Multimodal System Based on Electrooculography and Voice Recognition to Control a Robot Arm

José A. Martínez1, Andrés Úbeda1, Eduardo Iáñez1,*, Jose M. Azorín1 and Carlos Perez-Vidal1

1 Biomedical Neuroengineering Group (nBio), Miguel Hernández University of Elche, Spain
* Corresponding author E-mail: eianez@umh.es

Received 5 Mar 2013; Accepted 25 Apr 2013

© 2013 Martínez et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract People with severe motor disorders cannot use their arms and/or hands to interact with the environment. In such cases, a human-machine interface can be used to allow these people to perform activities of daily life. Voice recognition would be the best method to control a device devoted to help with these tasks, but this method is very dependent on the background noise and, under certain circumstances, it cannot be used successfully. In this sense, other methods must be included in the system to assist the voice recognition module. In this project, electrooculography (EOG) has been selected due to its stability and robustness. This way, a multimodal system based on EOG and voice recognition has been developed to control a robotic arm. This paper presents the procedures designed to combine both methods to create a multimodal interface useful for disabled people. Tests presented in this document compare the skills of five different users while controlling the robotic arm to perform pick-and-place tasks. Task duration and accuracy have been measured to obtain specific scores that are used to evaluate both interaction methods independently and the multimodal combination of them. The system works successfully using both methods (EOG and voice recognition). In addition, the multimodal interface improves robustness and reduces the uncertainty generated by the environment when there is background noise.

Keywords Multimodal System, Electrooculography, Voice Recognition, Robot Control

1. Introduction

The interest in developing human-machine interfaces (HMI) for disabled people has increased over recent years. People suffering from severe motor disorders cannot use their arms or hands to interact with devices. In these cases, human orders must be obtained by using new human-machine systems, such as brain-computer interfaces [1–3], electrooculography (EOG) interfaces [4, 5] or voice recognition [6].

On the other hand, between people, speech is the main method of communication [7, 8]. It is the fastest and most efficient way to transmit commands and ideas, but it is not very much used in human-machine interaction. Although several improvements have been considered to
increase the reliability of speech recognition in HMIs, in some circumstances, the system cannot recognize commands from the human operator (voice recognition systems are very dependent on the background noise).

When the background noise is relevant and the human operator cannot use voice recognition, an alternative interface must be considered, and for this reason, a multimodal system is highly recommendable. Multimodal interfaces combine the use of one or more techniques to communicate with an external device [9–11].

An EOG method has been selected to be used in combination with voice due to its stability and robustness. EOG allows detecting the movement of the eyes by measuring the difference of potential between the cornea and the retina. In [4], we developed an EOG-based interface that allows obtaining the direction of the eyes' movement (left, right, up or nothing) from the EOG signals. This interface was used to control a 6 degrees of freedom (DoF) robot arm. However, the interface could only control the movements of the robot end effector in the horizontal plane (i.e., only 2 DoF). To solve this drawback, in this paper the combination of an EOG interface with a voice recognition system is proposed.

This multimodal system is able to control the movements of the robot end effector in a 3D space. The EOG interface will be used to control the movements of the robot end effector in the horizontal plane. On the other hand, the voice recognition system will be used to move up or down the horizontal plane where the end effector is placed or to generate other commands to control the robot (e.g., to open/close the robot gripper). In the paper, the multimodal system has been used to perform pick-and-place tasks by controlling the robot. In addition, both interfaces have been tested independently in order to compare them with the multimodal combination. Five volunteers have successfully tested the multimodal system. The results of the tests are reported in this paper. In spite of the fact that the multimodal system has been used in this project to control a robot, it could be applied to interact with other devices, such as a computer or a domotic system.

The rest of the paper is organized as follows. Section 2 describes the EOG-based interface of the multimodal system, explaining the processing algorithm that allows obtaining the direction of the users' eye movements from the EOG signals. The voice recognition interface included in the multimodal system is described in Section 3. Section 4 shows how the EOG-based interface and the voice recognition system have been used individually and in combination in order to control the movements of the end effector of a robot with 6 DoF in a 3D space. The experimental results are shown and discussed in Section 5. Finally, the main conclusions are summarized in Section 6.

2. EOG-based Interface

2.1 Electrooculography

Electrooculography (EOG) is one of the most common methods for registering eye movements [12]. It is based on the fact that the eye acts as an electrical dipole between the positive potential of the cornea and the negative potential of the retina. Thus, in normal conditions, the retina has a bioelectrical negative potential related to the cornea. For this reason, rotations of the ocular globe cause changes in the direction of the vector corresponding to this electric dipole. The recording of these changes requires placing some small electrodes on the skin around the eyes. Fig. 1 shows the location of the electrodes. In order to detect horizontal movements, two electrodes are placed on the right and the left of the eyes, respectively (HR: Horizontal Right and HL: Horizontal Left). Vertical movements are detected by placing two electrodes on the top and bottom part of the eye, respectively (VU: Vertical Up and VL: Vertical Low). The reference electrode (REF) is placed approximately on the forehead.

Figure 1. Electrodes position to register EOG signals (right) and example of placement of electrodes over the user (left).

Figure 2. Two views of the EOG hardware (top and down). In each view: printed circuit board (PCB) connected to the electrodes placed around the eyes (big box), and PCB to register the EOG signals into the computer (small box).
EOG values vary from 0.05 to 3.5 mV with a frequency range of about DC-100 Hz between the cornea and the Bruch membrane located at the rear of the eye [13]. Its behaviour is practically linear for gaze angles of ±50° horizontal and ±30° vertical [14].

2.2 EOG Hardware
The hardware described in [4] has been used to register the EOG signals of the users. It is a portable, cheap and small-sized device based on EOG with USB compatibility. It is divided into two parts: one is connected to the electrodes placed around the eyes of the user (big box, Fig. 2), and the other one is connected to the computer in order to obtain the EOG signals (small box, Fig. 2). Both parts are connected wirelessly.

The hardware uses dry electrodes in order to register the EOG signals. The electrodes used are the model E273 from Easycap, which are flat Ag/AgCl electrodes of 12-mm diameter with a light-duty cable and 1.5-mm touchproof safety sockets. The advantage of these types of electrodes is that they do not require conductive gel to operate, so there is only need for some cleaning abrasive gel on the skin before contact, which makes the placing of the electrodes on the user easier and faster.

2.3 Processing Algorithm
A processing algorithm has been developed in order to obtain the direction of the eyes’ movement (left, right, up, down or nothing) and the blink from the EOG signals. The users must perform a fast movement of their eyes in the desired direction and then return to the centre position. The hardware registers the EOG signals using a frequency of 30 Hz. The processing algorithm is applied independently for each channel (horizontal and vertical). A window of 1 s (30 samples) has been used to apply the processing algorithm, obtaining an ocular command every second. The EOG samples are saved in a vector of 30 samples (one vector for each channel). The vector follows several processing steps after being filled out: (see Fig. 3):

1. A moving average filter is applied to get a clearer signal.
2. Next, the derivative is performed to detect the abrupt changes of the signal when the user moves the eyes or performs a blink.
3. Then, it is verified if the signal exceeds a specific threshold in order to differentiate the signal from the rest state. If not, the zero value is assigned to the signal.
4. Afterwards, the vector of the EOG signal samples is analysed:
   - If there is a value that exceeds the blinking threshold, the algorithm output is a blink (value 2).
   - If there is a positive/negative sequence of values (i.e., an up or right eye movement is performed), the algorithm output is right or up (value 1), depending on the channel.
   - If there is a negative/positive sequence of values (i.e., a down or left eye movement is performed), the algorithm output is down or left (value -1), depending on the channel.
   - Otherwise, the algorithm output will be rest (value 0).

Thus, the algorithm output will be right, left, up, down, rest or blink. The thresholds of the algorithm will be adjusted for each user during the training phase.

3. Voice Recognition Interface
Artificial speech recognition systems can be classified according to whether they are speaker dependent or speaker independent [15, 16]. Speaker dependent speech recognition engines require the users to train a profile of their voices for the engine to use when performing recognition. Conversely, speaker independent systems do not require the user to train them before achieving high recognition accuracies. In this case, a pre-recorded corpus of words is compared to the input speech vectors to generate recognition results [17]. In general terms, speaker dependent systems obtain greater recognition accuracy than speaker independent systems, but the latter is ideally suited for applications with a large number of different users, as presented in this paper.

Some proprietary softwares available in the market to do speech recognition are (among others): AT&T WATSON (by AT&T Bell Laboratories), HTK (copyrighted by Microsoft), Voice Finger (by Robson Cozendeny), Dragon NaturallySpeaking (by Nuance Communications), e-Speaking (by e-Speaking.com) and ViaVoice (by IBM) [18,
In 1999, a piece of ViaVoice software was released and became free software. Nowadays it is one of the most extended libraries for voice/speech recognition. It has multilingual support including English, German, French, Italian, Japanese and Spanish (amongst others) and has a complete and useful manual for developers. Based on experiences presented in [20], ViaVoice has been selected for this project.

The vocabulary can be specified in one of two ways: either as a structured grammar or as a dynamic command vocabulary. In this work, structured grammar has been selected as the control of the robot is limited to a small number of commands. The grammar provides a language model for the speech engine, constraining the valid set of words to be considered, increasing recognition accuracy while minimizing computational requirements. A plain text editor can be used to create the grammar file. The grammar is specified using a specialized speech recognition control language (SRCL).

As Andrew P. Coates explains on the site ai-depot.com (site owned and maintained by Alex J. Champandard), the speech recognition control language (SRCL) is a grammar formalism produced by the Speech Recognition API Committee (www.srapi.com) and the Enterprise Computer Telephony Forum [19]. It is a particular type of the Backus-Naur form (BNF) of grammars, which is well understood among the formal language and linguistics communities. Linguist computer software usually uses grammar based on the constructs of a language (e.g., noun phrase and intransitive verb), but it is demonstrated that grammars with template mechanisms allow greater rates of recognition with a lower computational cost. Otherwise, the user needs some training to fit the speech to the templates and this is only possible in several applications (other applications cannot perform training). The grammar shown below (extracted from the site ai-depot.com) is an example of the SRCL format using such a template mechanism, and this will allow speech such as “follow me”, “may I see your passport” and “good evening may I see your invitation”.

```xml
<root> = <greeting>? <request>
  <greeting> = hello | good <time of day>
  <time of day> = moring | afternoon | evening
  <request> = may I see your <docue> | follow me
  <docue> = passport | &docType | invitation | &docType
</root>
```

In the SRCL format, each rule is defined using an equal sign `=` to separate the left and the right side. Nonterminals are defined using `~' and there is no delimiter for terminals. In this project, the (simple) set of rules used allows the recognition of a sentence like “hello robot, can you move down and increase speed, please”.

```xml
<root> = <greeting>? <request> <greeting>?
  <greeting> = hello robot | robot good <time of day>
  <time of day> = moring | afternoon | evening
  <request> = can you [move <movement>] |
    | <change mode> | H]
  <movement> = UP|DOWN|FORWARD|L
  <change mode> = STOP|INCREASE SPEED|
    | DECREASE SPEED|L
  <greeting> = please
</root>
```

On the other hand, the definition of rules in this case is not essential. The system can work properly using only the word recognition option because in this application the user does not use a “complex” natural language (due to the limited defined vocabulary). In this project, 16 commands (shown in table 1) are included in the vocabulary.

<table>
<thead>
<tr>
<th>Spoken command</th>
<th>User No. 1</th>
<th>User No. 2</th>
<th>User No. 3</th>
<th>User No. 4</th>
<th>User No. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No det.</td>
<td>Success</td>
<td>No det.</td>
<td>Success</td>
<td>No det.</td>
</tr>
<tr>
<td>UP</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>DOWN</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>FORWARD</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>BACK</td>
<td>0</td>
<td>100%</td>
<td>2</td>
<td>80%</td>
<td>0</td>
</tr>
<tr>
<td>RIGHT</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>LEFT</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>OPEN</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>CLOSE</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>INCREASE SPEED</td>
<td>3</td>
<td>70%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>DECREASE SPEED</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>STOP</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>HOME</td>
<td>0</td>
<td>100%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>LOWER FRAME</td>
<td>1</td>
<td>90%</td>
<td>0</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>RIGHT FRAME</td>
<td>3</td>
<td>70%</td>
<td>0</td>
<td>100%</td>
<td>1</td>
</tr>
<tr>
<td>FRONTAL FRAME</td>
<td>1</td>
<td>90%</td>
<td>0</td>
<td>100%</td>
<td>1</td>
</tr>
<tr>
<td>LEFT FRAME</td>
<td>1</td>
<td>90%</td>
<td>0</td>
<td>100%</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Success rate identification and number of no detections (No det.) of the spoken commands.
4. Control of a Robot Arm

The EOG-based interface and the voice recognition interface have been combined in order to control a Fanuc LR Mate 200iB robot arm, Fig. 4. It has six rotational joints, the first three (J1, J2 and J3) are used to determine the robot position, while the last three (J4, J5 and J6) are used to determine the orientation of the end effector. The robot has a reach of 700 mm and can support a maximum load of 5 kg. A gripper has been attached to the end effector in order to perform pick-and-place tasks. The Fanuc R-J3iB Mate controller commands the robotic arm. It has an Ethernet connection, which allows the communication between the human-interface controller and the robot controller.

The FrRobot library, developed by Fanuc, has been used to control the robot. This library, originally implemented to be used in Visual Basic, has been modified and encapsulated in an Active-X DLL. Then, this DLL has been used in our software developed in C++ programming language. Modifications performed in the FrRobot library were presented in [21].

The workspace has been designed to avoid robot singularities. The limits of the workspace (in millimetres) are: X = [360 650], Y = [-180 180] and Z = [80 250], see Figure 5. Within these limits, two boxes have been placed, therefore, when the robot approaches the boxes, if the height of the end effector is not adequate, it stops to avoid an impact. The size of the small and big box is 160x120x60mm and 160x120x120mm, respectively.

The aim is to control the robot arm using the EOG-based and voice recognition interfaces in order to perform pick-and-place tasks. The robot has been controlled by sending the Cartesian coordinates of the end effector from the interfaces.

The EOG interface controls the movements of the end effector in the horizontal plane. The eye movements to the right, left, up and down allow moving the robot end effector to the right, left, backwards and forwards, respectively. When the EOG-based interface detects a movement of the user’s eyes, the robot end effector starts its movement in the detected direction to the established limits of the workspace. To stop the robot, the user must perform a movement of their eyes in the opposite direction.

The voice recognition interface allows controlling the robot using the following commands:

- **UP**: the robot end effector starts moving up (or if it is moving down, it stops). The limit of the up movement is Z=250mm.
- **DOWN**: the robot end effector starts moving down (or if it is moving up, it stops). The limit of the down movement depends on where the robot end effector is situated (see Fig. 5). If there is no box below the robot end effector, the limit is Z=80mm. If there is a big box below the robot end effector, the limit is Z=200mm. If there is a small box below the robot, the limit is Z=140mm.
- **FORWARD**: the robot end effector starts moving forward (or if it is performing a backward movement, it stops). The limit of this movement is X=650mm.
- **BACK**: the robot end effector starts moving backwards (or if it is moving forward, it stops). The limit of this movement is X=360mm.
- **RIGHT**: the robot end effector starts moving to the right (or if it is moving left, it stops). The limit of this movement is Y=180mm.
- **LEFT**: the robot end effector starts moving to the left (or if it is moving right, it stops). The limit of this movement is Y=−180mm.
- **OPEN**: opens the robot end effector gripper.
- **CLOSE**: closes the robot end effector gripper.
- **INCREASE SPEED**: it increases the movement speed of the robot. The increment is 2% of the maximum speed to obtain accurate control.
- **DECREASE SPEED**: it decreases the movement speed of the robot. The decrement is 2% of the maximum speed to obtain accurate control.
- **STOP**: cancels all robot movements.
- **HOME**: the robot returns to a (previously established) start position.
- **LOWER FRAME**: the robot pose is changed to work with the gripper pointing down.
- **RIGHT FRAME**: the robot pose is changed to work with the gripper oriented to the right.
• FRONTAL FRAME: the robot pose is changed to work with the gripper oriented to the front.
• LEFT FRAME: the robot pose is changed to work with the gripper oriented to the left.

When the EOG and the voice recognition interfaces are combined, EOG controls the movement in the horizontal plane and voice recognition controls the remaining commands to perform the pick-and-place operations. The voice commands needed are: UP/DOWN, INCREASE/DECREASE SPEED and STOP the robot end effector. In Section 5, the operation modes of each individual interface and the multimodal combination will be detailed.

5. Experimental Results

The multimodal system has been used to control the robot arm in order to perform pick-and-place tasks. Several experiments have been designed to evaluate the performance of the multimodal system. In these experiments, the user must control the robot in order to pick two objects and place each of them over the box with the same colour, see Fig. 5. First, the users have been trained with the EOG interface and the voice recognition system. Afterwards, both interfaces have been tested alone in the designed environment. Finally, the multimodal interface is tested. Five healthy users, all men, took part in the tests. One of them (user 2) had previous experience with EOG interfaces. Each of them performed three repetitions of the test for each interface (EOG, voice recognition and multimodal).

![Image 5. Experimental environment with the boxes (big and small) and the objects (smaller boxes). A detail of the workspace is shown over the real environment.](image)

In order to evaluate the performance of the multimodal system, two parameters have been considered in the tests: duration of the test, and accuracy to place the objects over the box with the same colour. To evaluate the accuracy, each box has been divided into three zones and a score has been given depending on the zone where the object is placed, see Fig 6. The scores are as follows: 10 points (zone 1), 5 points (zone 2), 1 point (zone 3), and 0 points (outside previous zones).

5.1 Evaluation of the EOG system

First, training must be performed in order to adjust the thresholds of the EOG algorithm. Each user must perform specific movements of their eyes to the right, left, up and down, and they must perform a blink. From the analysis of the registered EOG signals, the thresholds are selected. It is not necessary to recalibrate the system in later sessions. Each training session lasts less than 10 minutes. In previous works, the EOG processing algorithm has been verified obtaining a success rate of eye movement detection of almost 100% if the thresholds are selected properly [4].

The EOG interface only allows moving the robot end effector in the horizontal plane by using the ocular movements: up, down, right and left. The robot arm moves at a fixed height over the objects. When the user is near an object, by blinking, the robot automatically picks it up and returns to the initial height. To place the object, the user must blink again when they stop the robot over the desired box.

![Figure 6. Schematic representation of the score system over one of the boxes. Scores are assigned depending on the distance to the centre of the box.](image)

Table 2 shows the duration and the scores obtained on each test for each user. The scores are related to the accuracy in placing the objects over the box with the same colour. It can be verified that all users were capable of performing the pick-and-place tasks using the EOG interface. However, the time taken to perform each test is very variable for each user, even after performing several repetitions. The scores obtained generally indicate that the users are able to put both objects in the corresponding box, but without obtaining a very high accuracy.
Moreover, time decreases with the performance of more tests. This way, the duration of the third test is less than the first and second tests for each user. Regarding the scores, there is no clear improvement because of the resolution of the tests. However, the scores are quite high for all the tests.

5.3 Evaluation of the Multimodal System

For the multimodal interface tests, the EOG and the voice recognition interfaces have been combined. EOG controls the movement in the horizontal plane of the robot end effector and the open/close action of the gripper. The remaining movements are controlled by the voice recognition interface: go up/down, increase/decrease speed and stop the robot.

Table 4 shows the duration and the scores obtained on each test for each user. The scores are related to the accuracy to place the objects over the box with the same colour. It can be verified that all the users were capable of performing the pick-and-place tasks using the multimodal system. Furthermore, in spite of the fact that the duration of the first test is quite different for each user, the duration of the tests is reduced with the training. This way, the duration of the third test is less than the first and second test for each user. In addition, the duration of the third test is similar for each user. It can be verified that the scores increase as users have more training. Thus, the scores of the last test are better than the scores of the first one.

5.4 Discussion

If the average of all tests (1, 2 and 3) performed by the five users is done, the voice recognition interface obtains an average time of 2m:02s and an average score of 12 points.

If only the last test (3) is taken into account, the results improve (time: 1m:43s; score: 13.2 points), suggesting that training is useful to obtain better results. The same procedure has been performed for the multimodal interface. Taking into account all tests performed by the five users, the average time is 4m:14s and the score is 9 points. In the same way, if only the last test is averaged, the results also improve (time: 2m:42s; score: 10.8 points). Finally, for the EOG interface, it can be seen that the improvement is irrelevant. For all the tests performed by the five users the average time is 8m:14s and 3 points, and for the final test the average time is 7m:56s and the average score is 3 points. Contrary to expectations, the multimodal system does not obtain the best time and score results, although it improves the results obtained with the EOG interface.

The voice recognition interface obtains the best results. Another conclusion that emerges from this study is that...
although the voice recognition obtains the best results, it cannot always be used in all the situations because it is very dependent on external noise and recognition accuracy decreases when the number of commands is very high. In these cases, the multimodal interface can solve these problems by decreasing the number of commands needed for controlling the system. The EOG interface obtained the worst results in this comparison. However, in cases where the users have a severe motor disability and they are not able to speak, it is the only way to perform the experiments.

6. Conclusion

People with motor disorders cannot use their arms to interact with the environment around them. For these cases, other kinds of interfaces can replace traditional control interfaces, like a mouse and a keyboard. In this paper, a pick-and-place application has been designed using a robot arm. This application has been controlled using an EOG interface, a voice recognition interface and a combination of both (multimodal interface).

The conclusion that emerges from this study is that the voice recognition interface is the best solution in terms of accuracy and time. However, voice recognition is very dependent on the background noise and the number of voice commands. For this reason, the EOG interface has been combined with the voice recognition interface to increase the robustness, creating a multimodal interface to control the robot arm. All the users were capable of finishing the tests and obtaining good score and time results using the multimodal interface.

The EOG interface was also tested individually. Although the results obtained are worse, in cases where users have a severe motor disability and are not able to speak, it is the only way to perform the experiments.

As future works, more realistic applications will be tested using this multimodal interface. For example, performing activities of daily life, such as grasping tasks with objects like a bottle or a glass. Moreover, more tests will be performed with a larger number of users to perform a statistical analysis of the performance of the system after training. Ergonomics issues will also be studied to improve the comfortability of the system and to obtain real feedback from users.

7. Acknowledgements

This research has been supported by grant DPI2011-27022-C02-01 of Ministerio de Economía y Competitividad of Spain and by Conselleria d’Educació, Cultura i Esport of Generalitat Valenciana of Spain through grants VALivd ACIF/2012/135 and ACOMP/2013/018. The authors thank Ernesto Carpintero for help with robot programming and tests.

8. References


