KEYBOARD FINGER GESTURE RECOGNITION USING TIME DELAY NEURAL NETWORKS

Abstract

Gesture recognition systems recognize hand movements. Time Delay Neural Networks (TDNN) is a widely used method for managing gesture procedures. This paper proposes a low cost artificial vision method for finger gesture recognition on the keyboard. The methodology is based on four aspects: preparation and capture of a video signal, hand segmentation, finger identification, feature vector exportation and finally classification using TDNN. This research is a step towards a more vision-guided user interface, rather than a peripheral-guided one.

1 INTRODUCTION

Computer Vision, or artificial vision, focus on determining the steps which allow computers "to comprehend" what it "sees" through a connected camera. The advantage of computer vision as a geture recognition method, is the interaction without contact.

Many computer devices (mouse, keyboard etc) require human-device contact in order to recognize a gesture. Virtual reality devices (gloves, sensors etc) are also based on contact, and have the additional disadvantage of being uncomfortable.

This paper presents an alternative method for retrieving information of a keyboard user's right hand, based on computer vision technology. It also studies the use of Time Delay Neural Networks as a recognition method for the finger gestures.

The aim of this research is to replace the computer's keyboard, respecting the human natural manipulative and communicative skills. A system, which could detect finger gestures, could indeed replace the keyboard; all that would be needed is a small camera mounted to see the hand. If the computer vision system could recognize the meaning of a finger gesture, it might suitably correspond.

2 RELATED APPROACHES

The task of developing systems that recognize large amplitude of gestures, adapted for different users, is difficult. Additionally, there are well known gesture recognition methods which are efficient and provide accurate results but it is considered costly. Sensor technology belongs to such methods. Without doubt sensors provide easier access in exceptionally precise information, while being uncomfortable due to the special equipments There are many works using a sensor data glove for gesture recognition, still the main disadvantage in them is the high cost (*Vicon Peak, 2005 ; Quad, 1990*).

On the other hand, artificial vision systems can have similar results as the sensor technology, costing less. A large number of researches are focused on hand gesture recognition, based on static hand postures (*Heckenberg, 2000, 2002*). They only recognize one object in each video frame. They don't take into consideration of the dynamic nature of the gesture. This paper presents a method for recognizing each finger's gesture on the video frame (*Boukir, 2004; Manitsaris, 2008; Palmer, 2000; Drouin, 2003*).

3 CONTRIBUTION

This research is a step towards a more vision guided interface, rather than a peripheral guided one; the computer retrieves information based on what it "sees", rather than from information forced in by a keyboard. This method could contribute to solve hand's kinesitherapy problems related to the use of computer peripherals.

Additionally, this study could be proven beneficial to HCI, by providing a "finger spelling" algorithm on specific problems, such the "fingering problem" in the field of the music interaction on a piano. Such a method can be used for educational purposes helping beginners to learn piano. It can also be applied for the composing of melodies without a musical instrument, as well as, for the modeling of the image of the

pianist's performance both in automatic music production and score generation.

4 METHOD OVERVIEW

A small low cost camera has been used so as to fully recover and record finger movements in two dimensions. Thus, it is enough to have the signal coming from the video camera in order to a) locate finger's position, b) estimate their orbit, c) extract the features of the gesture.

After the data capturing process, video signal processing techniques have been applied using appropriate MATLAB toolboxes (*DIPimage, Neural Network*) so as to construct our finger recognition model. Firstly, a skin color model has been created for detecting skin pixels on any video frame. The skin color model outputs binary masks.

Then, noise reduction techniques are applied so as to improve the hand's image for further evaluation. In the filtered hand images-video frames mathematical morphology techniques are applied for defining the hand's silhouette. This step is essential so as to retrieve the hand's contour by applying the Canny algorithm on the hand silhouette binary mask (*Radicioni, 2004; Burns, 2006; Traube, 2000*).

In order to define the fingers in each hand, the Euclidean distances between the hand's centroid and contour, must be computed. Then, the local maxima of these distances can lead to finger identification. The finally step is to compute the coordinates of each finger so as to be used as feature data in a neural network for classification and gesture recognition.

Time Delay Neural Networks (TDNN) is used for recognizing gestures. TDNN are frequently used in applications concerning voice, text or gesture. Finger movements constitute expressive movements of the human body, generally entailing both time and space variations. Gesture's classifications are effected with the use of TDNN.

In the following section, the main parts of the proposed methodology are presented.

4.1 Deployment phases

The methodology is divided in four main phases (Figure 1):

Phase 1: Capturing video signal

Phase 2: Finger recognition

Step 1**:** Skin model and detection

Step 2**:** Hand segmentation

Step 3**:** Finger extraction

Phase 3: Feature vectors exportation

Phase 4: Classification using TDNN

4.2 The algorithm

In the following subsections the phases and subsequent steps of the proposed methodology are presented along with particular methods and tools applied.

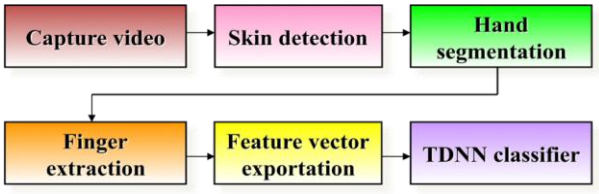

Figure 1: Deployment phases.

4.2.1 Capturing video signal

The ergonomic setup of the camera is an essential factor. Fingers move along three axes in space X, Y, Z. For better results, the hand should be at the center of the image. Due to difficulties encountered in the process of calibration and video synchronization, we only use one camera (*front view*) for capturing axes X and Y.

4.2.2 Finger recognition

In this phase the main model design and our gesture recognition approach are explained and presented. The second phase of the methodology involves three subsequent steps which are skin model and detection, hand segmentation and finger extraction (*Rehg, 1993*).

1) Skin model and detection

One of the most popular image segmentation methods is based on the color of the objects. The segmentation is made using only the chrominance components of the image and not the luminance ones (*Heikkilä, 2008; Jones, 1998*)

As a first step is the collection of pixels samples of the skin color and the determination of the Region of Interest (RoI). RoI is produced by a number of images UIL (*User Image Library*) in RGB format in order to create the appropriate model. The skin model is based on skin pixel samples, chosen from an image library.

Next the RoI normalization phase it follows. RoI normalization is the RGB transition to the normalized rg. During this procedure, RGB components (*chrominance and luminance*) are transformed only to chrominance components and are depicted on the two-dimensional space of chrominance rg. Every pixel of RoI^{RGB} is depicted into RoI^{rg} (Figure 2).

Last but not least, the skin model is defined. The ranges of the skin model are defined as

$$s_r = [r_{min}, r_{max}] \text{ and } s_g = [g_{min}, g_{max}].$$

These ranges describe the sample distribution in the 2D chrominance space.

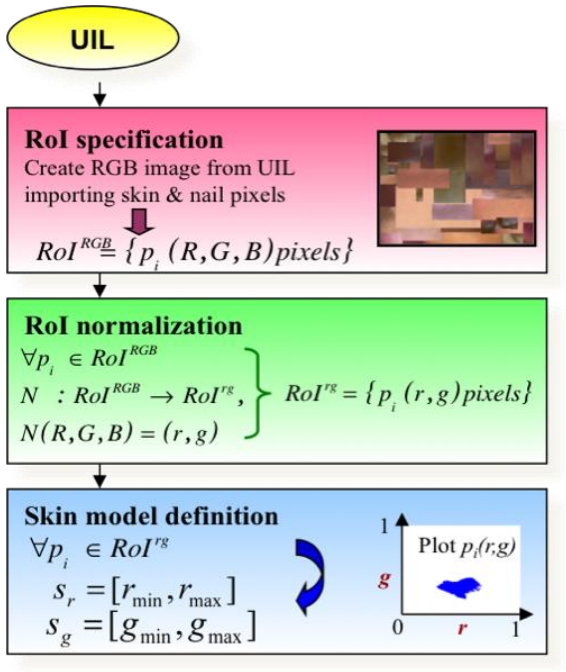

Figure 2: Skin model definition.

After having defined the skin model the detection of dermal regions in the video input can be easily achieved. In every pair of (r, g) we match the possibility that this pair belongs to a skin pixel for each one of the two models. Thus, binary masks are created, which define on the image regions that contain skin information.

The evaluation of a great number of frames shows that the skin model works fine. The algorithm for finger allocation in our method works on the basis of the allocation of the fingertips. Obviously, the fingertips are identified with the nails. We can conclude that it is preferable nail pixels to be included in the skin model. In case that this is not possible, the problem can be avoided by applying proper filters and increasing the complexity of the algorithm in real time applications.

2) Hand segmentation

Hand segmentation is the procedure via which the captured image is divided into regions including only a user's hand. This procedure is considered essential since the skin detection model is not perfect. In our experiment some shadowed dermal regions have been excluded by the skin detection algorithm, while other non-dermal regions have been assumed as dermal. As a consequence, the hand is perceived like one mass with a lot of noise.

Filters for noise reduction are applied taking as granted the specific position of the user's hand. An opening filter of a high value in elliptic form is applied by producing a morphological opening in binary images. By applying this filter, the resulted images indicate a much better separation of fingers and, also the decrease of the non-dermal regions that have been considered as dermal. The result of filter usage is impressing when the fingers are together.

Furthermore, a min filter is applied in order to exclude discreet points of the background, which is aimlessly considered as dermal regions. Following the min filter, a Gauss filter has been applied. This filter provides a smooth enough hand contour. Then, a double bilevel thresholding is applied in order to extract the hand silhouette while the fingers are distinguished clearly. At a final stage, a Canny filter is chosen for extracting the hand's silhouette. The result of this procedure is the exportation of a binary image including only the hand's contour.

3) Finger extraction

By having extracted the hand contour, then the centroid is calculated for every frame. Then, the Euclidean distance between every pixel of the hand contour and the centroid is calculated (Figure 3). From the graphic depiction of the curve of the distances local maxima are detected. After applying a low pass filter for the constriction of the maxima and minima we came up with five points matching to the five fingertips.

A classifier of five classes (*one for each finger*) has been used in order to estimate the finger position. This characteristic is very useful when a finger is not detected on the image, or a finger is put into another. The estimated position is the average of the last three positions.

4.2.3 Feature vectors exportation

Suitable vectors, created by these features, constitute the input of Time Delay Neural Networks (TDNN), modeling each movement. Features used for the training of the TDNN are comprised by coordinates on

the two axes (X, Y) of the five fingers (k) of the right hand.

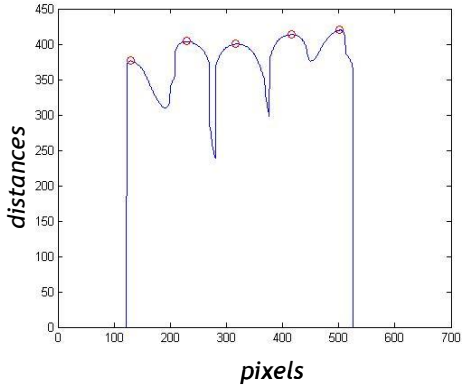

Figure 3: Finger extraction and identification.

These features are: X_k, Y_k, for k=1 to 5. For every frame (j), the position of each finger is located using all features (i) for each gesture (w) consequently. All the above is registered in 3D array A_{ixjxw}. Using the coordinates (*training features for the TDNN*), on the two axes for the five fingers, segregation for the gesture of each finger is achieved.

4.2.4 Classification using TDNN

The approach of neural networks is adapted on the classification of the models. TDNN transform the time into space parameter. A neural network of 12 inputs, 25 neurons and 6 outputs has been used. The inputs and outputs vary between -1 and +1. This network is already adapted to the Fourier descriptors, which offer high computational speed.

A set of 6 English words has been defined. The TDNN has been trained with a set of 10 video sequences from a user typing each of the 6 words on the keyboard. The "back propagation" algorithm is used for training the network (*training rate: 0.3*). The array descriptors are normalized between -1 and +1. Their standard deviation is 1 and the average is 0.

The TDNN classification output concerns the transformation of the recognized finger gestures data into commands, in order to control the software. For every gesture, method outputs pairs of values, which indicates fingers and position on the keyboard. These positions are compared with a matrix matching letters and key coordinates. So, letters replaces positions on the keyboard, while pairs of letters and fingers are exported.

5 EVALUATION

The aim of our research is to make the computer user feel and move freely in time and space so he/she will not be obliged to always keep in mind that he/she must keep his hand and fingers in a specific position.

By evaluating each phase of the proposed method, we came to the conclusion that the method works better for clearly defined finger movements. The major increase in the error percentage of our method is observed when the fingers of a user are put the one into the other. In such cases finger extraction is quite difficult because of blurring effects caused by applying the Gauss filter. This undesired outcome is avoided with the use of classifiers, which estimate the finger position for the next frame. Additionally, even if it is not fully examined yet, we believe that the application of an adaptive convolution scheme could solve this problem in a number of cases. On the other hand, the method doesn't work in cases of head existence in the captured video frames. Unfortunately, it is not studied in our work a method for cutting non-finger dermal regions.

Furthermore, the TDNN work really faster than other techniques, such as template matching. The processing time for every video frame is 1 second. Generally, the evaluation of the method concerning the fingerings retrieval via the video signal comes with satisfying conclusions, but it is also important to further evaluate the recognition of finger movements.

CONCLUSION

This paper discusses a new artificial vision method of capturing the finger gestures on the computer keyboard, using a low cost camera. We attempted an initial approach to recognize the finger gestures of the right hand of the user using content-based video analysis and TDNN. Although we succeeded in recognizing the finger gestures of the user, however, the method has only been tested on one person and its function has not yet been evaluated under different circumstances. The plan is to further investigate the calibration and synchronization of both cameras (*front and top view*), recording at the same time.

References

Boukir S., Cheneviere F. (2004). *"Conception d'un système de reconnaissance de gestes dansés"*, TS. Traitement du signal ISSN 0765-0019, vol. 21, no3, pp. 195-203.

Burns A.-M. and Wanderley M. (2006). "Computer vision method for guitarist fingering retrieval" *In Proc. of the Sound and Music Computing Conference*, Centre National de Création Musicale, Marseille, France.

Drouin, S., Hébert, P. & Parizeau, M. (2003). Simultaneous tracking and estimation of a skeletal model for monitoring human motion. *In*

Proceedings of Vision Interface Conference, p. 81-88. Halifax, Canada.

Heckenberg, D. and Lovell, Brian C. (2000). "MIME: A Gesture-Driven Computer Interface", *In Proc. Visual Communications and Image Processing (SPIE) V 4067*, 20-23.

Heckenberg, D. and Lovell, Brian C. (2002). "Low-Cost Real-Time Gesture Recognition", *in Proc. ACCV2002*, pp. 22-25.

Heikkilä, J. (2008, 11/2-9/5), Machine Vision [Online]. http://www.ee.oulu.fi/mvmp/courses/mv/?en

Jones, M. J. and Rehg, J. M. (1998). "Statistical Color Models with Application to Skin Detection", Tech. Rep. CRL 98/11, Cambridge, United Kingdom, Research Laboratory.

Manitsaris S. and Pekos, G. (2008). "Computer vision method for pianist's fingers information retrieval", *In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, Linz, Austria.

Palmer, C. and Pfordresher, P. Q. (2000). From my hand to your ear: the faces of meter in performance and perception. *In C. Woods, G. Luck, R. Brochard, F. Seddon & J. A. Sloboda (Eds.) Proceedings of the 6th International Conference on Music Perception and Cognition*. Keele, UK: Keele University.

Quam D. L. (1990). "Gesture recognition with a DataGlove, in Aerospace and Electronics Conference". *In Proc. of the IEEE National (NAECON)*, vol.2, pp. 755-760.

Radicioni D., Anselma L. and Lombardo V. (2004). "A segmentation-based prototype to compute string instruments fingering", *In Proc. of the Conference on Interdisciplinary Musicology*, Graz, Austria.

Rehg, J. M. and Kanade, T. (1993). "DigitEyes: Vision-Based Human Hand Tracking", Tech. Rep. CMU-CS-93-220, Carnegie Mellon University, Canada.

Traube C. and Smith J. O. (2000). "Estimating the plucking point on a guitar string", *In Proc. of the COST G-6 Conference on Digital Audio Effects*, Verona, Italy.

Vicon Peak (2005). *Vicon Motion Capture System*, Lake Forest, CA.