Gesture based Interaction with the Humanoid Robot ROMAN

Flávio Garcia Pereira[1], Norbert Schmitz[2], Raquel Frizera Vassallo[1], and Karsten Berns[2]

[1] Federal University of Espírito Santo, Vitória, Brazil,
flaviog@ele.ufes.br, raquel@ele.ufes.br,
WWW home page: www.ele.ufes.br
[2] Technical University of Kaiserslautern, Germany,
nschmitz@informatik.uni-kl.de, berns@informatik.uni-kl.de,
WWW home page: http://agrosy.informatik.uni-kl.de/

Abstract. The development of humanoid robots for natural interaction is a challenging research topic. The inter-human communication is very complex and offers a variety of interaction possibilities. Although speech is often seen as the primary channel of information, psychologists claim that 60% of the information is transferred non-verbally. Besides body pose, mimics and others gestures like pointing or hand waving are commonly used. In this paper the gesture detection and control system of the humanoid robot ROMAN is presented using a predefined dialog situation. The whole information flow from gesture detection till the reaction of the robot is presented in detail.

1 Introduction

Keyboards and mice are still the most common and most used interfaces between humans and computer systems, no matter if it is just a desktop PC, a notebook or a robot. However there is an increasing interest in developing additional interfaces such as speech recognition, handwriting, gesture recognition and emotion detection, since these ways of interaction may bring more naturalness into the human-computer interface. In addition most people tend to feel more comfortable if they can interact with computers and robots in the same manner humans are communicating with each other. The interesting thing is that psychologists claim that over 60% of interaction signals are transferred non-verbal [5].

This is one of the motivations for the work presented in this paper. Our main goal is to establish a communication interface through gesture recognition between a human and the robot ROMAN, that stands for Robot Human Interaction Machine. The robot must recognize some gestures made by a human and respond by speaking or making some body movements.

Traditionally there are some techniques for gesture recognition based on the shape of the hand or based on the movement of the human arm and hand. Also there is no tight definition for a gesture. Gestures can be viewed as a non-verbal interaction and may range from simple static signals defined by hand shapes,

actions like pointing to some objects, waving a hand or more complex movements to express ideas or feelings allowing the communication among people [8].

Thus for recognizing gestures it is necessary to find a way by which computers can detect dynamic or static configurations of the human hand, arm, and even other parts of the human body. Some methods used mechanical devices to estimate hand positions and arm joint angles such as the glove-based approaches [12]. The main drawback of this interface is that, besides being expensive, the user must wear an uncomfortable glove, having a lot of cables to connect the device to the system that restricts the workspace to a small area and limits the movements of the user. Therefore, one of the best options to overcome the disadvantages of the glove-based methods and to implement less restricted systems is the usage of computer vision to detect and track hands and arms.

Basically, the computer vision approaches concentrate on recognizing static hand shapes (pose gestures) or interpreting dynamic gestures and motion of the hands (temporal gestures). For the static gesture approaches the focus is to identify a gesture by the appearance of the hand, silhouettes, contours, 2D or 3D models [4, 11, 16], while the methods that consider dynamic gestures are concerned about motion analysis [9, 17]. There are also works like [15] that evaluates both pose gestures and temporal gestures.

Our approach, like many others, aims to support the interaction to a robot and control some movements. But besides that we intend to improve the interaction and communication with a robot, allowing the usage of gestures and dialogs. As an example of some related work, in [4] hand shapes are used to control a walking robot through gestures that indicate commands as stop, go forward, etc. Their system has four modules: a hand detection using skin segmentation, hand tracking, a hand-shape recognizer based on contours and the robot controller. In [19] an integration of gaze and gestures is used to instruct a robot in an assembling task. Basically, pointing hands are detected through skin color and splines for contour descriptions. In [13] an attention model for humanoid robots is defined using gestures and verbal cues. Human motion and gestures are captured using markers attached to the subject body.

In this work we focus on an appearance based method for recognizing gestures that will help improve the interaction with a robot allowing the usage of gestures and dialogs. Our approach also includes a hand tracking module, thus the user can make different sequences of gestures while the robot keeps looking at the user's hand, without having to process the entire image every time in order to detect a hand and then identify a gesture. The preliminary results are encouraging and for future work we expect to take the motion history into account for recognizing more complex and temporal gestures.

This paper is organized as follows. Section 2 presents the skin detection method applied for detecting the hands of the user. Section 3 describes the gesture recognition approach and the following section explains the hand tracking module. The integration of these gesture detection and tracking into the control system of the robot is described in Section 5. Our preliminary experiments are

described in Section 6 while Section 7 finally presents the conclusions and future work.

2 Skin Detection

Skin color is an important feature for face detection, gesture recognition, people identification and many more. The skin color detection is usually a pre-processing step to segment skin regions in the image. After this segmentation only the skin regions are processed further.

There are a lot of ways to implement skin color detection. Initially it is necessary to choose the color space and the method that must be used. Since the RGB color space is the most commonly used space for representing digital images, it is widely used for skin detection. The normalized RGB color space can be used in order to reduce the illumination disturbance [20]. In other applications the color spaces HSV, HSI, HSL or TSL are used instead of the RGB color representation [10, 18, 1, 3, 14]. The idea is to get invariant to high intensity at ambient lights and surface orientations relative to light sources, separating the color hue information from the intensity and saturation information. Also orthogonal color spaces such as YCbCr, YUV, YIQ and YES are commonly used because they can represent the color through luminance and chrominance statistically independent components which are good options for skin detection.

In this paper the RGB color space model and a mixture of Gaussians to determine the probability of skin for each pixel has been used. This probability is given by

$$P(x) = \sum_{i=1}^{N} \omega_i \frac{1}{\sqrt{(2\pi)^3 |\Phi_i|}} e^{-\frac{1}{2}(x-\mu_i)^T \Phi_i^{-1} (x-\mu_i)}, \quad (1)$$

where N is the number of Gaussians used, x is a color vector containing the RGB values for each pixel from the image, μ_i and Φ_i are, respectively, the mean value and the diagonal covariance matrix. The value ω_i is the contribution of each gaussian. The Gaussian parameters (μ_i, Φ_i and ω_i) are based on large set of training data as shown in [6]. An example of an RGB image and the respective skin detection by using this method can be seen in the Figure 1.

3 Gesture Recognition

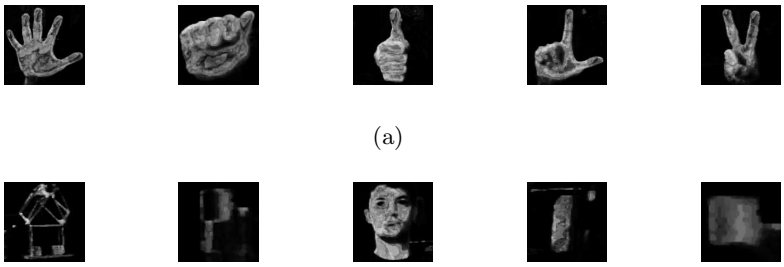
After the skin segmentation, some blobs containing skin color are generated. Only blobs whose area is greater than 200 pixels are processed by the gesture detector algorithm. The gesture detector algorithm uses the Principal Component Analysis (PCA) technique to generate an eigenvector base and classify the blobs that are found.

To build the eigenvector base five gestures from different people and some regions containing skin color that does not represent any gestures are selected.

(a) (b)

Fig. 1. (a) Input image captured from the camera in RGB color space. (b) Grey scale image after skin detection. A high brightness indicates a high probability.

The five gestures (open hand, closed hand, positive, L and V) used to train the PCA are shown in Figure 2 (a) and the false examples are depicted in Figure 2 (b).



(a)

(b)

Fig. 2. (a) The five gestures used to build the training set: open hand, closed hand, positive, L and V. (b) Skin regions that do not represent gestures.

The image set database has 90 images. Fifteen images for each gesture and 15 for unknown gestures. The base generated by the PCA technique has 22 eigenvectors. This number of eigenvectors was chosen based on the eigenvectors energy. The energy of a set of n eigenvectors is calculated based on the eigenvalues associated to each eigenvector

$$E(n) = \frac{\sum_{j=1}^{n} eigenvalue(j)}{\sum_{k=1}^{T} eigenvalue(k)}, \tag{2}$$

where n is the number of eigenvectors that are used to compute the energy and T is the total number of eigenvalues. A graphic containing the information

about that process appears on the Figure 3. In this graphic it can be seen that the energy of the eigenvectors grows very fast, because the energy is determined using the eigenvalues, and the first eigenvalues are big in comparison to the following ones.

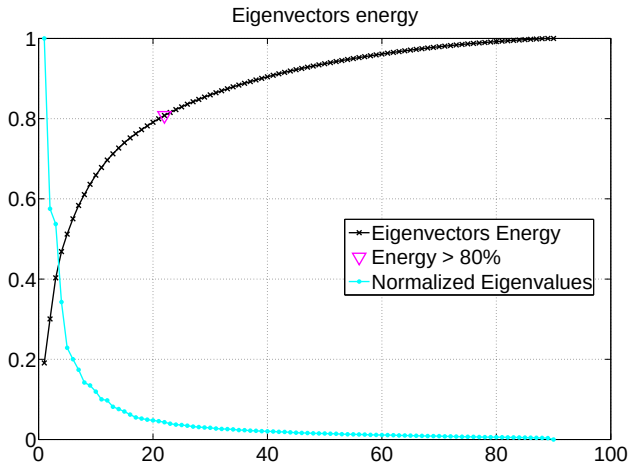


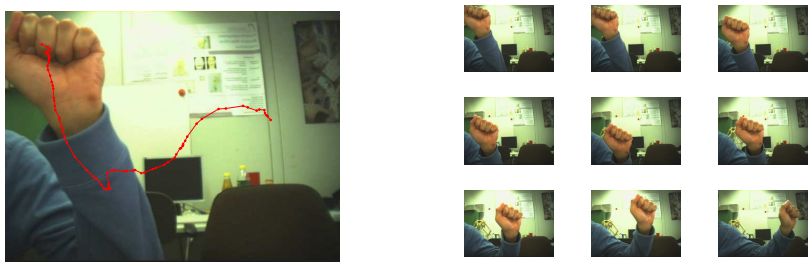
Fig. 3. Selection of number of vectors that will compose the base.

The implemented gesture recognition algorithm is able to correctly classify the gestures done in front of the camera. Thus, with the gesture recognition information, it is possible for a human to interact with the robot using an specific set of gestures. For example, the human can stop the robot with a simple gesture.

4 Hand Tracking

The interaction based on gestures requires the static and a dynamic detection of gestures. The tracking of a shown gesture can be used to present more complex commands like look at a point or turn in a direction. The tracking of gestures is realized using the camshift algorithm [2] which is a color based tracker using the histogram information concerning the region of interest. The initial tracking area is the position and size of a detected gesture.

Figure 4 shows the color based tracker in a cluttered test environment. The first image shows the initial position of a gesture and the path of the hand captured by the tracking algorithm. The following small images show some hand positions during the tracking process.



(a) (b)

Fig. 4. (a) Input image captured from the camera. The motion of the hand is marked with a red line. (b) Images of the moving hand during the tracking process.

5 Integration

The experiments of the presented approaches have been performed on the humanoid robot ROMAN [7]. The robot consists of a humanoid upperbody and head with 24 degrees of freedom: 7 for the movement of the body and neck, 6 for eye movement and 11 for emotional expressions.

The behavior-based control architecture of ROMAN consists of three main groups: the habits, the motives and the percepts. The habits are realizing all motions of the robot, the motives are providing the overall goals like communication and the percepts module is observing the environment to extract relevant information for the robot. The gesture detection and the tracking are both located within the perception module which is divided into 4 subgroups: capturing, fusion, memory and extraction.

The capturing group extracts information from a single image like skin color or face candidates. The gesture detector module is a member of the capturing group. In the fusion group the captured information is fused using the current and previous observations in time. The output of the fusion group is a more general description of the objects in the scene. This group contains the camshift tracker besides other fusion modules. The memory module stores the information of the fusion group in a short term memory. This short term memory is used by the extraction group which analyzes the output of the fusion over time to extract information like hand waving or nodding.

The complete architecture of the robot is implemented using the MCA-KL framework and the iB2C behavior definition [3]. Figure 5 shows the camera group as a part of the perception module in the mcabrowser (the browser is a tool of the MCA-KL framework). The FrameGrabber at the bottom of the image grabs the images from the cameras. Each capturing module as described above can have multiple input and output images. The SkinDetector for example fetches

[3] see http://rrlib.cs.uni-kl.de/ for details

the images from the camera and creates the skin color image which is at the same time the input image of the GestureDetector. Inside the gesture detector the gestures are stored in the element list handler. The handler realizes the communication between the capturing and fusion group. The information of the gestures together with the original input images are used in the CamShift module. The output of the fusion modules is then stored in the object list handler which realizes the communication between fusion and extraction group. Before the data can be passed to the robot the InformationExtraction module generates a high level view of the environment including objects and their properties like an object 'left hand' with the property 'shows open hand gesture'.

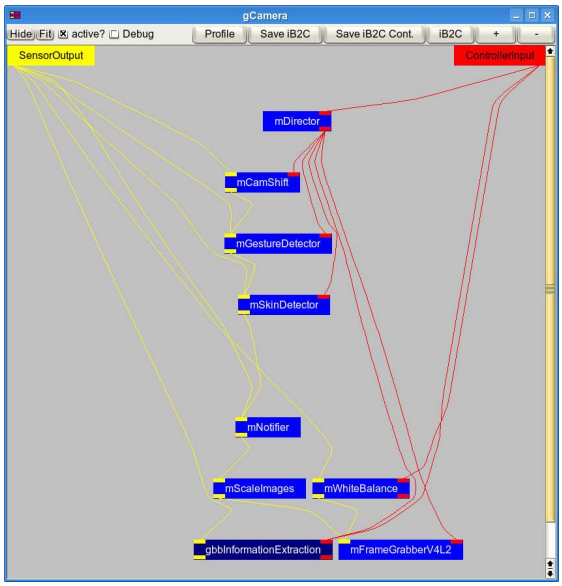


Fig. 5. Screenshot of the Camera group with GestureDetector and CamShift module. The image shows the tool mcabrowser which can be used for run-time analysis of the control architecture.

6 Experiments

The experiments are realized using a small dialog with non-verbal input signals. When the dialog is started the robot welcomes the user. During the dialog the user can present one of the five gestures: open hand, closed hand, L, V and positive. The open hand gestures immediately stops the robot whatever he is doing whereas the V gesture allows the robot to move again. Whenever the L gesture is detected the robot starts to play a song and stops playing when the closed hand gesture is detected. The positive gesture initializes the tracking and

(a) (b) (c)

Fig. 6. (a) Input Image of the FrameGrabber module (b) Output image of the gesture detector module (c) Image of the start of the CamShift tracking module

lets the robot follow the hand by constantly looking at the tracked object until the open hand gesture is shown.

The gesture detection tests are realized with four different people. Each person presents the five gestures that the algorithm is able to recognize in front of the camera several times while the gesture detection algorithm was running. The statistical data obtained during the test run can be seen in Table 1 while the Figure 7 shows the robot in the experimental setup.

Table 1. Results of Gesture Detection Algorithm.

	Open Hand	Closed Hand	Positive	V	L	Unknown Gesture
Open Hand	0.917	0.00	0.00	0.00	0.00	0.083
Closed Hand	0.00	0.850	0.00	0.00	0.00	0.150
Positive	0.00	0.00	0.883	0.017	0.00	0.100
V	0.00	0.00	0.050	0.900	0.00	0.050
L	0.00	0.050	0.00	0.017	0.867	0.066

Figure 8 shows an interaction between the robot and a human during the experiments.

7 Conclusions and Future Work

Non-verbal interaction capabilities of humanoid robots are getting more and more important. Besides speech-based interaction non-verbal signals can be used to improve the human-robot communication. Commonly used non-verbal interaction signal are besides others gestures like pointing or hand waving. In this paper the gesture detection and the control system of the humanoid robot ROMAN has been presented.

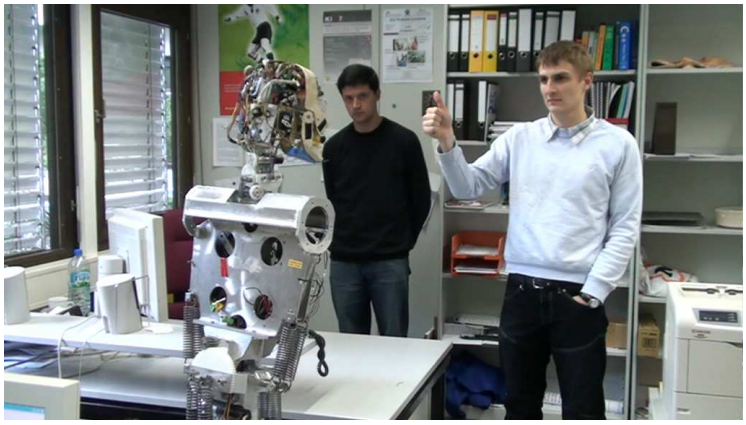


Fig. 7. Image of the robot and two simultaneous interaction partners in a typical dialog situation with a person showing the positive gesture.

Fig. 8. Interaction between the robot and a human showing the L-gesture during an experiment.

Starting from the skin color image generated by a trained color segmentation module the gesture detection has been realized using the principle component analysis. The position and size of the detected gestures is used to initialize the camshift tracking algorithm. The information of gestures and the tracking are captured, fused and transferred to the robot where the information is used to trigger a small predefined dialog.

Future work concerning the gesture based interaction will focus on the integration of verbal and non-verbal interaction signals into multimodal dialog situations.

Acknowledgment

The authors would like to thank CAPES(Brazil) and DAAD(Germany), through the bi-national cooperation project PROBRAL 282/07, for their financial support. This cooperation project allowed Norbert Schmitz to spend one month in Brazil, while Flávio Garcia Pereira is currently realizing part of his PhD project at the Technical University of Kaiserslautern, Germany.

References

1. Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. In *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 1998.
2. G.R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (Q2):15, 1998.
3. David Brown, Ian Craw, and Julian Lewthwaite. A som based approach to skin detection with application in real time systems. In *in Proc. of the British Machine Vision Conference*, 2001.
4. J. S. Chang, E. Y. Kim, and H.J. Kim. Mobile robot control using hand-shape recognition. In *Transactions of the Institute of Measurement and Control*, volume 30, pages 143–152, 2008.
5. Edward Hal. *The Silent Language*. B and T, 1990.
6. M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46((1)):81–96, 2002.
7. K. Mianowski, N. Schmitz, and K. Berns. Mechatronics of the humanoid robot roman. In *Sixth International Workshop on Robot Motion and Control (RoMoCo)*, Bukowy Dworek, Poland, June 11-13 2007.
8. V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, pages 677–695, 1997.
9. F. Quek. Gesture, speech and gaze cues for discourse segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 247–254, 2000.
10. Leonid Sigal, Stan Sclaroff, and Vassilis Athitsos. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *CVPR*, pages 2152–2159, 2000.
11. B. Stenger, P. Mendonca, and R. Cipolla. Model-based 3d tracking of an articulated hand. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 310–315, 2001.
12. D. J. Sturman and D. Zeltzer. A survey of glove-based input. In *IEEE Computer Graphics and Applications*, volume 14, pages 30–39, 1994.
13. O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. Three-layered draw-attention model for humanoid robots with gestures and verbal cues. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2423–2428, 2005.
14. Jean-Christophe Terrillon, Hideo Fukamachi, Shigeru Akamatsu, and Mahdad N. Shirazi. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *FG 00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, pages 54–61, Washington, DC, USA, 2000. IEEE Computer Society.

15. S. Waldherr, S. Thrun, and R. Romero. A gesture-based interface for human-robot interaction. In *Autonomous Robots*, volume 9, pages 151–173, 2000.
16. Y. Wu, J. Lin, and T. Huang. Capturing natural hand articulation. In *International Conference on Computer Vision*, pages 426–432, 2001.
17. M. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.
18. Benjamin D. Zarit, Boaz J. Super, and Francis K. H. Quek. Comparison of five color models in skin pixel classification. In *RATFG-RTS 99: Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 58–63, Washington, DC, USA, 1999. IEEE Computer Society.
19. J. Zhang, T. Baier, and M. Hueser. Integration of gaze and gesture detection in nature language instructing of robot in a assembly scenario. In *IEEE Int. Workshop on Robot and Human Interactive Communication*, 2002.
20. Jian-Hua Zheng, Chong-Yang Hao, Yang-Yu Fan, and Xian-Yong Zang. Adaptive skin detection under unconstrained lighting conditions using a bigaussian model and illumination estimation. *Image Analysis and Stereology*, 2005.