Zernike Velocity Moments for Description and Recognition of Moving Shapes

Jamie D. Shutler and Mark S. Nixon
Dept. of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
{jds1,msn}@ecs.soton.ac.uk
http://www.isis.ecs.soton.ac.uk/

Abstract

New Zernike velocity moments have been developed to describe an object, not only by its shape , but also by its motion throughout an image sequence. These are an extended form of the orthogonal Zernike moment set and include velocity information introduced via centralised moments. Initial analysis shows that they perform well when applied to analysing gait sequences resulting in a good recognition rate and a compact description. They have exhibited promising attributes when applied to occluded data, which is reflected in the method of describing a complete temporal sequence and not single images. Further, they appear to provide measures intimately related to the moving and/or morphing shape within the sequence. Their invariance properties suggest that they will be useful in real situations where poor quality or camera zoom problems are apparent.

1 Introduction

The application of classical moments to two dimensional images was first shown in the early sixties by Hu [4]. Hu tested their validity using a simple experiment to recognise written characters. Hu was only concerned with images without noise, but further work [2], showed that traditional moment performance degrades where the view is occluded or noisy. A survey of moment based techniques with respect to computer vision [10] details many of the current techniques regarding representation and recognition. However, all of these are only interested in processing single images. Little [6] used moment based features to characterise optical flow for automatic gait recognition, thus linking adjacent images but not the complete sequence. Further work analysed the complete sequence by extending the centralised moments to include velocity, [5]. However, the features produced were highly correlated, due to the non-orthogonality of the original Cartesian moments on which the description is based. These moments were invariant to scale changes between sequences, but not to changes within each sequence.

Here, a new method is proposed which aims to describe motion, through a time varying sequence, producing less correlated and more compact descriptions. This new technique includes the advantages of invariance to both translation and scale (both between

BMVC 2001 doi:10.5244/C.15.72

and within sequences), making it applicable to a real situation where camera zoom is apparent. The method studied is an extended form of the established Zernike moment set which is well proven in both pattern recognition and in the presence of image noise, [10]. This new approach has its basis in a standard technique for region description, enabling its use as a general method for describing moving objects by holistic measures.

1.1 Non-orthogonal Cartesian Moments

Moments, when applied to images, describe the image content with respect to its axes. They are designed to capture global information about the image. Here we are using them to characterise a grey level image so as to extract properties which have analogies in statistics or mechanics. The moment expressions use basis functions which have a range of useful properties that are passed onto the moments. This produces descriptions that have rotation, scale, translation and orientation invariance properties. Early work [4] applied statistical moments to image analysis using the Cartesian moments, which in discrete form are:

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

Extending them to include translation invariance produces the Centralised moments:

$$\mu_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} (x - \overline{x})^p (y - \overline{y})^q P_{xy} \tag{2}$$

Where M and N are the image dimensions, while $(x-\overline{x})^p(y-\overline{y})^q$ and $x^p y^q$ are the basis functions. P_{xy} is the pixel value at position (x, y), while $\overline{x}$ and $\overline{y}$ are the x and y centres of mass (COMs) respectively.

1.2 Orthogonal Complex Zernike Moments

Complex Zernike moments [12] are constructed using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc $x^2 + y^2 \leq 1$. They are expressed as:

$$A_{mn} = \frac{m+1}{\pi} \int\int_{x^2+y^2 \leq 1} f(x,y)[V_{mn}(x,y)]^* \, dx \, dy \tag{3}$$

where $m = 0, 1, 2 \ldots \infty$, $f(x, y)$ is the function being described, $*$ denotes the complex conjugate and n is an integer, subject to the conditions:

$$m - |\, n \,| \;\; = \; even \;\; , \quad |\, n \,| \;\; \leq \;\; m \tag{4}$$

The Zernike polynomials [3] $V_{mn}(x, y)$, expressed in polar coordinates are:

$$V_{mn}(r, \theta) \;\; = \;\; R_{mn}(r) \exp(jn\theta) \tag{5}$$

where (r, θ) are defined over the unit disc and $R_{mn}(r)$ is the orthogonal radial polynomial, defined as:

$$R_{mn}(r) = \sum_{s=0}^{\frac{m-|\, n \,|}{2}} (-1)^s \;\; F(m, n, s, r) \tag{6}$$

where:

$$F(m,n,s,r) = \frac{(m-s)!}{s!(\frac{m+\lfloor n \rfloor}{2} - s)!\,(\frac{m-\lfloor n \rfloor}{2} - s)!}\, r^{m-2s} \tag{7}$$

The first six orthogonal radial polynomials are:

$$\begin{aligned} R_{00}(r) &= 1 & R_{11}(r) &= r \\ R_{20}(r) &= 2r^2 - 1 & R_{22}(r) &= r^2 \\ R_{31}(r) &= 3r^3 - 2r & R_{33}(r) &= r^3 \end{aligned} \tag{8}$$

Figure 1 shows eight such radial responses, where it can been seen that the polynomials become more grouped, as they approach the edge of the unit disc. So for a discrete image,

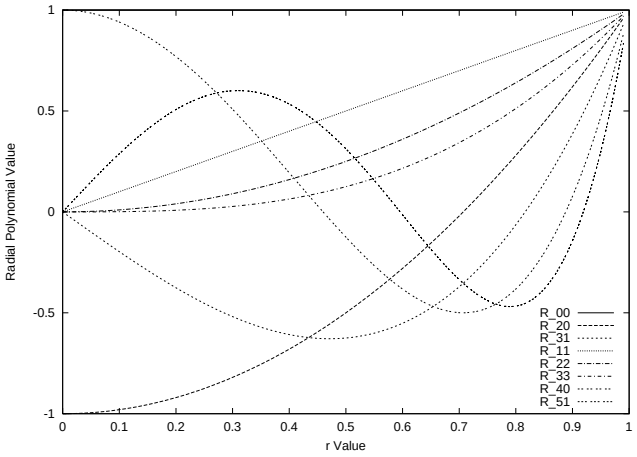


Figure 1: Eight orthogonal radial polynomial plots.

if $P(x,y)$ is the current pixel then Equation 3 becomes :

$$A_{mn} = \frac{m+1}{\pi} \sum_x \sum_y P(x,y)[V_{mn}(x,y)]^* \quad \text{where } x^2 + y^2 \leq 1 \tag{9}$$

To calculate the Zernike moments of an image $f(x,y)$, the image (or region of interest) is first mapped to the unit disc using polar coordinates, 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. The coordinates are then described by r which is the length of the vector from the origin to the coordinate point and θ which is the angle from the x axis to the vector r, by convention measured from the positive x axis in a counter clockwise direction. The mapping from Cartesian to polar coordinates is:

$$x = r\ \cos\theta \qquad y = r\ \sin\theta \tag{10}$$

where

$$r = \sqrt{x^2 + y^2} \qquad \theta = \tan^{-1}\left(\frac{y}{x}\right) \tag{11}$$

However, $\tan^{-1} A$ in practice is often defined over the interval $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$, so care must be taken as to which quadrant the Cartesian coordinates appear in. Translation and

scale invariance can be achieved by normalising the image using the Cartesian moments prior to calculation of the Zernike moments [1]. Translation invariance is achieved by moving the origin to the centre of the image by using the centralised moments, causing $m_{01} = m_{10} = 0$. Following this, scale invariance is produced by altering each object so that its area (or pixel count for a binary image) is $m_{00} = \beta$, where β is a predetermined value. Both invariance properties can be achieved using :

$$g(x, y) = f\left(\frac{x}{a} + \overline{x}, \frac{y}{a} + \overline{y}\right) \qquad \text{where } a = \sqrt{\frac{\beta}{m_{00}}} \tag{12}$$

and $g(x, y)$ is the new translated and scaled function. The error involved in the discrete implementation can be reduced by interpolation. If the coordinate calculated by Equation 12 does not coincide with an actual grid location, the pixel value associated with it is interpolated from the four surrounding pixels. As a result of the normalisation, the Zernike moments $\mid A_{00} \mid$ and $\mid A_{11} \mid$ are set to known values. $\mid A_{11} \mid$ is set to zero, due to the translation of the shape to the centre of the coordinate system. This however will be affected by a discrete implementation where the error will decrease as the image size increases. $\mid A_{00} \mid$ is dependent on m_{00}, and thus on β:

$$\mid A_{00} \mid \; = \frac{\beta}{\pi} \tag{13}$$

Further, the absolute value of a Zernike moment is rotation invariant as reflected in the mapping of the image to the unit disc. The rotation of the shape around the unit disc is expressed as a phase change, if ϕ is the angle of rotation, A^R_{mn} is the Zernike moment of the rotated image and A_{mn} is the Zernike moment of the original image then:

$$A^R_{mn} = A_{mn} \exp(-jn\phi) \tag{14}$$

1.3 Cartesian Velocity Moments

The Cartesian velocity moments [5] are computed from a sequence of images as:

$$vm_{pq\mu\gamma} = \sum_{i=2}^{Images} \sum_{x=1}^{M} \sum_{y=1}^{N} U(i, \mu, \gamma)\, C(i, p, q)\, P_{i_{xy}} \tag{15}$$

$C(i, p, q)$ arises from the centralised moments:

$$C(i, p, q) = (x - \overline{x_i})^p (y - \overline{y_i})^q \tag{16}$$

$U(i, \mu, \gamma)$ introduces velocity as:

$$U(i, \mu, \gamma) = (\overline{x_i} - \overline{x_{i-1}})^\mu (\overline{y_i} - \overline{y_{i-1}})^\gamma \tag{17}$$

$\overline{x_i}$ is the current COM in the x direction, while $\overline{x_{i-1}}$ is the previous COM in the x direction, $\overline{y_i}$ and $\overline{y_{i-1}}$ are the equivalent values for the y direction. 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 $\overline{x_i} = \overline{y_i} = 0$). The zero order velocity moments for which $\mu = 0$ and $\gamma = 0$ are then:

$$vm_{pq00} = \sum_{i=2}^{Images} \sum_{x=1}^{M} \sum_{y=1}^{N} (x - \overline{x_i})^p (y - \overline{y_i})^q P_{i_{xy}} \tag{18}$$

which are the averaged centralised moments. The zero order components for which $p = 0$ and $q = 0$ are:

$$vm_{00\mu\gamma} = \sum_{i=2}^{Images} \sum_{x=1}^{M} \sum_{y=1}^{N} (\overline{x_i} - \overline{x_{i-1}})^{\mu} (\overline{y_i} - \overline{y_{i-1}})^{\gamma} P_{i_{xy}} \tag{19}$$

which is a summation of the difference between COMs of successive images (ie the velocity). The structure of Equation 15 allows the image structure to be described together with velocity information from both the x and y directions. These results are averaged by normalising with respect to the number of images and the average area of the object. This results in pixel values for the velocity terms, where the velocity is measured in pixels per image. The normalisation is expressed as:

$$\overline{vm_{pq\mu\gamma}} = \frac{vm_{pq\mu\gamma}}{A \cdot I} \tag{20}$$

where A is the average area (in numbers of pixels) of the moving object, I is the number of images and $\overline{vm_{pq\mu\gamma}}$ is the normalised Cartesian velocity moment.

2 Zernike Velocity Moments

The new Zernike velocity moments are expressed as:

$$A_{mn\mu\gamma} = \frac{m+1}{\pi} \sum_{i=2}^{Images} \sum_{x} \sum_{y} U(i,\mu,\gamma)\, S(m,n)\, P_{i_{xy}} \tag{21}$$

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

$$S(m,n) = [V_{mn}(r,\theta)]^* \tag{22}$$

Velocity is introduced as before (Equation 17), while normalisation is produced by:

$$\overline{A_{mn\mu\gamma}} = \frac{A_{mn\mu\gamma}}{A \cdot I} \tag{23}$$

The coordinate values for $U(i,\mu,\gamma)$ are calculated using the Cartesian moments and then translated to polar coordinates. If we consider first the x direction case only, from Equation 10 the angle 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:

$$x = r\ \cos\theta = r\ \cos(0) = r \tag{24}$$

where r is the length of the vector from the previous COM to the current COM, ie the velocity in pixels/image. Alternatively, if the movement is right to left then:

$$x = r\ \cos\theta = r\ \cos(\pi) = -r \tag{25}$$

The mapping to polar coordinates results in a sign change which could be used to detect the direction of motion. Similarly for the y direction velocity, the values of θ are either $\frac{\pi}{2}$ or $\frac{3\pi}{2}$ radians, and using Equation 10 produces:

$$y = r\ \sin\theta = r\ \sin\left(\frac{\pi}{2}\right) = r \tag{26}$$

and

$$y = r\ \sin\theta = r\ \sin\left(\frac{3\pi}{2}\right) = -r \tag{27}$$

3 Application to Gait Recognition

3.1 Gait

Gait is the manner by which we walk and is primarily determined by our musculature and joint structure. There are two main computer vision approaches to gait recognition [8]: either model-based, or statistical holistic descriptions. The latter has been used by Huang [9] where temporal changes in gait are detected using optical flow techniques. This method produces 100% recognition on a small database, it shows that statistical methods indeed look promising, however this approach lacks the intimacy of gait as none of the measures are specifically linked to gait itself. Another statistical approach [6] has taken moments of optical flow, again producing encouraging results. A set of features, derived from moments are used to produce a model free description for recognition. The method exploits the periodic variations in a person's motion, achieving recognition rates of over 90% for five features. Here we apply the Zernike velocity moments to the problem of gait recognition, to produce a compact invariant description of both shape and motion.

3.2 Database Characteristics

The subject database used here is identical to that used by Little [6] and Huang [9], consisting of six subjects, with seven sequences per subject. In each sequence the subjects are walking from right to left, along a slight incline in front of a static background, an example of which can be seen in Figure 2a. The subjects are all walking at similar speeds, however variations in speed exist both within a sequence and between sequences. The distance between the camera and the subject varies between some sequences thus the need for scale invariance within the feature description. The variation in distance leads to interaction between both the ground and the background causing shadows to appear. This is in addition to the shadows appearing (on most sequences) from reflections on the floor, Figure 2b. There is also evidence of interaction between some subjects' clothes and the background affecting the feature (subject) extraction.

3.3 Feature Extraction

A statistical based subject-extraction method [11] was used to produce a small database of silhouettes. (However, alternative model-based extraction techniques exist, [7]). The extraction method analyses the statistics of the sequence, and uses both luminance values and edges to determine the background and foreground information. Silhouette extraction is based on scene variance and standard deviation. The silhouette is then windowed using the average velocity to produce the final spatial templates (128×160 pixels), Figure 2b.

a b

Figure 2: Example image from the subject database and windowed spatial templates.

3.4 Subject Mapping

Prior to the translation and scale invariance mapping detailed in Section 1.2, the COMs are calculated to adjust the velocity calculations for differences between the average velocity and the actual velocity between successive images. This difference is then added to (or removed from) the extracted average velocity during the velocity moment calculation. The subject is then mapped onto the unit disc. The value of β for Equation 12 is set, so that the mapped pixels coordinates are within 90% of the unit disc's radius. This is done to reduce the effect of the converging polynomials as r approaches unity, Figure 1.

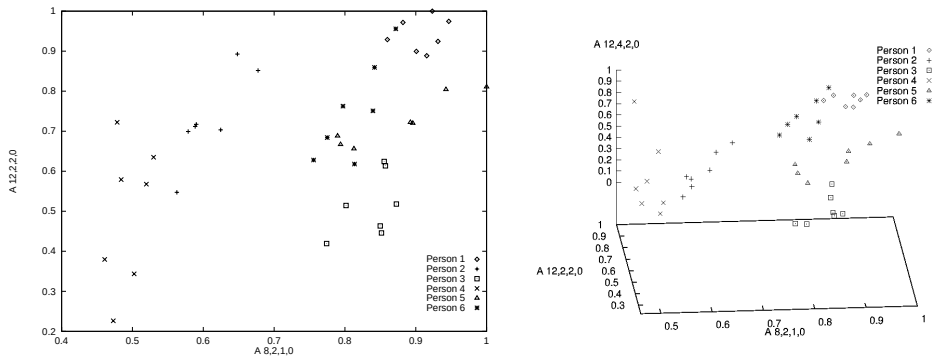
3.5 Gait Recognition

Zernike velocity moments up to order $m, n = 12$, $\mu = 4, \gamma = 0$ were calculated for all the sequences of spatial templates in the database. The new moments were calculated for one complete gait cycle, from heel strike to heel strike (or two steps). To further reduce the size of the selection problem only the magnitudes of the velocity moments were studied. Suitable moments for classification were then selected using the one-way ANOVA technique, where ANOVA is a general method for studying linear models. The purpose is to test for differences in class means, where here a class is a specific human subject. Using this technique, features which produce discriminatory capabilities can be selected. Classification (or recognition) was then achieved with a k-NN approach with $k = 1$ and 3, using the $leave\ one\ out$ rule with $cross\ validation$. Prior to classification the velocity moments were normalised by their maximum values, to ensure that moments with larger average values did not bias the results. Figure 3 details the recognition results. It can be seen that a recognition rate of over 80% with $k = 1$ is achieved using only two features. 100% recognition is achieved using just five features, for both $k = 1$ and 3. Figure 4a

Zernike Velocity Moments order - $mn\mu\gamma$	**Recognition Rate**	
	$k = 1$	$k = 3$
8210	61.90%	52.38 %
8210,12220	80.95%	76.19 %
8210,12220,12420	85.71%	88.10 %
8210,12220,12420,5100	97.62%	97.62 %
8210,12220,12420,5100,9900	100.00%	100.00 %

Figure 3: Gait recognition rates.

shows a scatter plot of the first two Zernike velocity moments listed in Figure 3, while



a b

Figure 4: Scatter plots of selected moments used for recognition.

Figure 4b shows a scatter plot of the first three moments illustrating both clustering and cluster separation. Naturally, a larger database would doubtless require more moments to separate subjects, but it is perhaps worth noting that only a basic classifier has been used.

3.6 Occlusion

The Zernike velocity moments were then re-run with increasing amounts of occlusion applied to one of the subjects. This analysis aimed to simulate to some extent the effects of a subject walking behind a lamp post or another such object. Figure 5 shows example images of the subject at different stages of occlusion. The effects of the occlusion were studied for a single subject picked at random. With the five velocity moments used for recognition (Figure 3), the normalised mean squared error (NMSE) between the original un-occluded and the occluded moments was calculated, as the occlusion increased. This

Figure 5: Images showing the 5% occluding strip (subject is walking right to left).

Figure 6: Increasing occlusion, 0%, 11%, 18%, 25% and 31%.

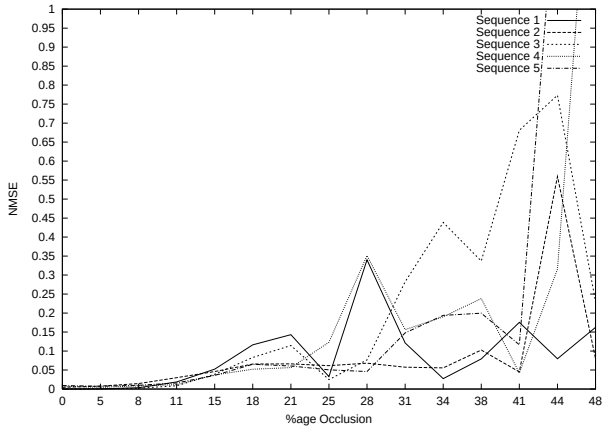


Figure 7: Normalised mean squared error with increasing occlusion

was repeated for five sequences of the same subject, Figure 7. The size of the occluding strip is expressed as a proportion of the distance over which the subject walks. The NMSE is below 0.05 with 15% occlusion applied. The descriptions can be seen to become noisy as the occlusion increases past 18%, which Figure 6 shows to occlude a large proportion of the spatial template. It must be noted that only one gait cycle has been used for the calculation, if however more than one gait cycle is analysed then the effects of the occlusion will essentially be further diluted, due to an increase in the spatial resolution. This would in turn increase the amount of occlusion that can be handled before the descriptions become noisy. This analysis was not compared with the recognition rate, as the results would be dependent on the database characteristics, (ie subject cluster compactness and separation).

4 Conclusions

A new description aimed at capturing both structural and temporal information of a time varying sequence has been proposed. It contains both scale and translation invariance. A recognition rate of 100% has been achieved on a gait database of 42 sequences. The technique has been shown to handle simple occlusion, the performance of which is partly due to the integration of complete sequences, rather than describing each image separately. The orthogonality property of the original Zernike moments means that the features produced are both smaller in magnitude than a Cartesian implementation and less correlated. They are however correlated in the sense that the images being described constitute a correlated sequence. Invariance to changes in scale within a sequence, is directly applicable to the problem of camera zoom on a piece of imagery, which is an area of future investigation. Further to this, we intend to investigate how this new technique performs on a larger and more varied database, including studying the effects of people's clothes and items that they might be carrying. We also aim to investigate how its generic capability translates to the analysis of arbitrary moving objects.

5 Acknowledgements

This material is based upon work supported by the European Research Office of the US Army under Contract No. N68171-01-C-9002. The subject database analysed here was kindly provided by Dr Jeffrey Boyd previously at the University of California, USA.

References

[1] A. Khotanzad and Y. H. Hong. Invariant image recogntion by zernike moments. *IEEE Trans. PAMI*, **12**(5):pp. 489–497, 1990.

[2] C. Teh and R. T.Chin. On image analysis by the method of moments. *IEEE Trans. PAMI*, **10**(4):pp. 496–513, 1988.

[3] F. Zernike. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode. *Physica*, **1**:pp. 689–704, 1934.

[4] M. Hu. Visual Pattern Recognition by Moment Invariants. *IRE Trans. on Information Theory*, **IT-8**:pp. 179–187, 1962.

[5] J. D. Shutler, M. S. Nixon and C. J. Harris. Statistical gait recognition via temporal moments. *Proc. SSIAI 2000 - Austin,Texas*, :pp. 291–295, 2000.

[6] J. J. Little and J. E. Boyd. Recognising people by their gait: The shape of motion. *Videre*, **1**(2):pp. 2–32, 1998.

[7] J. M. Nash, J. N. Carter and M. S. Nixon. Extraction of moving articulated-objects by evidence gathering. *Proc. BMVC98*, **2**:pp. 609–618, 1998.

[8] M. S. Nixon, J. N. Carter, D. Cunado, P. S. Huang and S. V. Stevenage. *Biometrics: Personal Identification in Networked society, A.K. Jain, R.Bolle and S. Pankanti*, chapter 11:Automatic gait recognition, pages 231–249. Kluwer Academic Publishers, 1999.

[9] P. S. Huang, C. J. Harris and M. S. Nixon. Recognising humans by gait via parametric canonical space. *AI in Eng*, **13**:pp. 93–100, 1999.

[10] R. J. Prokop and A. P. Reeves. A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP GMIP*, **54**(5):pp. 438–460, 1992.

[11] S. Jabri, Z. Duric, H. Wechsler and A. Rosenfeld. Detection and location of people in video images using adaptive fusion of color and edge information. *Proc. ICPR00*, **4**:pp. 627–630, 2000.

[12] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, **70**(8):pp. 920–930, 1979.