2016 14th International Conference on Control, Automation, Robotics & Vision
Phuket, Thailand, 13-15th November 2016 (ICARCV 2016)

Su23.2

A Time-of-Flight-Based Hand Posture Database for Human-Machine Interaction

Thomas Kopinski
ENSTA ParisTech
858 Blvd des Maréchaux
91762 Palaiseau, France
thomas.kopinski@ensta-paristech.fr

Alexander Gepperth
ENSTA ParisTech
858 Blvd des Maréchaux
91762 Palaiseau, France
alexander.gepperth@ensta-paristech.fr

Uwe Handmann
Computer Science Institute
Hochschule Ruhr West
Lützowstrasse 5, 46236 Bottrop
thomas.kopinski@hs-rw.de

***Abstract*—We present a publicly available benchmark database for the problem of hand posture recognition from noisy depth data and fused RGB-D data obtained from low-cost time-of-flight (ToF) sensors. The database is the most extensive database of this kind containing over a million data samples (point clouds) recorded from 35 different individuals for ten different static hand postures. This captures a great amount of variance, due to person-related factors, but also scaling, translation and rotation are explicitly represented. Benchmark results achieved with a standard classification algorithm are computed by cross-validation both over samples and persons, the latter implying training on all persons but one and testing on the remaining one. An important result using this database is that cross-validation performance over samples (which is the standard procedure in machine learning) is systematically higher than cross-validation performance over persons, which is to our mind the true application-relevant measure of generalization performance.**

I. INTRODUCTION

Recognizing (static) hand postures or (dynamic) gestures from sensor data is a popular research field due to its many areas of application in Human-Machine Interaction (HMI). Interpreting user input in a non-immersive way not only allows for intuitive interaction techniques, but moreover overcomes limits presented by the prevalent technologies. Application scenarios where the user is unable to interact with capacitive touch screen technology due to her/him wearing gloves, such as a sterile operation room, become feasible with contactless gesture interaction. In order to achieve such a means of communication one usually relies on either a model-driven approach, where an underlying hand model is assumed as the basis for the developed algorithms, or an appearance-based approach. The latter relies on many data samples optimally taken from many different sources as to capture variance and the many possible exceptional cases. For the specific case of hands this variance in appearance can occur due to many reasons such as difference in hand size, various positioning of the fingers/hand/person and, above all, the many different ways in which humans express one and the same posture. In order to be able to develop robust software capable of reliably interpreting all these various cases, large databases form the backbone of these kinds of systems. In this contribution we present the REHAP (Recognition of Hand Postures) database, comprising 10 different static hand postures (cf. Figure 1) with over 1 million data samples taken from 35 individuals and demonstrate its applicability in a specific scenario[1]. REHAP comes in two separate sets: REHAP-1 consists of 600.000 samples recorded solely with a ToF sensor from 20 persons while REHAP-2 is recorded with a calibrated ToF and RGB sensor comprising 450.000 samples from further 15 persons. Not only is this the only publicly available dataset of such magnitude, it moreover allows for applicability in any kind of indoor and outdoor scenario, simultaneously capturing the variance in rotation, translation and most importantly scaling - a crucial issue in the field of hand posture recognition. This contribution is laid out as follows: Section II gives an overview over the most important related work in this field, explaining the novelties coming with our contribution. Section III briefly describes the hardware employed for the recording of the database followed by the description of the database itself (Section IV). Section V describes a possible application scenario within the field of Automotive HMI and gives a brief outlook on further fields of application. Section VI shows the performance of a standard classification technique to outline the idea of the challenges and the possibilities of this database. We conclude with a summary in Section VIII coupled with the most significant insights as well as future work to be conducted in this area.

II. RELATED WORK

Over the course of the last few years the body of work on hand gesture/posture recognition has increased significantly. Therefore the number of publicly available datasets has also grown. One can distinguish between 2D [1], [2], [3] and 3D datasets, the former consisting of image and/or video data recorded by RGB cameras and the latter consisting of three-dimensional information, or possible hybrid sets. Ruffieux et al. [4] provide an extensive overview of the currently publicly available datasets.

Large-scale datasets recorded from ToF sensors, however, are sparse. The Dexter 1 dataset presented by Sridhar et al. [5] focuses mainly on RGB data however also contains depth information coming from a Kinect as well as the Creative Gesture Camera (CGC) which provides ToF-capabilities. While the set contains seven dynamic gestures performed repeatedly by

[1]download link: www.gepperth.net/alexander/postures.html

978-1-5090-3549-6/16/$31.00 ©2016 IEEE

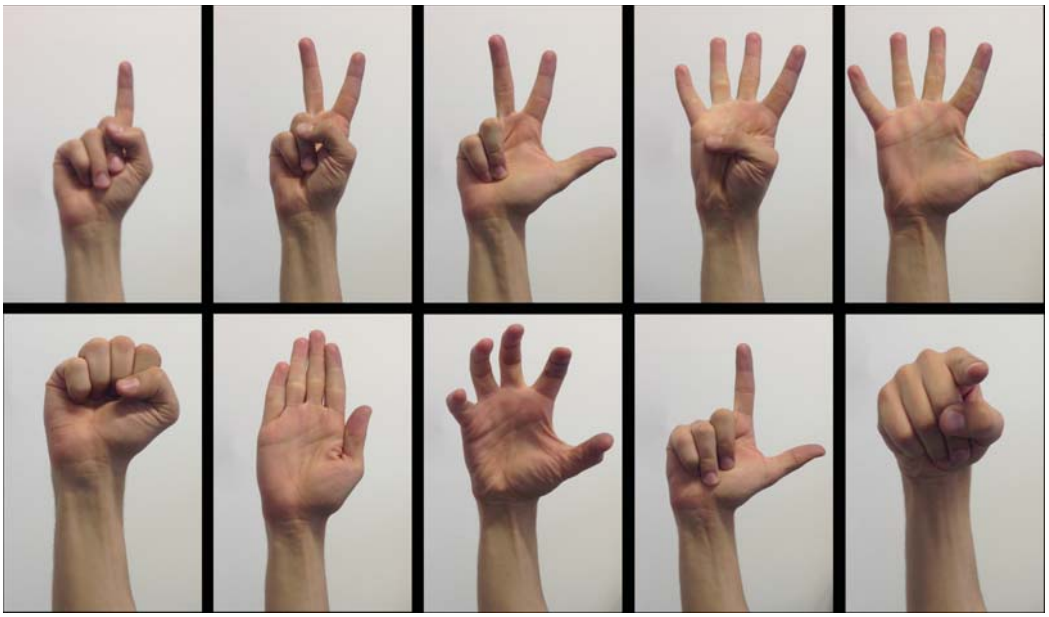

Fig. 1. The hand posture database. From left to right, top to bottom: *ONE*, *TWO*, *THREE*, *FOUR*, *FIVE*, *FIST*, *FLAT*, *GRAB*, *PINCH*, *POINT*

one person, this does not exhaustively address the main issues of data variance, translation, rotation and most importantly scaling.

Simion et al. [6] present a ToF-based 3D database recorded with a PMD[vision]®CamCube 3.0. It contains 6 different gestures coming from 10 different subjects, however it currently does not seem to be publicly accessible. Datasets recorded with the Kinect sensor often comprise a reasonable number of classes, however are taken from few subjects only and do not contain many samples. The ChAirGest dataset [7] contains 1200 samples from 10 different subjects. It combines the Kinect sensor with the XSens system to precisely record the user's dynamic hand movement. The SKIG [8] dataset contains 2160 hand gesture sequences recorded with a Kinect (1080 RGB sequences and 1080 depth sequences) and collected from 6 subjects which makes in-depth algorithm validation difficult from the viewpoint of sample variance. The ICVL Hand Posture Data presented by Tang et al. [9] set contains samples taken with the CGC describing the joint locations of the participants' hand. The Microsoft Research hand tracking dataset presented by Qian et al. [10] was also recorded with the CGC and contains 2400 samples from six persons labeled manually. Supancic et al. [11] present a dataset of various hand postures in everyday scenery holding different objects. Snapshots were taken with the CGC and the dataset aims at creating a standard format for evaluating algorithms for hand detection in difficult scenery. The NYU dataset [12] contains 72757 training set samples and 8252 test set samples with RGB-D data taken from 3 Kinect sensors (1 front view, 2 side views). Sharp et al. [13] present the FingerPaint dataset containing a simultaneously captured video of painted hands using both a prototype ToF sensor and a standard RGB camera. The HandNet dataset presented by Wetzler et al. [14] comprises a training set of 200.000+ samples of hand postures, a test set of 10.000 samples and further validation data of 2700+ samples. Each sample was taken with the Realsense RGB-D camera and shows the 6D postures of the hand as well as the position and orientation of each fingertip.

Our contribution differs in many aspects from publicly available datasets, most notably in that there exists no publicly available set of comparable size and diversity, as our dataset comprises $>$1million samples taken from 35 different individuals. Moreover, using hardware with ToF technology allows to test the developed algorithms regarding robustness vs. various kinds of illumination interferences. Furthermore, we provide data coming from a small, low-cost ToF sensor in order to support flexible setups in any kind of environment. As any large-scale solution will be forced to work with cheap hardware (with higher noise on depth measurements), our dataset models any such situation very closely and is thus well suited for benchmarking realistic applications and products. Moreover, by establishing two separate datasets, we provide the possibility of directly comparing the same gesture set recorded by two sensors of different depth resolution. Any algorithm developed can be benchmarked with respect to its versatility regarding a multitude of parameters: scaling, rotation, translation, person- and sensor-related and depth vs color dependence.

III. The Hardware

REHAP-1 was recorded with the Camboard Nano. It is comparatively small in size (37 x 30 x 25 mm) and able to capture depth data with up to 90 fps with a resolution of 160×120. Its high frame rate suppresses motion artifacts caused by rapidly moving objects. This is of importance during the recording of the database as the participants were asked to move their hand and change their hand posture in order to induce variance into the data. Moreover, its small dimensions make it possible to quickly integrate the sensor into any kind of environment as part of a demonstrator system (cf. Section V). It furthermore has an integrated chip for the suppression of background illumination (SBI) allowing for flexible use in various lighting environments. REHAP-2 was recorded with the CGC which provides twice the lateral resolution in depth (320×160) while additionally recording the RGB information of the environment. RGB pixels can be mapped onto the depth pixels (voxels) resulting in an RGB-D point cloud. The depth sensor of each camera operates with the ToF-principle: The wavelength of light emitted from the infrared chip is modulated with a fixed frequency making precise disambiguation possible against any other kind of light source, therefore allowing for robust data samples to be recorded in real-time for indoor as well as outdoor applications.

Depending on the distance of the object, the precision of the pixel-wise measurements suffers. Measurement errors vary between 1cm-3cm if the object is several meters away and additionally has a small reflection coefficient (i.e. absorbs most of the light), nonetheless the reflection coefficient of hands typically yields very satisfactory measurements. Uncropped point clouds result in 19.200 and 38.400 voxels respectively for REHAP-1 and REHAP-2.

The complete scene with the RGB values mapped onto the voxels is depicted in Figure 2, before any form of hand-background segmentation is applied.

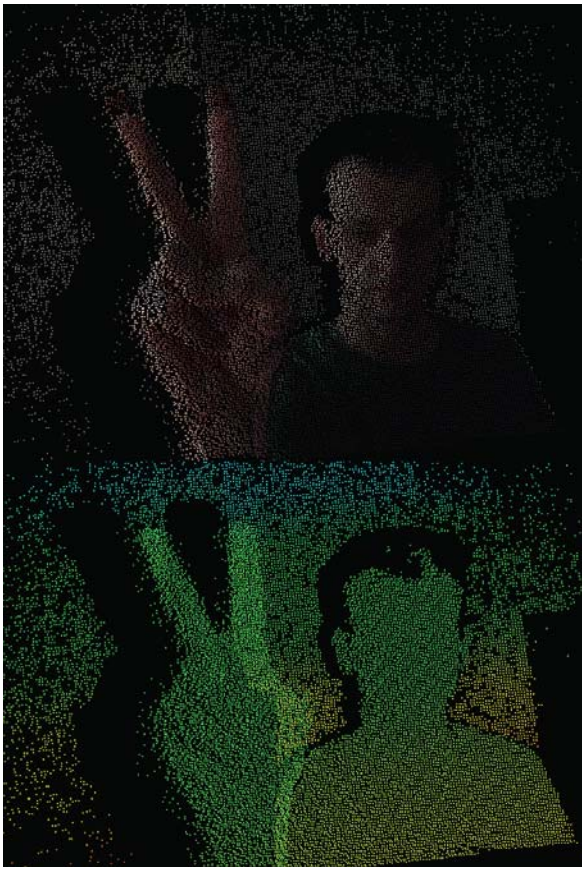

Fig. 2. Full scene recording - RGB and depth info merged (top) and depth info only (color coded, bottom): Depth thresholding allows for rapid segmentation of the hand-arm region from the background. Cropping irrelevant arm parts is achieved via PCA-segmentation.

IV. The database

When referring to the databases, REHAP-1 comprises depth information obtained from recordings with the Camboard nano while REHAP-2 also carries RGB-information mapped onto the voxels obtained from the CGC (cf. Figure 3 showing this for a sample gesture). Data is recorded from 35 different individuals, nevertheless both sets contain the same gesture set and are labeled accordingly. Currently 20 persons are contained in REHAP-1 yielding 600.000 data samples while 15 persons are contained in REHAP-2 consequently yielding 450.000 data samples, however we aim at complementing the second set with 5 more persons to be able to better compare the developed algorithms.

To only capture the relevant data points which are part of the user's right hand, distance thresholding is introduced during the recording. Points recorded by the sensor are simply cropped if above a certain threshold value Θ. Furthermore, the recording takes place in a predefined Volume of Interest (VOI) to ignore irrelevant data points to the sides of the user's hand. The resulting data is denoted a point cloud (PC) of a posture and is saved in the point cloud data format (.pcd file format), carrying the information about the data and relevant meta information.

Our database comprises 10 different static hand postures as recorded and described in Sec. III. The individual postures are denoted *ONE*, *TWO*, *THREE*, *FOUR*, *FIVE*, *FIST*, *FLAT*, *GRAB*, *PINCH*, *POINT* (as shown in Figure 1). These hand postures were chosen with respect to a trade-off between meaningfulness and complexity (in terms of disambiguation). Regarding the meaning of the postures, all of them can be facilitated to represent typical functions useful in various HMI scenarios. Pointing, e.g. can be used to direct a robot or a drone to a certain point of interest. The flat hand typically denotes the halt of a system while grabbing is commonly employed in VR environments to pick up and move an object. Counting from one to five, as indicated by the number of fingers, are very generic gestures applicable to various scenarios like selecting channels or levels. The difficulty in disambiguation results from the fact that the difference between some postures is defined by one finger only (e.g., *ONE* vs *TWO*) which, depending on the distance to the sensors, is equivalent to as few as 20-40 voxels.

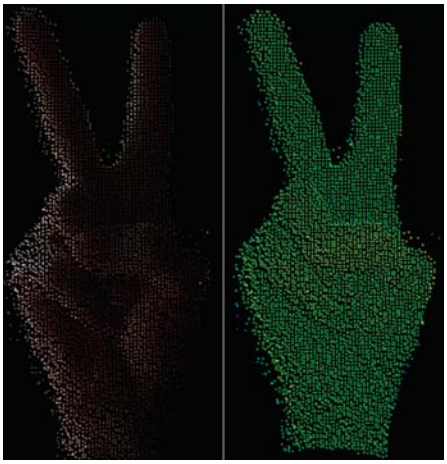

Fig. 3. Depth recording of hand posture *TWO* (right) and with the RGB information mapped onto the depth pixels (left).

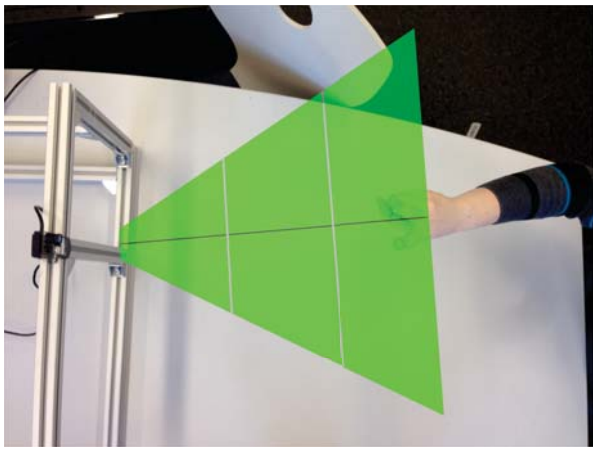

Fig. 4. Setup for the recording with the Camboard nano. The three zones of the recording: Near, intermediate and far.

The sensor is mounted in front of the user recording the nearby environment from an orthogonal angle. Each posture is performed and recorded 3000 times. In order to induce some variance into the data, each participant is asked to translate and rotate her/his hand during the recording. Furthermore, the recording area is divided into three zones: near (15-30cm), intermediate (30-45cm) and far (45-60cm) with respect to the distance between sensor and hand (cf. Figure 4).

The result of such a recording can be seen in Figure 5. The resulting PC is depicted for two different snapshots in subsequent movements (top vs. bottom) of the same participant from two different angles (left vs. right). Points closer to the sensor are depicted in yellow color, points further distant in a dark green color. Depending on the angle the user postures her/his hand toward the sensor, more or less light is reflected back and hence the precision of the measurement suffers. Another possible source for noise is the fact that depth measurement relies on the amount of light reflected from the object, however too much light reflected over-saturates the measurement. This is visible by the amount of noise (or

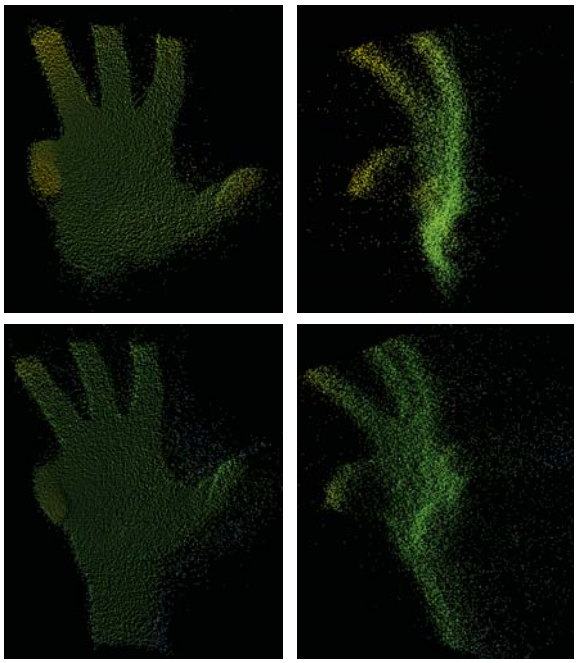

Fig. 5. Sample recording of a hand posture. Top row: The same posture from the front view (left) and the side view (right). Bottom row: The same hand posture in a subsequent state taken after the snapshot in the top row (same angles). Noise and outliers resulting from errors in measurement are clearly visible when seen from the side view (right column).

outliers) existent in the image. In the upper row, the user poses in a rather orthogonal angle towards the sensor, therefore there are less outliers visible towards the edges of the object. As compared to the bottom row, more outliers are recognizable as can be seen in the front view (left) and the side view (right) of the same posture. Dealing with noise is an important factor for the task of hand posture recognition in particular as depth sensors typically have a lower resolution than RGB cameras and therefore data samples suffering from much noise tend to strongly impede the employed algorithms. Consequently, no filtering or noise reduction techniques have been utilized to remove said outliers. However, due to the movement performed by each individual during the recording, the amount of data points belonging to the forearm differs strongly as can be seen in Figure 5 (top left vs. bottom left). Data points belonging to the forearm carry no information necessary to distinguish any of the posture in the database therefore we employed a cropping algorithm relying on a Principal Components Analysis (PCA) of the hand-arm object. Automatically removing most of the forearm results in a smaller first principal component, and more relevant information included in each sample, leaving only the palm and the fingers.

The following chapters demonstrate a sample infotainment application as well as the experiments and results of a standard pattern recognition algorithm. This is included as a performance baseline and to establish a well-defined experimental procedure, in order to allow other algorithms to be compared meaningfully.

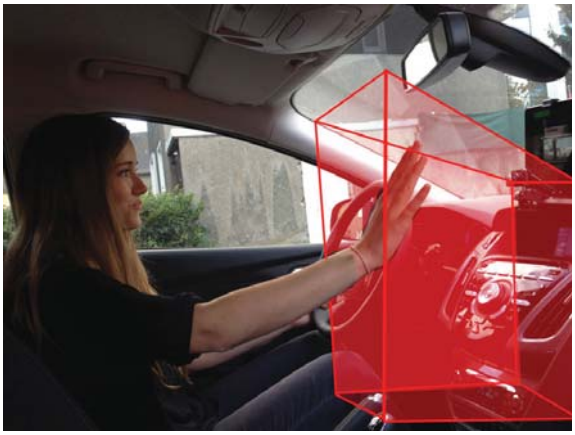

Fig. 6. Demo of our system integrated into the vehicle interior for infotainment control.

Part.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MSE	13,2	46,7	26,8	39,7	14,9	35,4	37,7	30,3	11,1	26,5
Part.	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
MSE	10,4	22,6	3,5	25,8	6,3	17,3	3,9	13,6	23,6	6,1

TABLE I
GENERALIZATION PERFORMANCE OF THE MLP ON REHAP-1.

	0	1	2	3	4	5	6	7	8	9	score
0	2717	56	10	13	4	36	1	4	74	39	0.92
1	74	2748	104	45	1	20	0	10	29	19	0.90
2	23	104	2767	62	29	4	9	20	60	4	0.90
3	9	20	50	2715	120	28	52	41	19	15	0.88
4	6	4	28	97	2620	18	81	86	21	14	0.88
5	25	8	6	28	5	2903	24	8	15	18	0.95
6	9	3	7	55	56	30	2751	9	1	3	0.94
7	1	10	13	40	52	10	7	2788	8	11	0.95
8	78	29	101	25	24	23	6	24	2638	30	0.89
9	47	21	12	11	0	21	5	2	26	2843	0.95

TABLE II
CONFUSION MATRIX FOR THE 30,000 RANDOMLY SELECTED SAMPLES

V. DEMO SETUP

Being able to interpret hand postures in a non-immersive way yields many possible fields of application, within a broad range of scenarios such as healthcare (sterile operating room), home entertainment control, robotics (e.g. drone control) or any kind of HMI scenario. Within an automotive environment the most obvious scenario is for infotainment control (cf. [15]). Figure 6 shows a demo setup of an infotainment system realized on a tablet.

The ToF-sensor is mounted in the vehicle interior on the front console recording the VOI depicted in red. The VOI is defined to allow for full freedom of expression in a similar manner to the database recording. Data is cropped and processed as described in Section IV by a standard laptop. This is one of the many possible scenarios becoming feasible due to the nature of our setup, hence this demonstrator can easily be setup into any of the aforementioned scenarios allowing for a quick analysis of e.g. any HMI-related questions. The following section describes the design of the hand posture recognition module and the most important results of a number of experiments conducted on our database.

VI. EXPERIMENTS AND RESULTS

We ran a number of interesting experiments on both datasets, however, due to space-related reasons, we confine ourselves to the most important ones and outline the chosen parameters only briefly. Due to the multiclass nature of the problem, the number of persons and data present in the dataset, we opt in favor of multilayer perceptrons (MLPs) over support vector machines (SVMs) as our tests have shown this reduces training time drastically and allows for a single model to be utilized instead of training $\frac{n(n-1)}{2}$ SVMs. We utilized the FANN library [16] to realize the training and test runs for a and deal with the problem of multiclass classification. Training was conducted for a maximum of 150 epochs with standard parameters and one hidden layer of 100 hidden neurons (chosen empirically). Input is presented in the form of a feature histogram, in turn based on the normals calculated from a cloud (see [17] for in-depth explanation).

We perform n-fold cross-validation on the REHAP-1 database in order to test the performance of each model on unseen data. We do this in two ways which we compare here: sample-based and person-based.

a) Person-based cross-validation: 20 different MLPs are trained, each on all the data samples from all persons except one which is retained for testing. This amounts to 570.000 data samples for the training of one model and 30.000 data samples for testing. During the test phase, each sample is presented to the MLP and processed layer-wise to generate the MLP's output, determined by the neuron with the highest activation. The results for the generalization of each model can be seen in Table I. Results are presented as the classification error (CE) for each person. For 5 persons we achieve a CE of approx. 10% and less which is remarkable for a simple MLP without parameter fine-tuning. However, the worst results range at around a CE of 46,7% and overall we achieve an averaged CE of 20,8%. Here we can clearly see one advantage of our set: person-based comparison shows strong variance in the overall performance of our models, however strong scores can be achieved rapidly.

b) Normal sample-based cross-validation: For this approach, we train 20 MLPs as before, each on a randomly selected set of 570.000 samples, not taking into account the person they are coming from. This gives an average classification performance of ~92%. Table II shows the averaged confusion matrix for this approach. The numbers 0-9 denote the hand postures in the same order as presented in Section IV. One row adds up to ~3000 samples therefore row **0** shows the number of correctly classified samples (2717) for posture *ONE* as well as the misclassifications per other class. The last column show the amount of correctly classified samples for this class.

VII. Discussion of results

Overall, performances reported in Sec. VI are very acceptable but not perfect. It has to be kept in mind that these experiments just serve to establish a procedure for comparing results using this database, and are not itself intended to be extremely competitive. It is evident that there are many methods to improve classification performances in a real application: parameter optimization, other classification methods, temporal stabilization by, e.g., Kalman filtering, fusion with RGB data or introducing a reject option, just to name a few.

We furthermore observe that the averaged CE for sample-based cross-validation is markedly lower than the averaged CE for person-based cross-validation. The average CE per posture class ranges from 12% to 5% which is also better for most of the cases compared to the average values per class and person. Experiments conducted on the REHAP-2 dataset yield similar scores, albeit with differences across persons, naturally. Our descriptor is normal-dependent which in turns means we have to adjust the parameters (e.g. radius) for the normal calculations, as the clouds in REHAP-2 are denser than the counterparts in REHAP-1.

We believe that person-based cross-validation is the appropriate measure to estimate generalization performance for applications. First of all, because it is more conservative which is always preferable in applications, but above all because only person-based cross-validation guarantees a test on completely unknown samples. Despite variations in distance or rotation, different samples coming from the same person and gesture class may be similar in nature, and thus not suited for representing unknown test samples. We therefore argue that the results of person-based cross-validation as stated in Table I are the ones that other algorithms should compare their performance to. It is a very interesting result of our work on this benchmark database that these two measures differ to such an extent, a fact that should always be kept in mind for gesture recognition.

VIII. Conclusion and Outlook

We present the new publicly available REHAP-1 and REHAP-2 database of $>$1million samples (grouped into 10 classes) for the specific purpose of hand posture recognition from ToF data. As machine learning algorithms rely on the availability of a large number of data samples representing all possible variations, this contribution allows training and reliable benchmarking of own algorithms according to well-defined evaluation criteria. The experiments have shown that significant performance differences can occur, depending on the precise way of performing cross-validation, and we propose that person-based cross-validation (a kind of leave-one-out scheme operating on individual persons instead of samples) is in fact the most accurate estimate of generalization performance for applications. Our database allows for algorithm comparison for a multitude of important parameters, sensor-related questions and human-related factors. Due to its magnitude, any developed algorithm can be evaluated by an in-depth statistical analysis. We aim at launching a number of experiments, optimizing our approached developed so far for this problem which will be made publicly available along with the results and the database itself.

References

[1] Jochen Triesch and Christoph Von Der Malsburg. Robust classification of hand postures against complex backgrounds. In *fg*, page 170. IEEE, 1996.

[2] Jochen Triesch and Christoph Von Der Malsburg. A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1449–1453, 2001.

[3] Sébastien Marcel and Olivier Bernier. Hand posture recognition in a body-face centered space. In *Gesture-Based Communication in Human-Computer Interaction*, pages 97–100. Springer, 1999.

[4] Simon Ruffieux, Denis Lalanne, Elena Mugellini, and Omar Abou Khaled. Gesture recognition corpora and tools: A scripted ground truthing method. *Computer Vision and Image Understanding*, 131:72–87, 2015.

[5] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[6] Georgiana Simion and Catalin-Daniel Caleanu. A tof 3d database for hand gesture recognition. In *Electronics and Telecommunications (ISETC), 2012 10th International Symposium on*, pages 363–366. IEEE, 2012.

[7] Simon Ruffieux, Denis Lalanne, and Elena Mugellini. Chairgest: A challenge for multimodal mid-air gesture recognition for close hci. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 483–488, New York, NY, USA, 2013. ACM.

[8] Li Liu and Ling Shao. Learning discriminative representations from rgb-d video data. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1493–1500. AAAI Press, 2013.

[9] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3786–3793. IEEE, 2014.

[10] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1106–1113. IEEE, 2014.

[11] James S Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1868–1876, 2015.

[12] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (TOG)*, 33(5):169, 2014.

[13] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim Christoph Rhemann Ido Leichter, Alon Vinnikov Yichen Wei, Daniel Freedman Pushmeet Kohli Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. Accurate, robust, and flexible real-time hand tracking. In *Proc. CHI*, volume 8, 2015.

[14] Aaron Wetzler, Ron Slossberg, and Ron Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. *arXiv preprint arXiv:1507.05726*, 2015.

[15] Thomas Kopinski, Stéphane Magand, Alexander Gepperth, and Uwe Handmann. A light-weight real-time applicable hand gesture recognition system for automotive applications. In *Intelligent Vehicles Symposium (IV), 2015 IEEE*, pages 336–342. IEEE, 2015.

[16] Steffen Nissen. Implementation of a fast artificial neural network library (fann). *Report, Department of Computer Science University of Copenhagen (DIKU)*, 31, 2003.

[17] Thomas Kopinski, Stefan Geisler, Louis-Charles Caron, Alexander Gepperth, and Uwe Handmann. A real-time applicable 3d gesture recognition system for automobile hmi. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 2616–2622. IEEE, 2014.