Universities of Leeds, Sheffield and York
http://eprints.whiterose.ac.uk/

A Comparative Study on using Zernike Velocity Moments and Hidden Markov Models for Hand Gesture Recognition

Moaath Al-Rajab, David Hogg and Kia Ng

Computer Vision Group, School of Computing, University of Leeds
7.27 E.C. Stoner Building, School of Computing, University of Leeds, Leeds, LS2 9JT, UK
{moaath, dch, kia}@comp.leeds.ac.uk

Abstract. Hand-gesture recognition presents a challenging problem for computer vision due to the articulated structure of the human hand and the complexity of the environments in which it is typically applied. Solving such a problem requires a robust tracking mechanism which in turn depends on an effective feature descriptor and classifier. Moment invariants, as discriminative feature descriptors, have been used for shape representation for many years. Zernike moments have been particularly attractive for their scale, translation and rotation invariance. More recently, Zernike moments have been extended to a spatio-temporal descriptor, known as the Zernike velocity moment, through combining with the displacement vector of the centre of mass of the target object between video frames. This descriptor has hitherto been demonstrated successfully in human gait analysis. In this paper, we introduce and evaluate the application of Zernike velocity moments in hand-gesture recognition, and compare with a bank of hidden Markov models with Zernike moments as observations. We demonstrate good performance for both approaches, with a substantial increase in performance for the latter method.

Key words: Spatio-temporal description, hand gesture recognition, skin-colour segmentation, Zernike velocity moments, HMM

1   Introduction

Interest in hand gesture recognition has increased in recent years motivated in large part by the range of potential applications for human-machine interaction mediated by hand gestures. More generally, the human hand poses a substantial challenge for tracking and action recognition due to the way in which it deforms and self-occludes and the range of different semantically distinguishable actions it can perform.

Within the wide range of application scenarios, hand gestures can be classified into several categories [1], including conversational gestures, controlling gestures, manipulative gestures, and communicative gestures.   For instance, sign language is an important example of a domain involving communicative gestures [2]. Our work is

aimed at the use of controlling gestures within multimedia applications similar to [3-5].

Recognizing gestures automatically from visual input is a complex task that typically involves several stages, including signal processing, tracking, shape description, motion analysis, and pattern recognition. Machine learning methods have been used widely, particularly in recent years, and inspiration has come from several quarters including psychology and human behaviour studies.

In selecting a method for shape description, a key requirement is to provide sufficient discriminative information for successful classification of hand gestures. The use of moments for shape description was introduced by Hu [6] in 1962. Hu introduced a set of six functions of standard central moments which provide a description that is invariant to scaling, translation and rotation, and a seventh function invariant to scaling, translation and skew. Another form of moments are the Zernike moments (ZM) where the kernel is a set of orthogonal Zernike polynomials defined over polar co-ordinates inside a unit circle. ZMs are the projection of the image function onto these orthogonal basis functions [14, 16]. This original idea has been extended recently by Shutler and Nixon [7] to produce the so-called Zernike Velocity Moments ZVMs. These are generated from sequences of images depicting objects in motion and were shown in [7] to be effective for human gait analysis. The current paper introduces the use of ZVMs coupled with an appropriate classifier for hand-gesture recognition. We evaluate and compare the performance of this classification method based on ZVM with Zernike Moments coupled with a bank of HMMs.

In Section 2 we review the background that is most relevant to the study. We explain the formulation of Zernike Velocity Moments in Section 3, and the dataset used in both sets of experiments in Section 4. Section 5 and 6 deal with the two experiments, the first on classification using Zernike Velocity Moments and the second on classification using a bank of HMMs over Zernike Moments. Finally, in Section 7 we draw conclusions from the experiments and briefly address further work.

2   Background and Related Work

The literature on hand gesture recognition using computer vision can be categorized into those approaches that use a 3-D model for tracking the hand, and those that use an entirely image-based representation. An early development in the former category used a 3D mesh with vertices embedded in a linear subspace to characterize allowable shape variations [8]. The linear subspace was obtained from a set of examples using principal component analysis, generalising earlier applications of the same technique to sets of point landmarks in the image plane. The model was used to detect the pose of a hand in the visual field and subsequently to track this hand from frame to frame.

Later, Athitsos and Sclaroff [9] used a 3D model to generate projections of a 3D hand in different shapes and from different viewpoints. The projected hands were then matched to hands depicted in the incoming video stream, both through edge matching and through the use of Hu moments. In related research Sudderth et al. [10] introduced probabilistic methods for visual tracking of a three-dimensional geometric hand model from monocular image sequences. Model components are represented by

their positions and orientations. A prior model is then defined which enforces the kinematic constraints implied by the joints of model. They enhanced matching between the 3D geometric model and the tracked hand using a redundant representation, that measures Chamfer distance for colour and edge-based likelihood features. They tracked the hand's motion using non-parametric belief propagation (NBP) algorithm.

Tracking methods that depend entirely on an image-plane representation of the hand have been worked on extensively. Typically such systems are computationally less expensive than those methods that use a 3D model. Skin colour segmentation is a common method for locating the hand due to its fast implementation, where usually skin colour is modelled as a Gaussian distribution in a suitable colour space. Static background subtraction and adaptive background models are also commonly used for segmentation. Shadows can be a problem for such algorithms, although the worst effects can be ameliorated to some extent; for example, by using infrared (IR) cameras as in Ahn [11, 12].

For recognizing gestures, Ng and Ranganath [13] divided the process into two stages. The first stage was to find the poses of the gesture in each frame and the second stage was to analyse the motion of the gesture. Zernike moments and Fourier descriptors were used to generate the feature vectors used to train a Radial Basis Function (RBF) neural network combined with an HMM. Park *et al.* [3] used HMMs for gesture classification on 13 different gestures. They divided the gesture into four distinctive states: start, intermediate, distinctive and the end of the gesture. They assumed that the gestures start and end from the same state. Skin-colour is used for hand segmentation.

3 Zernike Velocity Moments (ZVM)

The concept of moments comes from physics, where it relates to the force required to affect a given angular acceleration on an object of known mass distribution. For the purpose of image analysis, a set of generalized moments is defined for a 'density' distribution $f(x, y)$ derived in some fashion from an image and often a binary function denoting membership of a target region. The discrete centralized moment of order *(p, q)* is given in equation 1:

$$\mu_{pq} = \sum_x \sum_y (x - \overline{x})^p (y - \overline{y})^q f(x, y) \tag{1}$$

Different degrees of moments represent different features of the target shape. For example, moments with *p=q=0 (0, 0)* compute the summation of all density values of the image (e.g. the number of 'on' pixels for a binary density function). Moments *(1, 0)* and *(0, 1)* represent the pixel distribution along *X* and *Y* axis respectively. Moments of degree two represent the variance of the density function, moments of degree three represent skewness, and moments of degree four represent kurtosis of the distribution.

The central moments are invariant to translation in the image plane and can be normalized for variation in scale by forming the so-called normalized central moment η_{pq} derived from the corresponding central moment as follows (for p and q both greater than zero):

$$\eta_{pq} = \frac{\mu_{10}}{\mu_{00}^{\gamma}} \quad \text{and} \quad \gamma = \frac{p+q}{2} + 1 \tag{2}$$

The original definition has been extended and combined with the theory of algebraic invariants to become the mathematical device used today in image analysis.

Zernike moments are a class of orthogonal complex moments, which in contrast to Hu moments can be computed for higher degrees, giving more discriminative potential – for example, higher order moments characterize the detailed shape of an object. The magnitudes of Zernike moments are rotation and reflection invariant [14] and can be easily constructed to an arbitrary order [15]. By projecting the image function onto the basis set, the Zernike moment A_{pq} of order p and repetition q is defined by:

$$A_{pq} = \frac{p+1}{\pi} \sum_x \sum_y f(x,y)[V_{pq}(x,y)]^* \qquad x^2 + y^2 \leq 1 \qquad |q| \leq p \tag{3}$$

$$V_{pq}(\rho,\theta) = R_{pq}(\rho)e^{jq\theta} \quad \text{where } \theta = \tan^{-1}\frac{y}{x} \qquad \rho = \sqrt{x^2 + y^2} \tag{4}$$

$$R_{pq}(\rho) = \sum_{s=0}^{(p-|q|)/2} (-1)^s \left(\frac{(p-s)!}{s!\left[\frac{p+|q|}{2}-s\right]!\left[\frac{p-|q|}{2}-s\right]!} \right) \rho^{p-2s} \tag{5}$$

Zernike Velocity Moments (ZVM) [7, 16] are essentially a weighted sum of Zernike moments over a sequence of frames (length T), weighted by a real-valued scalar function of the displacement of the centre of mass (CoM) between consecutive frames:

$$A_{pq\beta\lambda} = \alpha \sum_{i=2}^{T} \sum_x \sum_y f_i(x,y)U(i,\beta,\lambda)[V_{pq}(x,y)]_i^* \qquad \text{where } x^2 + y^2 \leq 1 \tag{6}$$

$$U(i,\beta,\lambda) = (\bar{x}_i - \bar{x}_{i-1})^{\beta}(\bar{y}_i - \bar{y}_{i-1})^{\lambda} \tag{7}$$

U is the series of weights derived from the displacements of the CoM, and * donates the complex conjugate. Usually (β,λ) are set to (0, 1) or (1, 0) to avoid zero

weights derived from the displacement of the CoM when there is only horizontal or vertical motion of the hands.

Normalized Zernike velocity moments $\overline{A}_{mn\beta\lambda}$ are defined in equation 8:

$$\overline{A}_{pq\beta\lambda} = \frac{A_{pq\beta\lambda}}{A.T} \qquad (8)$$

Where A is the average area (in pixels) of the moving object, *T is the length of the video segment.*

4   Dataset and Experiment Setup

For our comparative study of classification performance, we captured a video stream depicting a series of hand gestures using a normal webcam (30 fps, 320x240 pixels) in an office environment, as illustrated in figure 1. The video consists of 80 instances for each of five distinct gestures (referred to as A-E), performed in total by 10 different people (i.e. 8 instances of each gesture by each person). The five gestures used are shown in figure 2. More information about our dataset is available online[1]. We marked the start and end of each gesture manually for the entire dataset to provide a ground-truth for training and testing, experiments were conducted on a machine of 2.2 GHz dual-core cpu speed, 2 GB RAM.

A Gaussian skin-colour model [17] is used to produce a map of skin likelihood values for each frame, and we then track the hand using a CAMShift tracker [18, 19]. The CAMShift output is cleaned up automatically to remove small isolated regions and finally the result is binarised to give a final segmented hand region. For our hand gesture dataset, the procedure gives a well segmented hand region in each frame.

The start of gesture "A"   *The end of gesture "A"*

Figure 1. Start and end of gesture "A" (see figure 2) as an example of gesture. It shows the webcam input and the output of the CAMShift tracker in office environment after skin-colour segmentation. First frame is input stream and the second is the projection of HSV colour after weighting.

[1] http://www.comp.leeds.ac.uk/moaath/gHand/DS/

5 Classification using CvR and ZVM

In the first experiment, we used a ZVM descriptor to characterise the video segment corresponding to each gesture instance. The speed at which a gesture is performed results in video segments of different lengths within the dataset (between 45 and 150 frames). We did not attempt to interpolate gestures to a fixed number of frames, since the weighted mean computed by the ZVM is invariant to the overall speed of a gesture, although not to non-linear variations in the temporal execution. Computing the ZVM on a typical video segment takes up to 1 sec.
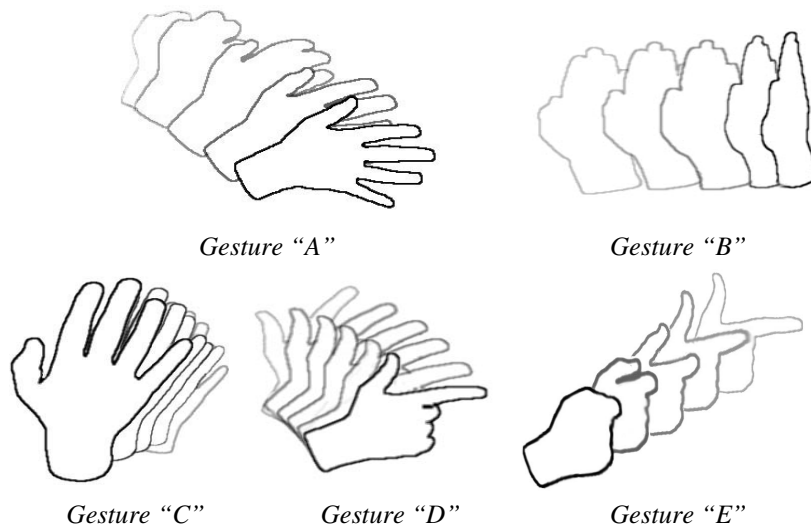

Gesture "A"   *Gesture "B"*

Gesture "C"   *Gesture "D"*   *Gesture "E"*

Figure 2. Training and testing gestures. Each of these gestures can be utilized to perform different actions. For example, in photo albums application, *Gesture "A"* may be used to change the album; *Gesture "B"* for rotating a photo; *Gesture "C"* for zooming a photo by moving the hand towards/away from the camera; *Gesture "D"* for navigating to the next photo; *Gesture "E"* for switching the application on/off.

For this experiment, we used a feature vector obtained from three Zernike Velocity Moments, defined by setting the four parameters (p, q, β, λ) to $(12,4,0,0)$, $(12,4,0,1)$, and $(12,4,1,0)$ (see equation 6).

We experimented with a number of standard classifiers and obtained the best overall performance using the 'Classification via Regression' (CvR) classifier reported in [20] and implemented as part of the WEKA package [21]. This builds a decision tree with a linear regression function at the leaves. We used ten-fold cross validation on our entire dataset and obtained the results laid out in the confusion matrix shown in Table 1, giving a mean accuracy of 84.22%. It can be seen that the greatest confusion is between gestures "A" and "E" where hand poses in the first part of the gesture and the horizontal displacement are similar.

A potential problem with the ZVM is that low displacement components of the centre of mass will result in clustering of feature vectors around the origin in feature space. For classifiers, this may impact on the ability to discriminate between gestures that involve little horizontal or vertical motion – for example, gesture B.

		Predicted Gesture				
		"A"	"B"	"C"	"D"	"E"
Actual Gesture	"A"	**66**	2	3	1	8
	"B"	5	**64**	3	5	3
	"C"	5	2	**70**	1	2
	"D"	0	1	2	**72**	5
	"E"	6	2	2	4	**66**

Table 1. The confusion matrix using CvR classifier and ZVM descriptor with (p, q, β, λ) set to (12,4,0,0), (12,4,0,1), (12,4,1,0), see equation 6.

6   Classification using HMMs and ZM

In the second experiment, we used hidden Markov models (HMMs) for classification [23], with observations defined by a vector of Zernike moments with parameters (p, q) set to (10, 2), (10, 4), (12, 2) and (12, 4) - see equation 3. We used Matlab HMM code from [22]. Prior to training and testing, we linearly interpolate each gesture sequence to a fixed number of frames. The reported results are for HMMs with four hidden states and a Gaussian observation density.

We train a separate HMM for each of the five gestures. Given an observation sequence O, the likelihood of the observation sequence given the model λ is obtained by summing the likelihood over all possible state sequences S:

$$P(O \mid \lambda) = \sum_{all\_S} P(O \mid S, \lambda) P(S \mid \lambda) \tag{9}$$

We then select the model with maximum likelihood as the chosen gesture:

$$\lambda = \arg \max_{k} (P(O \mid \lambda_k)) \tag{10}$$

The optimal number of hidden states chosen for each of the five HMMs was four-see figure 3(a). In training the HMM (as in gesture A), the increase in observation likelihood with each iteration is shown in figure 3(b). As for the first experiment, we use ten-fold cross validation giving a mean accuracy of 94.45%. Table 2 shows the

confusion matrix using Hidden Markov Models which can be compared to table 1. Computing the ZMs on a typical frame of the video segment takes 30 ms.
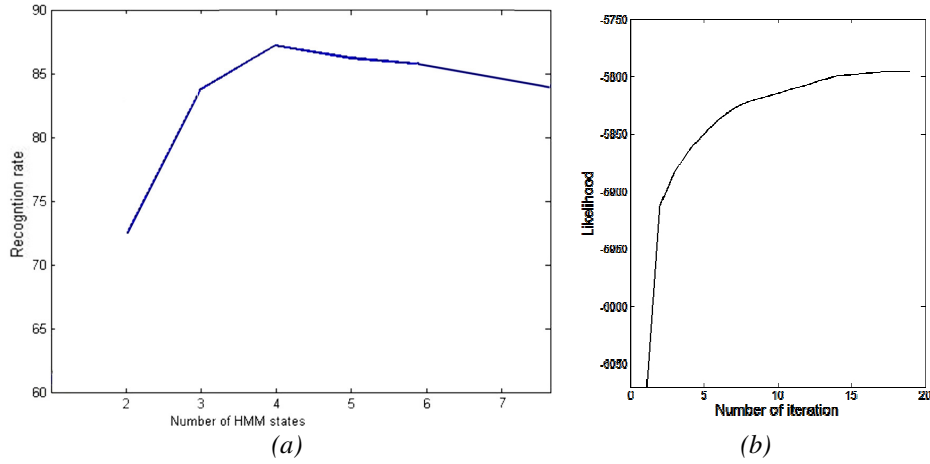

<div align="center">(a)　　　　　　　　　　　　(b)</div>

Figure 3. *(a)* shows that 4 is the optimal number of states for each of the HMM models. *(b)* shows the increasing likelihood of gesture A against the number of iterations during training.

		Predicted Gesture				
		"A"	"B"	"C"	"D"	"E"
Actual Gesture	"A"	**70**	8	0	2	0
	"B"	0	**78**	0	2	0
	"C"	0	1	**79**	0	0
	"D"	4	1	0	**75**	0
	"E"	0	0	0	0	**80**

Table 2. Confusion matrix presents the obtained results using the ZMs as shape descriptor to generate the training and testing sequences that been used to train and test 5 HMM models for the gestures.

7 Discussion and Conclusion

From the confusion matrices and relative accuracies obtained in the two experiments, it is clear that the ZM+HMM combination used in the second experiment has given substantially better results than the ZVM+CvR combination used in the first. However, it seems intuitively plausible that the displacement of the centre of mass between frames carries discriminative information on the set of gestures. To explore this further, we carried out a third experiment in

which this displacement is added to the feature vector of Zernike moments used in the HMM. This increased the mean accuracy obtained to 98.3%, although this only represents a small number of additional examples being correctly classified.

In this paper, we have introduced and investigated hand gesture recognition with Zernike Velocity Moments, previously used successfully for human gait analysis. The results yield a potential use for their simplicity, but not as good as Zernike moments coupled with a bank of HMMs. We aim to explore further the use of ZVMs spanning a short time interval to replace the ZMs in the HMM-based classifier. This would in principle provide an alternative way of introducing displacement into the HMM+ZM formation together with the discriminative abilities of an HMM for sequential data.

References

1. Y. Wu and T. S. Huang, Human Hand Modeling, Analysis and Animation in the Context of HCI, in *IEEE International Conference Image Processing* Kobe, Japan, 1999.

2. Q. Yuan, S. Sclaroff, and V. Athitsos, "Automatic 2D Hand Tracking in Video Sequences," in *IEEE Workshop on Applications of Computer Vision*, 2005.

3. H. Park, E. Kim, S. Jang, and H. Kim, An HMM Based Gesture Recognition for Perceptual User Interface, *Advances in Multimedia Information Processing,* pp. 1027-1034, 2004.

4. S. Carbini, J. E. Viallet, and O. Bernier, "Pointing Gesture Visual Recognition for Large Display," in *International Conference of Pattern Recognition*: IEEE Computer Society Press, 2004.

5. A. Sepehri, Y. Yacoob, and L. S. Davis, Parametric Hand Tracking for Recognition of Virtual Drawings, in *Proceeding of the fourth IEEE International Conference on Computer Vision Systems*: IEEE Computer Society Press, 2006.

6. M. Hu, "Visual Pattern Recognition by Moment Invariants," *IEEE Transactions On Information Theory,* vol. 8, pp. 179-187, 1962.

7. J. D. Shutler and M. S. Nixon, Zernike velocity moments for sequence-based description of moving features, *Image and Vision Computing,* vol. 24, pp. 343-356, 2006.

8. A. Heap, Learning deformable shapes models for object tracking, in *Computing School* Leeds: University of Leeds, 1998.

9. V. Athitsos and S. Sclaroff, An Appearance-based Framework for 3D Hand Shape Classification and Camera Viewpoint Estimation, in *Fifth IEEE International Conference Proceedings on Automatic Face and Gesture Recognition*. vol. 1: IEEE Computer Society Press, 2002.

10. E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, Visual Hand Tracking Using Nonparametric Belief Propagation, in *Conference on Computer Vision and Pattern Recognition Workshop*: IEEE Computer Society Press, 2004.

11. S. C. Ahn, T. Lee, I. Kim, Y. Kwon, and H. Kim, Large Display Interaction Using Video Avatar and Hand Gesture Recognition, in *ICIAR'04*: Springer-Verlag Berlin Heidelberg, 2004.

12. Q. Pham, L. Gond, J. Begard, N. Allezard, and P. Sayd, Real-Time Posture Analysis in a Crowd using Thermal Imaging, in *IEEE Conference on Computer Vision and Pattern Recognition CVPR '07.*: IEEE Computer Society Press, 2007.

10

13. C. W. Ng and S. Ranganath, Gesture Recognition via Pose Classification, in *15th International Conference on Pattern Recognition* vol. 3: IEEE Computer Society Press, 2000.

14. R. Bailey and M. Srinath, Orthogonal Moment Features for Use with Parametric and Non-Parametric Classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 389-399, 1996.

15. H. Hse and A. R. Newton, Sketched Symbol Recognition using Zernike Moments, in *17th International Conference on Pattern Recognition (ICPR'04).* vol. 1: IEEE Computer Society Press, 2004.

16. J. D. Shutler and M. S. Nixon, Zernike velocity moments for the description and recognition of moving shapes, in *Proceeding of British Machine Vision Conference BMVC.* vol. 2 Manchester, 2001.

17. M. J. Jones and J. M. Rehg, Statistical Color Models with Application to Skin Detection, *International Journal of Computer Vision,* pp. 81- 96, 2002.

18. G. R. Bradski, Computer Vision Face Tracking For Use in a Perceptual User Interface, *Intel Technology Journal,* 1998.

19. A. R. J. François, CAMSHIFT Tracker Design Experiments with Intel OpenCV and SAI, Institute for Robotics and Intelligent Systems IRIS-04-423, 2004.

20. E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, Using Model Trees for Classification, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

21. University of Waikato, Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/~ml/weka/, Accessed April 2008.

22. K. Murphy, Hidden Markov Model (HMM) Toolbox for Matlab, in *http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html*, Accessed April 2008.

23. L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *IEEE Transactions on Information Theory,* vol. 77, pp. 257-286, 1989.