Hand gesture understanding for human-computer interaction.

Ruben Posada-Gomez, J. I. Rivalcoba-Rivas, Giner Alor-Hernandez,
Albino Martínez-Sibaja, Juan Antonio Quintana Silva
Departamento de Postgrado e Investigación, Instituto Tecnológico de Orizaba
Av. Ote. 6 No 852, Col. Emiliano Zapata, Orizaba, Veracruz, México
{rposada, jrevalcova, galor, amartinez, jquintana} @itorizaba.edu.mx

Abstract

Body gesture recognition and understanding is an important task in computer vision for machine and human interaction. This paper shows a first system for hand gesture understanding (to identify several manual signs). The system will be able to understand hand gestures and once identified, the sign will be associated with a command, using a database recorded previously in the computer.

The system can be used for communicate some instructions to a personal computer through a common webcam. Future works would allow using this kind of system as an intuitive communication interface for people with speech disorders.

To achieve the sign recognizing, some geometric classifier has been chosen. Under this approach is possible to model each element of a sign universe. Hu invariants are the geometric classifier chosen which derive of geometric moments and are constant under changes of translations, rotations and scale.

1. Introduction

The interactions between humans and computers needs of in/out interfaces. Mouse and keyboard are the most common input interfaces, but typing in a keyboard is not natural or intuitive. When anybody acquires the typing ability, it does not have much meaning, because there is not an explanation sensible why the alphanumeric keys are located in a certain way [1]. A more natural communication way (after the oral communication) is the signs language; in fact gestures and body language communicate as effectively as words [2], For instance, it has been demonstrated that young children learn to communicate with gesture before they learn to talk [3,4].

The roots of gestural communication become obvious watching a human baby grows, it is easy to see that we learn much more about the world by manipulating it than by simply observing.

The motivation of this work is developing a new communication way between humans and computers, looking for a more intuitive and natural way to transmit commands. This work consists in the automatic sign language identification (recognizing six different signs that could mean six different commands) with a personal computer. A web cam will acquire the images using an acquisition system that facilitate the signs recognition.

For the first experiences, each used sign is represented by a different hand surface. Then the Hu invariants are determined for this image, finally a distance measure leaves to know which sign of a database is the closest. It is important to remark that this work could help many individuals with speech problems to having a better insertion in their environment. Actually silent people have problems to access a great number of public services because of their communication characteristics; even if silent translators and hand writing can help to communicate, these solutions are expensive and slow (mainly for hand writing) [5,6].

As a first approach, it is proposed to recognize at least six different patterns. However this communication system can be improve for detection of more different signs and detecting the movement of the manual signs using matching algorithms. The goal is to determine which of previously recorded sign correspond to actual generated hand sign.

2. Methodology

One of the objectives of this work is to provide a support for developing (in a future work), the hardware system implementation, that is why the simplicity of methodology is essential. The hand gesture identification system HGIS (Fig. 1.) is divided in the next stages: Image acquisition and representation, Color processing, Image processing and segmentation, Extraction of geometric characteristics and Euclidean distance classifier.

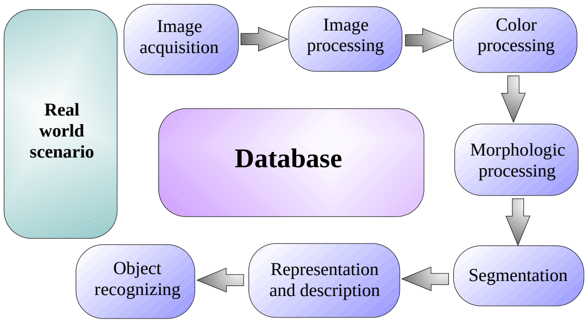

Fig. 1 Block diagram of the HGIS, each one of this component could be in contact with the database, the database contain the universe of signs that the computer is enabled to recognize**.**

2.1. Image acquisition and representation

Let f(c,y) the two dimensional mathematically representation of an image and their spectral components ($f_1, f_2, \ldots, f_n$), where each one of those components represents the intensity from the image to different wave longitudes in the coordinates (x,y). This combination of three wave longitudes is called tri–stimuli:

$$f(x,y) = [\, f_{red}(x,y),\ f_{green}(x,y),\ f_{blue}(x,y)] \qquad (1)$$

The acquisition of the image is carried out through the webcam, so an image matrix representation for this image I then obtained:

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \ldots & f(0,N-1) \\ f(1,0) & f(1,1) & \ldots & f(1,N-1) \\ \ldots & \ldots & & \ldots \\ f(M-1,0) & f(M-1,1) & \ldots & f(M-1,N-1) \end{bmatrix} \qquad (2)$$

A prototype has been designed for the images acquisition; the proposed design of this acquisition system is shown in fig 2. A webcam is located at the top of the structure. The camera is setting on two supports that give the user the possibility of place his hands easily at the bottom at the camera field of vision. To simplify the subsequent segmentation stage, a black background is located behind the hand. In future works algorithms were added for achieve the skin detection, for avoid this background.

Finally, it is necessary that a black bracelet be employed by user, this bracelet serve as divider between the hand and the arm, this division will be useful in later stages of segmentation.

2.2. Color processing

The discretization process is necessary for having a matrix in which the gesture identification could be done, but this process is completed by web-cam and PC interface. The mathematic representation of spatial sampling is:

$$I(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x-x_0, y-y_0)dxdy \qquad (3)$$

The camera used in the HGIS is a color web cam, which produces a color image represented by 3 matrixes. The next task is obtaining a gray level image, using Grassman's Laws given by

$$I_{gs}(x,\ y) = 0.3I_r(x,\ y) + 0.59I_g(x,\ y) + 0.11I_b(x,\ y) \, (4)$$

2.3. Image processing and segmentation

For simplify the image recognition process, the gray scale image is converted into a binary image. The binary image is obtained using an appropriate thresholding operation that sets all pixels at 0 or 1 according to their value. Thresholding is frequently used technique in the segmentation process in computer vision systems, especially when there is a great quantity of data to process.

The black background presents a disadvantage in the acquisition because it increases the noise of image. To reduce this effect a mean filter was used, which is usually recommended for this kind of noises, then given a gray scale image $I_{gs}(x,y)$ their corresponding filtered image is given as:

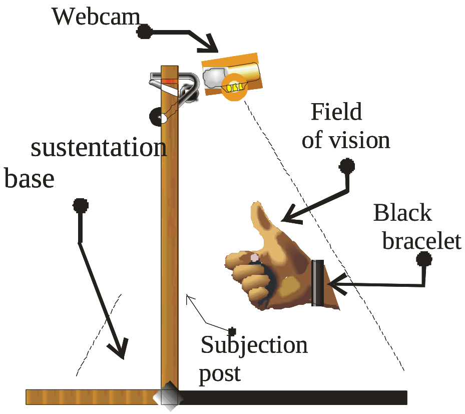

Fig. 2 Parts of acquisition system.

$$I_{fil}\left(x,\ y\right) = mean\left\{N_8\left[I_{gs}\left(x,\ y\right)\right]\right\} \qquad (5)$$

Where $N_8(P)$ it represents the 8 neighboring pixels of a P pixel.

Given the filtered image $I_{fil}(x,y)$, it is necessary to detect the area of interest that will be useful for the recognition of the sign, for do it, a threshold level “U” that can value 0 to L-1 is chosen; being L the levels of gray of acquired image. The procedure to find the U value is:

1. An initial U value is selected as the average between max and min values of I_{fil}.
2. The image in segmented in two regions using U, two groups are obtained, G1 with pixels >= U and G2 with pixels < U.
3. u_1 and u_2 are obtained as the average of G1 and G2 respectively, then a new U is given by $U=0.5(u_1+u_2)$.
4. Repeat the steps 2 and 3 until the difference in U will be smaller than 0.5.

Having the definitive value of U, the image is binarized:

$$Iu\left(x,y\right) = \begin{cases} 1; I_{fil}\left(x,\ y\right) \geq U \\ 0; I_{fil}\left(x,\ y\right) < U \end{cases} \qquad (6)$$

With the segmented image Iu(x,y), it is possible to separate the arm of the hand in the image using the back bracelet, after the segmentation they are two objects one is the hand, which is the object of interest and the other one is the forearm the object that is wanted to eliminate, to achieve the selection of interest object, the next algorithm is implemented:

1. Search in Iu(x,y), the first pixel that values 1, and to assign him a new label “E”.
2. In a recursive algorithm, to assign the label E to all their 4 neighbors.
3. Stop if they are not more pixels of object type.
4. Go to step 1.

Fig. 3 shows the result of apply the algorithm for obtaining the hand in an image.

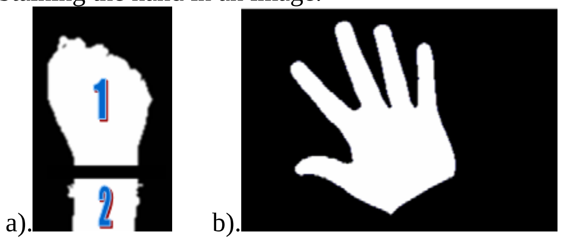

Fig. 3. Image segmentation, a) two objects in an image, b) Selection of the hand in the image.

2.4. Extraction of geometric characteristics

When the hand is isolated of the scene, the next stage is to obtain a vector of patterns or characteristic that represent to each sign, these will be called discriminants, like the purpose of this work is to recognize signs done with the hand, it is necessary to have two fundamental factors for the selection of thee discriminants:

1. The segmented form can change depending of distance between the camera and the hand.
2. As the user's hand has not restriction of space, the main axes in the segmented form can vary in small angles and the center of mass will never be fixed.

Considering the last points, it is possible to take the Hu invariants like discriminants, because are very useful in the recognition for two dimensional images, since they are robust under transformations like scale changes, rotations and translations, That is the reason why it were chosen as discriminates that modeled each class of signs. The Hu invariants derive of the geometric moments, for the case of a continuous function f(x, y), the one first order moment (p+q) is given as:

$$m_{pq} = \int_{-\infty}^{\infty} x^p y^q f\left(x,y\right) dxdy \qquad (7)$$

For p,q = 1,2,3,4… the infinite group of moments determines in unique form each function f(x,y), in a reciprocal way to the group of moments (m_{pq}) determine an only function f(x,y), it is possible to take advantage of the moments occupying their central form, this is that the center of mass of the object coincides with the coordinate (0,0). Then the central moments was defined as:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(x-\overline{x}\right)^p \left(y-\overline{y}\right)^q dxdy \qquad (8)$$

In this case:

$$\overline{x} = \frac{m_{10}}{m_{00}}; \overline{y} = \frac{m_{01}}{m_{00}} \qquad (9)$$

This corresponds to the center of mass of the object. And for the case of a digital image they are

$$\mu_{pq} = \sum_x \sum_y \left(x - \bar{x}\right)^p \left(y - \bar{y}\right)^q f\left(x, y\right) \qquad (10)$$

In the case of a binary image the equation 10 takes the form:

$$m_{pq} = \sum_x \sum_y \left(x - \bar{x}\right)^p \left(y - \bar{y}\right)^q \qquad (11)$$

Where ***(x, y)*** $\in$ ***R***, being R the region of the object of interest. The most used central moments are until those of third order which are expressed in function of the geometric moments as:

$$\mu_{00} = m_{00} \qquad (12)$$

$$\mu_{10} = 0 \qquad (13)$$

$$\mu_{01} = 0 \qquad (14)$$

$$\mu_{11} = m_{11} - \bar{y}m_{10} \qquad (15)$$

$$\mu_{20} = m_{20} - \bar{x}m_{10} \qquad (16)$$

$$\mu_{02} = m_{02} - \bar{y}m_{01} \qquad (17)$$

$$\mu_{30} = m_{30} - 3\bar{x}m_{20} + 2\bar{x}^2 m_{10} \qquad (18)$$

$$\mu_{03} = m_{03} - 3\bar{y}m_{02} + 2\bar{y}^2 m_{01} \qquad (19)$$

$$\mu_{21} = m_{21} - 2\bar{x}m_{11} - \bar{y}m_{20} + 2\bar{x}^2 m_{01} \qquad (20)$$

$$\mu_{12} = m_{12} - 2\bar{y}m_{11} - \bar{x}m_{02} + 2\bar{y}^2 m_{10} \qquad (21)$$

From the central moments it is possible to obtain its normalized version, which is defined as $\boldsymbol{\eta_{pq}}$:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu^{\gamma}{}_{00}} \qquad (22)$$

Where:

$$\gamma = \frac{p+q}{2} + 1 \qquad (23)$$

Is finally from these equations that the Hu invariants are defined as:

$$\phi_1 = \eta_{20} + \eta_{02} \qquad (24)$$

$$\phi_2 = \left(\eta_{20} - \eta_{02}\right)^2 + 4\eta^2{}_{11} \qquad (25)$$

$$\phi_3 = \left(\eta_{30} + 3\eta_{12}\right)^2 + \left(3\eta_{21} - \eta_{03}\right)^2 \qquad (26)$$

$$\phi_4 = \left(\eta_{30} + \eta_{12}\right)^2 + \left(\eta_{21} + \eta_{03}\right)^2 \qquad (27)$$

$$\begin{aligned}\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \\ &+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]\end{aligned} \qquad (28)$$

$$\begin{aligned}\phi_6 &= (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \\ &+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \qquad (29)$$

$$\begin{aligned}\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} - \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \\ &(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]\end{aligned} \qquad (30)$$

$$\mu_{03} = m_{03} - 3\bar{y}m_{02} + 2\bar{y}^2 m_{01} \qquad (31)$$

$$\mu_{21} = m_{21} - 2\bar{x}m_{11} - \bar{y}m_{20} + 2\bar{x}^2 m_{01} \qquad (32)$$

$$\mu_{12} = m_{12} - 2\bar{y}m_{11} - \bar{x}m_{02} + 2\bar{y}^2 m_{10} \qquad (33)$$

The next step is to take "n" samples of the invariants of each sign; these samples will be stored in a vector that represented each class of signs.

2.5. Euclidean distance classifier

The recognition based on the Euclidean distance comes from two hypotheses [7], in the first place the classes will be of deterministic nature, and in second place that all the necessary information for their design is available a priori. Also the classes should be lineally separable. In the case of this classifier the three conditions are completed. The stages for the classifier will be:

1. Determination of the number of classes: In this case there are five classes, each one represented for their corresponding sign.
2. Selection of features: In this case the Hu invariants are taken as features for objects identification.
3. Prototypes calculus: To calculate each one of those prototypes that the classifier will store in memory the equation (34) will be used, this stage is also called in the literature like training of the classifier, because the computer will be aided of the information of the array Z_i for carry out a correct classification.

$$Z_i = \frac{1}{M} \sum_{j=1}^{M} X_{ij} \qquad (34)$$

4. Test of the classifier: Once the prototype array is calculated, it is necessary to test the classifier performance, introducing to the classifier a test array X, preferably a different to the employed in the train.

Finally the form for determine to that class belongs an X entrance pattern is through:

$$\textbf{Arg min } d(\boldsymbol{x}, \boldsymbol{Z}_i) \qquad (35)$$

Where :

$$d\left(x,\ Z_i\right) = \left[\sum_{i=1}^{N}\left(x_j - Z_{ij}\right)^2\right]^{\frac{1}{2}} \qquad (36)$$

3. Test and results

The developed system is not independent of the user, it is necessary to begin a phase of training for each new user. The test has been carried out in conditions of controlled light, and against a black background, the Fig. 4 shown the signs universe for the first tests.

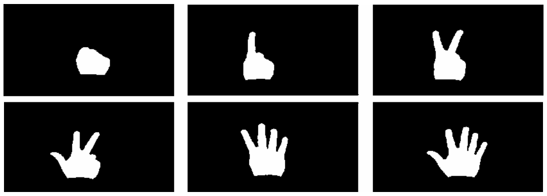

Fig. 4. Signs universe for test of the classifier.

Once the classifier is training with the signs universe, test were done with different positions of signs for determining is the classifier can detect a signs in different conditions of rotation, translation of scale factor (Fig. 5). From Table 1 it is possible to see that results have been quite satisfactory because the system recognizes the sign perfectly made by the user to 99% once this it has been trained.

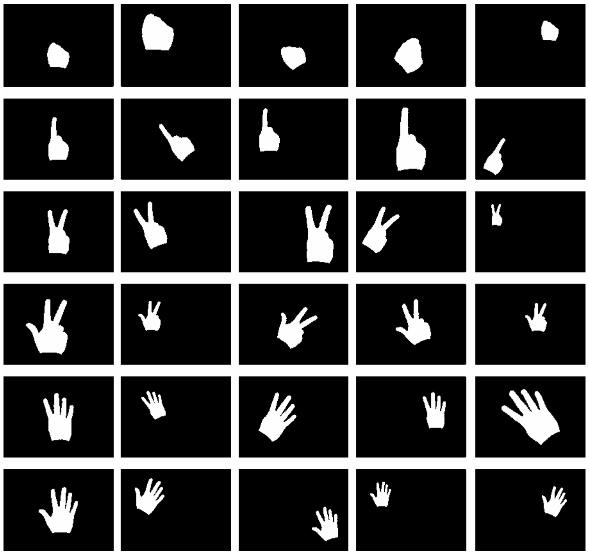

Fig. 5 Different acquisitions of the signs used in the classifier test.

The system to future is planned to prove in the one set real, controlling an object at distance.

SIGN	% OF SUCCES
SIGN 0	99%
SIGN 1	99%
SIGN 2	99%
SIGN 3	99%
SIGN 4	99%
SIGN 5	99%

Table 1. Results of identification of signs

4. Related Works

In [8], some issues involved in controlling computer applications via gestures composed of both static symbols and dynamic motions are discussed. This approach models each gesture from either static model information or a linear-in-parameters dynamic system. The hand tracking method is based primarily on color matching and is performed in several distinct stages. Each stage analyzes and reduces the information into more easily interpretable and relevant representation. A vision-based hand tracking system for gesture based man-machine interactions and a statistical hand detection method is presented in [9]. The hand tracking system employs multiple cameras to reduce occlusion problems. To detect hand regions in images, a statistical method by using geometrical structures involved in the appearances of the target objects is proposed. This method can describe and recognize the appearances of hands based on geometrical structures. In [10], a robust hand tracking and gesture recognition method for wearable visual interfaces are proposed. The method integrates shape and depth information for robust hand tracking. Gesture recognition is realized through the maximum posterior estimation of several pre-defined gestures.

Recent works have proposed approaches more sophisticated. A framework for 3D hand tracking and dynamic gesture recognition by using a single camera is presented in [11]. Hand tracking is performed in a two step process: (1) a 3D hand posture hypothesis using geometric and kinematics inverse transformations is generated and; (2) the hypothesis is validated by projecting the postures on the image plane and comparing the projected model with the ground truth using a probabilistic observation model. Dynamic gesture recognition is performed by using a Dynamic Bayesian Network model. In [12], a method for the simultaneous localization and recognition of dynamic hand gestures is proposed. For doing this, a dynamic space-time warping (DSTW) algorithm which aligns a pair of query and model gestures in both space and time is presented. This approach includes translation invariant recognition of gestures, a desirable property for many HCI systems. The performance of the approach is evaluated on a dataset of hand signed digits gestured by people wearing short sleeve shirts, in front of a background containing other non-hand skin colored objects. An approach for deaf-people interfacing using computer vision is presented in [13]. The recognition of alphabetic static signs of the Spanish Sign Language is addressed. The proposed approach combines a number of norms to evaluate the distance of the current sign, to the sign models stored in a dictionary.

5. Conclusion

An automatic recognition of signs system has been presented, it has been shown than the use of HU invariants of HU could be an advisable election in the tasks of recognition of plane objects.

The success in the recognition of the system depends in a certain way of the segmentation stage, so that if the system is under conditions of to much light, the camera will be saturated causing a bad segmentation of the scene.

In future works a stage of contrast improvement will be incorporated and detection of human skin that will make the system more robust.

6. References

[1] Larry Long, *Introducción a las computadoras y al procesamiento de información*, Ed. Prentice Hall Hispanoamericana S.S 1990.

[2] Axtell, Roger E. *Gestures: The Do's and Taboos of Hosting International Visitors*. John Wiley & Sons, 1990.

[3] Acredolo, L., & Goodwyn, S., *Baby Signs*, Contemporary Books, Inc., 1996.

[4] Kendon, A. "*Some relationships between body motion and speech*" In ***A.** W. Siegman and B Pope* (Eds.), Studies in Dynamic Communications, new York, Pergamon Press.

[5] Elena Sanchez Nielsen "*An autonomous and userindependent hand posture recognition system for vision-based interface task*" Departamento de Computacion de la Laguna, Spain**,** *revista del Instituto de Sistemas Inteligentes* 2004.

[6] Jorg Zieren; Karl-Friedrich Kraiss "*Non-Intrusive Sign Language Recognition For Human-Computer Interaction" WISDOM (Wireless Information Services for Deaf people On the Move) Chair of Technical Computer Science*, *RWTH* Aachen University, Germany, 2004.

[7] Jan Flusser "*Moment invariants in Image Analysis*" *Transaction on Engineering, computing and technology*, Febrary 2006.

[8] Charles J. Cohen, Glenn Beach, Gene Foulk. "A Basic Hand Gesture Control System for PC Applications". In Proceedings of the 30th Applied Imagery Pattern Recognition Workshop (AIPR'01). IEEE Press.

[9] Akira Utsumi, Nobuji Tetsutani and Seiji Igi. "Hand Detection and Tracking using Pixel Value Distribution Model for Multiple-Camera-Based Gesture Interactions". In Proceedings of the IEEE Workshop on Knowledge Media Networking (KMN'02). IEEE Press.

[10] Yang Liu and Yun de Jia. "A Robust Hand Tracking and Gesture Recognition Method for Wearable Visual Interfaces and Its Applications". In Proceedings of the Third International Conference on Image and Graphics (ICIG'04). IEEE Press.

[11] Ayman El-Sawah, Chris Joslin, Nicolas D. Georganas, Emil M. Petriu. "A Framework for 3D Hand Tracking and Gesture Recognition using Elements of Genetic Programming". In Proceedings of the Fourth Canadian Conference on Computer and Robot Vision (CRV'07). IEEE Press.

[12] Jonathan Alon, Vassilis Athitsos, Quan Yuan, and Stan Sclaroff. "Simultaneous Localization and Recognition of Dynamic Hand Gestures". In Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05). IEEE Press.

[13] Isaac García Incertis, Jaime Gómez García-Bermejo, Eduardo Zalama Casanova. "Hand Gesture Recognition for Deaf People Interfacing". In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06). IEEE Press.