

Region Based Hand Gesture Recognition

Ahmet Birdal
Cankaya University
Computer Engineering Department,
Ankara Turkey
ahmetbirdal@gmail.com

Reza Hassanpour
Cankaya University
Computer Engineering Department,
Ankara Turkey
reza@cankaya.edu.tr

ABSTRACT

Evolution of user interfaces shapes the change in the human-computer interaction. With rapid emerge of three-dimensional (3-D) applications, the need for a new type of interaction device arises since traditional devices such as mouse, keyboard, and joystick become inefficient for interaction within these virtual environments. A better interaction in virtual environments requires a natural and suitable device. "Hand Gesture" concept in human computer interaction which has become popular in recent years can be used to develop such an interaction device. This study reports the development of a real-time, low cost, vision based hand gesture recognition system that works precisely on a relatively small restricted gesture space for single user robot control and human-computer interaction in such an environment that the lighting is relatively stable and the background is not complex.

Keywords

Hand gesture recognition, human computer interaction, tracking

1. INTRODUCTION

Evolution of user interfaces shapes the change in the human-computer interaction devices. One of the most common human-computer interaction devices is the keyboard which has been the ideal choice for text-based user interfaces. Graphical user interfaces brought mouse into the desktops of the users. As three-dimensional (3-D) applications take place the need for a new type of interaction device arises since traditional devices such as mouse, keyboard, joystick, etc... become inefficient for interaction within these virtual environments. A better interaction in virtual environments requires a natural and suitable device. "Hand Gesture" concept in human computer interaction which has become popular in recent years can be used to develop such an interaction device. The motivation for this study is to develop a real-time, low cost, vision based hand gesture recognition system that works precisely on a relatively small restricted gesture space for single user robot control and human-computer interaction in such an environment that the lighting is relatively stable and the background is not complex. Our main contribution are developing a region based algorithm for simplifying gesture recognition and designing a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright UNION Agency – Science Press, Plzen, Czech Republic.

hand gesture interpretation method using the combination of hand posture and position.

2. CLASSIFICATION OF THE METHODS

Hand gesture recognition is a relatively new field for the computer science. Applications for hand gesture recognition in machine learning systems have been developed approximately for 20 years. The methods used in these systems can be categorized into two groups. Generally the earlier systems used gloves for gesture recognition. That method was unpractical since the gloves were limiting moving abilities of the user. Recent studies like Malima et. al. [MALI] concentrated on vision-based systems since they provide relatively cost-effective methods to acquire and interpret human hand gestures while being minimally obtrusive to the participant.

2.1 Vision Based Hand Gesture Recognition Systems

A typical vision based hand gesture recognition system consists of a camera(s), a feature extraction module, a gesture classification module and a set of gesture models. In the feature extraction process the necessary features are extracted from the captured frames of the camera(s). This process can be divided into three sub categories:

- High-level features, generally based on three dimensional models,
- The image itself as a feature by view-based approaches,
- Low-level features measured from the image.

High level features can be inferred from the joint angles and pose of the palm. For this feature set, generally the anatomic structure of the hand is used as a reference. For precision purposes colorful gloves

can be used. View-based approaches are alternatives to the high-level modeling and they model the hands as a set of two dimensional intensity images. Low level features are based on the thought that the full reconstruction of the hand is not essential for gesture recognition. Therefore only some cues like the centroid of the hand region, the principle axes defining an elliptical bounding region of the hand, the optical flow/affine flow of the hand region in a scene, etc can be chosen as features. One of the most popular areas is recognition of a local sign language. Similarly finger alphabet is a popular field. The general purpose of these applications is either helping the deaf-dumb people for their communication with others or completely translating from a sign language into a normal one. Another type of application about sign languages is man-computer interaction, in other words, using sign languages as input, the information conveyed by the gestures is transferred to the computer via camera(s). Eisenstein and Davis [EISE] controlled a display in their application. Bretzner [BRET] developed a prototype system, where the user can control a TV set and a lamp. Robot control is the aim of the works of Malima et. al. [MALI], Starner et. al. [SORR]. Malima et. al. [MALI], propose an algorithm for automatically recognizing a limited set of gestures from hand images for a robot control application. The algorithm enables the robot to identify a hand pose sign in the input image, as one of five possible commands . The identified command is then used as a controller input for the robot to perform a certain task. Fujisawa et. al. [FUJI] developed an HID device as an alternative for mouse for physically handicapped persons. Human-Building interaction is considered at Malkawi and Srinivasan [MALK]. Marschall [MARS] has an interesting application which provides a visual sculpture. While some of the works mentioned above uses a complete sign language, some of them uses just a part of a sign language or develops an application-specific sign language for human-computer interaction.

2.1.1 High Level Features

High level features are extracted by model based approaches. A typical model based approach is creating a 3D model of a hand and projecting its edges onto 2D space by using some kinematics parameters. Ueda et. al. [UEDA] used such a method that estimates all joint angles to manipulate an object in the virtual space. In the method, the hand regions are extracted from multiple images obtained by the multi-viewpoint camera system. By integrating these multi-viewpoint silhouette images, a hand pose is reconstructed as a “voxel model”. Then all joint angles are estimated using three dimensional model fitting between hand model and voxel model. An experiment was performed in which the joint angles were estimated from the silhouette images by the

hand-pose simulator. Utsumi et. al. [UTSU] used multi-viewpoint images to control objects in the virtual world. Eight kinds of commands are recognized based on the shape and movement of the hands. Bray et. al. [BRAY] proposed a tracker based on ‘Stochastic Meta-Descent’ for optimizations in such high dimensional state spaces. The algorithm is based on a gradient descent approach with adaptive and parameter-specific step sizes. The Stochastic Meta-Descent tracker facilitates the integration of constraints, and combined with a stochastic sampling technique, can get out of spurious local minima. Furthermore, the integration of a deformable hand model based on linear blend skinning and anthropometrical measurements reinforce the robustness of the tracker.

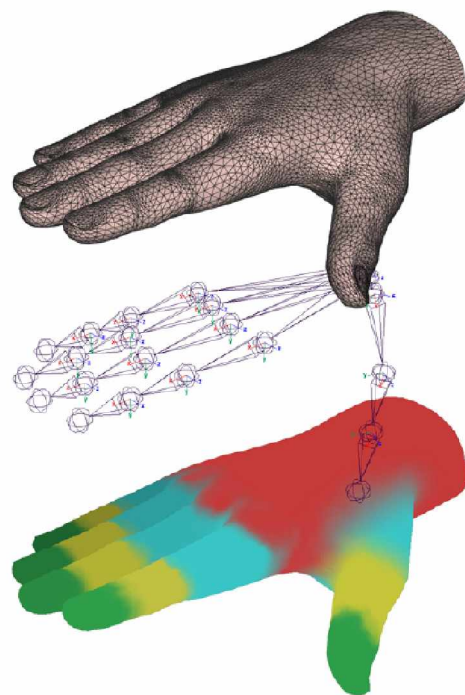


Figure 1.1 Anthropometrical Measurements [BRAY]

Bettio et. al. [BETT] presented a practical approach for developing interactive environments that allows humans to interact with large complex 3D models without them having to manually operate input devices. The system provides support for scene manipulation based on hand tracking and gesture recognition and for direct 3D interaction with the 3D models in the display space if a suitably registered 3D display is used. Being based on markerless tracking of a user’s two hands, the system does not require users to wear any input or output devices. In model based approaches the initial parameters have to be close to the solution at each frame and noise is a real problem for fitting process. Another problem is the textureless nature of the human hand difficulties

to detect the inner edges of the hand. Davis and Shah [DAVI] used a glove with markers in order to make the feature extraction process easier. Similarly manual parameter instantiation or placing user hands in a specific position were also used for the ease of initialization process. Processes on these features may be relatively slower than the other feature approaches due to the 3D structure complexity of high-level features.

2.1.2 View-based Approaches

These approaches are also called appearance-based based approaches. These approaches model the hand by a collection of 2D intensity images. At the same time, gestures are modeled as a sequence of views. Eigenspace approaches are used within the view-based approaches. They provide an efficient representation of a large set of high dimensional points using a small set of orthogonal basis vectors. These basis vectors span a subspace of the training set called the eigenspace and a linear combination of these images can be used to approximately reconstruct any of the training images. These approaches were used in many of the face recognition approaches. Success of them in face recognition made them attractive for other recognition applications like hand gesture recognition. Black [BLAC] demonstrated their approach by tracking four hand gestures with 25 basis images and provided three major improvements to the original eigenspace approach formulation:

- i) A large invariance to occlusions
- ii) Some invariance to differences in background from the input images and the training images
- iii) The ability to handle both small and large affine transformations of the input image with respect to the training images

Zahedi et. al. [ZAHE] showed how appearance-based features can be used for the recognition of words in American Sign Language from a video stream. The features are extracted without any segmentation or tracking of the hands or head. Experiments are performed on a database that consists of 10 words in American Sign Language with 110 utterances in total. The video streams of two stationary cameras are used for classification. Hidden Markov Models and the leaving one out method are employed for training and classification. Using the appearance-based features, they achieved an error rate of 7%. About half of the remaining errors are due to words that are visually different from all other utterances. Although these approaches may be sufficient for a small set of gestures, with a large gesture space collecting adequate training sets may be problematic. Another problem is the loss of compactness in the subspace required for efficient processing.

2.1.3 Low Level Features

Starner et. al. [STAR] noticed that prior systems could recover relatively detailed models of the hands from video images when given some constraints. However, many of those constraints conflicted with recognizing American Sign Language in a natural context, either by requiring simple, unchanging backgrounds; not allowing occlusion; requiring carefully labeled gloves; or being difficult to run in real time. Therefore they presented such a new and relatively simple feature space that assumes detailed information about hand shape is not necessary for humans to interpret sign language. They found that all human hands have approximately the same hue and saturation, and vary primarily in their brightness. By using this color cue they used the low level features of hand's x and y position, angle of axis of least inertia, and eccentricity of the bounding ellipse. This feature set is one of the first low-level features in the literature for hand gesture concept of computer vision. They combined the low-level feature set by HMM network and achieved the accuracy of %97 per word on a 40 word lexicon. Gökner and Yıldırım [GOKN] presented a hand gesture recognition system using an inexpensive camera with fast computation time. They used skin tone density and eccentricity of the bounding ellipse low level features and Multilayer Perceptron and Radial Basis Function neural networks for classification. They achieved the success of %78.3 on 3 layered structures and %80 for 4 layered structures. Lee [YANG] used low level feature, the distance from the centroid of the hand region to the contour boundary. The method obtains the image through subtract one image from another sequential image, measuring the entropy, separating hand region from images, tracking the hand region and recognizing hand gestures. Through entropy measurement, they have got color information that have near distribution in complexion for region that have big value and extracted hand region from input images. They could draw hand region adaptively in change of lighting or individual's difference because entropy offer color information as well as motion information at the same time. Detected contour using chain code for hand region that is extracted, and present centroidal profile method that is improved little more and recognized gesture of hand. In the experimental results for 6 kinds of hand gesture, the recognition rate was found more than 95%. Malima et. al. [MALI] proposed a fast algorithm for automatically recognizing a limited set of gestures from hand images for a robot control application. They considered a fixed set of manual commands and a reasonably structured environment, and developed a procedure for gesture recognition. The algorithm is invariant to translation, rotation, and scale of the hand. The low-level feature used in the algorithm is the center of the gravity and the distance from the most extreme point in the hand to the center

which is the farthest distance from centroid to tip of the longest active finger in the particular gesture.

Yang [YANG] presented an algorithm for extracting and classifying two-dimensional motion in an image sequence based on motion trajectories. First, a multi-scale segmentation is performed to generate homogeneous regions in each frame. Regions between consecutive frames are then matched to obtain two-view correspondences. Affine transformations are computed from each pair of corresponding regions to define pixel matches. Pixels matches over consecutive image pairs are concatenated to obtain pixel-level motion trajectories across the image sequence. Motion patterns are learned from the extracted trajectories using a time-delay neural network. They applied the proposed method to recognize 40 hand gestures of American Sign Language. They approximated the human head and hand shapes by ellipses. Roy and Jawahar [ROY] presented a feature selection method for hand geometry based person authentication system. They used lengths of four fingers and widths at five equidistant points on each finger as raw features. Since the localization of hands in arbitrary scenes is difficult, one of the major difficulties associated with low level features is that the hand has to be localized before feature extraction.

2.2 GESTURE CLASSIFICATION

The hand gesture classification approaches in the literature can be categorized into two main categories: rule-based approaches, in which the gestures are classified according to manually encoded rules and machine learning based approaches those are using a set of exemplars to infer models of gestures.

2.2.1 Rule-Based Approaches

In these approaches features of the input features are compared to the manually encoded rules. If any of the features or feature sets matches a rule, the related gesture will be given as output. As an example Cutler and Turk [CUTL] used a rule-based technique to identify an action based on a set of conditions in their view-based approach to gesture recognition. They defined six motion rules for corresponding six gestures. When the hands trace a motion path like a predefined rule, corresponding gesture is selected as output. The gestures and the rules used in the work are defined as the following:

2.2.2 Learning Based Approaches

As indicated in the previous section, the rule-based approaches depend on the ability of humans to find rules to classify the gestures. Learning-based approaches are alternative solutions to this problem when finding rules between features is not applicable. In this approach mappings between high-

dimensional feature sets and gestures are done by machine learning algorithms. The most popular method for this approach is using Hidden Markov Models in which gestures are treated as the output of a stochastic process. Many of the recent works like Nair and Clark [NAIR], Stamer et. al. [STAR] focused on Hidden Markov Models for gesture recognition. Russell and Norvig [RUSS] defines the HMM as “a temporal probabilistic model in which the state of the process is described by a single discrete random variable.” The possible values of the variable are the possible states of the world. Haberdar [HABE] used HMM for gesture recognition in his thesis study which is about Turkish Sign Language recognition. In the study 172 signs are used. Model parameters are determined by training G different HMM and using training data. For recognition of a sequence of observations, forward and backward algorithms are used by calculating the probabilities of HMM to generate the related observation.

3. PROPOSED METHOD

Two features of the hand, posture and position of the hands, are used in the proposed system. The “boundary rectangle” structure used for finding posture and position of the hands are described at the following section.

3.1 Boundary Rectangles

When the hands are segmented, an imaginary rectangular region is drawn outside of each hand. The boundaries are simply calculated by finding the positions of the minimum and maximum skin pixels in both vertical and horizontal. Figure 3.1 illustrates a sample boundary rectangle with a green mark.

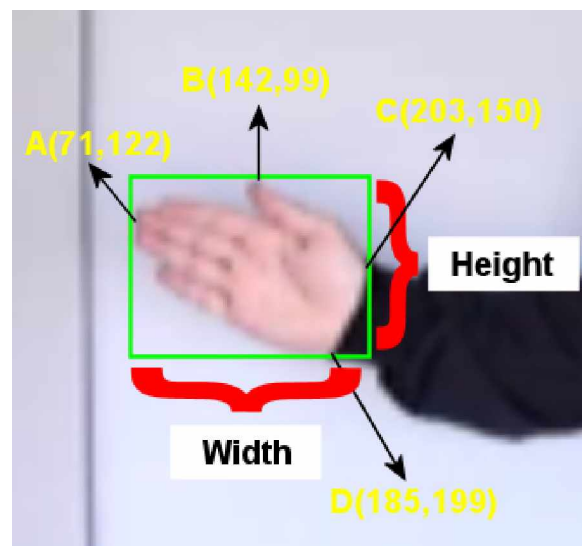


Figure 3.1 Boundary Rectangle

If the origin is at the top left corner of the image, the leftmost pixel that contains a skin color, which is

point A in Figure 3.1, determines the horizontal start point component of the boundary rectangle. The uppermost pixel that contains the skin color, which is point B in Figure 3.1, determines the vertical start point component of the boundary rectangle. Similarly points C and D define the left and lower boundaries of the region. For the ease of learning system usage and performance purposes two general groups of postures are used in the system. The longest edge of the boundary rectangle of each hand determines the type of the corresponding posture. If the longest edge is the height as shown in the Figure 3.2, this posture is called a “vertical posture”. Similarly, if the longest edge is the width, as shown in the Figure 3.3, the posture is called a “horizontal posture”.



Figure 3.2 Vertical Posture



Figure 3.3 Horizontal Posture

The type of the hand shape can be formulized as the following:

$$type = \begin{cases} \text{if } width \geq height, & \text{horizontal} \\ \text{else} & \text{vertical} \end{cases}$$

3.2 Hand Positions

In the system, the center point of a boundary hand rectangle is accepted as the position of the corresponding hand. In Figure 3.4, some positions of each hand are demonstrated and it can be seen that they are relative to the position of elbows those nearly have the same positions. We are interested in the gestures staying near the regions marked with red circles

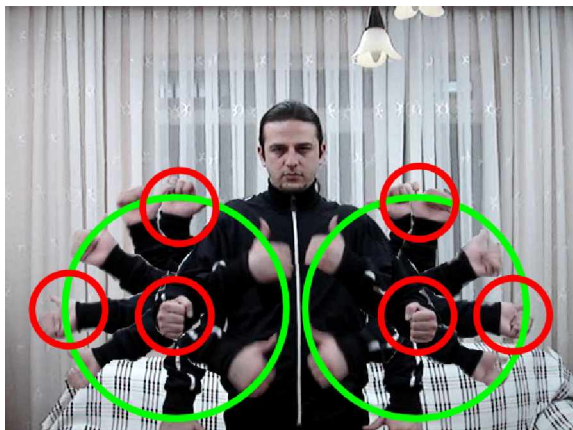


Figure 3.4 The regions considered in the proposed system

The system requirements have been established by taking consideration of practical usability and computer vision constraints. General algorithm of the system consists of three parts:

- Background Modeling
- Initialization
- Main Loop

3.2.1 Background Modeling

The first step for the system to work properly is the background modeling. This model will be used as a reference to subtract the captured frames during the hand gesture recognition process. The background is modeled by computing the mean and standard deviation of each pixel from a sequence of images taken while nobody exists in the scene. The background model works on HSV color model. Therefore, the mean of the squares sum and the square of the mean pixel values of these images are calculated after converting them into HSV color format.

3.2.2 Initialization

After background modeling, the performer should make a predefined special gesture. This gesture is used for initializing parameters for the trackers. Figure 3.5 shows the initializing gesture.

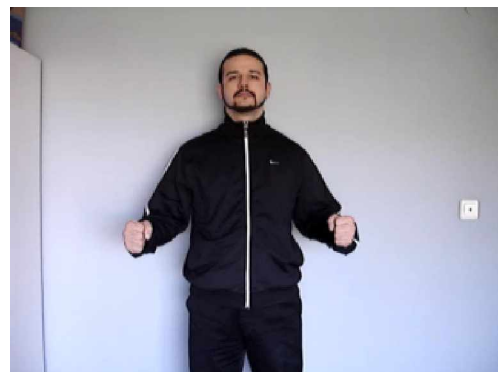


Figure 3.5 Special Gesture

Initialization process consists of three steps:

- Image Acquisition
- Segmentation
- Calculating Initial Parameters

After a successful image acquisition, segmentation process filters out the pixels which do not belong to the head and hands in the current image. This process is based on skin color and background color information. Segmented hand and head regions are used for calculating initial parameters for the system. Segmentation process consists of three parts:

- Background Subtraction
- Skin Filtering
- Noise Removal

Once an initial background model is created, the current frame captured by the device is subtracted from the reference image. Some noise may be observed in the resulting image because of the lighting and focusing problems. Hue and Saturation components of the HSV color model is used for skin color modeling since these components are not affected by the change in illumination so much. The skin color model is put into Gaussian model by $N(\mu, \Sigma)$ where μ is the mean and Σ is the covariance.

3.2.3 Calculating Initial Parameters

This process simply consists of locating the head and two hands and drawing boundary rectangles around these parts. Once a skin pixel is found region labeling algorithm is used to group the pixel together. The head is assumed to be near the center of the image and the hands are expected to be below the head. The parameters which are output of this process and also the entire initialization step are the center points of the head's and both hands' boundary rectangles and also the height and width of the head boundary rectangle.

4. Tracking

Tracking process is basically a mean-shift algorithm application. The initial parameters are used for determining the initial locations of the mean-shift search windows. The size of the initial tracking windows of the hand regions is the same as the boundary rectangle of the head. The head boundary rectangle is also used for the mean-shift tracker window without change. The segmentation step of the initialization process is applied here to have clear results from the mean-shift tracking algorithm. Some frames from application in which the tracking windows demonstrated with green rectangles are shown in the Figure 4.1.

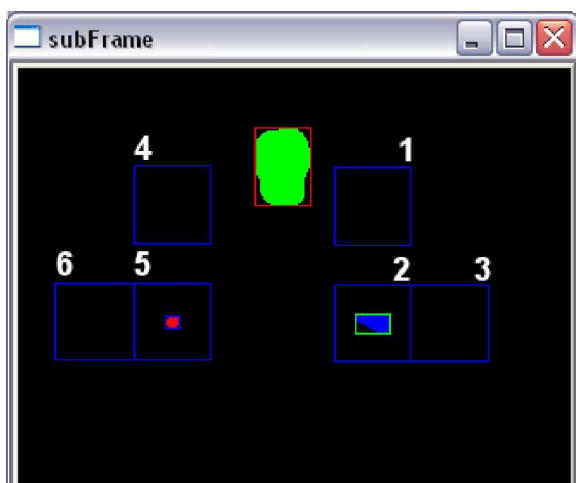


Figure 4.1 Hand Gesture Region Locations

4.1 Region checking

In this step the center of the tracking windows are checked whether they are in the predefined hand gesture regions. If any of the hands' centers of tracking windows are within any hand gesture regions, the region is reported to the Feature Extraction step, otherwise the system turns back to the first step of the main loop. The hand gesture region locations and sizes are calculated with the outputs of the initialization step. The shape of the hand gesture regions is square and the edges of these squares have the length of face rectangular region height.

4.2 Feature Extraction

As it is indicated in section 3, two types of features used in the system; hand positions and hand postures. For determining hand posture type a boundary rectangle for the related hand is drawn and the width and height properties of this rectangle are used.

4.3 Classification

A rule-based approach, as mentioned in section 2.3.1 is used in the system. The combinations of position and posture data of two hands give the value of the current state. An action classification only occurs when the left hand is inside any of the regions 1,2,3 and at the same time the right hand is inside any of the regions 4,5,6 and head is found. Some of the other cases such as one hand found in a hand gesture region may considered as a application related command such as "stop the application", "start the application", "undo the last operation", etc...

5. SAMPLE APPLICATION

In order to present the ease of use and efficiency of our system, we implemented a 3D application. The application consists of controlling a teapot shown in a window and it is developed by using Microsoft Visual Studio .Net , C++ with OpenGL library (Figure 5.1). The teapot can be translated along any X, Y, Z axis in both directions or rotated around any of these axes by hand gestures of a user.

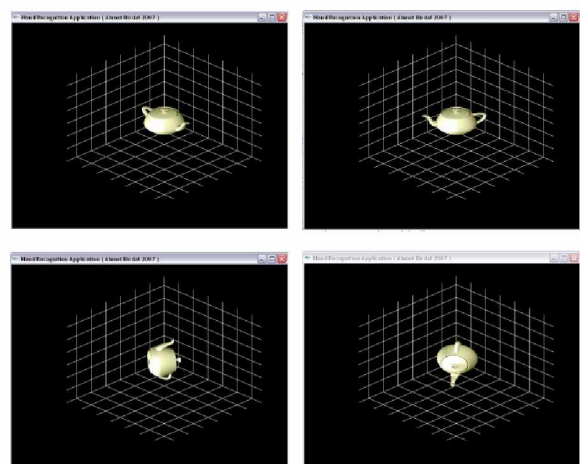


Figure 5.1 Rotation

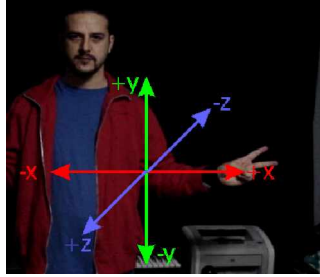


Figure 5.2 Axis

6 CONCLUSION

To measure the success of the proposed method we randomly chose 6 frames per action which makes 72 frames in total. From this set, 63 of the gestures are recognized successfully. The success rate of the system is calculated as %87,5 . The failure of the system to recognize the gesture is due to the unstable lighting conditions, performer's failure to move the hand to the proper region and initialization problems due to the skin-like colors present at the background. The system described here, offers a solution to some basic problems of hand gesture recognition systems: transition between gestures, precision, unintended gesture prevention, learning the system usage, feedback.

REFERENCES

- [BETT] BETTIO F. et. al.(2007).A practical vision based approach to unencumbered direct spatial manipulation in virtual worlds, In Eurographics Italian Chapter Conference. Eurographics Association.
- [BLAC] BLACK M. J. (1996). *EigenTracking : Robust Matching and Tracking of Articulated Objects Using a View-Based Representation*, International Journal of Computer Vision, 26(1), pp. 63-84, 1998. also Xerox PARC. Technical Report P95-000515.
- [BRAY] BRAY M. et.al (2004). *3d Hand Tracking By Rapid Stochastic Gradient Descent Using A Skinning Model*, Visual Media Production., (1st European Conference publication) 59- 68.
- [BRET] BRETZNER L., et.al.(2001).A Prototype System for Computer vision based Human Computer Interaction, Technical report, ISRN KTH/NA/P-01/09-SE, Stockholm, Sweden.
- [CUTL] CUTLER R., TURK M. *View-based Interpretation of Real-time Optical Flow for Gesture Recognition*, Third IEEE Conference on Face and Gesture Recognition, Nara, Japan, April 1998."
- [DAVI] DAVIS J. and SHAH M. (1994). *Visual Gesture Recognition*, Vision, Image and Signal Processing, 141(2):101-106.
- [EISE] EISENSTEIN J., DAVIS R., (2003) *Natural Gesture in Descriptive Monologues*, Proc. ACM Symp. User Interface Software and Technology (UIST 2003), ACM Press, , pp. 69-70.
- [FUJI] FUJISAWA S. Et.al.(1997). *Fundamental research on human interface devices for physically handicapped persons*, IECON 97. 23rd International Conference, New Orleans.
- [GOKN] GÖKNAR, G. and YILDIRIM, T. (2005) *Hand Gesture Recognition Using Artificial Neural Networks*, Signal Processing and Communications Applications Conference, 2005. (Proceedings of the IEEE)13th Volume , Issue , 2005 Page(s): 210 - 213
- [HABE] HABERDAR H.(2005). *Real Time Isolated Turkish Sign Language Recognition from Video Using Hidden Markov Models with Global Features*, MSc. Thesis., Istanbul: Computer Engineering, Yıldız Teknik University.
- [MALI] MALIMA A. et.al.(2006).A fast algorithm for vision-based hand gesture recognition for robot control, SIU06, , Antalya, Turkey
- [MALK] MALKAWI AM and SRINIVASAN RS (2005) *A new paradigm for Human-Building Interaction: the use of CFD and Augmented Reality*, Automation in Construction .Journal 14 (1): 71-84.
- [MARS] MARSCHALL M., *Virtual Sculpture - Gesture-Controlled System for Artistic Expression*, ConGAS Symposium on Gesture Interfaces for Multimedia Systems, AISB2004, Leeds, UK.
- [NAIR] NAIR V., CLARK J. J.(2002) *Automated Visual Surveillance Using Hidden Markov Models*, In VI02, pp 88.
- [ROY] ROY V. and JAWAHAR C.V.(2005). *Feature Selection for Hand-Geometry based Person Authentication*, in proceedings of International conference on advanced computing and communication, Coimbatore, India.
- [RUSS] RUSSELL S., NORVIG P (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ.
- [SORR] SORRENTINO A. et. al.(1997). *Using Hidden Markov Models and Dynamic Size Functions for Gesture Recognition*, BMVC.
- [STAR] STARNER T. et.al.(1997). *Wearable Computing Meets Ubiquitous Computing: Reaping the Best of Both Worlds*. ISWC 1999: 141-149
- [UEDA] UEDA E. et.al.(2003) *A Hand-Pose Estimation for Vision-Based Human Interfaces*, IEEE Transactions on Industrial Electronics, Vol. 50, No. 4, pp.676-684.
- [UTSU] UTSUMI A. et. al.(1999) *Multiple-Hand-Gesture Tracking using Multiple Cameras*, In Proc. of International Conference on Computer Vision and Pattern Recognition, pp.473-478,.
- [YANG] YANG M.-H.. *Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence archive.
- [ZAHE] ZAHEDI M.et.al.(2005). *Appearance-Based Recognition of Words in American Sign Language*, In IbPRIA 2005, (2nd Iberian Conference on Pattern Recognition and Image Analysis), LNCS volume 3522, pp511-519, Estoril, Portugal.