A Gesture Recognition System Using 3D Data

S. Malassiotis, N. Aifanti and M. G. Strintzis
Informatics & Telematics Institute, Thessaloniki, Greece,
malasiot@iti.gr

Abstract

In this paper a gesture recognition system using 3D data is described. The system relies on a novel 3D sensor that generates a dense range image of the scene. The main novelty of the proposed system, with respect to other 3D gesture recognition techniques, is the capability for robust recognition of complex hand postures such as those encountered in sign language alphabets. This is achieved by explicitly employing 3D hand features. Moreover, the proposed approach does not rely on colour information, and guarantees robust segmentation of the hand under various illumination conditions, and content of the scene. Several novel 3D image analysis algorithms are presented covering the complete processing chain: 3D image acquisition, arm segmentation, hand-forearm segmentation, hand pose estimation, 3D feature extraction, and gesture classification. The proposed system is tested in an application scenario involving the recognition of sign-language postures.

1 Introduction

Although, the last two decades we have witnessed a rapid evolution of computing, communication and display technology, the physical human-computer interface remains unchanged since the first workstations. Traditional interface devices, keyboards and mice, are inadequate for modern applications such as interaction with complex three-dimensional environments and sign language recognition. Recently several innovative controllers and sensors have been investigated towards a more “natural” interaction with the machine. Several of these new systems, such as glove-based devices, compromise convenience by requiring the user to be instrumented with encumbering devices, in order to achieve high expressiveness.

The use of gesture recognition provides an attractive alternative to cumbersome interface devices for human-computer interaction. Applications of this technology include advanced user interfaces for the interaction with virtual objects and interpretation of gestures from the sign-language alphabet to aid natural interaction of hearing impaired people with computing devices.

Vision-based recognition of hand gestures, is an extremely challenging task due to the complexity of the human hand structure and motion. Traditional single camera sensors are sensitive to environmental conditions e.g. background, illumination, activity etc. and suffer from distortion and ambiguities introduced by the perspective projection of 3D points on the sensor plane. To cope with the above difficulties several researchers have proposed using more than one camera, or exploiting 3D information acquired by passive stereo sensors.

In [9] a gesture recognition system based on a range sensor is proposed. The algorithm is capable of recognising a limited set of simple manipulative gestures, containing static finger configuration, while the hand segmentation problem is not addressed. To cope with the problem of occlusions, a multi-viewpoint hand gesture tracking system is proposed in [10]. The best viewpoint is selected based on the estimated hand rotation in 3D. Then, 2D shape Fourier descriptors are extracted and used to recognize a limited set of simple gestures. A finger tracking scheme relying on a multiple camera configuration is proposed in [7]. The system combines various cues such as color, edges, 3D shape and motion for robust detection of fingertip position and orientation. A similar system, relying however on dense stereo measurements is described in [8]. The hand pose estimation problem is also addressed in [3]. A 3D model of the hand is used that is iteratively fitted to dense 3D data. A stereo-based gesture recognition technique is described in [6]. The orientation of the arm and location of the hand is estimated by means of sparse 3D data. This is subsequently used to drive a colour-based hand segmentation algorithm. Moment-based 2D shape gesture classification is finally performed on the perspectively un-warped hand images.

In this paper a gesture recognition system using 3D data is described. The system relies on a novel 3D sensor that
generates a dense range image of the scene. The main novelty of the proposed system, with respect to the 3D gesture recognition techniques mentioned above, is the capability for robust recognition of complex hand postures such as those encountered in sign language alphabets. This is achieved by explicitly employing 3D hand features. Moreover, the proposed approach does not rely on colour information, and guarantees robust segmentation of the hand under various illumination conditions, and content of the scene. Several novel 3D image analysis algorithms are presented in the following sections, covering the complete processing chain: 3D image acquisition, hand segmentation, hand-forearm segmentation, hand pose estimation, 3D feature extraction, and gesture classification.

2 3D data acquisition

A 3D & colour camera acquiring 3D as well as color images is used. This is based on an active triangulation principle, making use of an improved and extended version of the well-known Coded Light Approach (CLA) for 3D data acquisition [5]. The average depth accuracy achieved, for object located about one meter from the camera, is less than 1mm. In the experiments, a range-only acquisition mode is preferred to avoid the annoying flickering of the projected light pattern. The acquired range images contain artifacts and missing points mainly over areas that cannot be reached by the projected light. Instead of filtering or interpolating 3D data, a process that may lead to further artifacts, we prefer making subsequent processing stages robust to the above artifacts.

3 Arm Segmentation

An important step in vision-based gesture recognition is the segmentation of user hands from the background, i.e. the user’s body and other objects in the scene. The problem is usually simplified by several assumptions on the scene content, illumination, motion and camera configuration, by controlling the environment and by limiting the working space. Under these constraints skin colour-based segmentation, motion detection or background subtraction based on previously trained background model have been proposed that demonstrate relatively good performance. The benefit of using depth information is that robust image segmentation may be achieved without posing any constraints to the environment or the users of the system. This is a very important requirement for natural human-computer interaction.

Segmentation of the scene by thresholding the depth values is not particularly robust especially when the user’s body is slanted with respect to the camera plane. Therefore, segmentation of the subject’s arms by means of a hierarchical unsupervised clustering procedure is proposed in this paper. This is based on the observation that the various parts of the body, such as the arms, torso and head, form compact 3D clusters in space. The proposed segmentation algorithm comprises the following steps:

1. An initial clustering is obtained by sequentially scanning the depth image, and classifying each pixel according to the distance from previously classified pixels in its neighborhood. The distance measure used is the Euclidian metric in the Z component. This procedure leads to a large number of small regions.

2. The aim of this step is to favor larger regions, by merging small regions into larger ones. The smallest cluster is selected and merged with the cluster that minimizes the between cluster variance $S_B$. The iterative procedure is terminated as soon as a certain number of clusters is reached.

3. Finally, hierarchical merging of adjacent clusters is performed. Two clusters are merged if the total scatter of the combined cluster $S_T$ is minimized. The procedure is terminated when a specific number of clusters is reached.

The clusters corresponding to the arms may by subsequently selected by employing prior knowledge. From the 3-4 clusters closer to the camera, we select those that have a relatively elongated shape, corresponding to the arms.

4 Hand - Forearm Segmentation

The segmentation of the palm and fingers from the forearm is important for accurate estimation of the hand pose and subsequent feature extraction procedure. Our approach relies on statistical modelling of the arm points in 3D space.

The probability distribution of a 3D point $x$ is modelled as a mixture of two Gaussians:

$$P(x) = P(\text{hand})P(x|\text{hand}) + P(\text{forearm})P(x|\text{hand})$$

$$= \pi_1 N(x; \mu_1, \Sigma_1) + \pi_2 N(x; \mu_2, \Sigma_2)$$

where $\pi_1, \pi_2$ are prior probabilities of the hand and forearm respectively, and $N(x; \mu, \Sigma)$ is the 3D Gaussian distribution.

Maximum-likelihood estimation of the unknown parameters $\pi_k, \mu_k, \Sigma_k, k = 1, 2$ from the 3D data is obtained by means of the Expectation-Maximisation algorithm. The convergence of the iterative procedure relies on good initial parameter values. In our case these may be obtained by exploiting prior knowledge of the arm geometry.
Classification of a 3D point \( x_n \) to the class \( k \) is performed by the maximum likelihood criterion. Experimental results demonstrate robustness of the algorithm under various orientations of the palm and finger configurations.

5 3D Pose Estimation and Compensation

Classification algorithms are prone to rigid transformations of the input pattern. In our case, availability of 3D information leads to efficient estimation of the orientation of the hand, thus making a 3D pose estimation-compensation approach more appropriate. To alleviate the effect of small errors in the 3D pose estimation process, the training set used for gesture classification is augmented by a small set of rigidly transformed views of the patterns around the canonical position.

An estimate of the 3D orientation of the hand is obtained by computing the principal direction of the 3D data, given by the eigen-vector \( u_1 \) of the scatter matrix \( S_{\gamma} \), corresponding to its larger eigen-value. To limit the effects of the fingers to the estimated orientation vector, each 3D point is weighted by its distance \(|d|\) from the center of the palm, estimated using the algorithm proposed in [2]. To cope with outliers that are due to 3D sensor noise, a robust LMS (Least Median of Squares) algorithm was used.

6 Feature Extraction and Gesture Classification

Each normalised depth image is represented by a feature vector, that is extracted using the following procedure. Let \( s(x, y) \) be the binary segmentation mask designating “hand” pixels in the normalised depth image. The distance transform image \( dt \) is computed from \( s \) using the approximate chamfering algorithm [1]. The distance transform gives for each pixel the minimum distance from the shape boundary and may fully describe the 2D shape. The redundancy among the values of \( dt \) and the depth values of the normalised depth image is removed by exploiting the circular structure of the hand. The center of the palm \( c \) is located by searching for a maximum in \( dt(x, y) \) [2]. This is equivalent to locating the maximal circle that is totally within the 2D shape in \( s \). Then, the distance function as well as the depth image are uniformly sampled along \( k_r \) circles with center \( c \) and increasing radius \( r_i, i = 1, \ldots, k_r \) and the corresponding values are sequentially stored in a vector \( f = \{ f_i \}, i = 1, \ldots, d = k_r l_c \) where \( l_c \) is the circle sampling rate. Typical values for \( k_r = 6 \) – 8 and \( l_c = 100 – 200 \) lead to feature vectors with high dimensionality \( d \). This may be significantly reduced by performing a Fourier transform to the values corresponding to each circle, and keeping the low frequency coefficients. This has the additional advantage of achieving rotation invariance on the x-y plane.

Classification of an input depth image, to one of the predefined hand posture classes represented by the training set is performed using the k-nearest-neighbor rule [4]. This rule classifies a new input image, and extracted feature vector \( a \) by assigning it the label most frequently represented among the \( k \) nearest training samples. The proximity of input vector to training sample vectors is computed using the normalised cross correlation measure.

7 Experimental Results

A set of 20 hand postures are selected from the Greek sign language alphabet. Ten of these postures correspond to the numbers 0 to 9, while the other ten are generic ones that are used in many gesture sequences. Currently our training set contains 10 images for each posture.

The training images are generated in an semi-automated manner using a 3D hand model. This model is controlled by a set of parameters that correspond to the configuration and structure of the fingers. A data-glove device is used to capture the above parameters from the user in real-time. Then, small perturbations are introduced to the model parameters, in order to generate rigid and non-rigid variations of the given hand posture and simulate user specific attributes such as finger length. Then a sub-set of the generated models is selected, and depth images are generated for each of them.
by defining a virtual camera and measuring the distance of the 3D model surface from the camera plane (see fig. 3).

The obvious advantages of the above training process is the efficiency in adding new postures and direct control over the parameters of variability. Also, the training samples are free from any errors that may be introduced by the 3D sensor, such as occluded areas.

The proposed gesture recognition system has been implemented on a PC Platform with Pentium III 1 Ghz processor. The computation time for the complete chain is less than 1 second, without any optimisations, operating on sub-sampled images ($180 \times 144$). The most intensive task (consuming 80% of the processing time) appears to be the initial segmentation step. This may be sufficiently reduced if some constraints are posed to the distance of the hands from the body.

Hand segmentation and normalisation results demonstrated robustness of the proposed algorithms, giving correct results in all but a few tested images (5 from 200 total).

To test the performance of the classification algorithm we have performed two tests. The first test is performed on the synthetic data, by excluding each time one of the training samples, and training the system with the remaining samples. Then, the excluded sample is classified. The recognition rate achieved is 97%. The second test was performed using real data. The test set consisted of 4 different postures with 50 images for each posture. The achieved recognition rate was found to be equally high (95%).

Due to lack of space, extensive results are given in (http://server-2.iti.gr/sotiris/gesture.html)

8 Conclusions and Future Work

We have demonstrated a complete system for the recognition of static hand postures based on a 3D sensor. The system relies only on range data, therefore is invariant to the content and illumination of the scene. This makes it suitable for operation in unconstrained environments. Also, it is tolerant to the 3D pose of the hand by including a pose compensation procedure. The classification of hand postures is achieved by representing the range images by a discriminative feature vector that incorporates 3D shape information. Experimental results demonstrate that improved efficiency and robustness may be achieved through the use of 3D information.

Extensive testing, with a larger number of images and test cases remains to be performed in order to judge the limitation and breakdown points of the proposed system. Also, adaptation of the system to a real-time application environment is currently under investigation.

9 Acknowledgement

This work has been supported by the EU project HIS-CORE “High Speed 3D and Colour Interface to the Real World” (IST-1999-10087).

References