OUHANDS database for hand detection and pose recognition

Matti Matilainen, Pekka Sangi, Jukka Holappa and Olli Silvén
Center for Machine Vision and Signal Analysis (CMVS) Oulu, Finland
e-mail: mattimat@ee.oulu.fi, psangi@ee.oulu.fi, jukkaho@ee.oulu.fi, olli@ee.oulu.fi

***Abstract*— In this paper we propose a publicly available static hand pose database called OUHANDS and protocols for training and evaluating hand pose classification and hand detection methods. A comparison between the OUHANDS database and existing databases is given. Baseline results for both of the protocols are presented.**

***Keywords*— Database, Computer vision, Hand pose estimation, Hand detection, Pattern recognition**

I. Introduction

Part of the communication between humans is done via hands and gestures (e.g. the sign languages and the natural way people move their hands while having a verbal conversation). To train a machine capable of understanding these gestures – and ultimately replace or complement other I/O devices – has been the goal of many research projects to date (e.g. [1], [2], [3]).

The key problems when developing these kinds of methods are to locate the hand, classify the hand pose, estimate the hand keypoint locations and track the motion of the hand from the input video data stream.

In order to facilitate the development of solutions to these problems, many databases containing hand images with various data modalities have been released during the past 20 years. Most of them are static hand pose databases. We define a static hand pose database as follows: a database of images where at least one hand is shown in a pose belonging to a finite set of pose classes. The other types of hand databases are hand keypoint databases ([4], [5], [6], [7], [8]) and dynamic hand gesture databases ([9], [10], [11], [12], [13]). In hand keypoint databases there are no hand pose classes, but the locations of the hand keypoints (usually the joints or fingertips) in 2-D or 3-D are provided for each of the images. The evaluation is performed by calculating the mean error distances to the ground truth keypoint locations. In dynamic hand gesture databases the gesture classes are defined as specific changes of the hand keypoint locations and translations of the whole hand in the spatiotemporal domain.

Many of the static hand pose databases released to date lack either in the image data quality, data modalities, annotations or the number of subjects. In this paper we propose a new publicly available static hand pose database called OUHANDS. It was collected in order to have more data publicly available containing any data modalities or annotations a hand pose classification or hand detection researcher might need. The following data modalities are provided in the training part of the database for each of the hand samples: RGB, depth, binary segmentation mask for the hand area, bounding box and orientation normalisation by the annotation of the wrist and the middle finger. The same is true for the testing part, except that no normalisation information is provided. There are ten hand pose classes in the database.

A comparison between OUHANDS and other similar databases is given in Section II. Protocols for evaluating the performance of hand pose classification and detection methods are proposed in Sections III and IV, respectively. The baseline hand pose classfication and detection methods and results are presented in Section V.

II. OUHANDS database

The OUHANDS[1] is a database of static hand pose images and non-hand images (see the Figures 1 and 2), that are captured in a setting where the user is giving gesture commands to a hand-held mobile device. The hand images can be used to train models for hand pose classification and hand detection. The non-hand data is intended to be used only when training and testing hand detection methods.

The database was collected using the Intel RealSense F200 camera, which was held in hand during the capturing process in order to get more variations to the appearances and viewing angles of the backgrounds and the hands in the data. In addition to the RGB data, the depth images were also captured. The depth data was used to produce the binary hand segmentation masks. The segmentation masks were enhanced by hand when necessary. The resolution for all image data modalities is 640 x 480 pixels.

The non-hand samples do not contain segmentation masks, bounding boxes and orientation normalisations. There are 3150 hand samples (2150 are in the train database and 1000 in the test database) and 5288 non-hand samples (3412 are in the train database and 1876 are in the test database) in total in the database.

The size and data modalities available in the OUHANDS database are compared with other freely downloadable databases in Tables 1 and 2. The OUHANDS database compares favourably to the other databases – with the exception of the Marcel database, which contains more data samples –

[1] The database and associated evaluation software can be downloaded from http://www.ouhands.oulu.fi

978-1-4673-8910-5/16/$31.00 ©2016 IEEE

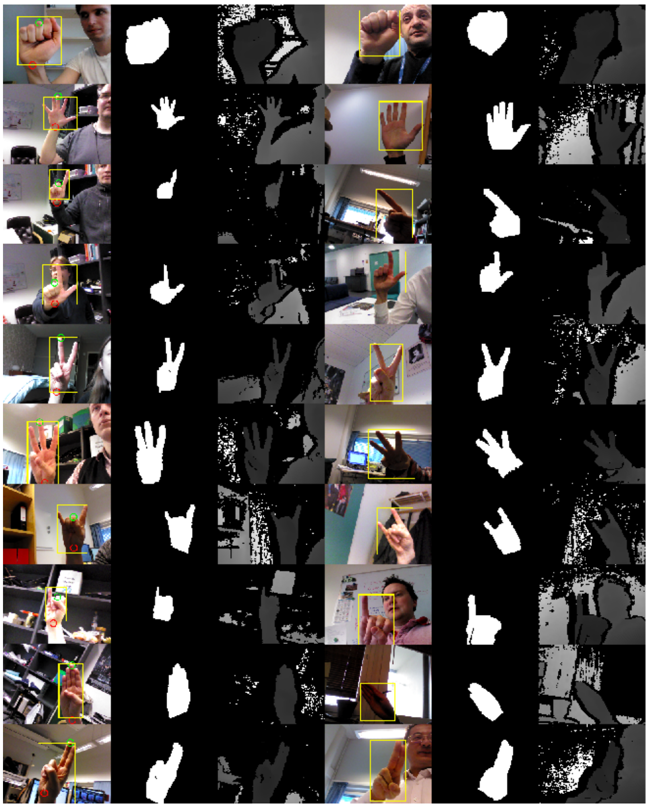

Fig. 1. Samples from the OUHANDS train (the image set on the left) and test (the image set on the right) databases. All the ten different hand poses are performed by different subjects. The first column shows the RGB data that has the bounding box and orientation normalisation information superimposed, the binary hand segmentation mask is in the middle column and the depth data is in the right column. Note that there is no orientation normalisation in the test set.

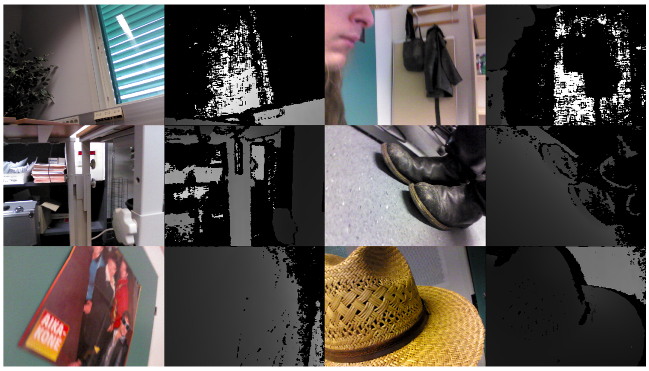

Fig. 2. Non-hand RGB and the corresponding depth map samples from the OUHANDS train database.

when the number of samples and the available data modalities are considered. The majority of the databases published do not provide depth data, segmentation, bounding boxes or any hand keypoint or orientation information. Most of the databases – including OUHANDS – contain ten different hand poses. A vastly superior count of poses is found in the HGR1 and HGR2B: 27 and 32, respectively. As the OUHANDS was intended to be used as a testing tool for HCI methods (e.g. hand pose classification and hand detection), it was decided that a smaller set of hand poses will suffice. A higher number of poses is difficult to utilise in a user interface without making the hand gestures difficult to remember and reproduce.

The number of subjects in the reviewed databases varies from 3 to 40. It can be argued that a higher number is always better, unless one wishes to train a system to recognise only the few specific subjects.

Table 1. A comparison of the database properties related to data quantity. In this Table, c is the number of channels in the images, N_f is the number of image frames, r is the image resolution, N_d is the number of depth frames, r_d is the depth image size.

Name	c	N_f	r	N_d	r_d
HGR1 [18]	3	899	174 x 131	0	N/A
HGR2A [18]	3	85	4672 x 3104	0	N/A
HGR2B [18]	3	574	3264 x 4928	0	N/A
Marcel [19]	3	5819	66 x 76	0	N/A
Marin [20]	3	1400	1280 x 960	1400	640 x 480
Memo [17]	3	1320	640 x 480	1320	320 x 240
OUHANDS	3	3000	640 x 480	3000	640 x 480
Triesch [24]	1	717	128 x 128	0	N/A
Triesch II [25]	3	1143	128 x 128	0	N/A
NTU [23]	3	1000	640 x 480	1000	640 x 480
NUS I [21]	3 (1)	240	160 x 120	0	N/A
NUS II [22]	3	2000	160 x 120	0	N/A

Table 2. A comparison of the database properties not related to data quantity. $N_{classes}$ is the number of hand pose classes, $N_{subjects}$ is the number of subjects and the columns seg, BB and $keypoint$ denote wether or not the segmentation masks of the hands, bounding boxes and keypoint locations are provided, respectively.

Name	$N_{classes}$	$N_{subjects}$	seg	BB	keypoint
HGR1 [18]	27	12	yes	no	yes
HGR2A [18]	10	3	yes	no	yes
HGR2B [18]	32	18	yes	no	yes
Marcel [19]	6	N/A	no	no	no
Marin [20]	10	14	no	no	no
Memo [17]	11	4	no	no	no
OUHANDS	10	23	yes	yes	yes
Triesch [24]	10	24	no	yes	no
Triesch II [25]	12	19	no	yes	no
NTU [23]	10	10	no	no	no
NUS I [21]	10	N/A	no	no	no
NUS II [22]	10	40	no	no	no

III. Hand pose classification protocol

The data for the hand classification protocol is in the training database. There are three image data modalities (RGB, depth and binary segmentation masks) and two plain-text data files (orientation and bounding box information) for each data sample.

Models trained with data other than the training data provided are allowed. The use of external training data must be noted when publishing results on the OUHANDS database.

The final score for the classification task is the number of correctly classified hand pose samples from the test database divided by the total number of hand pose samples, which is 1000. Any or all the data modalities can be used in training and testing. There is no orientation normalisation in the test database. Note that because one of the modalities is the bounding box of the hand, the classification methods do not have to be able to detect the hand. Also, if one wishes to train and test a classification method without using the hand location, it is possible. One should also consider reporting the confusion matrix.

IV. Hand detection protocol

The training hand data and the rules for using any external data for the hand detection protocol are the same as in the hand pose classification protocol with the exception that the detection supplemental data (150 samples where both of the hands of the subject are shown) can be used in training. In addition to the hand data the non-hand data can also be used in training.

The detection ground truth is defined by the bounding box of the hand. A detection is defined the same way as in the Pascal VOC [15] challenge:

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{1}$$

where a_0 is the amount of overlap, B_p is the estimated bounding box and B_{gt} is the ground truth bounding box. The $area(B_p \cap B_{gt})$ is the intersection of the estimated and the ground truth bounding boxes and the $area(B_p \cup B_{gt})$ is the union of the estimated and ground truth bounding boxes.

The overlap criterion used in OUHANDS is 70%. The overlap criterion in the Pascal VOC database was 50% because the ground truth annotations are noisy. As the ground truths in the OUHANDS database are more accurate (see the ground truth bounding boxes in Figure 1), a higher threshold was chosen.

For the hand detection protocol, choosing only one way of evaluating the performance of a method is difficult. We propose that the area under the ROC curve, maximum F1 score (and the precision and recall values that were used to obtain the score) and the precision-recall curve are reported.

V. Baseline methods and results

As a baseline method for classification, we used a technique which uses specific orientation histogram features for neural network based classification. The orientation histogramming method [27] evaluates image gradients, whose orientation is normalized based on the relative angular pixel location within the image patch. The pairs of normalized local orientation and gradient magnitudes provide input to the aggregation step, where several histograms are obtained using a circular ring based spatial binning strategy. In this way, the feature computation is made tolerant to in-plane rotations [27].

The histogram features provide the input to a neural network that has two hidden layers with 960 and 60 neurons, respectively. The neural network was trained to classify five different rotations (-45°, -22.50°, 0°, 22.50°and 45°) for each of the ten hand poses. The amount of training data was increased by scaling and rotating the data. The rotations were done with 11.25°increments in the range of [-56,25°; 56,25°].

The final classification result was determined by selecting the hand pose class with the highest neural network output and the estimated hand rotation information was ignored.

Histogram of Oriented Gradients (HOG) method [14] was used as the baseline detection method. A SVM [26] model was trained for each of the hand pose classes with the Dlib machine learning toolkit [16]. These detectors were applied separately to the validation data and the overlapping detections were filtered by selecting the instance with the highest confidence value.

In this Section the baseline results using the data split in the directory `OUHANDS/data_split_for_the_intermediate_tests/` are presented. In classification only the RGB data converted to gray scale images was used. The detection method used RGB data and depth data. When training the detection method, the detection supplemental data was not used.

A. Baseline classification results

The result for classification is 83.25%. The confusion matrix is as follows (the column index and row index denote the classification result and the target class, respectively):

$$\begin{bmatrix} 40 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 40 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 25 & 2 & 1 & 0 & 8 & 0 & 3 & 1 \\ 2 & 2 & 0 & 33 & 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 2 & 34 & 2 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 5 & 28 & 0 & 0 & 0 & 4 \\ 1 & 1 & 0 & 3 & 0 & 1 & 25 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 37 & 3 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 38 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & 3 & 33 \end{bmatrix}$$

B. Baseline detection results

The precision-recall curves for the detection task when using the RGB and depth are shown in Figure 3. For the detection from the RGB data task, the area under the ROC curve is 0.79 and the maximum F1 score is 0.50 (precision 0.52 and recall 0.48). For the detection from the depth data task, the area under the ROC curve is 0.83 and the maximum F1 score is 0.71 (precision 0.70 and recall 0.72).

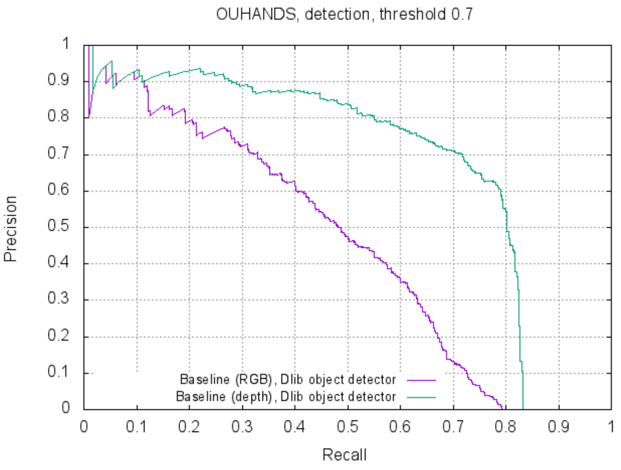

Fig. 3. The precision-recall plot.

The effect of varying the overlap threshold is illustrated in Figure 4.

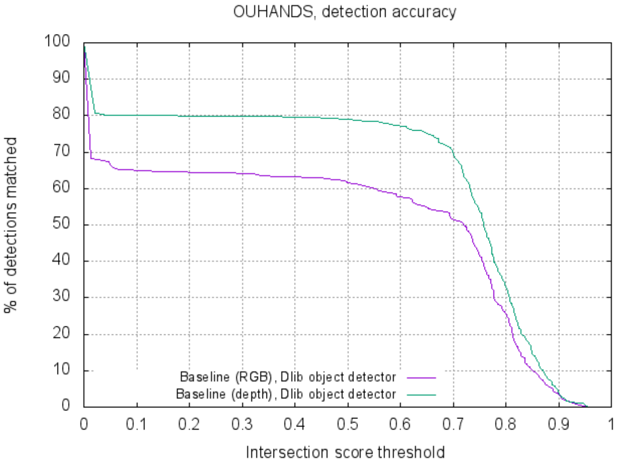

Fig. 4. The effect of varying the overlap threshold value.

VI. Discussion

Often the hand pose databases are captured using fixed camera and lighting. The OUHANDS database proposed in this paper contains images captured in a HCI-like setting: the camera was hand-held and the lighting and backgrounds in the various capturing locations were uncontrolled. This makes the data hard to analyse.

When considering our early experiments (see [27]) lower hand pose classification rates with OUHANDS compared to ones obtained using the Triesch [24] database can be expected; the OUHANDS data is more difficult to classify. Also, the hand detection results produced with the widely used HOG features and SVM classifier are on the level that encourages more testing to be done on the database. A convolutional neural network, trained with external data, is yet to be tested.

Acknowledgment

The authors would like to thank everyone who participated in the data capturing process.

References

[1] P. Barros, S. Magg, C. Weber and S. Wermter. A multichannel convolutional neural network for hand posture recognition. *ICANN*, 403 – 410, 2014.

[2] Y.T. Li and J.P. Wachs. HEGM: A hierarchical elastic graph matching for hand gesture recognition. *Pattern Recognition*, 14, 80 – 88, 2014.

[3] J. Song, G. Sörös, F. Pece, S.R. Fanello, S. Izadi, C. Keskin and O. Hilliges. In-air gestures around unmodified mobile devices. *UIST*, 319 – 329 , 2014.

[4] S. Sridhar, A. Oulasvirta and C. Theobalt. Interactive markerless articulated hand motion tracking using RGB abd depth data. *IEEE International Conference on Computer Vision*, 2013.

[5] A. Wetzler, R. Slossberg and R. Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. *British Machine Vision Conference*, 2015.

[6] D. Tang, H. J. Chang, A. Tejani and T. K. Kim. Latent regression forest: Structured estimation of 3D articulated hand posture. *IEEE Conference on Computer Vision and Pattern Recognition*, 3786 – 3793, 2014.

[7] C. Qian, X. Sun, Y. Wei, X. Tang and J. Sun. Realtime and robust hand tracking from depth. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[8] J. Tompson, M. Stein, Y. Lecun and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics 33*, 2014.

[9] T. K. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8), 1415 – 1428, 2009.

[10] I. Guyon, V. Athitsos, P. Jangyodsuk and H. J. Escalante. The ChaLearn gesture dataset (CGD 2011). *Machine Vision and Applications*, 25(8), 1929 – 1951, 2011.

[11] Z. Jiang, Z. Lin and L. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3), 533 – 547, 2012.

[12] X. Shen, G. Hua, L. Williams and Y. Wu. Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields. *Image and Vision Computing*, 30(3), 227 – 235, 2012.

[13] L. Liu and L. Shao. Learning discriminative representations from RGB-D video data. *International Joint Conference on Artifical Intelligence*, 1493 – 1500, 2013.

[14] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 366 – 371, 2005.

[15] M. Everingham, L. Van Gool, C.L.I. Williams, J. Winn and A. Zisserman. The pascal visual object classes (voc) challenge *International Journal of Computer Vision*, 88(2), 303 – 338, 2010.

[16] D.E. King Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, Volume 10, 1755 – 1758, 2009.

[17] A. Memo, L. Minto and P. Zanuttigh. Exploiting silhouette descriptors and synthetic data for hand gesture recognition. *Eurographics Italian Chapter Conference, The Eurographics Association*, 2015.

[18] M. Kawulok, J. Kawulok and J. Nalepa. Spatial-based skin detection using discriminative skin-presence features. *Pattern Recognition Letters*, 41:3 – 13, 2004.

[19] S. Marcel and O. Bernier. Hand posture recognition in a body-face centered space. *Gesture-Based Communication in Human-Computer Interaction*, 77–100, 1999.

[20] G. Marin, F. Dominio and P Zanuttigh. Hand gesture recognition with Leap Motion and Kinect devices. *IEEE International Conference on Image Processing (ICIP)*, 1565 – 1569, 2014.

[21] P.P. Kumar, P. Vadakkepat and A.P. Loh. Hand posture and face recognition using a fuzzy-rough approach. *International Journal of Humanoid Robotics*, 7(3): 331 – 356, 2010.

[22] P.K. Pisharady, P. Vadakkepat and A.P. Loh. Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101(3) 403 – 419, 2013.

[23] Z. Ren, J. Yuan and Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. *Proceedings of the 19th ACM International conference on Multimedia*, 1093 – 1096, 2011.

[24] J. Triesch and C. Malsburg. Robust classification of hand postures against complex backgrounds. *International conference on Automatic Face and Gesture Recognition*, 170 – 175, 1996.

[25] J. Triesch and C. Malsburg. A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12): 1449 – 1453, 2001.

[26] D. E. King. Max-Margin object detection. *CoRR* 2015.

[27] P. Sangi, M. Matilainen and O. Silvén. Rotation tolerant hand pose recognition using aggregation of gradient orientations. *International Conference on Image Analysis and Recognition*, LNCS 9730, pp. 257 – 267, 2016.