DACTYL ALPHABET GESTURE RECOGNITION IN A VIDEO SEQUENCE USING MICROSOFT KINECT

S. G. Artyukhin[a] , L. M. Mestetskiy [a]

[a] MSU, Faculty of Computational Mathematics and Cybernetics, Russia, 119991, Moscow, GSP-1, 1-52, Leninskiye Gory

KEY WORDS: Gesture recongniton, Kinect, Morphological skeleton, Dactyl sign, Segmentation

ABSTRACT:

This paper presents an efficient framework for solving the problem of static gesture recognition based on data obtained from the web cameras and depth sensor Kinect (RGB-D - data). Each gesture given by a pair of images: color image and depth map. The database store gestures by it features description, genereated by frame for each gesture of the alphabet. Recognition algorithm takes as input a video sequence (a sequence of frames) for marking, put in correspondence with each frame sequence gesture from the database, or decide that there is no suitable gesture in the database. First, classification of the frame of the video sequence is done separately without interframe information. Then, a sequence of successful marked frames in equal gesture is grouped into a single static gesture. We propose a method combined segmentation of frame by depth map and RGB-image. The primary segmentation is based on the depth map. It gives information about the position and allows to get hands rough border. Then, based on the color image border is specified and performed analysis of the shape of the hand. Method of continuous skeleton is used to generate features. We propose a method of skeleton terminal branches, which gives the opportunity to determine the position of the fingers and wrist. Classification features for gesture is description of the position of the fingers relative to the wrist. The experiments were carried out with the developed algorithm on the example of the American Sign Language. American Sign Language gesture has several components, including the shape of the hand, its orientation in space and the type of movement. The accuracy of the proposed method is evaluated on the base of collected gestures consisting of 2700 frames.

1. INTRODUCTION

A gesture is a form of non-verbal communication or non-vocal communication in which visible bodily actions communicate particular messages, either in place of, or in conjunction with, speech. Gestures include movement of the hands, face, or other parts of the body. In the last decade, more and more attention is paid to the automatic recognition of gestures. This is because the gestures is a convenient way to enter information into a computer. The appearance of depth sensors such as Kinect, and the computing power of personal systems provide an opportunity to solve the problem of gesture recognition in real time. Important in the field of gesture recognition is the problem of recognition of sign language. Gestures of such languages are divided into two types: dynamic, in which the important movement and change hands posture, and static, which are determined only by the hand posture. In sign language most gestures are dynamic. Their diversity makes it difficult to recognize. More simple language recognition is dactyl letters and numbers. Dactyl sign language - is the language in which each letter and number corresponds to a gesture usually static. There are more and more devices in the market which help to solve the problem of gesures recognition. One such device is the Kinect, developed by Microsoft. Kinect consists of two cameras and an infrared projector and allows to obtain both color image and depth map. Also, Microsoft has developed a library that allows to recognize human posture, but there is no standard solutions for hand gestures. We solve the problem of static gesture recognition based on data obtained from the web cameras and Kinect depth sensor. Static gesture given by a pair of images: color image and depth map. The database gestures stored in one or more frames for each gesture. At the stage of the recognition algorithm is applied to the input video sequence (a sequence of frames) which we want to mark, that is put in correspondence each frame of the sequence with bases gesture, or decide that there is no appropriate gesture in the database. Classification of frames of a video sequence is independent, without interframe information. Then, a sequence of successful marked frames in equal gesture is grouped into a single static gesture.

2. RELATED WORK

Many approaches for gesture recognition are described in the literature. Most articles describe the methods that work with a small set of highly different gestures. Some methods are not applicable to real-time. To Some techniques required plain background or special gloves.

The first group of methods are defined by geometrical model for each hand gesture, then it can be reconstructed from the image. For example, in (Wang2009) multi-colored glove is used. Glove arranged so that in most cases it is possible to restore the hand model. Glove helps to accurately segment the image for almost all type of background, but recovery is performed by storing a large database and finding a suitable gesture. It lead to a great deal of time and memory. In (Stenger2006) also the three-dimensional model of the hand is restored, the method is based on storage the large number of examples for each pose hands. The method proposed in (Malik2003), based on skin color segmentation and searching for fingers. Selecting only ends of the fingers narrows the set of gestures for which method is applicable.

The second group includes methods based on direct feature generation from the image. For example, in (Suryanarayan2010) the image is normalized with the palm width and height, a uniform grid is introduced, and the degree of filling of each cell is used as feature. For qualitative classification requires stored for each rotation gesture, which leads to increase the storage.

The third group of methods is based on a comparison of the hand shape with the standards of the base. In (Beristain2010) method for comparing the shape bassed on discrete skeletonization and the method of comparing the skeletons are described. Skeletonization is powerful method for comparing shapes. It reduces

the amount of information with minimum loss and let to use effective methods to extract special points. Methods of Discrete skeletization have higher complexity, in contrast by continuous methods. Discrete skeletons can be discontinuous that complicates the solution of the problem of classification.

Method of gesture classification based on continious skeletons proposed in (Kurakin2012). Authors condsider dynamic gestures in this work. Each frame is described by the position of several special points, that is not enough to detect a large number of static gestures.

In this paper we solve the problem of image and depth maps segmentation in order to select the hand on the frame. Segmentation is the process of dividing image into several segments. The result of image segmentation a set of segments that together cover the whole image. Several algorithms and universal methods is developed for image segmentation. We list the main types of segmentation methods:

- based on clustering
- based on the analysis of the histogram
- based on edge detection
- based on the calculating minimum graph cut
- method of watershed and others.

Most of them are described in (Shapiro, 2001). Since the general solution for the problem of segmentationtation image does not exist, these methods often have to combine with knowledge of the subject area to effectively solve this problem.

3. PROPOSED METHOD

In this paper we propose a method combined segmentation frame for depth map and images. The primary segmentation is based on the depth map and provides information about the position and fuzzy hand border. Then color image is analyzed to clarify the boundaries. Feature generation for hand shape is based on continuous skeleton. Introduceing of clear interpretational features allows to create classifications based on simple rules.

3.1 Segmentation

The image of segmentation is an unsolved problem in the general case. Usually the problem narrowed to a specific area. In this paper we want to select a hand for further features generation. By itself, the hand has a clear spatial structure and color, but because of the different lighting, color and complex background necessary to use additional tools for segmentation. In this paper, the problem is solved by a combination of segmentation by the depth map and the color image.

The target object of segmentation is hand. Let with the image of I (Figure 1) algorithm is applied to the input depth map D (Figure 2) — matrix of the same size as the I, whose elements are the distance from the camera to the object corresponding pixel in the image I. Assume that the hand is the foreground. This makes it possible to localize the position of the hand using threshold binarization. Because of the large error of the depth map, we use image I to find the exact boundaries.

Combined segmentation algorithm is follows. The first step is the sellection of S_{front} — foreground depth map (Figure 3). For this we introduce the parameter t_f and select those pixels whose depth is different from the minimum no more than t_f:

$$S_{front} = \{s_{ij} = [d_{ij} < \ mind_{ij} + t_f]\}. \quad (1)$$

Apply the morphological erosion operation (Figure 4). Set the resulting region of $S_{erosion}$

$$S_{erosion} = S_{front} \ominus S_0^r, \quad (2)$$

where S_0^r circle with radius r. This operation is necessary because the boundary of the foreground could get background pixels. In the region of $S_{erosion}$ contains only the pixels of the target, so they can be used to calculate the average color of the target avg_color. Then we search the edge of the image by algorithm Canny. The final step is to run the search algorithm BFS on image pixels starting with the field $S_{erosion}$. In BFS used 4-connected system of neighborhood pixels. Stop criterion of the wide search is that pixel belonging to the edge or condition $[\rho(current\_color, avg\_color) > t_c]$, where ρ — is the sum of RGB component distance. All pixels that we visited mark 1 in the matrix S (Figure 5).

Figure 1: Source RGB image

Figure 2: Source depth map

Figure 3: Foreground

Figure 4: Foureground core

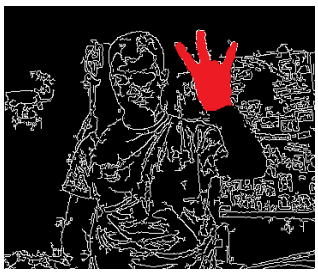

Figure 5: Segmentation result

3.2 Feature space

This paper proposes a method of generating features based on the shape obtained after segmentation image and depth map. Continious skeleton (Mestetskiy2008) is build by the two-dimensional projection of the hand. We consider skeleton as graph with width function on it's edges.

If there is no hole in the figure, the skeletal graph is a tree, otherwise it has cycle. The presence of cycles in a connected graph

can be determined by running the search algorithm in width from each vertex.

With a special terminal branches classification algorithm described in the next section, we can determine the number of fingers and end points of terminal branches that correspond to the fingers. We also search for a branch forming the wrist and its end point, searching a point of the skeleton, in which the width function is maximum. Call it base point. Then we can calculate the angle between the end point of the branch of the wrist and each endpoint branches of fingers relative to a bas point. On the Figure 6 highlighted circle by red color correspond to the end points of the fingers, yellow — end points of the wrist, green — base point.

Using binary search method will find the minimum param of regularization in which the skeletal graph is transformed into a chain. This chain is called a main. If none of the branches of the skeleton was not classified as a finger, a good feature that describe the shape is the width of the main chain.

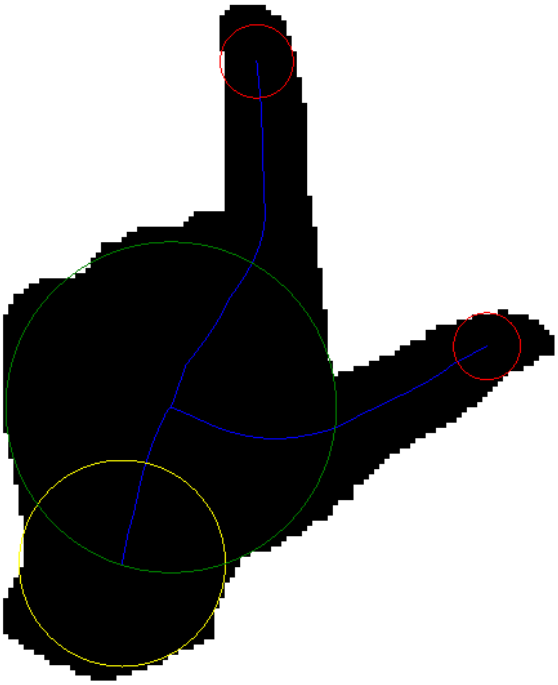

Figure 6: Terminal branches

Thus, the feature space will be as follows

- N — number of fingers,
- M — number of loops in skeletal graph (number of holes in hand shape),
- $A = (\alpha_1, \ldots, \alpha_N)$ — angles between the end point of the wrist and each endpoint of fingers
- $W = (w_1, \ldots, w_{50})$ — value of width function of main chain in uniform 50 points.

3.3 Terminal branch classification

In this section we describe an algorithm classification of terminal branches. Each branch of the skeleton is described by its position and width function. Normalize the length of each branch so that it is equal to 1 and normalize width function relative to the maximum width. If we consider plot or resulting width functions of terminal branches that match fingers, we will notice that they all behave the same way. The functions of the remaining terminal branches are very different from them. Given a training set consisting of a set of functions corresponding to fingers $\{W_i(x)\}_{i=1}^{N}$, we construct two functions

$$W_{max}(x) = \max_i W_i(x) + \epsilon, \tag{3}$$

$$W_{min}(x) = \min_i W_i(x) - \epsilon. \tag{4}$$

If the width function falls between W_{min} and W_{max} we assign a terminal branch to fingers, Figure 7 is an example of the resulting functions, built for 10 points.

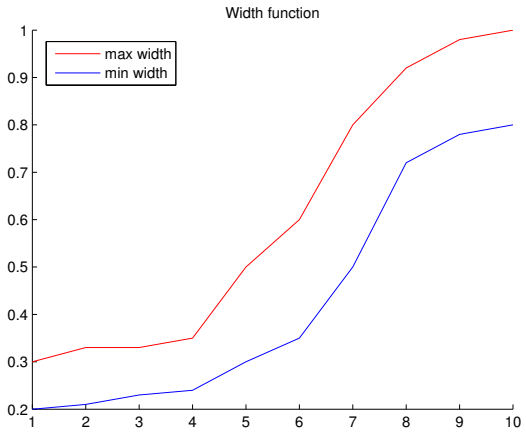

Figure 7: Fingers brunch width function

3.4 Gesture classification

Given a database of gestures G_b, the set of labels Y and the target function $f^* : G \longrightarrow Y \cup \{\lambda\}$, where λ is empty gesture. We need to construct an algorithm for approximation f^*.

In the problem of gesture recognition hypothesis of compact is true, so we can use the classification method of nearest neighbor. For each gesture from the database will hold segmentation and generation features will continue to keep the base in the form of features. In the same type of gesture should be the same number of holes and fingers. If the first two features do not match, then this is exactly different gestures, otherwise you need to compare the function of the width of the main chain and the angles between the fingers and the wrist. Find closest gesture from the database, if the proximity is less than a predetermined threshold, then the gesture is associated with the gesture from the database, otherwise λ (empty gesture).

Introduce $\rho_W(g_1, g_2)$ and $\rho_A(g_1, g_2)$:

$$\rho_W(g_1, g_2) = \sum_{i=1}^{|g_1.W|} (g_1.W_i - g_2.W_i)^2 \tag{5}$$

$$\rho_A(g_1, g_2) = \frac{\sum_{i=1}^{g_1.N} (g_1.A_i - g_2.A_i)^2}{g_1.N} \tag{6}$$

Than the proposed classification method is a method of nearest neighbor with the two measures $\rho_W(g_1, g_2)$ and $\rho_A(g_1, g_2)$, which are used defferent cases. $\rho_W(g_1, g_2)$ is used when number of fingers is equal to 0 ot 1, $\rho_A(g_1, g_2)$ otherwise.

4. EXPERIMENT

4.1 Gesture base

To check the quality of the resulting methods we formed the gesture base. There are involved 10 persons. Each was asked to show 27 sign of the ASL dactyl alphabet, each gesture has 10 frame. Thus, the set of gestures consists of a 2700 gestures, which are divided into 27 classes. All gestures were recorded on a webcam and Kinect depth sensor. To check the quality of the segmentation problem solution 25 gestures was manually processed and hand was selected. To assess the classification quality of terminal branches was marked 30 frames, which was attended by 100 terminal branches. Examples of the gestures from base are shown in Figure 8.

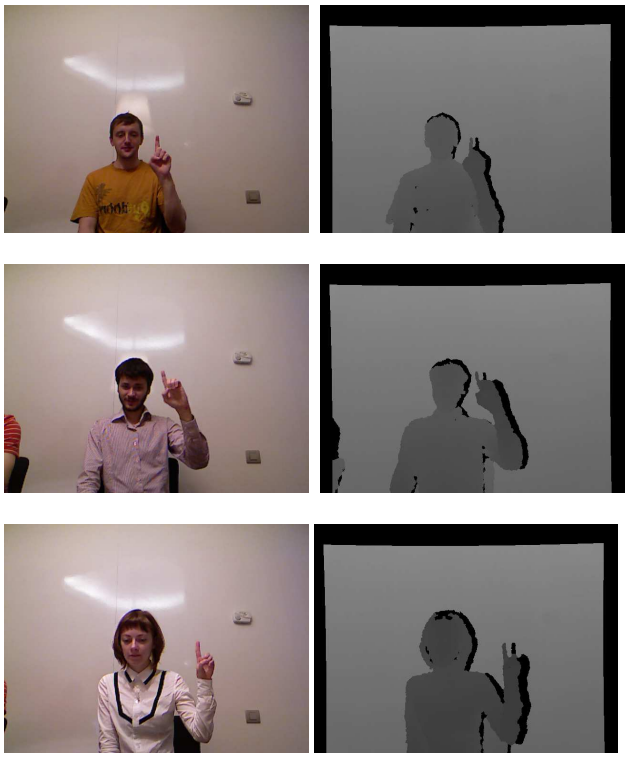

Figure 8: Gesture base, sign D

4.2 Results

Quality of classifying gesures preblem solution is the number of correctly recognized gestures, divided by the sample size. Quality solution to the basic problem of gesture recognition for the training set is 0.94, 0.93 on control (Table 1). The results show a good generalization performance of the method. Average time recognition of gesture is 0.2 seconds, which allow to solve the problem in real time. As a rule, the process of displaying a single gesture extended in time, so you can use several consecutive frames and classify gesture by voted majority. For a group of 10 gestures, the quality increases to 0.97. In this experiment, a sign was based on one representative per class. Increasing the base will improve the quality of recognition.

Dataset	Quality	Empty class	Wrong class
train	0.941	0.023	0.036
test	0.929	0.024	0.047

Table 1: Quality of gestures classification

5. CONCLUSIONS

In this paper we propose a method for gesture recognition of sign language. For gesture recognition is necessary to set the base. For each gesture is sufficient presence of the one example in the database. Described method is very efficient in terms of execution time and resources requiriment. It can recognize up to 5 gestures per second. Base stored as attributes and it saves memory. Thus, the system can be used in real-time. The method consists of several independent algorithms that can be optimized separately. This allows us to find the optimal set of parameters which give us maximum of proposed functional quality. Method proved it's quality on collected gesture base. The method can be used in practice for any of hand gestures for which the condition of compact is true.

6. ACKNOWLEDGEMENTS

This work was supported by Russian Foundation for Basic Research, research projects 14-01-00716, 14-07-00965 and 12-07-92695-IND_a.

REFERENCES

Beristain A., Grana M., 2010. A stable skeletonization for tabletop gesture recognition. Computational Science and Its Applications ICCSA.

Kurakin A., Zhang Z., Liu Z., 2012. A real-time system for dynamic hand gesture recognition with a depth sensor, Proc. of EUSIPCO.

Malik S.,2003. Real-time hand tracking and finger tracking for interaction. CSC2503F Project Report.

Mestetskiy L., and Semenov A., 2008. Binary image skeleton - continuous approach. VISAPP (1), pages 251258. INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

Shapiro L. G., Stockman G. C.,2001. Computer Vision. New Jersey, Prentice-Hall.

Stenger B., Thayananthan A., 2006. Model-based hand tracking using a hierarchical. IEEE Transactions on pattern analysis and machine intlelligence, vol. 28, no. 9.

Suryanarayan P., Subramanian A., Mandalapu D., 2010. Dynamic hand pose recognition using depth data. 20th International Conf. on Pattern Recognition (ICPR).

Wang R. Y., Popovic J., 2009. Real-time hand-tracking with a color glove. ACM Transactions on Graphics. Vol. 28, no. 3.