PLATFORM TO DRIVE AN INTELLIGENT WHEELCHAIR USING FACIAL EXPRESSIONS

Pedro Miguel Faria, Rodrigo A. M. Braga, Eduardo Valgôde and Luís Paulo Reis
LIACC – Artificial Intelligence and Computer Science Lab. – University of Porto, Portugal

FEUP – Faculty of Engineering of University of Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal

Keywords: Human-computer interface, computer vision, image processing, artificial intelligence, intelligent wheelchair.

Abstract: Many of the physically injured use electric wheelchairs as an aid to locomotion. Usually, for commanding this type of wheelchair, it is required the use of one's hands and this poses a problem to those who, besides being unable to use their legs, are also unable to properly use their hands. The aim of the work described here, is to create a prototype of a wheelchair command interface that do not require hand usage. Facial expressions were chosen instead, to provide the necessary visual information for the interface to recognize user commands. The facial expressions are captured by a digital camera and interpreted by an application running on a laptop computer on the wheelchair. The software includes digital image processing algorithms for feature detection, such as colour segmentation and edge detection, followed by the application of a neural network that uses these features to detect the desired facial expressions. A simple simulator, built on top of the known (Ciber-Mouse) was used to validate the approach by simulating the control of the intelligent wheelchair in a hospital environment. The results obtained from the platform provide strong evidence that it is possible to comfortably drive an intelligent wheelchair using facial expressions.

1 INTRODUCTION

1.1 Motivation

Several physical disabilities or diseases result in severe impairments to mobility. Spinal chord damage, cerebral palsy and several other conditions can result in the loss of movement of some or all four limbs. Usually people in these situations are either completely dependant on others, or have suitable technological help for moving from place to place. The most common aid for this kind of disability is the electric wheelchair, which is quite good in what concerns motion, but frequently inadequate in the user interface. The standard command method available is the joystick that can only be used by people with relatively good hand dexterity, and impossible to be used by people who have limited control of the body and only can move some muscles of the face, and those movements will be the only possible interaction available with an intelligent wheelchair. Facial expressions may be interpreted basically as a communication language, used voluntarily but also, quite often, involuntarily, by people, to show their emotions. It is considered that they are inborn and cultural independent. This can be observed in blind people since birth, as they show the same relationship between facial expressions and emotions, as normal people.

Although there is a strong relationship between emotion and facial expression, there is a wide range of face movements not connected with emotions, such as eye blinking or raising only one eyebrow.

Ekman and Friesen (Ekman, 1978) developed the Facial Action Coding System (FACS), which describes thoroughly all the visually perceptible face movements. This system defines all individual face movements and facial expressions as their possible combinations and was used to select some basic expressions for the command language.

1.2 Objectives and Restrictions

The main objective of this work was to create an interface that allows a person to drive a wheelchair, using only facial expression in an easy, practical and robust fashion. The hardware used to achieve these goals, a digital camera and a PC, and the software

Miguel Faria P., A. M. Braga R., Valgôde E. and Paulo Reis L. (2007).
PLATFORM TO DRIVE AN INTELLIGENT WHEELCHAIR USING FACIAL EXPRESSIONS.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - HCI*, pages 164-169
DOI: 10.5220/0002394101640169

developed in C++, was the main tools used to put this idea into practice. The prototype developed had to meet certain restrictions in order to achieve good results:

- **Time:** the software development process bared in mind that it should be as fast as possible (it is mandatory that the time interval between the command and the action is as small as possible).
- **Lighting conditions:** image processing, with varying illumination, demands more complex algorithms in order to achieve good results (one restriction was that the system should only operate indoors in good light conditions).
- **Regular background:** the face detection is done through colour detection and therefore the background must be of a different colour.
- **Face Type:** in order to use colours as much as possible in feature detection, the face must have dark hair, dark eyes and somewhat lighter skin colour.

This paper is organized as follows: Section 2 presents related work. Section 3 shows the developed work, distributed by 6 items: system architecture; data acquisition; pre-processing; identification; interface; simulator.

2 RELATED WORK

Most of the work done in this area (facial processing) concerns facial recognition or facial expression recognition for facial modelling. These two main fields differ mainly in the fact that one aims to identify a face regardless of the expression, while facial expression recognition's target is to identify facial expressions regardless of identity. Human-Computer interaction and psychological tools are the areas where most of the facial expression recognition work was performed.

FACS was used as base of work in many projects (Cohen, 2003a). The first step is, in most cases, to separate background from areas of interest or to detect the face in the image. Two approaches to this problem are common. One is to recognize the face as a whole, using models or templates (Aas, 1996), (Bartlett, 2003), (Lyons, 1998) and the other is to detect relevant details (eyes, mouth, skin) and from that information obtaining the whole face or simply gathering information about those details. On other situations, some authors have chosen to build face models (Alekic, 2005). Once gathered the image information, there is a wide variety of methods that can be used. One of the most frequent is to build a database or eigenspace where images or characteristic data, typical of each expression to detect, can be stored followed by the computing of the expression to be identified, using comparison (Aas, 1996), (Frank, 2003). Other method is to process the information in a network (Neural, Bayesian, etc) whose outputs are used to determine the expression (Cohen, 2003b), (Franco, 2001). Another approach uses local or global classifiers (K nearest neighbour, kernel, Gabor), (Bartlett, 2003). Finally, there are also authors that base their work in Markov Models (Hidden Markov Models, Pseudo 3D Hidden Markov Models) (Alekic, 2005).

Some projects supported on interaction interfaces based on movements of the head (Matsumoto, 1999), (Wei, 2004) are being developed. But, those systems don't have the necessary robustness to be used in the real world, and need several improvements and adaptations to situations like: rapid colour change of the face; some involuntary movements of the head; different characteristics of the users (use of glasses, moustache). It urges the development of robust methodologies of facial expressions detection and his integration with systems to configure and control intelligent wheelchairs.

3 DEVELOPED WORK

3.1 System Architecture

The architecture that was chosen for implementing the interface is resumed in the following logical blocks:

- **Data acquisition:** the process of capturing the image with a digital camera and pre-processing at the camera's embedded software level.
- **Pre-Processing:** after acquiring the digital picture it is necessary to extract useful information from the picture.
- **Identification:** as soon as the feature information is gathered, an intelligent algorithm does the facial expression deduction.
- **Interface:** the command language is implemented on this level, based on facial expressions like "opened mouth" or "raised left eyebrow".

3.2 Data Acquisition

The system is ready to operate with any webcam connected to an USB port. The camera used in the experimental setup was the Logitech Sphere webcam, using a resolution of 320 x 240 pixels and

24 bits of colour. Logitech Sphere has pan/tilt motion capability and zoom. The driver provides automatic histogram and white-balance correction.

3.3 Pre-processing

There are basically two types of information in an image, which are relevant for describing the facial expression: low spatial frequency components and high frequency components. Low spatial frequency information is used to determine where important sets are located, mainly the face, while the high frequency components provide information about contours and shapes. The face detections and segmentation are achieved through colour classification and the contours are extracted using the Canny Algorithm (Haykin, 1999).

3.3.1 Noise Filter

The main types of noise that affect the image are: thermal noise, sensibility variations in each pixel and random noise. Usually these kinds of noise have a relatively high spatial frequency, so in order to attenuate it, a Gaussian filter was implemented in the space domain.

3.3.2 Face Tracking and Segmentation

In order to identify interest zones in an image it was decided proceed with colour segmentation (Gonzalez, 2002). This method allows, for instance, identifying all face pixels as being of the same category. In this way, it is possible to extract from the image all the relevant zones in a fast and efficient way as we can see on Fig. 1 representation.

Figure 1: Segmentation results.

To improve the efficiency of this technique, a "Lookup Table" that associates each group of colours (usually 8 bits leap for memory and time consuming efficiency), to a given category with a given trust level, was created.
Once the table is built, segmenting an image is done by analysing it and, for each pixel, matching the category with the colour and then replacing the colour value for the category code. To enhance the system's performance in a brilliance varying environment, an intensity correction method was developed.
Once the image is segmented, the next step is to track the face. This is done in a progressive way. First the center of the face is computed simply as being the average of the points of the category associated to the face.

$$Center(x,y) = \left(\sum_{i=1}^{n} \frac{x_i}{n}, \sum_{y=1}^{n} \frac{y_i}{n} \right) \quad (1)$$

At this point we have the center of the face, so what is done next is expanding in four directions (axis oriented) to detect face limits (upper, lower, right and left). Now, relevant zones of the face have to be defined, aiming to extract information to input the neural network. This is achieved dividing the height of the face in order to obtain the zones where eyebrows, eyes and mouth are expected to be found. The visual result of this approach is shown on figure below.

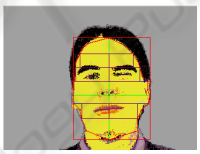

Figure 2: Relevant zones of the face.

On the next step the face, eyebrows, mouth and no category points are counted in each zone. The relationship between those values and the total points of each zone is a characteristic of a given facial expression. Once tracked the face and established the 6 interest zones, a simple count of the points of all categories (eyebrow, mouth, face and others) is performed in each zone. These zones are obtained by dividing the height of the face in equal parts in a successive way, and next, chose the upper and lower limits of the zones (Table 1).

Table 1: Face zones (values are fractions of the face height).

	Eyebrows	Eyes	Mouth
Upper	7/8	3/4	3/8
Lower	3/4	5/8	1/4

If no points of the expected category are found in a given zone, eyebrow (black points for example) points are not found in the eyebrow zone, for instance, this is also a characteristic input of an expression for the neural network. This happens if the head is turned up, down, left or right.

3.3.3 Edge Detection

Edge detection is made in order to capture the shapes inside the face (Fig. 3). This process consists in extracting the high frequency components and for this purpose the Canny Operator is used. The Canny Operator use criteria that increase precision and accuracy to speed processing. Using a Gaussian mask of 5 by 5 pixels of dimension, a Sobel operator of 3 by 3 and 8 connectedness for thresholding, gives about 135 operations per pixel. In this implementation it took about 200ms per image of 320 by 240 pixels to produce the final result. Since the real time constraint here is not very strict, a better relationship between the extracted information and the facial expression was preferred to a faster but no so accurate method.

Figure 3: Canny results.

In order to better describe the image, several patterns for each expression are taken during the configuration phase. During the execution cycle the contour image is compared with each saved pattern. The comparison consists in counting the number of coincident points with each saved pattern.

3.4 Identification

Here the measures taken in the pre-processing are translated into facial expressions. The data deployed to this block is relatively large, 23 inputs, and there is no guarantee of a linear relationship with the outputs. Establishing a relationship manually, even using fuzzy logic, would be very inefficient. A multilayer perception (MLP) with a continuous activation function can be used to compute the probability of an expression, for a given input pattern, after learning automatically with a training set. An Artificial Neural Network (Franco, 2001) (Haykin, 1999) takes some advantage over other methods with automatic learning facing this problem. The MLP type of network was chosen because it is possible to achieve good results. At this stage only a good solution was sought and optimizations to the identification process are left for future development. The network uses 22 inputs, 11 outputs (expressions) 1 hidden layer, 10 hidden neurons all with sigmoid activation function, trained with simple backpropagation and 110 training patterns are used. Concerning the number of hidden layers, one is enough for universal approximation, and it is also less prone for local minima to occur than using 2 hidden layers. There is no exact method for calculating the necessary number of hidden neurons, but the following heuristic helps: $h \geq \frac{p-1}{n+2}$, in which p is the number of training examples and n the number of input neurons. The Vapnic Chervonenkis dimension was also taken in consideration. The VP dimension is less than 4533.5 given by $2|W|\log_2(eN)$ – with W being the number of weights and N the number of neurons – and greater than 271 given by $n_1 n_2 + n_2(n_3 - 1)/2 + 1$, in which n_i is the number of neurons in the respective layer (Kasabov, 1998). The number of training examples used is 110 which are within range. Training was done so that the final training error would be less than 6E-5 (experimentally it was found to be a good value), and it takes about 5 minutes (approx. 100000 iterations) to achieve that result. The network must be trained for each user.

3.5 Interface

After identifying the facial expression, the corresponding command is checked in the command language definitions and sent to the (simulated) control system. Only in the instruction mode the facial expressions with commands associated are validated. Once out of the instruction mode, all commands other than the "instruction mode" command are ignored. In this way, the necessary level of concentration, while driving, can be reduced. Finally, it is possible to create a new training set for the Neural Network.

3.6 Simulator

The Ciber-Mouse simulator was used in order to evaluate the final results as shown below on Fig. 4.

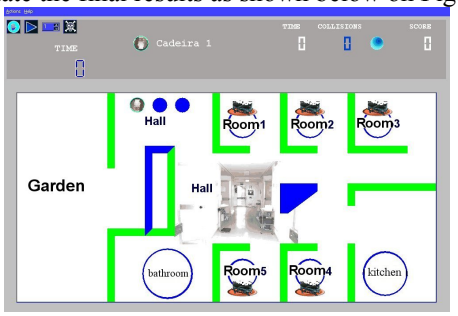

Figure 4: Ciber-Mouse Interface.

Ciber-Mouse is a competition among virtual robots, which takes place in a simulation environment running in a network of computers. The simulation system creates a virtual arena with a starting grid, a target area, signalled by a beacon, and populated of obstacles. It also creates the virtual bodies of the robots. Participants must provide the software agents which control the movement of the virtual robots, in order to accomplish some goals. All virtual robots have the same kind of body (Fig. 5). It is composed of a cylindrical shape, equipped with sensors, actuators, and command buttons. The simulator estimates sensor measures which are sent to the agents. Reversely, receives and apply actuating orders coming from agents. The Ciber-Mouse simulator was configured for wheelchair representation moving in a hospital environment. It was taken into consideration the dynamic behaviour of the wheelchair detecting collisions with objects as well.

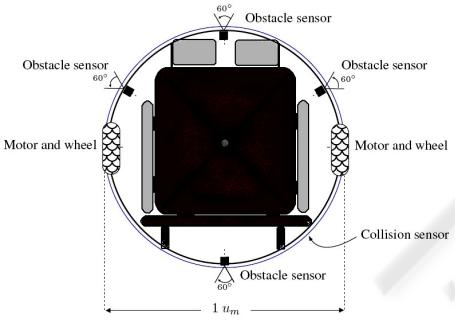

Figure 5: Body of virtual wheelchair.

At the end of the whole process, all the wheelchair commands were send for the control unit, which receives commands and processes it with sensor information. After control unit calculate the signal of control, he sends directly on the simulator who interprets them and makes the correct correspondence to the wheelchair model movements. The system architecture allows understanding the information cycle information.

4 RESULTS

The image processing and identification takes about 200 ms (using an Intel Centrino 1.8 GHz processor). The facial expression identification results of one good test were putted on confusion matrix which analysis is presented below. A satisfactory average accuracy was obtained in the results presented here.

Table 2: Confusion Matrix analysis: acc – accuracy, tpr – true positive rate, fpr – false positive rate, tnr – true negative rate, fnr – false negative rate, p – precision.

	ACC	TPR	FPR	TNR	FNR	P
Opened mouth	94,8	100,0	1,0	99,0	0,0	90,0
Frowned	54,7	100,0	6,5	93,4	0,0	30,0
Frowned & wrinkled nose	89,4	100,0	1,9	98,0	0,0	80,0
Leaned right	100,0	100,0	0,0	100,0	0,0	100,0
Leaned left	100,0	100,0	0,0	100,0	0,0	100,0
Normal	95,3	90,9	0,0	100,0	9,1	100,0
raised right eyebrow	54,8	100,0	6,5	93,6	0,0	30,0
raised left eyebrow	94,9	100,0	1,0	99,0	0,0	90,0
Raised eyebrows	95,5	90,9	0,0	100,0	9,1	100,0
turned right	100,0	100,0	0,0	100,0	0,0	100,0
turned left	89,4	100,0	2,0	98,0	0,0	80,0
AVERAGE	88,1	98,3	1,7	98,3	1,7	81,9

It is clear that the frowned expression had a weak accuracy and precision. This is not unexpected since two very similar expressions were chosen on purpose to determine how well the system can discriminate between slightly different expressions. “Frowned” and “Frowned and wrinkled nose” are quite similar and judging from the results the system can tell the difference between them but not in a very reliable way. As for the “Raised Right Eyebrow”, it’s below average results are probably due to some casual error during the extraction of training patterns, which can happen naturally. In order to have an estimate of the performance of each type of features, colour segmentation and edge comparison, the identification process using the same network architecture (apart from the number of input neurons) was done using them separately. The average results are shown in table 3.

Table 3: Identification results using only colour segmentation or contours.

	ACC	TPR	FPR	TNR	FNR	P
Col. Seg. Average	82,6	90,9	2,2	97,8	9,1	77,3
Contour Avg	79,6	85,7	2,2	97,8	5,2	76,4

Table 4: Training and testing results.

	segmentation only	contours only	Both
training error	4,65E-03	5,40E-03	5,80E-05
test error	1,50E-01	1,65E-01	8,57E-02

Using only colour segmentation, the identification results are quite good, and it is less time consuming (less than 150ms). The contour measures alone are not as good as the colour segmentation ones (table 4) though they are more robust to lighting variations. The combination of both types is better than only one set.

5 CONCLUSIONS

The main conclusion of this work is that it will be possible to drive a wheelchair, by using facial expressions, in a very comfortable way, using the system here presented. This conclusion is based on the very good results achieved by our facial expression detection process, together with the general empirical feel attained by the project members and volunteers when driving our simulated wheelchair. The main limitations to the good performance of our system are presently located in the pre-processing stage. The colour segmentation is much too sensitive to large light variations and slight colour shifts, and the shape extraction should have better precision without increasing the processing time. These problems affect the robustness of the system and its response time. All the other parts of the system have not posed any practical limitation to the system's performance. In order to overcome the problems mentioned, future developments will be first focused on the pre-processing aiming to allow the system to better perform in outdoor environments or rough lightning conditions.

Other future developments will include improvements on the wheelchair simulator and the next steps will be the introduction of commands for the wheelchair using the keyboard, and compare the behaviours with the ones that are obtained by the use of the platform using facial expressions. Then, it will be necessary to obtain objective indicators of the global performance, and this can be achieved by including on the simulator some simple skill tests such as obstacle avoiding. In the context of this project the system will be improved and implemented on a real wheelchair driven by quadriplegia and cerebral palsy handicapped people, using our intelligent wheelchair platform.

REFERENCES

Aas, Kjersti 1996. Audio-Visual Person Recognition: A Survey, Report no. 911, Norwegian Computing Center

Alekic, Peter et al. 2006. *Automatic Facial expression recognition using facial animation parameters and multi-stream HMMs*. IEEE Signal Processing Society

Bartlett, M. et al 2003. *Real Time Face detection and facial expression recognition: Development and Applications to Human Computer Interaction*. CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction

Cohen, Ira et al 2003a. *Facial expression recognition from video sequences*. Computer Vision and Image Understanding, Volume 91, Issues 1-2, Pages 160-187, Special Issue on Face Recognition

Cohen, Ira et al. 2003b. *Semi-Supervised Learning for Facial Expression Recognition*, 5th ACM SIGMM International Workshop on Multimedia Information Retrieval

Ekman & Friesen 1978. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press

Franco, Leonardo 2001. *A Neural Network Face Expression Recognition System using Unsupervised Local Processing*. Proceedings of the Second International Symposium on Image and Signal Processing and Analysis (ISPA 01), Croatia, pp. 628-632, June 19-21

Frank, Carmen & Nöth 2003, E., Automatic Pixel Selection for Optimizing Facial Expression Recognition using Eigenfaces, Pattern Recognition 25th DAGM Symposium

Gonzalez, Rafael C. & Woods, Richard E. 2002. *Digital Image Processing*. Prentice-Hall

Haykin, Simon 1999. *Neural Networks - A Comprehensive Foundation*. Prentice-Hall

Matsumoto, Y. and Zelinsky, A., 1999. *Real-time Face Tracking System for Human-Robot Interaction*. Proceedings of the 1999 IEEE International Conference on Systems, Man and Cybernetics (SMC99), Vol. 2, pp. 830-835.

Wei, Y., 2004. *Vision-based Human-robot Interaction and Navigation of Intelligent Service Robots*. PhD Dissertation, Institute of Automation, Chinese Academic of Sciences, Beijing, China.