AUTOMATIC STATIC HAND GESTURE RECOGNITION USING TOF CAMERAS

Serban Oprisescu[1], Christoph Rasche[1], Bochao Su[2]

[1]LAPI, University Politehnica from Bucuresti, Romania, soprisescu@imag.pub.ro
[2]Harbin Institute of Technology, Harbin, HeiLongJiang Province, P. R. China, subc97@gmail.com

ABSTRACT

This paper presents an automatic algorithm for static hand gesture recognition relying on both depth and intensity information provided by a time-of-flight (ToF) camera. The combined depth and intensity information facilitates the segmentation process, even in the presence of a cluttered background (2 misses out of 450 images). Gesture classification is based on a decision tree using structural descriptions of partitioned contour segments. Classification was tested on 9 different gestures. The final mean recognition rate is satisfactory, of about 93.3%.

Index Terms— ToF camera, static hand gesture recognition, curve partitioning

1. INTRODUCTION

Human-Computer Interaction (HCI) has evolved rapidly in the recent years toward a more intuitive, natural and suitable way of communication, adapted to human behavior. Hand gestures are natural ways for humans to communicate commands, actions, sign language etc. The research field of gesture recognition is expanding due to the development of intelligent devices such as smartphones, interactive displays, robots etc. Gesture recognition makes HCI easier in a wide range of applications, such as sign language recognition for hearing impaired, monitoring patients, navigation systems, distance learning etc.

Gestures can be static (for instance a certain hand pose or finger configuration), dynamic (hand motion on a certain trajectory), or a combination of two [1]. There exists different tools for gesture recognition [1], belonging to the field of statistical modeling, computer vision, pattern recognition, image processing etc. In the following we will focus on static hand gesture recognition.

In [2] a hand gesture recognition system is proposed to solve the problem of Chinese sign language alphabet. As in most methods relying on natural RGB images, the authors use a skin color model to segment the hand area, and then apply a Locally linear embedding (LLE) method in order to reduce the dimensionality of the data while preserving the spatial relation between neighboring pixels. The same idea of dimensionality reduction was applied in [3] where, the authors proposed a nonlinear manifold learning and representation approach based on an Isometric Self-Organizing Map (ISOSOM), and a hierarchical version of it, H-ISOSOM for hand pose estimation. In [4] the LLE algorithm is modified to a new distributed locally linear embedding (DLLE) method, used to discover the inherent properties of the input data. In [5] a neural network is trained to recognize hand gestures transformed in vectors by applying orientation histograms on the hand contours. Hidden Markov Models (HMM) represent one of the most frequent techniques used for gesture recognition [6], and in [7] HMMs are applied on an observation sequence consisting of angles extracted from the contour of each gesture. Recently, in [8] an algorithm based on skin color and angle, combined with Hu invariant moments and followed by an Euclidean distance template matching technique is used for hand gesture classification.

One of the main weaknesses of gesture recognition from color images is the low reliability of the segmentation process, if the background has color properties similar to the skin. This weakness is overcome by the use of the new 3D range sensors, such as the time-of-flight (ToF) camera, whose 3D information of the scene makes the hand region segmentation much easier. In [9] a range sensor is used and several classification algorithms are applied. In [10] a PMD ToF camera is used to obtain a reliable hand segmentation and then a chamfer distance matching is employed.

In this paper we use the depth and intensity information from time-of-flight (ToF) camera to obtain a reliable hand segmentation by means of a modified region growing algorithm. The goal is to discriminate between nine static hand gestures, whereby a structural description approach is used as it represents a robust method for recognition that does not necessarily rely on a closed contour.

The paper is organized as follows: §2 reviews the contour partitioning method, §3 describes the proposed gesture recognition algorithm and depicts the obtained results, §4 contains conclusions.

2. CONTOUR PARTITIONING

Hand silhouettes were analyzed using a local/global analysis [11], which we can only sketch here due to its complexity. The analysis is most similar to the curvature-scale space but

is crucially different that 1) it does not perform any low-pass filtering of the curve, 2) amplitudes are measured (instead of curvatures). More specifically, the silhouette (or any open or closed planar curve) is investigated with different window sizes at each point. For a given curve point, a window (local neighborhood) of length ω along the arclength variable v is selected and the amplitude of the selected segment determined. The window is shifted through the curve, and the amplitude is taken for different sizes, hence creating an amplitude-scale space $\mathbf{B}(\omega,v)$. The amplitude is taken only for windows, in which the segment appears as an arc, meaning all segment points need to lie on either side of the segment's chord; otherwise the amplitude is set to 0. Window sizes were generated with an increment of $\sqrt{2}$, starting from 5 pixels to ca. 70% of the silhouette diameter. Then, the amplitude space is searched for consistent arcs, that is arcs that appear also on adjacent scales (for more local or global window sizes). A similar analysis is done for inflexions (instead of arcs). In summary, this process reliably detects arcs (curved or straight) corresponding to human interpretation; it does not correspond to high-curvature detection, but to a detection of segments between high curvatures (see Figure 1). It was shown that this partitioning is relatively homologous across class instances in the MPG7 collection (under review).

3. PROPOSED METHOD AND RESULTS

We used for our experiments the SR-3000 ToF camera, which concomitantly delivers depth and intensity images at 176x144 pixels. Nine different types of static hand gestures were displayed in front of the camera (at a distance ranging from 30cm to 1.5m), and 50 frames per gesture were recorded, whereby the hand was slightly moving introducing therefore some variability in the silhouette (450 test images in total). In Figure 2 are shown distance images for each of the nine types of gestures, and in Figure 3 is an example of variability within the same gesture.

The first step in the image processing chain is hand segmentation, starting with the ToF distance image, as one can see in Figure 4. To obtain robust segmentation, a region growing algorithm was applied: basically, the algorithm starts growing a region with similar properties (gray level, or in our case distance) from an initial point called seed. The seed was chosen to be the nearest point to the ToF camera, assuming that any hand gesture points towards the camera.

The region growing stops at the boundaries of the region (in our case the silhouette borders of the hand image) and these boundaries are detected using three thresholds: the classic region growing threshold which compares the gray level (or distance in our case) of the current point with the gray level of the seed, a local distance threshold within a neighborhood of the current pixel, and an luminance threshold within the intensity ToF image. This intensity threshold is consistent, since the hand, being close to the ToF camera, appears much brighter. We used three thresholds because we preferred, in the worst case obtaining an incomplete hand segmentation, rather than a region expanding into other image parts, like the face, which in some images is close to the hand. The overall successful (complete) hand segmentation rate was about 99.55% (two incomplete segmentations from 450 test images). After

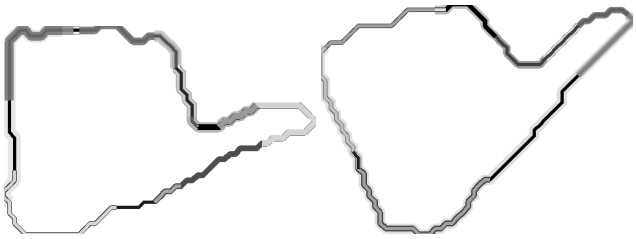

Figure 1. Contour partitioning example (1 pixel wide)

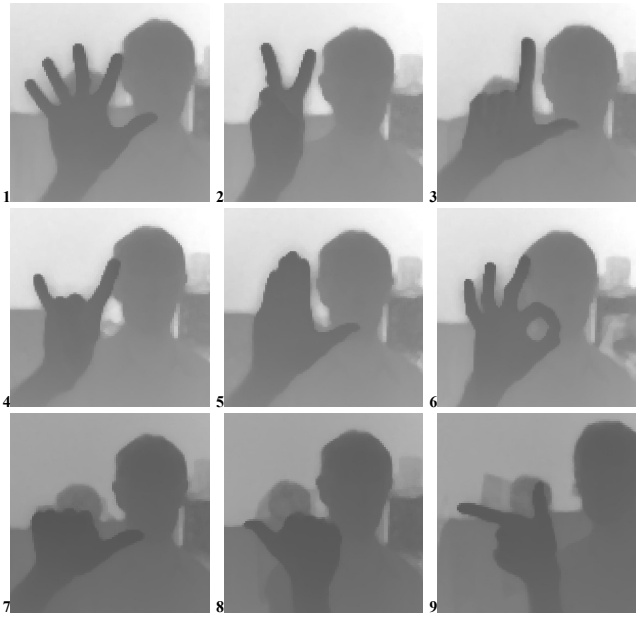

Figure 2. The nine static hand gestures (1-9 rowwise)

Figure 3. Different hand rotation angles

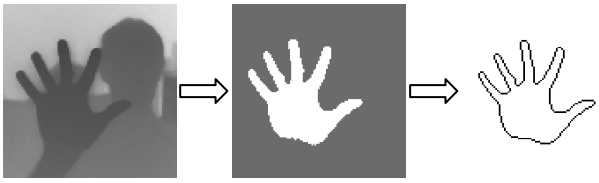

Figure 4. First processing steps: hand segmentation and contour extraction

segmentation, contour extraction was applied. The algorithm was designed such as to obtain closed contours (with no gaps), one pixel thin (some morphological operations were performed). A hand contour was then partitioned as mentioned in §2, which returns the segmented contour along with some parameters extracted from each segment (such as curvature, length, endpoints etc.).

Gesture identification was carried out with a structural description mostly, but also with global descriptors such as the aspect ratio. The identification process is organized as a decision tree, which has the advantage of being fast. Gesture identification started with finger detection, that is fingers that are extended and separated. Fingers were identified with three parameters. 1) high curvature, which corresponded to the maximal amplitude (distance between segment chord and segment points) divided by the image-normalized segment length. 2) high peakness: segment chord length / arc length. 3) convexity, which is determined with reference to the shape center (or centroid; average of all silhouette points). Convexity is determined by determining an angle α between two vectors: one vector points from the peakpoint A_i of the segment (where the amplitude is maximal) to the segment's chord midpoint H_i; the other vector points from H_i to the centroid C. For a finger segment α is smaller than 90°, for an inter-finger segment - a V feature for some gestures - α is larger than 90° (see also Figure 5).

The root node of the decision tree simply counted the number of (separated) fingers, and gestures 1 and 6 can be discriminated by a count of 5 and 3 respectively.

For one-finger gestures (no. 5, 7 and 8) a decision node is introduced, which observes an aspect ratio and the spatial positioning of the finger. Gesture no. 8 can be distinguished from the other two by its *x* coordinate of the peak of the thumb, which is smaller than the centroid's *x* coordinate (see Figure 6) and larger for gestures 5 and 7. In order to decide between the latter two, the aspect ratio of the bounding box (width/height) is used: it is smaller for gesture no. 5 and larger for gesture no. 7.

For two-finger gestures (no. 2, 3, 4 and 9) a decision node is introduced that also analyses structural relations. First of all, by looking at Figures 7 and 8, one notices that the angle between the two fingers is smaller for gestures 2 and 4 than for gestures 3 and 9. Also, the fourth gesture can be easily distinguished from the other three (2, 3 and 9) because the distance between the finger's bases (*L'* in Figure 7) is much bigger than *L*. Within the algorithm, to be scale independent one uses the *L/l* or *L'/l* ratio (*l*=chord length). In order to increase the recognition rate and eliminate any possible confusion between, for instance the gestures 2 and 3, the ratio between the *Y* coordinates of the two finger tips (which has a value of approximately 1 for gesture 2 and much larger for gesture 3) are used. Again, the aspect ratio of the boundary box can be used to discriminate between gesture no. 2 and gestures no. 3 and 9.

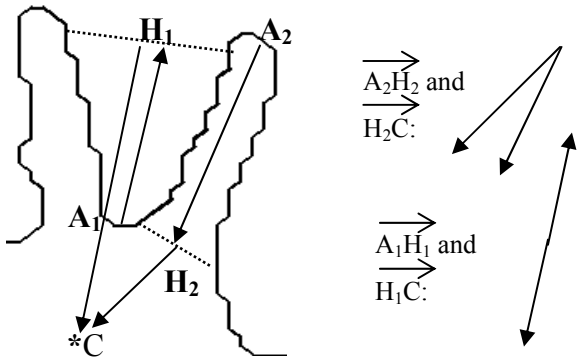

Figure 5. Determining convexity and concavity of finger and inter-finger segments (around A_2 and A_1 resp.)

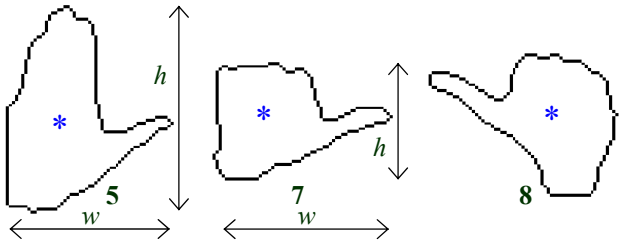

Figure 6. Aspect ratio (w/h) and finger positioning in reference to centroid (*) for gestures no. 5, 7 and 8

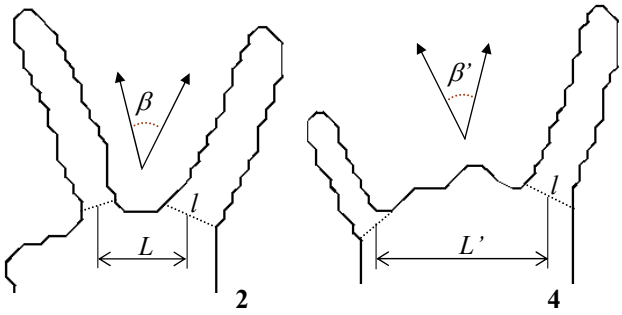

Figure 7. Distinguish between gestures 2 and 4

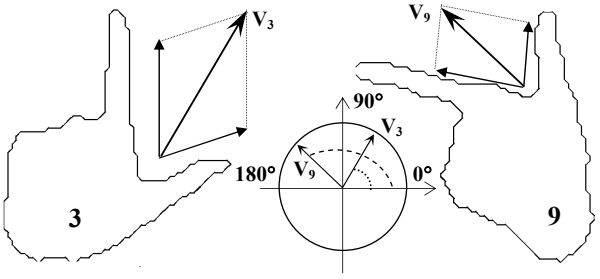

Figure 8. Distinguish between gestures 3 and 9

Another geometric property of the gestures 3 and 9 is, as one can see in Figure 8, the sum of the vectors corresponding to the direction of the fingers, denoted V_3 and V_9. One can easily observe that V_3 belongs to the quarter between 0° - 90° and V_9 belongs to the quarter 90° -

180°. This information permits the distinction between V_3 and V_9. In Table 1 is shown the confusion matrix obtained after processing all 450 frames, 50 frames per gesture. The "?" sign means un-identified gesture. In Table 2 is presented the recognition rate for each individual gesture. The low score obtained by the gesture 4 is caused by the fact that at some rotation angles, the small finger becomes too small to be identified, and the system decides that there is only one finger. Hence, in this case, the confusion with the gesture 5, which has one finger to the right, is obvious. Overall, the mean recognition rate is 93.3%, which is satisfactory. One could increase the recognition rate by considering several consecutive frames and remove the possible outliers.

Table 1 Confusion matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**?**
1	**46**	0	0	0	0	2	0	1	0	1
2	0	**50**	0	0	0	0	0	0	0	0
3	0	0	**48**	0	0	0	1	1	0	0
4	0	0	0	**43**	5	0	0	0	0	2
5	0	0	0	0	**46**	0	4	0	0	0
6	0	3	0	0	0	**45**	0	0	0	2
7	0	0	0	0	2	0	**46**	2	0	0
8	0	0	0	0	0	0	0	**48**	0	2
9	0	0	0	0	1	0	0	0	**48**	1

Table 2 Recognition rate per gesture

Gesture number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Recognition rate (%)	92	100	96	86	92	90	92	96	96

4. CONCLUSIONS

An automatic method for static hand gesture recognition using a ToF camera was presented. Hand segmentation is straightforward using a region-constrained region growing algorithm which considers both depth and intensity image data. Gesture classification occurred with a decision tree exploiting structural descriptions of partitioned segments. The classification is in principal robust to contour fragmentation and is translation invariant, which however was not particularly exploited here. The original method described in [11] offers to the classifier a very good description of the contour structure which enables the subsequent identification of the fingers. If more than 9 gesture classes were used, then the decision tree had to be certainly more complex. However the recognition principle demonstrated here (classification rate of over 93%, which is better than 90% ([5]), 90.6% ([2]) or 91% ([8]), and similar to 93.2 ([4]) or 94.6% ([3])) is certainly promising. Recently we have tested the algorithm also with a Kinect camera, and the results are a bit worse because of the Kinect's depth image instability and poor depth resolution.

5. ACKNOWLEDGMENT

The work has been co-funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

6. REFERENCES

[1] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Trans. on Syst., man, and cybernetics*, Part C: Applications and Reviews, pp. 311-324, vol. 37, no. 3, may 2007.

[2] X. Teng, B. Wu, W. Yu and C. Liu, "A hand gesture recognition system based on local linear embedding," *Journal of Visual Languages & Computing*, vol. 16, no. 5, pp. 442-454, 2005.

[3] H. Guan, *Vision-based 3D hand posture estimation using hierarchical-ISOSOM*, PhD thesis, University of California, Santa Barbara, 2007.

[4] S.S. Ge, Y. Yang, and T.H. Lee, "Hand gesture recognition and tracking based on distributed locally linear embedding," *Image and Vision Computing*, vol. 26, no. 12, pp. 1607-1620, 2008.

[5] T.H.H. Maung, "Real-Time Hand Tracking and Gesture Recognition System Using Neural Networks," *World Academy of Science, Engineering and Technology*, 50, pp. 466-470, 2009.

[6] A. Just, S. Marcel, "A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition," *Comp. Vis. and Image Und.*, Elsevier, 113, pp. 532-543, 2009.

[7] L.R. Vieriu, B. Goras, and L. Goras, "On HMM Static Hand Gesture Recognition," *Proceedings of ISSCS 2011*, Iasi, Romania, pp. 221-224, 2011.

[8] L. Yun, Z. Lifeng, and Z. Shujun, "A Hand Gesture Recognition Method Based on Multi-Feature Fusion and Template Matching," *Procedia Engineering,* vol. 29, pp. 1678-1684, 2012.

[9] S. Malassiotis and M.G. Strintzis, "Real-time hand posture recognition using range data," *Image and Vision Computing,* vol. 26, no. 7, pp. 1027-1037, 2008.

[10] Z. Li, R. Jarvis, "Real time Hand Gesture Recognition using a Range Camera," *Australasian Conf. on Robotics and Automation (ACRA)*, Sydney, Australia, December 2-4, 2009.

[11] C. Rasche, "An Approach to the Parameterization of Structure for Fast Categorization," *International Journal of Computer Vision,* vol. 87, no. 3, pp. 337-356, 2010.