402}$, A_{2200} A_{4202}, A_{7700}, A_{4402}	95.00	96.00	92.00	87.00

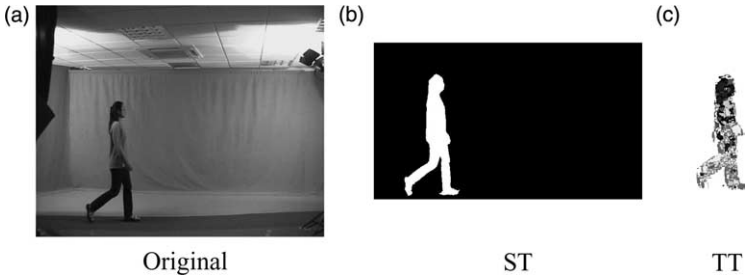

Fig. 4. Example image from the SOTON database, its corresponding cropped ST and TT (computed from image n, $n+1$ and histogram equalised for visualisation).

is normal to the camera. The sequences contain the subjects walking from left to right for one and a half gait cycles (three consecutive heel strikes). The subjects are walking in a relaxed manner, achieved by letting them settle into their walk before filming. The surface of the track was flat, while the loops at either end of the track are out view of the camera, allowing the subject to be walking in a straight trajectory when normal to the camera. Due to the background and the controlled lighting conditions, chroma-key extraction was possible enabling clean ST generation. A full description of the laboratory and extraction techniques can be found in [33]. TTs were then computed for this large database. Fig. 4 shows an example image from the database along with its corresponding cropped ST and TT. Due to the increased complexity of the TTs dataset, the scaling element of the Zernike mapping was disabled as its application would not produce meaningful data. (In a practical scenario the original image sequence should be scaled and then new TTs generated). The TT images were instead mapped to appear visually central to the unit disc, i.e. the thresholded image's COM was used in the mapping in-place of the greyscale COM. A reduced list of $\sim$200 Zernike velocity moments, up to and including orders m, $n=0\ldots12$; μ, $\gamma=0\ldots3$ were computed on the STs and TTs. The list was manually constructed using the results from previous database analyses reducing the computation time in place of computing an exhaustive list. The k-nn classification results for this list of moments for the STs and TTs can be seen in Table 4. These classification results are low, suggesting the need for feature selection.

Results for a subset of eight ST moments selected using the ANOVA technique, are shown in Table 5. The F statistic values for the eight selected ST moments all satisfy $F>15$ with many satisfying $F\gg15$ where the associated F_{crit} are 1.44(5%) and 1.67(1%). The five selected TT moments satisfied $F>47$ (with the same F_{crit} values as the STs). All of these results allow the rejection of the null hypothesis. A comparatively high classification of 69.50% ($k=3$) is achieved on this large database using just eight ST velocity moments, as shown in Table 5 and using just two velocity moments achieves 31% ($k=3$) discrimination capability. Table 6 shows the classification rates for the five selected TT velocity moments which are relatively low in comparison. The results of combining the STs and TTs feature spaces is shown in Table 7, resulting in a 93.00% ($k=3$) classification rate.

4.5. Case study

Through the use of the SOTON database we can illustrate the advantage of using a descriptor that includes both shape and motion. Firstly, one extra subject was added to the SOTON database, using the same laboratory conditions as used to capture the original database. The subject's data consisted of three sequences of them walking in a normal relaxed manner, achieved by asking them to walk normally around the continuous track. For their fourth sequence they were asked to walk in an abnormal manner. This fourth sequence produced the subject walking with more vertical motion throughout the sequence, along with variations in stride length and arm

Table 4
The classification results for the SOTON database using 234 velocity moments

SOTON template	Classification $k=1$	$k=3$
ST	74.00%	57.50%
TT	52.50%	28.50%

Table 5
The SOTON classification results for the spatial templates (STs)

Zernike velocity moments	Classification $k=1$	$k=3$
A_{6000}, A_{8200}	39.50%	31.00%
A_{6000}, A_{8200}, A_{8810}	63.00%	47.50%
A_{6000}, A_{8200}, A_{8810}, $A_{(12)(12)20}$, A_{7110}	64.00%	50.00%
A_{6000}, A_{8200}, A_{8810}, $A_{(12)(12)20}$, A_{7110}, A_{2200}, A_{8410}, $A_{(12)400}$	87.00%	69.50%

Table 6
The SOTON classification results for the temporal templates (TTs)

Zernike velocity moments	Classification $k=1$	$k=3$
A_{5100}, A_{6200}	24.50%	17.00%
A_{5100}, A_{6200}, A_{9900}, $A_{(10)(10)00}$	52.50%	38.00%
A_{5100}, A_{6200}, A_{9900}, $A_{(10)(10)00}$, A_{6610}	61.50%	50.50%

Table 7
The SOTON classification results for combining the template feature spaces

Zernike velocity moments	Classification $k=1$	$k=3$
(STs) A_{6000}, A_{8200}, A_{8810}, $A_{(12)(12)20}$, A_{7110}, A_{2200}, A_{8410}, $A_{(12)400}$ (TTs) A_{5100}, A_{6200}, A_{9900}, $A_{(10)(10)00}$, A_{6610}	95.50%	93.00%

motion. The subject swung their arms considerably more than usual and walked in a 'jerky' manner. A set of Zernike velocity moments for all four sequences were then calculated, allowing the subject to be added to the database. Fig. 5 shows the results of adding this new subject to the database. To allow for visualisation this extra subject is compared with nine randomly chosen subjects from the SOTON database. The plot shows that the feature point corresponding to subject 10's abnormal walking has drifted considerably away from their other three sequences. However, it is difficult to see how much of an effect that the abnormal walk has had on the subject's features by viewing just Fig. 5. Table 8 shows the mean μ, standard deviation σ and coefficient of variation $\sigma/\mu\%$ (indicating the percentage spread) of two example moments generated for subject 10. It can be seen that the variation is very small (low intra-class variation) when analysing the first three sequences (corresponding to their normal walk). This reflects the clean extraction of the silhouettes and the consistent manner of their walk. Once the fourth (abnormal walk) sequence is introduced, the purely spatial velocity moment (A_{7300}) variation increases

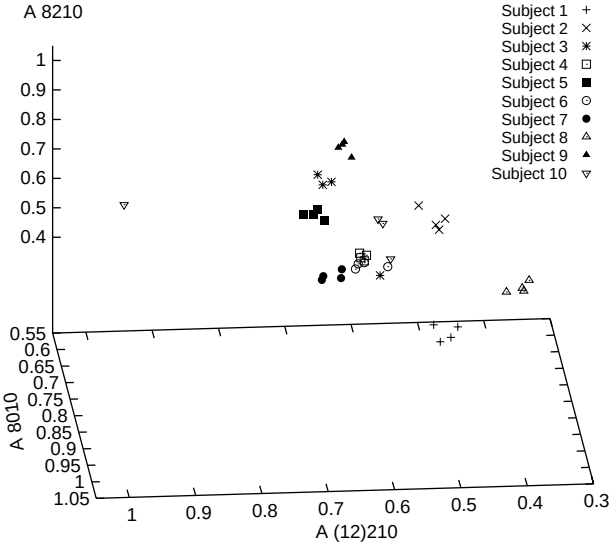

Fig. 5. Subject 10's fourth sequence is an outlier (the upside down triangle on the left) as is one of subject 3's sequences.

Table 8
The coefficient of variation for example velocity moments for subject 10

Subject 10's sequences	Normal walk		Normal + abnormal	
	A_{7300}	A_{7320}	A_{7300}	A_{7320}
1	0.259506	29.436546	0.259506	29.436546
2	0.266527	30.672182	0.266527	30.672182
3	0.259402	29.971671	0.259402	29.971671
4	–	–	0.290316	36.455319
μ	0.261812	30.026800	0.268938	31.633930
σ	0.003334	0.505950	0.012676	2.817905
$\sigma/\mu\%$	1.27	1.68	4.71	8.91

to 4.71% (from 1.27%) reflecting that the subject's silhouette has changed to some extent, i.e. arms are held higher, greater stride length. However, including motion (A_{7320}) further increases the variation to 8.91% (from 1.68%). The separation between the subject's normal and abnormal walk has increased through the inclusion of motion in the descriptor. However, it must be noted that projecting an alternative pair of moments A_{8010} and A_{8210} from Fig. 5 allows good clustering of all of subject 10's feature points.

Aside from this, in Fig. 5 one of subject 3's feature points is shown as being separated from the other sequences. This was due to the subject scratching their chin while walking in this particular sequence. A similar result to that of subject 10. This has altered their gait through reducing arm swing (as one arm is now stationary), which in turn has reduced their shoulder motion. Their silhouette structure has also altered with a change in pixel distribution around the upper body, effectively altering any symmetry (or asymmetry) characteristics in both their motion and spatial structure.

Both of these results (those of subjects 10 and 3) illustrate that the selection of velocity moments can allow separation and/or clustering of feature points within a dataset, dependent on the features chosen. This firstly shows potential for robust gait classification as moments that are independent of small variations in motion such as a subject scratching their chin are available. There is also potential for detecting a change in a subject's normal gait pattern through the selection of an alternative set of moments.

4.6. Discussion

Applying the velocity moments to the CMU and SOTON gait databases has produced good classification results using relatively few moments. Due to the use of a treadmill in the CMU database, none of the ST sequences have any apparent consistent forward velocity. This is reflected in the ANOVA selected velocity moments (refer to Table 3), as none include a forward velocity term, i.e. A_{**0*}. A subject's vertical motion information is visible in treadmill data, appearing as a vertical 'bobbing' motion as they walk. This richness of vertical motion information is reflected in the ANOVA selected moments, as many of them include y velocity information (mostly magnitude information, i.e. A_{***2}). In both cases (Table 3, $k=3$), the fast walk sequences have lower classification results

as compared with the slow walk sequences, reflecting a loss in temporal resolution as less images are describing the gait cycle.

The subjects in the SOTON database were filmed walking around a continuous track enabling both ST and TT generation. The ANOVA selected TT velocity moments favour those holding solely spatial information (i.e. A_{**00}). A similar result was found upon analysis of a TT database generated from data collected at the University of California San Diego (UCSD) [39], reflecting the optical flow technique describing a subject's limb motion, while the STs hold global shape/motion information. The limb motion alone of each subject (TTs) within this large dataset does not appear to provide sufficient information to discriminate between subjects. This is shown by the lower classification results achieved with just the TTs. However, they complement the STs increasing the classification rate to 93.00% ($k=3$), Table 7 (up from 69.59% ($k=3$) for the STs a lone). Doubtless a more complex classifier would further improve these classification results. It is interesting to note that the individual results for the STs and TTs consistently show the $k=1$ classification results to be greater than $k=3$. This suggests that the feature space is closely packed (with respect to subject clusters). By combining the STs and TTs feature spaces the $k=1$ and 3 results are more similar, suggesting a less packed feature space with respect to inter-subject differences and improved the cluster compactness. The difference between the 93.00% ($k=3$) and 95.50% ($k=1$) classification results in Table 7 are due to only five sequences (out of a possible 200). Furthermore, it is noted that $k=3$ is a very tight constraint given that there are only four sequences of each subject in each database.

The case study has illustrated an advantage of including motion in the descriptor. In this case, the motion information has enhanced the features to allow easy discrimination between a subject's walking styles. Fig. 5 shows alternative Zernike velocity moments to those already presented for the classification of the SOTON database, as does Table 8, illustrating the availability of features that produce tight class (intra-subject) clustering.

5. Performance analysis

This section details performance evaluation of the Zernike velocity moments as applied to the complete SOTON ST database. The analysis is intended to provide an insight into the robustness of the technique under a selection of simulated application scenarios. This analysis has been applied to

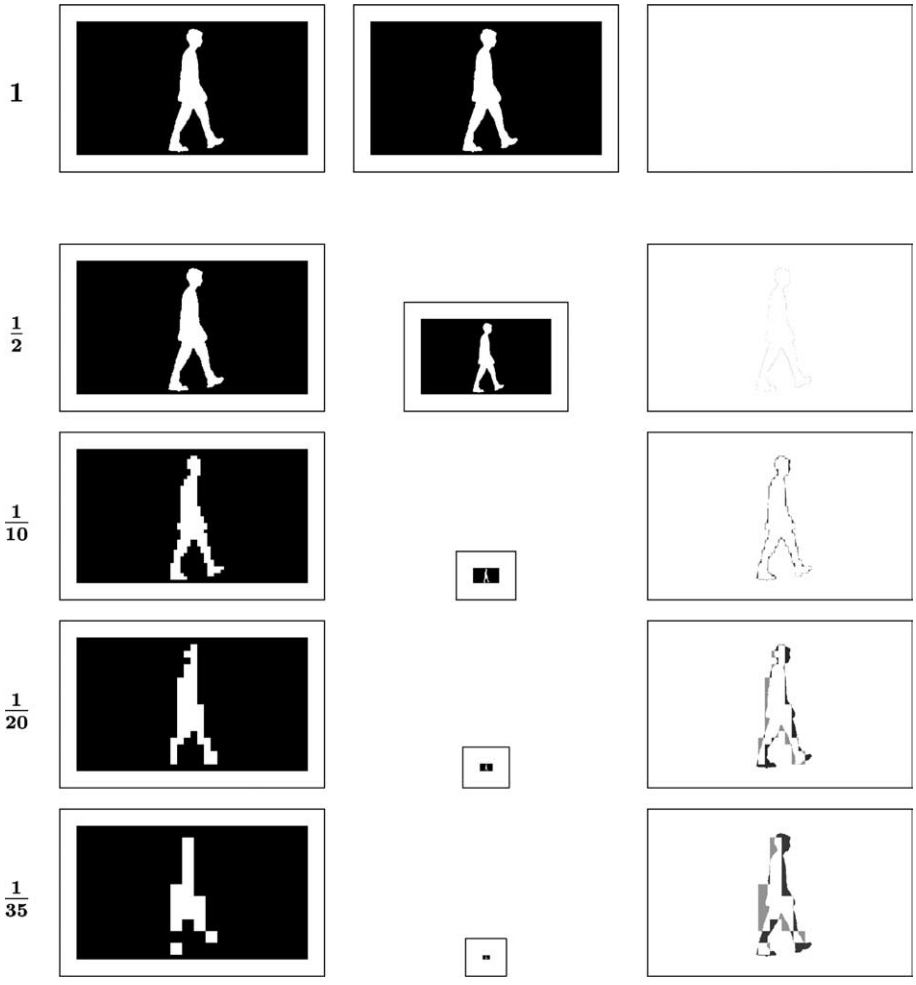

Fig. 6. ST resolution degradation, original at the top, (showing from left to right) the re-sampling scalar, resultant re-sampled image, their actual relative sizes and the difference image between the original resolution and the re-sampled.

the SOTON STs as their chroma-key extraction provides a suitable ground truth. To characterise the velocity moments it is important to describe the performance characteristics in terms of an error-rate, rather than a classification rate. This ensures that the results are independent of the characteristics of the database. If the classification rate is used, the results may be dependent on cluster compactness and separation. Alternatively, the subject clusters may all shift in the feature space relative to each other, potentially representing no change in the classification rate, even though the features themselves have altered. Primarily, this analysis is concerned with how the features behave under a series of different conditions; therefore, we concentrate on error-rates.

For each sequence of STs, the Zernike velocity moments used for classification (detailed in Table 5) were re-calculated for each increment step of the performance analysis. The normalised mean square error (NMSE) was then calculated between the original velocity moment values (O_i) and the new 'altered' values (W_i), for each incremental step. The NMSE is defined as

$$\mathrm{NMSE} = \frac{\sum_{i=1}^{K} (O_i - W_i)^2}{\sum_{i=1}^{K} O_i^2} \tag{21}$$

where K is the number of features, or moments and a NMSE value of 1 indicates 100% variation from the original values.

5.1. Image resolution

Camera resolutions vary considerably between different manufacturers, while the distance from the camera to the point/area of interest will also vary, dependent on the application. Gait as a biometric has the unique advantage of being potentially detectable from a distance (unlike for example, iris or fingerprint analysis). By analysing the effects on the velocity moments of reducing the image resolution, an insight can be gained into how the technique may translate to lower resolution imagery. Also, we can gain insight into the minimum resolution needed before the moments diverge grossly from their original value, effectively becoming overrun by noise due to loss of image detail.

Assuming that the original image is the highest resolution available, the images were progressively re-sampled to reduce their resolution. Sub-pixel estimation is allowed, enabling any re-sampling size to be achieved. A detailed description of the image re-sampling algorithm can be found in [39]. Eleven different resolutions were analysed from 1/2 the original resolution, through to 1/50. Fig. 6 shows an original ST and its reduced resolution versions, shown both expanded (to the size of the original resolution in the first column) and their relative reduced resolution sizes (in the second column). Fig. 7a shows the NMSE plotted against decreasing image resolution, for one randomly chosen subject (four sequences). The x-axis is the relative pixel size n, where $1/n$ is the new resolution. Fig. 7b shows that all subjects within the ST database exhibit a similar

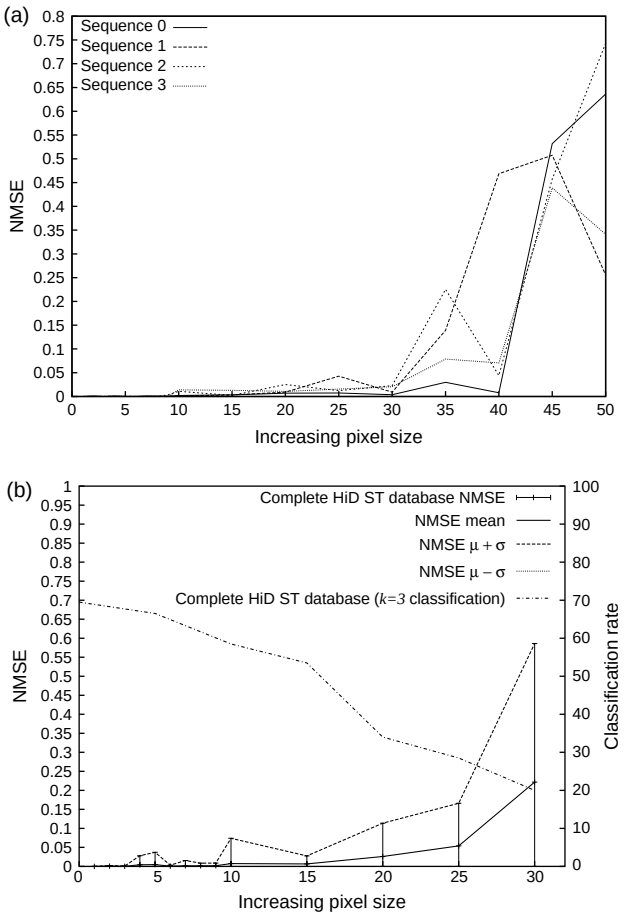

Fig. 7. NMSE with decreasing resolution for (a) one subject and (b) the complete ST SOTON database with the classification rates overlain.

trend. The mean μ result for the complete database is shown, with error bars indicating the standard deviation σ. The errors begin to diverge (the NMSE $|\mu \mp \sigma|$ increases) for $n > 10$, however, the NMSE is still low at less than 0.02. To illustrate that the NMSE analysis translates to the actual applied use of the Zernike velocity moments, the classification rates are also shown in Fig. 7b. The line corresponds to the $k = 3$ classifier results, beginning with 69.5% from Table 5. As the pixel size increases beyond $n = 10$, over 60% classification is maintained. Even at the lowest resolution (corresponding to a pixel size of $n = 30$) a classification rate well above that of chance is still achieved. A slight increase and subsequent decrease in the standard deviation σ can be seen around $n = 5$ in Fig. 7b. This may be due to moving decision boundaries in the re-sampling algorithm, effectively a rounding error. Similarly, the σ at $n = 15$ is thought to be due to the same issue, as the gradually decreasing classification rate (as n increases) supports the notion of an increasing trend in the NMSE (this trend is shown by the mean μ NMSE).

5.2. Occlusion

Occlusion is a common problem when applying computer vision techniques to real-world data. For example, optical

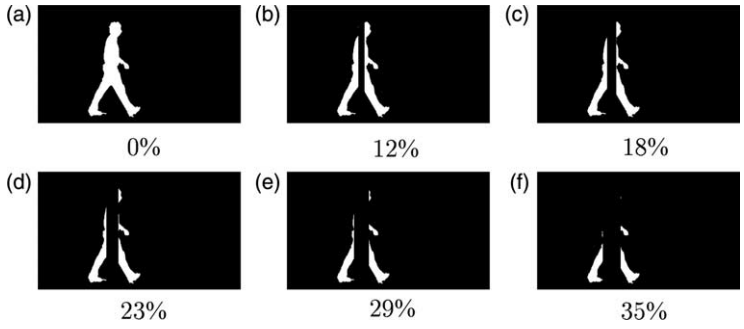

Fig. 8. STs with increasing amounts of occlusion (percentage of gait cycle occluded) shown for the mid gait cycle image.

remotely-sensed satellite data of an oil slick will invariably include areas occluded by cloud. This analysis aims to simulate the effects of a subject walking behind a lamp-post or another such static occluding object. If the occluding object is stationary and a background subtraction technique is used, then the occlusion will remove a strip from the ST as they pass. Fig. 10 shows an example ST sequence of a subject walking through a stationary occluding strip. As before, at each occlusion increment the Zernike velocity moments used for classification were re-calculated for the complete SOTON ST database and the NMSE calculated. The increment was determined in pixels, expressed here as a proportion of the average distance over which the subjects walked.

Fig. 9a shows the results for one subject (four sequences), whereas the results from the complete database can be seen in Fig. 9b. The NMSE is below 0.1 with 6% occlusion applied. The descriptions can be seen to become noisy and diverge (the NMSE $|\mu \mp \sigma|$ increases) as the occlusion increases past 18%, which Fig. 8 shows can occlude a large proportion of the ST.

5.3. Discussion

The Zernike velocity moments descriptions have been shown to degrade when analysing sequences at reduced resolutions and those containing occlusion. However, even after a considerable reduction in resolution ($n=30$), the NMSE is relatively low (Fig. 7b); there are two possible reasons for this. The first is with reference to the selected velocity moments themselves, which give measures of average pixel distribution in both the x- and y-directions. These properties will steadily degrade, however they will still be present until just before the image becomes one large pixel, refer to Fig. 6. This means that although the image sequence spatial resolution is being degraded, the overall x velocity will stay relatively consistent, as this is calculated using the COMs. The second reason for the low NMSE is due to the mapping process. The re-sampled images are passed onto the Zernike velocity moments for calculation at the same unit disc resolution, while the data itself is 'grainier'. Thus, even though the image resolution has been reduced the accuracy of the calculation has not. Degrading the resolution is effectively adding noise to the perimeter of the silhouette, up to the point where each image loses its overall shape. This can be seen in Fig. 6, where each re-sampled image has been subtracted from the original resolution image (shown in the top left), producing the difference images (in the third column). The dark grey areas are the remains of the original silhouette after the differencing operation. The light grey areas are the remnants of the re-sampled image. It can be seen that the re-sampling has both added and removed pixels from the perimeter of the original silhouette. Even at the 1/20 resolution very little specific spatial detail will be available (as illustrated in Fig. 6), whereas the NMSE error is still relatively low at <0.05 (Fig. 7b). The descriptor is exploiting the temporally correlated image

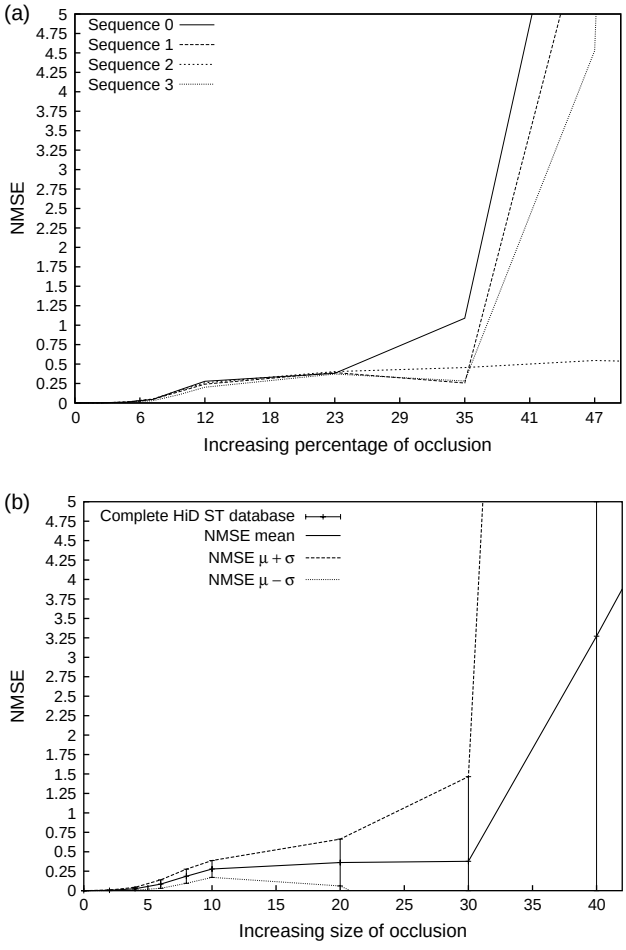

Fig. 9. NMSE with increasing occlusion for (a) one subject and (b) the complete SOTON database.

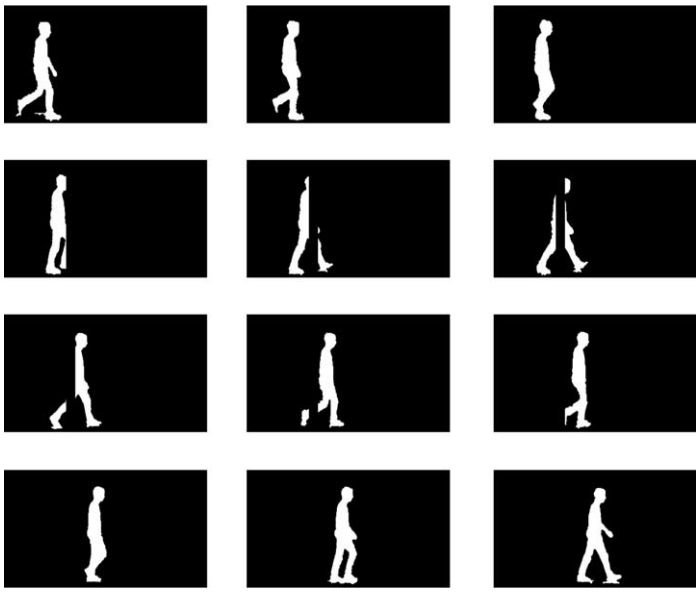

Fig. 10. Part of a sequence of STs showing the 18% occlusion case. The subject is walking left to right and the sequence runs from the top left to bottom right.

sequence to reduce the effects of noise. Perimeter noise around a shape can result from poor extraction, suggesting the velocity moments would perform favourably when analysing poorly extracted data.

The performance of traditional moments degrades where the shape is occluded due to the loss of spatial information. This is, in part, due to the moments being calculated from a single image. They are a global descriptor, so if a portion of the object is missing, it does not seem unreasonable to expect the result to be different from that of the original un-occluded object. Depending on the size of the occlusion, even the lowest order moments (i.e. mass) will be altered. Occlusion within the image sequence will degrade both the spatial and the motion information. This is illustrated in Fig. 9 where the Zernike velocity moment values diverge from their original values as the occlusion is introduced.

Only one gait cycle (approximately 30 images per sequence) has been used for these analyses. If, however, longer image sequences were analysed then the effects of the occlusion and resolution reduction will essentially be further diluted, due to an increase in the temporal resolution. This increases the amount of noise and distortion that can be handled before the descriptions diverge.

6. Conclusions

A new moment descriptor structure that includes spatial and temporal information is proposed. This allows the application of statistical moments to motion based time series analysis. Thus, classification of an image sequence can be based on moments describing spatial characteristics and/or motion information, while retaining both scale and translation invariance. For example, similar objects moving with different motion can be statistically discriminated. The Cartesian velocity moments are simplistic, although they will produce highly correlated features due to their non-orthogonal basis. This can become a problem, dependent on the features of interest and the size of the database being analysed. The Zernike velocity moments over come this as they are formed around the established and proven orthogonal Zernike basis. The single-image orthogonality condition of the Zernike velocity moments produces features that are both smaller in magnitude than the Cartesian implementation and less correlated. This also reduces the need for high accuracy in the calculation (in comparison to the Cartesian method). Furthermore, the Zernike velocity moments produce single-image scale invariant features, a property which is directly applicable to the problem of camera zoom on a piece of imagery.

The velocity moments compress a temporal image sequence into a set of features that enable description through both spatial and/or motion information. The use of an image sequence, in place of single images enables the exploitation of temporal correlation within the sequence, allowing the possibility of refining the description as the sequence length increases. Using a large database the performance of the Zernike velocity moments has been studied under simulated occlusion and reduced resolution scenarios with good results. The advantages of including motion in the descriptor have been illustrated, and the structure of the velocity moments allows motion free descriptors if desired. The theory behind this new technique is presented, while its performance has been further analysed using deforming-shapes, through the application to human gait classification.

Acknowledgements

Our thanks to Dr Michael Grant for his help in capturing and preparing the data and we gratefully acknowledge the partial support from the European Research Office of the US Army, Contract No. N68171-01-C-9002.

References

[1] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (3) (1999) 428–440.
[2] J.M. Nash, J.N. Carter, M.S. Nixon, Dynamic feature extraction via the velocity hough transform, Pattern Recognition Letters 18 (1997) 1035–1047.
[3] M.G. Grant, M.S. Nixon, P.H. Lewis, Extracting moving shapes by evidence gathering, Pattern Recognition 35 (2002) 1099–1114.
[4] M.R. Teague, Image analysis via the general theory of moments, Journal of the Optical Society of America 70 (8) (1979) 920–930.
[5] R. Mukundan, K.R. Ramakrishnan, Moment Functions in Image Analysis—Theory and Applications, World Scientific, Singapore, 1998.
[6] J. Hoey, J. Little, Representation and Recognition of Complex Human Motion, Proceedings of Computer Vision and Pattern Recognition (CVPR2000), vol. 1, 2000, pp. 752–759.
[7] J.J. Little, J.E. Boyd, Recognising people by their gait: the shape of motion, Videre 1 (2) (1998) 2–32.
[8] R. Rosales, Recognition of Human Actions Using Moment-based Features, Boston University Computer Science Technical Report, vol. BU 98-020, 1998.
[9] J.W. Davis, A.F. Bobick, The Representation and Recognition of Action Using Temporal Templates, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR97), 1997, pp. 928–934.
[10] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory (1962) 179–187 IT-8.
[11] C. Teh, R.T. Chin, On image analysis by the method of moments, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (4) (1988) 496–513.
[12] A.B. Bhatia, E. Wolf, On the circle polynomials of Zernike and related orthogonal sets, Proceedings of Cambridge Philosophical Society 50 (1954) 40–48.
[13] A. Khotanzad, Y.H. Hongs, Invariant image recognition by Zernike moments, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (5) (1990) 489–497.
[14] R.J. Prokop, A.P. Reeves, A survey of moment-based techniques for unoccluded object representation and recognition, CVGIP Graphical models and Image Processing 54 (5) (1992) 438–460.
[15] F.A. Sadjadi, E.L. Hall, Three-dimensional moment invariants, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-2 (2) (1980) 127–136.
[16] J.D. Shutler, M.S. Nixon, C.J. Harris, Global Statistical Description of Temporal Features, Proceedings of International Society for Photogrammetry and Remote Sensing Congress (ISPRS00), 2000, pp. 720–726.
[17] J.D. Shutler, M.S. Nixon, Zernike Velocity Moments for the Description and Recognition of Moving Shapes, Proceedings of British Machine Vision Conference (BMVC01), vol. 2, 2001, pp. 705–714.
[18] B.C. Li, A new computation of geometric moments, Pattern Recognition 26 (1) (1993) 109–113.
[19] D. Thompson, The Oxford Dictionary of Current English, second ed., Oxford University Press, Oxford, 1992.
[20] G. Johansson, A multi-view method for gait recognition using static body parameters, Scientific American 232 (1976).
[21] D. Cunado, M.S. Nixon, J.N. Carter, Automatic Gait Recognition via Model-based Evidence Gathering, Proceedings of AutoID '99: IEEE Workshop on Identification Advanced Technologies, 1999, pp. 27–30.
[22] S.A. Niyogi, E.H. Adelson, Analyzing and Recognising Walking Figures in XYT, Proceedings o IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR94), 1994, pp. 469–474.
[23] C. BenAbdelkader, R. Cutler, H. Nanda, L. Davis. EigenGait: Motion-based Recognition of People Using Image Self-similarity, Proceedings of Audio-and Video-Based Biometric Person Authentication (AVBPA01), 2001, pp. 284–294.
[24] A.Y. Johnson, A.F. Bobick, A Multi-view Method for Gait Recognition Using Static Body Parameters, Proceedings of Audio-and Video-Based Biometric Person Authentication (AVBPA01), 2001, pp. 301–311.
[25] J.P. Foster, M.S. Nixon, A. Prugel-Bennet, Automatic gait recognition using area-based metrics, Pattern Recognition Letters 24 (2003) 2489–2497.
[26] J.B. Hayfron-Acquah, M.S. Nixon, J.N. Carter, Automatic gait recognition by symmetry analysis, Pattern Recognition Letters 24 (2003) 2175–2183.
[27] A. Kale, A. Rajagopalan, N. Cuntoor, V. Kruger, Gait-based Recognition of Humans Using Continuous HMM, International Conference on Automatic Face and Gesture Recognition (FG02), 2002, pp. 336–341.
[28] L. Lee, W. Grimson, Gait analysis for Recognition and Classification, Proceedings of Automatic face and gesture recognition (FGR02), 2002, pp. 155–162.
[29] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1505–1518.
[30] C.Y. Yam, M.S. Nixon, J.N. Carter, Automated person recognition by walking and running via model-based approaches, Pattern Recognition 37 (5) (2004) 1057–1072.
[31] S. Sarkar, J. Phillips, Z. Liu, I.R. Vega, P. Gother, K.W. Bowyer, The HumanID gait challenge problem: data sets, performance and analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 162–177.
[32] R. Gross, J. Shi, The CMU Motion of Body (MoBo) Database, Technical Report, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2001.
[33] J.D. Shutler, M.G. Grant, M.S. Nixon, J.N. Carter, On a Large Sequence-based Human Gait Database, Proceedings of Recent Advances in Soft Computing (RASC02), 2002, pp. 66–71.
[34] H. Bulthoff, J. Little, T. Poggio, A parallel algorithm for real-time computation of optical fbw, Letters to Nature 337 (9) (1989) 549–553.
[35] P.S. Huang, C.J. Harris, M.S. Nixon, Recognising humans by gait via parametric canonical space, Artificial Intelligence in Engineering 13 (1999) 93–100.
[36] P.S. Huang, C.J. Harris, M.S. Nixon, Comparing Different Template Features for Recognising People by Their Gait, Proceedings of British Machine Vision Conference (BMVC98), vol. 2, September 1998, pp. 639–648.
[37] G.M. Clarke, D. Cooke, A Basic Course in Statistics, Arnold (1998) 520–546 (Chapter 22).
[38] P.R. Cohen, Empirical Methods for Artificial Intelligence, MIT Press, Cambridge, MA, 1995. pp. 185–287 (Chapters 6 and 7).
[39] J.D. Shutler, Velocity Moments for Holistic Shape Description of Temporal Features, PhD Thesis, University of Southampton, UK, 2002.