Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 9 (2014) 19 – 24

AASRI Procedia

www.elsevier.com/locate/procedia

2014 AASRI Conference on Circuits and Signal Processing (CSP 2014)

Comparison of SIFT and SURF Methods for Use on Hand Gesture Recognition Based on Depth Map

Peter Sykora, Patrik Kamencay, Robert Hudec*

Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Univerzitna 8215/1, 01026 Zilina Slovakia

Abstract

In this paper a comparison between two popular feature extraction methods is presented. Scale-invariant feature transform (or SIFT) is the first method. The Speeded up robust features (or SURF) is presented as second. These two methods are tested on set of depth maps. Ten defined gestures of left hand are in these depth maps. The Microsoft Kinect camera is used for capturing the images [1]. The Support vector machine (or SVM) is used as classification method. The results are accuracy of SVM prediction on selected images.

Keywords: SIFT; SURF; SVM; Kinect; depth map; hand gesture; recognition

1. Introduction

Gesture recognition is one of the directions in the non verbal machine-human communication. Non verbal communication can be useful in many life situations (e.i. in situations where human can't use speech). Several methods were publicized on topic of gesture recognition [2][3][4]. Many of them focus on feature extraction from color image of hand and classification as next step. For this, it is important to get the most accurate

* Peter Sykora. Tel.: +421-41-513-2238.
E-mail address: peter.sykora@fel.uniza.sk.

doi:10.1016/j.aasri.2014.09.005

feature vectors. The most common used local visual descriptors are SIFT and SURF [5]. The theory about these descriptors as well as experiment description is presented in first chapter of this paper. The results of this experiment are shown next and conclusion at last.

2. Feature extraction and classification methods

In this experiment the database of depth images is created. SIFT and SURF are applied on images of this database. Resulted feature vectors are divided in to train set and test set. For creation of SVM model the train set is used in process called training. Next the test set is used for prediction. Results are prediction accuracies of all tested images for SIFT and SURF.

2.1. SIFT algorithm

Scale invariant feature transform is one of the mostly used local visual descriptors. The method works in two steps. Detection of feature point as the first step. Feature description as the second step. At the beginning of procedure the computing of gradient magnitudes and orientations of pixels are computed. This is done in neighborhood of key point by using the scale of the point. This will make choice on what Gaussian kernel will be used for blur the image. Feature vector is combination computed from the orientation of histograms within the sub-regions around the feature point. The feature vector is normalized at least. For more information about SIFT see [6].

2.2. SURF algorithm

Speeded up robust features descriptor detects feature points by using determinant of the Hessian matrix $H(X, \sigma)$ which is defined as follows

$$H(X,\sigma)=\begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{yx}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \tag{1}$$

Where L is the convolution of the Gaussian second order derivation of image at point $X(x,y)$ in scale σ and similarly for L_{xy} and L_{yy}. For the classification the maxim and minim of the function the discriminant value is used. The description starts by constructing the window around detected feature point. Orientation of the window is same as reproducible orientation. From pixel in this region the resulted feature vector is calculated. For more information about SURF see [7].

2.3. SVM

Support vector machine as part of the model based classifiers uses model for prediction. This model is created in procedure called training. Model represents each class as pattern in vector space. Each feature vector is represented as point in feature space. In Fig. 1 only two dimensions are shown for simplicity. Class *A* is represented as pattern of gray dots in fig and class *B* as green dots.

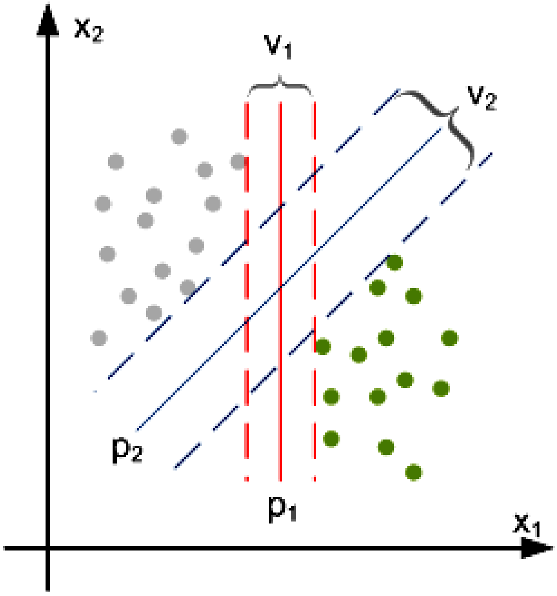

Fig.1. Calculation of best separation.

Separation line is searched in training process. The wider gap between two closes point to line the better train process. It is seen on fig that line p_2 represents better train process as line p_1 because the gap of line p_2 the v_2 is bigger as gap v_1 of line p_1. If two patterns can't be linearly separated, the kernel method is used. This will transform vector in to higher dimension space in which they are separable. This process will allow classification of multiple classes [mata6]. In this experiment the RBF kernel method is used. For more information about SVM see [8].

3. Experimental results

Used input database of depth images and experimental results are described in this chapter. Programming environment Matlab was used to execute this experiment.

3.1. Image database

Capturing the color image of hand, its followed segmentation can take significant calculation time as well as processing power. Some methods for segmentation aim for color of skin, as to detect the region of hand. Results of such process can vary by the light condition or color tone of a particular person. Microsoft Kinect camera has the advantage that it uses infrared spectrum of light. As such it is invariant to light conditions and color of skin. Kinect system can track parts of detected human body.

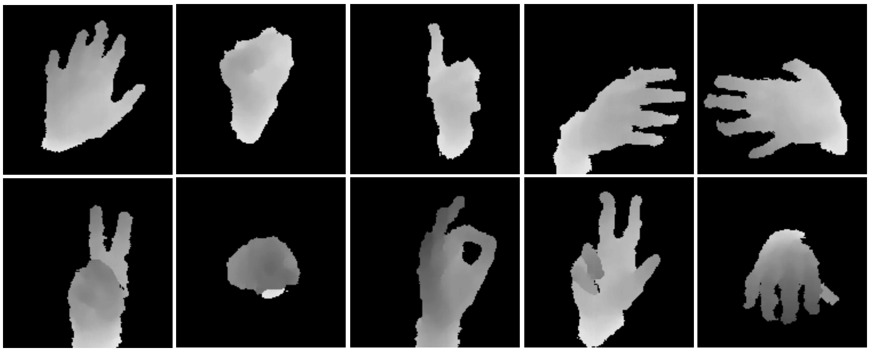

Fig.2. Representation of gestures from cass 1 to 5 (first row) and from 6 to 10 (second row).

Tracked body (or Sceleton) has as its part left and right hand. By locating left hand in the image, the region of hand can be created. This region is defined as 150x150 pixels square centered on the position of hand. This acts as simple position segmentation. The resulted depth image (of resolution 150x150 pixels) go throe threshold filter. Here all pixels with value lover than threshold are set to zero. With this procedure 1500 pictures are captured. There are 150 pictures for 10 classes (see Fig. 2). For the SVM method the database is divided in to 100 pictures per class for train set and 50 pictures per class for test set. This gives overall 1000 train pictures and 500 test pictures.

3.2. Experimental results

Table 1 contains performance matrix for descriptor SIFT and Table 2 contains data for descriptor SURF. Each field in this matrix contains the sum of pictures of class represented by its column number, to be recognized as class represented by row number. For example, in Table 1 for row 8, that is for input images of class 8, two pictures were wrong recognized as pictures of class 1 and forty-eight pictures were recognized as pictures of class 8.

Table 1. Performance matrix of resulted accuracy for SIFT descriptor

Output class/Target class	1	2	3	4	5	6	7	8	9	10
1	24	0	0	0	2	0	2	2	1	0
2	0	41	0	0	0	0	25	0	0	0
3	0	2	50	0	0	0	0	0	0	0
4	1	0	0	50	1	0	0	0	0	12
5	21	2	0	0	45	0	3	0	0	1
6	0	0	0	0	0	47	0	0	1	0
7	0	2	0	0	0	0	20	0	0	2
8	2	3	0	0	1	0	0	48	1	0
9	0	0	0	0	1	3	0	0	47	1
10	2	0	0	0	0	0	0	0	0	34

From results it is clear that the invariance to rotation of SIFT descriptor is a disadvantage here. Some hand gestures, mainly 1 and 5, wave similar shape and occurs as one, only in different orientation. For human mind it is clear that gesture represented by class 1 has different interpretation that of class 5. Another major error occurs for class 7 and 2. On both pictures the shape of hand is not the same but it is very similar shape.

For SURF (table 2) the results are similar to the SIFT method. The error occurs in classes with too similar shape, such as pictures of class 2 and class 7. Another error occurs for classes with the same shape but with rotation as pictures of class 4 and class 10.

Overall accuracy for SIFT is 81.2% and for SURF it is 82.8%. It is clear, that accuracy can be better if non-invariant to rotation feature extraction method was used.

Table 2. Performance matrix of resulted accuracy for SURF descriptor

Output class/Target class	1	2	3	4	5	6	7	8	9	10
1	40	0	0	0	0	0	0	2	4	0
2	0	46	0	0	0	0	27	0	0	1
3	0	0	48	0	0	0	3	0	0	0
4	0	0	0	50	0	0	0	0	0	16
5	9	1	1	0	50	0	8	0	0	3
6	0	0	0	0	0	50	0	0	4	0
7	0	1	0	0	0	0	12	0	0	0
8	1	2	1	0	0	0	0	48	2	0
9	0	0	0	0	0	0	0	0	40	0
10	0	0	0	0	0	0	0	0	0	30

4. Conclusion

In this paper the comparison between two feature extraction methods was presented. SIFT as the first method and SURF method as second. They were applied on set of depth map images of left hand gestures. There were 10 gestures. For capturing these images the Microsoft Kinect camera was used. For image classification the Support vector machine was used. The experimental results are prediction accuracies of SVM method on test set images for each descriptor. From the obtained experimental results is evident that best result using SURF method with accuracy of 82.8% was achieved. Two images capturing the same hand shape, but with different orientation can be interpreted as two gestures. These methods, SIFT and SURF are invariant to orientation and thus they are not suited for recognition system of such gestures.

Comparison of others visual descriptors is essential for finding the best candidate for our real-time gesture recognition system [Sykora2013]. In future work we plan to test these methods on larger test database and modified feature extraction methods (SIFT, SURF) so they will be non-invariant to orientation.

Acknowledgment

This contribution is the result of the project implementation: Project No. 1/0705/13 "Image elements classification for semantic image description", supported by the Slovak Science project Grant Agency.

References

[1] Yujie Shen, Zhonghua Hao, Pengfei Wang, Shiwei Ma, Wanquan Liu, A Novel Human Detection Approach Based on Depth Map via Kinect. Computer Vision and Pattern Recognition Workshops (CVPRW); 2013 IEEE Conference on , vol., no., pp.535,541, 23-28 June 2013.

[2] Panwar M, Hand gesture recognition based on shape parameters. Computing, Communication and Applications (ICCCA); 2012 International Conference on , vol., no., pp.1,6, 22-24 Feb. 2012.

[3] Jalal A, Uddin M Z, Kim T–S, Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. Consumer Electronics; IEEE Transactions on , vol.58, no.3, pp.863,871; August 2012.

[4] Wenjun T, Chengdong W, Shuying Z, Li J, Dynamic hand gesture recognition using motion trajectories and key frames. Advanced Computer Control (ICACC); 2010 2nd International Conference on , vol.3, no., pp.163,167, 27-29 March 2010.

[5] S Matuska, R Hudec, M Benco, M Zachariasova, Opponent colour descriptors in object recognition. 15th International Conference on Research in Telecommunication Technologies; Senec Slovakia; ISBN 978-80-227-4026-5; 11-13 Sep. 2013.

[6] Han X, Wenhao H, Kui Y, Feng W, Real-time scene recognition on embedded system with SIFT keypoints and a new descriptor. Mechatronics and Automation (ICMA); 2013 IEEE International Conference on , vol., no., pp.1317,1324, 4-7 Aug. 2013.

[7] Zhang H. Hu Q, Fast image matching based-on improved SURF algorithm. Electronics; Communications and Control (ICECC), 2011 International Conference on , vol., no., pp.1460,1463, 9-11 Sept. 2011.

[8] Soliman O S, Mahmoud A S, A classification system for remote sensing satellite images using support vector machine with non-linear kernel functions. Informatics and Systems (INFOS), 2012 8th International Conference on , vol., no., pp.BIO-181,BIO-187, 14-16 May 2012.

[9] Sykora P, Hudec R, Benco M, 3D Shape-Motion Detection. TRANSCOM 2013; Zilina; ISBN: 978-80-554-0692-3; pp.111,114; 24-26 June 2013.