Hand posture analysis for visual-based human-machine interface

Abdolah Chalechale, Farzad Safaei
Smart Internet Technology CRC
University of Wollongong
Wollongong, NSW, 2522, Australia
{ac82,farzad}@uow.edu.au

Golshah Nagdy, Prashan Premaratne
School of Electrical, Computer and Telecom. Eng.
University of Wollongong
Wollongong, NSW, 2522, Australia
{golshah,prashan}@uow.edu.au

Abstract

This paper presents a new scheme for hand posture selection and recognition based on statistical classification. It has applications in telemedicine, virtual reality, computer games, and sign language studies. The focus is placed on (1) how to select an appropriate set of postures having a satisfactory level of discrimination power, and (2) comparison of geometric and moment invariant properties to recognize hand postures. We have introduced cluster-property and cluster-features matrices to ease posture selection and to evaluate different posture characteristics. Simple and fast decision functions are derived for classification, which expedite on-line decision making process. Experimental results confirm the efficacy of the proposed scheme where a compact set of geometric features yields a recognition rate of 98.8%.

1. Introduction

Human-machine interface (HMI) has become an essential part of our technological revolution. It offers both consumers and providers enormous opportunities for expanded access. However, as with any burgeoning technological innovation, HMI faces a wide array of possibilities. More generally, virtual reality, as an artificial creation of interactive environment resembling real life, is attracting more attention among researchers. Furthermore, in many telemedicine applications such as remote patient care and smart home-based health care devices, patients are remotely monitored. In such applications, ambient intelligence is integrated into the monitoring devices such as cameras in order to measure patients' gestures and postures.

The technology for on-line interaction in all of above applications over the Internet is maturing due to advances in communication tools and modern video transcoding expertise. Users usually interact with machines using keyboard, mouse, joystick, trackball, or wired glove. Most of these are special devices that, by and large, are designed to suit computer hardware rather than human user. Nevertheless, humans use gestures in daily life as a means of communication, for example hand shaking, head nodding, and hand gestures are widely used in friendly communications. Using machine vision algorithms, a computer can recognize the user's gesture/posture and perform appropriate actions required in virtual reality environments or in computer and video games. This paper aims at application of posture-based interaction in the areas like telemedicine, sign language recognition, virtual reality, and computer and video games.

Although several aspects of directing computers using human gestures/postures have been studied in the literature gesture/posture recognition is still an open problem. This is due to significant challenges in *response time*, *reliability*, *economical constrains*, and *natural intuitive gesticulation* restrictions [9]. The MPEG-4 standard has defined Facial Animation Parameters to analyze facial expressions and convert them to some predefined facial actions [6]. Principal component analysis has been used for hand posture recognition [2]. Jian *et al.* [8] 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 [3].

A neural network-based computing system has been used in [14] to extract motion qualities from a live performance. 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.

Davis and Shah [4] have developed a method for recognizing hand gestures applying a model-based approach. 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 broken to postures and represented as a list of vectors and are then matched to some stored vectors using table lookup.

Invariant moments have been widely used for gesture/posture detection. Ng *et al.* [11] have proposed a system for automatic detection and recognition of human head gestures/postures. 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/postures are involved. As a result, the system can only recognize "YES", "NO", and "PO" head gestures.

In some circumstances it is necessary to ignore motion path analysis of the gestures for fast processing. This kind of analysis is referred to as *posture analysis*. In this paper we propose a new discipline on how to depict a set of appropriate hand postures for applications aiming at visual-based interface. This is to find simple but robust postures which could be easily recognized and have distinguishing features. This study addresses two aspects of posture recognition for human-machine interface. First, which postures are more recognizable, and second how to extract features which incorporate both recognition power and speed requirements in such applications. Towards these goals, we have developed a novel methodology based on recognition rates and introduce two matrices: *cluster-property* and *cluster-features*. The former is a structure to save single-valued properties of the postures while the latter is for multiple-valued feature vectors describing posture images.

The rest of the paper is organized as follows: next section explains our approach in detail. Section 3 presents experimental results and finally Section 4 concludes the paper and poses some new research directions.

2 Hand Posture Analysis

One of the most important aspects of HMI in virtual reality, telemedicine, and computer games, where user communicates with the program's engine using his/her hand gestures/postures, is to reasonably select (or design) appropriate gestures/postures. This section presents a general scheme on how to assess several possibilities. To explain the proposed scheme we utilize a collection of 2080 hand postures [2, 12], and show how the approach works on this collection. The procedure can be adopted for other collections without any need to change its general structure.

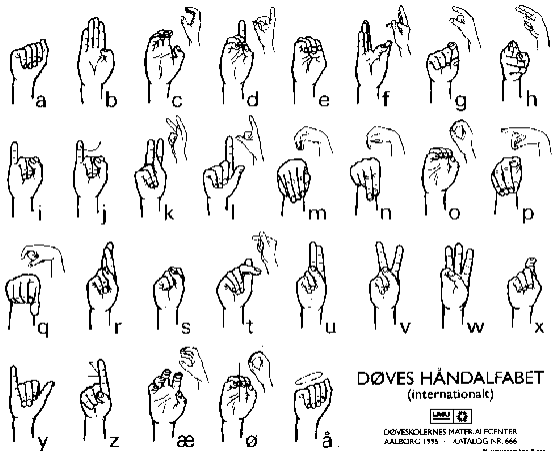

Figure 1. International sign language hand alphabet [2]

Initially, the collection is grouped into 25-hand alphabet. The images are 255-level gray scaled generated by a hand in black sleeve in a dark background. Figure 1 shows representative postures and Figure 2 depicts some examples of the images. Due to varying lighting conditions of the images within the database using a unique threshold to binarize images is inadequate. Figure 3 shows instances where a unique threshold cause inappropriate segmentation of the hand shape. For this, K-mean clustering is employed for binirization in the pre-processing stage. This successfully segments hand postures from the background (see Figure 3).

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

Next, for each segmented-normalized posture g belonging to a posture group G_i, $i = 1 \ldots I$, we extract J shape properties P_j, $j = 1 \ldots J$. Currently, for the hand collection, I is 25 and J is chosen to be 14 corresponding to 25 posture clusters and 14 predominant posture properties respectively. The properties include seven geometric and seven invariant moment-based functions. Geometric properties are: area (ar), perimeter (pr), major axis length (mj), minor axis length (mi), eccentricity (ec), and the ratio of ar/pr, and mj/mi. The invariant moment-based functions have been widely used in a number of applications [7, 13, 10]. The first six functions ($\phi_1 - \phi_6$) are invariant under rotation and the last one ϕ_7 is both skew and rotation invariant. They are based on the central i, j-th moments

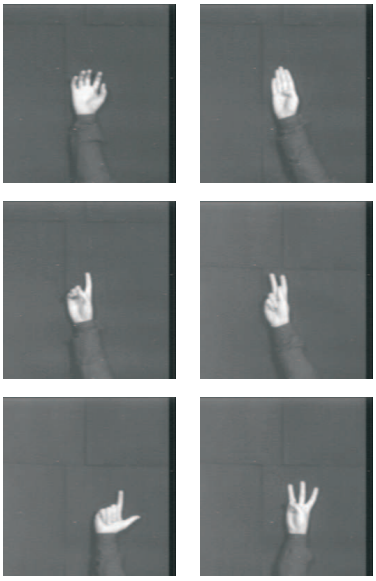

Figure 2. Hand posture samples

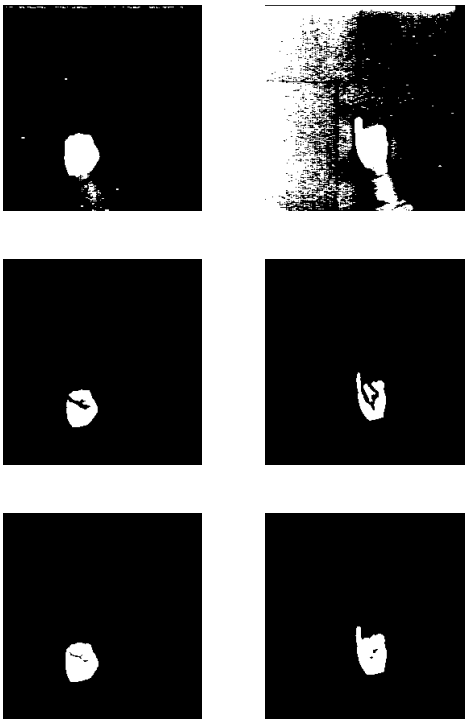

Figure 3. Instances where lower thresholds make many unwanted noisy regions (upper two images) and higher thresholds destroy the hand region (middle two images), while K-mean clustering segments hand region properly (lower two images)

(μ_{ij}) of a 2D image $f(x, y)$, which are defined as follows:

$$\mu_{ij} = \sum_x \sum_y (x - \bar{x})^i (y - \bar{y})^j f(x, y) \tag{1}$$

Then, the invariant moment-based functions are defined as

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \\
&\quad \cdot \left[3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad \cdot \left[3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \\
\phi_6 &= (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) \\
&\quad \cdot \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \\
&\quad \cdot \left[3(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right]
\end{aligned} \tag{2}$$

where $\eta_{ij} = (\mu_{ij})/(\mu_{00}^{\gamma})$ and $\gamma = (i + j)/2 + 1$.

To determine the recognition power of each G_i cluster, we exploit a classification scheme using the properties P_j. Initially; we try to classify 500 randomly selected postures (20 of each group) into the associated groups. Recognition rates R_{ij} for $i = 1 \ldots I$ and $j = 1 \ldots J$ are obtained and saved in appropriate entries in an *cluster-property matrix*. The classification is based on Bayesian rule assuming Gaussian distribution for the hand posture patterns [1, 2]. To extract a decision function for our classifier, we consider J number of 1D probability density functions. Each function involves I pattern groups governed by Gaussian densities, with means m_{ij} and standard deviation σ_{ij}. Therefore, the Bayes decision function have the following form [5]:

$$d_{ij}(g) = p(g/G_i)P(G_i) \tag{3}$$

that is identical as

$$d_{ij}(g) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{\left[-\frac{(g-m_{ij})^2}{2\sigma_{ij}^2}\right]} P(G_i) \tag{4}$$

for $i = 1 \ldots I$ and $j = 1 \ldots J$, where $p(g/G_i)$ is the probability density function of the posture 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., $P(G_1) = P(G_2) \cdots = P(G_i) \cdots = P(G_I) = 1/I$), 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. 4 can be modified to a more convenient form. In other words, based on the aforementioned assumption and facts, we can use the following decision function, which is less computationally expensive and much faster for the classification of hand postures:

$$\begin{aligned} d_{ij}(g) &= \ln\left[p(g/G_i)P(G_i)\right] \\ &= \ln p(g/G_i) + \ln P(G_i) \end{aligned} \quad (5)$$

considering Eq. 4, it can be written as

$$d_{ij}(g) = -\frac{1}{2}\ln 2\pi - \ln\sigma_{ij} - \frac{(g-m_{ij})^2}{2\sigma_{ij}^2} + \ln P(G_i) \quad (6)$$

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

$$d_{ij}(g) = -\ln\sigma_{ij} - \frac{(g-m_{ij})^2}{2\sigma_{ij}^2} \quad (7)$$

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

Utilizing the above classification approach we calculate recognition rates R_{ij} for each single-valued property P_j and for each posture 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 posture recognition, a set of $K = 18$ different combinations of the geometric properties and invariant moment-based functions is generated and recognition rates are obtained. Here, since the properties are multiple-valued, the decision function for the classification is obtained differently. In the multiple-valued case, the Gaussian density of the vectors in the ith posture class has the form

$$p(\xi/G_i) = \frac{1}{(2\pi)^{n/2}|C_{ik}|^{1/2}} e^{\left[-\frac{1}{2}(\xi-m_{ik})^T C_{ik}^{-1}(\xi-m_{ik})\right]} \quad (8)$$

for $k = 1, 2, \ldots K$, where ξ is the extracted feature vector of an unknown posture and n is the dimensionality of the feature vectors, $|\cdot|$ 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

$$m_{ik} = E_{ik}\{\xi\} \quad (9)$$

and

$$C_{ik} = E_{ik}\{(\xi - m_{ik})(\xi - m_{ik})^T\} \quad (10)$$

where $E_{ik}\{\cdot\}$ denotes the expected value of the argument over the postures 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

$$m_{ik} = \frac{1}{N_i}\sum_{\xi \in G_i} \xi \quad (11)$$

and

$$C_{ik} = \frac{1}{N_i}\sum_{\xi \in G_i} (\xi\xi^T - m_{ik}m_{ik}^T) \quad (12)$$

where N_i is the number of posture 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 multiple-valued case, considering that the logarithm keeps numerical order of its argument, substituting Eq. 8 in $d_{ik}(\xi) = \ln\left[p(\xi/G_i)P(G_i)\right]$ yields

$$\begin{aligned} d_{ik}(\xi) &= -(n/2)\ln 2\pi - (1/2)\ln|C_{ik}| - \\ &\quad (1/2)\left[(\xi - m_{ik})^T C_{ik}^{-1}(\xi - m_{ik})\right] - \\ &\quad \ln P(G_i) \end{aligned} \quad (13)$$

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

$$d_{ik}(\xi) = -\ln|C_{ik}| - (\xi - m_{ik})^T C_{ik}^{-1}(\xi - m_{ik}) \quad (14)$$

for $i = 1 \ldots I$ and $k = 1 \ldots K$. Note that C_{ik} values are independent of the input ξ, 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_{rr} is the variance of the rth element of the posture 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. Noteworthily, this fact renders the multivariate Gaussian density function to the product of univariate density of each element of ξ 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. 14 and saved in appropriate entries in another structure called *cluster-features matrix*. This

represents not only the distinguishably of the isolated hand postures but also the recognition power of different sets of features to describe postures.

The general paradigm explained above provides a straightforward method to select distinguishable postures 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 posture, which is an important clue to the selection of postures for the application in use.

3 Experimental Results

As stated before, a database of 2080 hand postures is used for the experiments. The database is publicly available in [12]. There are 25 sets of postures having number of members from 40 to 100. In the training stage the statistical model parameters are obtained. These include means and standard deviations (scalars) for individual properties and means (vectors) and covariance matrices for combined features. In the recognition stage 500 randomly selected postures (20 in each of 25 groups) from the database were applied and tried to do classification using the approach explained in Section 2.

For each test posture the singular properties and the feature vectors are obtained. These are to evaluate a specific posture 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 postures divided by the number of inputs. For example, if 12 out of 20 number of input postures in the cluster G_{10} are correctly classified by the decision function given in Eq. 7 using perimeter property into the same 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 (7 geometric and 7 invariant-based functions) are examined for the 25 posture 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 *ar/pr*. The top five best distinguishable postures, which are explored using row-wise averaging of the recognition rates in the *cluster-property matrix* are depicted in Figure 4.

Next, we tried to classify test postures using 18 combinatory feature sets. The recognition rates are obtained using the decision function in Eq. 14 and the results are saved in the *cluster-features matrix*, which currently in our experiments has 18 columns. The rows corresponds to hand posture clusters and the columns corresponds to a variety combination of features (feature vectors). The number of entries in the feature vectors varying from two to seven. There are a massive number of different combinations but we chose only those properties which previously showed to have better discriminating power. These properties have tentatively been chosen based on their independent characteristics using covariance matrices. The *cluster-property* and cluster-features matrices are relatively large and space limitation preclude us to represent them here.

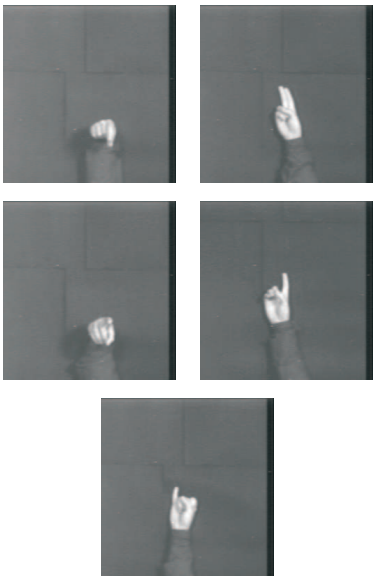

Figure 4. The top best five postures,in row-wise order, based on the data in the cluster-property matrix

Moment-invariant functions showed lack of efficacy while different combination of geometric properties exhibit higher recognition rates. The overall recognition rate of 98.8% is obtained using a five-entry feature vector $\{mj, mi, ec, ar, pr\}$.

4 Conclusion and Further Work

We proposed a novel paradigm to select efficient hand postures using *cluster-property* and *cluster-features matrices*. The former includes recognition rates for different postures using singular properties and the latter deals with multiple-valued features. The recognition rates are obtained utilizing two simplified decision functions. The proposed approach can be used in telemedicine, virtual reality, video games and sign languages aiming at visual-based interface. Moreover, we have examined several features to discriminate hand postures in a simple, fast, and robust way, which

is necessary in real-time applications. The results explicitly show discrimination rank of individual hand postures, which can be used to reasonably select appropriate postures in different applications. Moreover, the combination of features have been examined and a small feature vector containing only five simple features yields an overall recognition rate of 98.8%.

The proposed approach can be applied on other postures including limb, head, and whole body postures. Shape features extracted from the posture image can be easily evaluated for efficacy using the proposed scheme. Moreover, we intend to employ the proposed approach in immersive distributed environments, where several users using a distributed system communicate through their hand or body gestures/postures. For further improvements, objective criteria for user satisfaction can be defined and a time-based comparison can be accomplished.

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

References

[1] H. Birk and T. B. Moeslund. Recognizing gestures from the hand alphabet using principal component analysis. Master's thesis, Laboratory of Image Analysis, Aalborg University, 1996.

[2] H. Birk, T. B. Moeslund, and C. B. Madsen. Real-time recognition of hand gestures using principal component analysis. In *Proc. 10th Scandinavian Conf. on Image Analysis (SCIA'97)*, 1997.

[3] C.-C. Chang and W.-H. Tsai. Vision-based tracking and interpretation of human leg movement for virtual reality applications. *IEEE Trans. Circuits and Systems for Video Technology*, 11(1):9–24, 2001.

[4] J. Davis and M. Shah. Visual gesture recognition. *IEE Proc. Vision, Image and Signal Processing*, 141(2):101–106, 1994.

[5] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley, 1992.

[6] ISO/IEC JTC 1/SC 29/WG 11 N 2502. Information technology-generic coding of audio-visual objects-part 2: visual. Technical report, ISO/IEC, Atlantic City, Oct. 1998.

[7] A. J. Jain and A. Vailaya. Shape-based retrieval: a case study with trademark image databases. *Patt. Recog.*, 31(9):1369–1390, 1998.

[8] Z. Jian, M. N. Kaynak, A. D. Cheok, and K. C. Chung. Real-time lip tracking for virtual lip implementation in virtual environments and computer games. In *Proc. IEEE Int. Conf. Fuzzy Systems*, volume 3, pages 1359–1362, 2001.

[9] H. Kang, C. W. Lee, and k. Jung. Recognition-based gesture spotting in video games. *Pattern Recognition Letters*, 25(15):1701–1714, 2004.

[10] D. Mohamad, G. Sulong, and S. S. Ipson. Trademark matching using invariant moments. In *Proc. second Asian Conf. Comput. Vision, [ACVV'95]*., volume 1, pages 439–444, Singapore, 1995.

[11] P. C. Ng and L. C. D. Silva. Head gestures recognition. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, volume 3, pages 266–269, 2001.

[12] Thomas Moeslund's Gesture Recognition Database. http://www.vision.auc.dk/%7etbm/gestures/database.html/. The URL has been visited on 10/2/2005.

[13] S. J. Yoon, D. K. Park, S. Park, and C. S. Won. Image retrieval using a novel relevance feedback for edge histogram descriptor of MPEG-7. In *Proc. IEEE Int. Conf. Consumer Electronics*, pages 354–355, Piscataway, NJ, USA, 2001.

[14] L. Zhao and N. I. Badler. Acquiring and validating motion qualities from live limb gestures. *Graphical Models*, 67(1):1–16, 2005.