Hand gesture selection and recognition for visual-based human-machine interface

Abdolah Chalechale
University of Wollongong

Farzad Safaei
University of Wollongong, farzad@uow.edu.au

Golshah Naghdy
University of Wollongong, golshah@uow.edu.au

Prashan Premaratne
University of Wollongong, prashan@uow.edu.au

http://ro.uow.edu.au/engpapers/5332

Publication Details
HAND GESTURE SELECTION AND RECOGNITION FOR VISUAL-BASED HUMAN-MACHINE INTERFACE

Abdolah Chalechale¹, Farzad Safaei¹, Golshah Naghdy², and Prashan Premaratne²

¹ Smart Internet Technology CRC
² School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
NSW, Australia
{abdolah,farzad,golshag,prashan}@uow.edu.au

ABSTRACT
A new paradigm has been proposed for gesture selection and recognition. The paradigm is based on statistical classification, which has applications in teledmedicine, virtual reality, computer games, and sign language studies. The aims of this paper are (1) how to select an appropriate set of gestures having a satisfactory level of discrimination power, and (2) comparison of invariant moments (conventional and Zernike) and geometric properties in recognizing hand gestures. Two-dimensional structures, namely cluster-property and cluster-features matrices, have been employed for gesture selection and to evaluate different gesture characteristics. Moment invariants, Zernike moments, and geometric features are employed for classification and recognition rates are compared. Comparative results confirm better performance of the geometric features.

1. INTRODUCTION
Visual-based interface plays an important role in new technologies including virtual reality, robotics, computerized medicine, and digital games. The Internet and world wide web dramatically extend the scope of this domain and make these technologies accessible to a rapidly growing audience. As a result, researchers are facing with web-based disciplines known as distributed virtual reality, telerobotics, telemedicine, and networked games. Applying visual-based interface in web-based applications needs special considerations mostly in the areas of (a) swift detection and recognition of the scene, where user’s input (gesture) is created, and (b) fast communications through the network nodes, where the result is visualized. This paper focuses on the former task, that is design and analysis of a set of visual inputs for controlling a general web-based application.

Relying on computer vision-based methods, and new range of gesture transmitting devices a computer can recognize the user’s gesture and perform appropriate actions required. In this type of interface, gestures are used as commands in the virtual world, or computer games. That is, instead of pressing keys on a keyboard, moving a mouse, or using joystick or trackball, human’s body gestures are used to create appropriate commands for a running program. In fact, new proposed interface translates user’s gestures into the user’s intentions as the input of the system. Translation is either through images, which are motion pictures taken of the user’s gesture, or through a wearable device such as a marked glove or a wireless motion transmitter [1]. As an interesting and context sensitive characteristic, a specific gesture may invoke different actions in different environments based on the interpretation defined in the driving engine. For example, turning down of thumb may cause a gun shoot in a fighting game, while it can cause turning right the plane in a flight simulator.

Several aspects of directing computers using human gestures have been studied in the literature, however gesture recognition is still an open problem. This is due to significant challenges in response time, reliability, economical constrains, and natural intuitive gesticulation restrictions [2]. The MPEG-4 standard has defined Facial Animation Parameters to analyze facial expressions and convert them to some predefined facial actions [3]. Jian et al. [4] has developed a lip tracking system using lip contour analysis and feature extraction. Similarly, human leg movement has been tracked using color marks placed on the shoes of the user to determine the type of leg movement using a first-order Markov model [5].

Motion qualities have been extracted from a live performance using neural network-based computing system [1]. The inputs to the system are both 3D motion capture, where position and orientation sensors collect data from the whole body of the performer, and 2D video projections. This system, which has been used in an extended project at the Center for Human Modeling and Simulation, University of Pennsylvania, provides the capability of automating both
observation and analysis processes. Finally it produces natural gestures for embodied communicative agents. The performer wears a black cloth in a dark background to facilitate hand and face detection tasks.

A method for recognizing hand gestures applying a model-based approach have been developed by Davis and Shah [6]. Here, a finite state machine is employed to model four qualitatively distinct phases of a generic gesture. Binary marked gloves are exploited to track fingertips. Gestures are represented as a list of vectors and are then matched to some stored gesture vectors using table lookup.

Geometric invariant moments have been widely used for gesture detection. Ng et al. [7] have proposed a system for automatic detection and recognition of human head gestures. It combines invariant moments and hidden Markov model (HMM) for feature extraction and recognition tasks, respectively. The best advantage of this approach is that it can operate in a relatively complex background. However, the computational requirements arising from the invariant moments extraction and HMM’s application render the approach inappropriate for real-time applications where several gestures are involved. As a result, the system can only recognize “YES”, “NO”, and “PO” head gestures. Zernike moments, on the other hand, using orthogonal basis functions, are less sensitive to noise than geometric moments and are more powerful in discriminating objects with n-fold symmetries. They are exploited for building a region-based shape descriptor in the MPEG-7 standard [8].

This paper proposes a new paradigm on how to select a set of appropriate hand gestures for applications aiming at visual-based interface. This is to find simple but robust gestures which could be easily recognized and have distinguishing features. This study addresses two aspects of gesture recognition for human-machine interface. First, which gestures are more recognizable, and second how to extract features which incorporate both recognition power and speed requirements in such applications. Experiments have been conducted to compare recognition powers of 3 different approaches including geometric and Zernike invariant moments and new proposed statistical classification based on single and multi-valued features.

The rest of the paper is organized as follows: next section includes mathematical definition of the geometric and Zernike moments. Section 3 explains our approach in detail and Section 4 presents comparative results and discussion. Finally, Section 5 concludes the paper and poses some new research directions.

2. BACKGROUND

Invariant moments are widely used in pattern recognition and image analysis [8]. Two main moment approaches are geometric moments and Zernike moments.

In general, geometric moments describe an image as a numeric function with respect to a reference axis or frame and defined as

\[ M_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x,y) dx dy \]  

where \( p, q = 0, 1, 2, \ldots, \infty \), and \( f(x,y) \) is the density distribution function of the image. It is known that if \( f(x,y) \) is piecewise continuous and has nonzero values only in a finite region of the \( xy \) plane, then moments of all orders exist. Importantly, the moment set \( \{ M_{pq} \} \) is uniquely determined by \( f(x,y) \). Conversely, \( f(x,y) \) is uniquely determined by \( \{ M_{pq} \} \).

The shape of an image (e.g., a hand gesture) could be represented in terms of seven functions defined on invariant moments \( \phi_i - \phi_7 \). They have been widely used in a number of applications. The first six functions \( \phi_1 - \phi_6 \) are invariant under rotation and the last one \( \phi_7 \) is both skew and rotation invariant. They are based on the central \( i - j \) moments \( \mu_{ij} \) of a 2D image \( f(x,y) \), which are defined as follows [9]

\[ \mu_{ij} = \sum_x \sum_y (x - \bar{x})^i (y - \bar{y})^j f(x,y) \]  

where \( \bar{x} = M_{10}/M_{00} \) and \( \bar{y} = M_{01}/M_{00} \) are indicating the center of mass of the shape. Then, defining \( \gamma = (i+j)/2 + 1 \) and \( \eta_{ij} = \mu_{ij}/\mu_{00} \), the invariant functions are obtained by

- \( \phi_1 = \eta_{00} + \eta_{02} \)
- \( \phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \)
- \( \phi_3 = (\eta_{30} + \eta_{02})^2 + (3\eta_{21} - \eta_{03})^2 \)
- \( \phi_4 = (\eta_{30} + \eta_{02})^2 + (\eta_{21} + \eta_{03})^2 \)
- \( \phi_5 = (\eta_{30} + \eta_{02})^2 + (\eta_{21} + \eta_{03})^2 \)  

\[ \cdot \left[ 3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right] \]

\[ + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \]

\[ \cdot \left[ 3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right] \]

\[ \cdot \left[ (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\right] \]

\[ + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \]

\[ \cdot \left[ (\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\right] \]

\[ - (\eta_{30} - \eta_{02})(\eta_{21} + \eta_{03}) \]

\[ \cdot \left[ 3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right] \]

\[ \cdot \left[ (\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\right] \]

The \( \phi \) functions make for each gesture image a 7-entry vector, which serves as a feature vector and is used for measuring the similarity between images. Although the retrieval speed is very high due to compact feature vector, there are some disadvantages. First, the geometric moments have the form of projection onto monomial \( x^p y^q \) that is the basis set which is complete but not orthogonal. Second, geometric
moments are vulnerable to noise, and for all objects that have n-fold symmetries they are all zero [10].

Zernike orthogonal polynomials are employed to derive Zernike moment invariants of an image \( f(x, y) \) as follows

\[
A_{nl} = \frac{n + 1}{\pi} \int_0^{2\pi} \int_0^\infty [V_{nl}(r, \theta)]^* f(r \cos \theta, r \sin \theta) r \, dr \, d\theta
\]

where \( n = 0, 1, 2, \ldots, \infty \), and \( l \) takes on positive and negative integer values subject to the condition that \( n - |l| \) is even and \( |l| \leq n \). The Zernike polynomial is defined as:

\[
V_{nl}(x, y) = R_{nl}(r)e^{j\theta}
\]

and the radial polynomial is:

\[
R_{nl}(r) = \sum_{s=0}^{(n-|l|)/2} (-1)^s \frac{(n-s)!}{s!(n+s)!} \frac{(n-s)!}{(n-s-|l|)!} r^{n-2s}
\]

Next, for each segmented-normalized gesture \( g \) belonging to a gesture group \( G_i \), \( i = 1, \ldots, I \); \( J \) shape properties \( P_j \), \( j = 1, \ldots, J \) are extracted. Currently, for the hand collection, \( I = 25 \) and \( J = 14 \) are chosen to be 14 corresponding to 25 gesture clusters and 14 predominant gesture properties respectively. The properties include seven geometric and seven invariant moment-based functions. Geometric properties are: area (pr), perimeter (pr), major axis length (mi), minor axis length (mi), eccentricity (ec), and the ratio of (mi) and standard deviation on how to evaluate possible gestures for recognition. A collection of 3800 hand gestures, called Hand Gesture Bank (HGB), is created and used in this study. This section shows how the proposed approach works on the HGB collection. The procedure can be adopted for many other collection of gestures without any need to change its general structure.

Initially, the collection is grouped into 25-hand alphabet [11]. The images are color in JPEG format extracted from video sequences. Fig. 1 depicts some examples of the images. The full color frame is converted to a gray intensity image by eliminating the hue and saturation components while retaining the luminance. Due to varying lighting conditions of the images using a unique threshold to binarize images is inadequate. Fig. 2 shows instances where a unique threshold cause inappropriate segmentation of the hand shape. To avoid this, K-means clustering [12] is employed for binarization in the pre-processing stage. This successfully segments hand gestures from the background.

Size normalization using nearest-neighbor interpolation is applied next. This is to achieve scale invariance property, which allows different size gestures to have similar features. The bounding box of the region of interest is found first and then normalized to \( w \times h \) pixels (64 x 64 pixels in our experiments).

3. HAND GESTURE SELECTION AND RECOGNITION

Gesture selection is a primary task of gesture-based visual interface design. This section presents a general paradigm for how to evaluate possible gestures for recognition. A collection of 3800 hand gestures, called Hand Gesture Bank (HGB), is created and used in this study. This section shows how the proposed approach works on the HGB collection. The procedure can be adopted for many other collection of gestures without any need to change its general structure.

Initially, the collection is grouped into 25-hand alphabet [11]. The images are color in JPEG format extracted from video sequences. Fig. 1 depicts some examples of the images. The full color frame is converted to a gray intensity image by eliminating the hue and saturation components while retaining the luminance. Due to varying lighting conditions of the images using a unique threshold to binarize images is inadequate. Fig. 2 shows instances where a unique threshold cause inappropriate segmentation of the hand shape. To avoid this, K-means clustering [12] is employed for binarization in the pre-processing stage. This successfully segments hand gestures from the background.

Size normalization using nearest-neighbor interpolation is applied next. This is to achieve scale invariance property, which allows different size gestures to have similar features. The bounding box of the region of interest is found first and then normalized to \( w \times h \) pixels (64 x 64 pixels in our experiments).

Next, for each segmented-normalized gesture \( g \) belonging to a gesture group \( G_i \), \( i = 1, \ldots, I \); \( J \) shape properties \( P_j \), \( j = 1, \ldots, J \) are extracted. Currently, for the hand collection, \( I = 25 \) and \( J = 14 \) is chosen to be 14 corresponding to 25 gesture clusters and 14 predominant gesture properties respectively. The properties include seven geometric and seven invariant moment-based functions. Geometric properties are: area (pr), perimeter (pr), major axis length (mi), minor axis length (mi), eccentricity (ec), and the ratio of (mi) and standard deviation on how to evaluate possible gestures for recognition. A collection of 3800 hand gestures, called Hand Gesture Bank (HGB), is created and used in this study. This section shows how the proposed approach works on the HGB collection. The procedure can be adopted for many other collection of gestures without any need to change its general structure.
deviation $\sigma_{ij}$. Therefore, the Bayes decision function have the following form [14]:

$$d_{ij}(g) = p(g/G_i)P(G_i)$$

that is identical as

$$d_{ij}(g) = \frac{1}{\sqrt{2\pi}\sigma_{ij}}e^{-\frac{(g - m_{ij})^2}{2\sigma_{ij}^2}}P(G_i)$$

for $i = 1 \ldots I$ and $j = 1 \ldots J$, where $p(g/G_i)$ is the probability density function of the gesture pattern $g$ from cluster $G_i$ and $P(G_i)$ is the probability of occurrence of the corresponding cluster.

Assuming equally likely occurrence of all classes (i.e., $P(G_1) = P(G_2) = \ldots = P(G_i) = \ldots = P(G_I) = 1/I$), and because of the exponential form of the Gaussian density, which persuade the use of natural logarithm, and since the logarithm is a monotonically increasing function, the decision function in Eq. 9 can be modified to a more convenient form. In other words, based on the aforementioned assumption and facts, the following decision function, which is less computationally expensive and much faster for the classification of hand gestures, is being used:

$$d_{ij}(g) = \ln[p(g/G_i)P(G_i)]$$

considering Eq. 9, it can be written as

$$d_{ij}(g) = -\frac{1}{2}\ln 2\pi - \ln \sigma_{ij} - \frac{(g - m_{ij})^2}{2\sigma_{ij}^2} + \ln P(G_i)$$

Dropping the constant values $-\frac{1}{2}\ln 2\pi$ and $\ln P(G_i)$, which have no effect on numerical order of the decision function, an expeditious decision function is obtained as

$$d_{ij}(g) = -\ln \sigma_{ij} - \frac{(g - m_{ij})^2}{2\sigma_{ij}^2}$$

for $i = 1 \ldots I$ and $j = 1 \ldots J$, where $m_{ij}$ and $\sigma_{ij}$ are the mean and standard deviation of gesture group $G_i$ using property $P_j$, and $g$ is the corresponding scalar property of an unknown gesture.

Utilizing the above classification approach we calculate recognition rates $R_{ij}$ for each single-valued property $P_j$ and for each gesture group $G_i$ and save them in the crossing cells of the corresponding rows and columns of the cluster-property matrix.

Next, to appraise a combinatory analysis and depict an efficient feature vector to be used for gesture recognition, a set of $K = 20$ different combinations of the geometric properties, invariant moment-based functions, and a Zernike feature vector (as described in [8]) is generated and recognition rates are obtained. Here, since the properties are multi-valued, the decision function for the classification is obtained differently. In the multi-valued case, the Gaussian density of the vectors in the $i$th gesture class has the form

$$p(\xi/G_i) = \frac{1}{(2\pi)^n/2|C_{ik}|^{1/2}}e^{-\frac{\frac{1}{2}(\xi - m_{ik})^T C_{ik}^{-1}(\xi - m_{ik})}{2}}$$

for $k = 1, 2, \ldots, K$, where $\xi$ is the extracted feature vector of an unknown gesture and $n$ is the dimensionality of the feature vectors, $\cdot^T$ indicates matrix determinant. Note that each density is specified completely by its mean vector $m_{ik}$ and covariance matrix $C_{ik}$, which are defined as

$$m_{ik} = E_{ik}\{\xi\}$$

and

$$C_{ik} = E_{ik}\{(\xi - m_{ik})(\xi - m_{ik})^T\}$$

where $E_{ik}\{\cdot\}$ denotes the expected value of the argument over the gestures of class $G_i$ using multiple-valued property $P_k$. Approximating the expected value $E_{ik}$ by the average value of the quantities in question yield an estimate of the mean vector and covariance matrix as

$$m_{ik} = \frac{1}{N_i} \sum_{\xi \in G_i} \xi$$
where $N_i$ is the number of gesture vectors from class $G_i$ and summation is taken over those vectors for $k = 1, 2, \ldots, K$.

To obtain a simple decision function for the multi-valued case, considering that the logarithm keeps numerical order of its argument, substituting Eq. 13 in $d_{ik}(\xi) = \ln \left[ p(\xi | G_i) P(G_i) \right]$ yields

$$d_{ik}(\xi) = -(n/2) \ln 2\pi - (1/2) \ln |C_{ik}| - (1/2) \left[ (\xi - m_{ik})^T C_{ik}^{-1} (\xi - m_{ik}) \right]$$

(18)

Once again, the term $-(n/2) \ln 2\pi$ is the same for all cases and if all classes are equally likely to occur, then $P(G_i) = 1/I$ for $i = 1, 2, \ldots, I$ that is a constant and has no effect on the numerical order of the decision function. Hence, a simple and expeditious decision function is obtained as

$$d_{ik}(\xi) = -\ln |C_{ik}| - (\xi - m_{ik})^T C_{ik}^{-1} (\xi - m_{ik})$$

(19)

for $i = 1 \ldots I$ and $k = 1 \ldots K$. Note that $C_{ik}$ values are independent of the input $\xi$, which means they can be calculated off-line and saved in a look-up table. They are fetched from the look-up table at on-line stage to accelerate decision making process.

The diagonal element $c_{ii}$ is the variance of the $r$th element of the gesture vector and the off-diagonal element $c_{rs}$ is the covariance of $x_r$ and $x_s$. When the elements $x_r$ and $x_s$ of the feature vector are statistically independent, $c_{rs} = 0$. This property has been used to identify autonomous features and to pick them in the combination of features in multiple-valued properties. Remarkably, this fact renders the multivariate Gaussian density function to the product of univariate density of each element of $\xi$ vector when the off-diagonal elements of the covariance matric $C_{ik}$ are zero. This in turn expedites the generation of the look-up table.

The recognition rates $R_{ik}$ for $i = 1 \ldots I$ and $k = 1 \ldots K$ are calculated utilizing Eq. 19 and saved in appropriate entries in another structure called $cluster$-features matrix. This represents not only the distinguishably of the isolated hand gestures but also the recognition power of different sets of features to describe gestures.

The general paradigm explained above provides a straightforward method to select distinguishable gestures and has been shown to be effective in experimental results (next section). More importantly, column-wise summations in the $cluster$-property and $cluster$-features matrices indicate the recognition power of the simple properties and complex features, respectively. Row-wise summations exhibit the discrimination power of each gesture, which is an important clue to the selection of gestures for the application in use.

## 4. EXPERIMENTAL RESULTS

The HGB collection, which is currently includes 3800 hand gestures, is used for the experiments. There are 25 sets of gestures having number of members from 60 to 200. In the training stage the statistical model parameters are obtained. These include means and standard deviations (scalars) for the individual properties and means (vectors) and covariance matrices for the combined features. In the recognition stage 500 different gestures (20 in each of 25 groups) were applied and classified using the approach explained in Section 2.

For each test gesture the singular properties and the feature vectors are obtained. These are to evaluate a specific gesture based on its geometric properties and feature sets respectively. The recognition rate in each entry in the $cluster$-property matrix is the number of correctly classified gestures divided by the number of inputs. For example, if 12 out of 20 input gestures in the cluster $G_{10}$ are correctly classified by the decision function given in Eq. 12 using perimeter property into the correct cluster, then the recognition rate in row $G_{10}$, column $pr$ of the $cluster$-property matrix is calculated to be 12/20=60%. In this part, 14 individual properties including 7 geometric and 7 conventional invariant moment-based values are examined for the 25 gesture groups. To be able to compare recognition power of different properties, an overall recognition rate is obtained for each column of the matrix by simply averaging the recognition rates in that column. The overall results show that the top three best singular properties are $mj$, $mi$, and $arl/pr$. The top ten best distinguishable gestures, which are explored using row-wise averaging of the recognition rates in the $cluster$-property matrix are also determined.

Next, test gestures are classified using 20 combinatory feature sets. The recognition rates are obtained using the decision function in Eq. 19 and the results are saved in the $cluster$-features matrix, which currently in our experiments has 20 columns. The rows corresponds to hand gesture clusters and the columns corresponds to a variety combination of features including geometric properties, conventional moment invariants, and Zernike moments. For Zernike moments, the MPEG-7’s procedure for shape description [8] has been followed. The number of entries in the feature vectors varying from 2 to 36. There are a massive number of different combinations but only those properties which previously showed to have better discriminating power were selected. These properties have tentatively been chosen based on their independent characteristics using covariance matrix. As the 2D structures (matrices) used for experiments are relatively large they are not presented.
here.

Moment-invariant functions showed lack of efficacy while Zernike moments exhibited superiority over the regular moments. Different combination of geometric properties poses higher recognition rates. An overall recognition rate of 96.3% is obtained using a five-entry feature vector \( \{ mj, mi, ec, ar, pr \} \).

5. CONCLUSION AND FURTHER WORK

A novel paradigm to select efficient hand gestures using 2D structures has been proposed. This include two matrices, where the first is for organizing recognition rates extracted from single-valued properties while the second is for multi-valued features. The recognition rates are obtained utilizing two simplified decision functions. Moment invariants, Zernike moments, and geometric features are employed for classification; recognition rates are extracted and saved in appropriate entries in the matrices.

The proposed approach can be used in telemedicine, virtual reality, video games and sign languages aimed at visual-based interface. A range of singular properties have been examined to discriminate hand gestures in a simple, fast, and robust way, which is necessary in real-time applications. The results explicitly show discrimination rank of individual hand gestures, which can be used to reasonably select appropriate gestures in a range of applications. Moreover, the combination of features have been tested and a small feature vector containing only five simple features yields an overall recognition rate of 96.3%.

The proposed approach can be applied in other disciplines including limb, head, and whole body gestures. Shape features extracted from the gesture image can be easily evaluated for efficacy using the proposed scheme. The proposed approach is to be employed in an immersive distributed environment, where several users using a distributed system communicates through their hand or body gestures. For further improvements, objective criteria for user satisfaction can be defined. Also a time manner comparison for different algorithm should be designed to show how efficient they work.

Acknowledgments. This work is supported by the Smart Internet Technology Cooperative Research Centre (SITCRC), Australia.

6. REFERENCES


