Neural network based data fusion for hand pose recognition with multiple ToF sensors

Thomas Kopinski[1], Alexander Gepperth[2], Stefan Geisler[1] and Uwe Handmann[1]

1- University of Applied Sciences Bottrop - Computer Science Institute
Postfach 100755 - 45407 Mühlheim - Germany

2- ENSTA ParisTech- UIIS Lab
828 Blvd des Maréchaux, 91120 Palaiseau - France

Abstract. We present a study on 3D based hand pose recognition using a new generation of low-cost time-of-flight(ToF) sensors intended for outdoor use in automotive human-machine interaction. As signal quality is impaired compared to Kinect-type sensors, we study several ways to improve performance when a large number of gesture classes is involved. We investigate the performance of different 3D descriptors, as well as the fusion of two ToF sensor streams. By basing a data fusion strategy on the fact that multilayer perceptrons can produce normalized confidences individually for each class, and similarly by designing information-theoretic online measures for assessing confidences of decisions, we show that appropriately chosen fusion strategies can improve overall performance to a very satisfactory level. Real-time capability is retained as the used 3D descriptors, the fusion strategy as well as the online confidence measures are computationally efficient.

1 Introduction

As "intelligent" devices enter more and more areas of everyday life, the issue of man-machine interaction becomes ever more important. As interaction should be easy and natural for the user and also not require a high cognitive load, non-verbal means of interaction such as hand gestures will play a decisive role in this field of research. With the advent of low-cost Kinect-type 3D sensors, and more recently of low-cost ToF sensors (400-500€) that can be applied in outdoor scenarios, the use of point clouds seems a very logical choice. This presents challenges to machine learning approaches as the data dimensionality and sensor noise are high, as well as the number of interesting gesture categories. In this article, we confine ourself to optimize the categorization of static hand gestures (denoted "poses"), and investigate whether the addition of a second ToF sensor, viewing the hand from a different angle, may improve categorization performance if an appropriate fusion is performed. As the sensors we use are very cheap, this is not a barrier to wide-spread deployment in mass products. We will first discuss the related work relevant for our research (Sec. 2) and then go on to describe

the sensors and the used database in Sec. 3. Subsequently, in Sec. 4 we will give an account of the used different holistic point cloud descriptors and explain the meaning of the parameter variations we will test. The key questions we will investigate in Sec. 5 concern the proper **choice of parametrized descriptors**, furthermore the **added value of a second ToF sensor**, and lastly the issue of **efficient neural network based fusion strategies**. In Sec. 6, the obtained results will be discussed in the light of these questions.

2 Related Work

Depth sensors allow for an easy and robust solution for recognizing hand poses as they can easily deal with tasks as segmentation of the hand/arm from the body by simple thresholding as described in [1]. Several surveys have made use of this feature with various approaches to segmentation. Moreover it is possible to make use of the depth information to distinguish between ambiguous hand postures [2]. Nevertheless, it has not been possible to achieve satisfactory results utilizing only a single depth sensor. Either the range of application was limited or the performance results were dissatisfying. Usually a good performance result was achieved with a very limited pose set or if designed for a specific application [3]. ToF-Sensors - although working at stereo-frame rate - generally suffer from a low resolution which of course makes it difficult to extract proper features. Improved results can be achieved when fusing Stereo Cameras with Depth Sensors, e.g. in [4]. In [5] a single ToF-Sensor is used to detect hand postures with the Viewpoint Feature Histogram.
Various approaches make use of the Kinect's ability to extract depth data and RGB data simultaneously [6]. However this approach relies heavily on finding hand pixels in order to be able to segment the hand correctly. Moreover, approaches utilising the Kinect sensor will always suffer from changing lighting conditions which in our case is no drawback as ToF-sensors show robust results in such situations. [7] also make use of the Kinect sensor's ability to acquire RGB and depth data simultaneously albeit using a hand model as a basis for hand pose detection. Nevertheless this algorithm also relies on finding skin-colored pixels to allow for segmentation in 2D and 3D as well as tracking the hand.

Beneath the technology development research is conducted on how to design intuitive user interfaces. Bailly et al. investigate and compare different menu techniques in [8]. Wilson and Benko developed a system with several projectors and depth cameras named LightSpace [9].

In-car scenarios have been developed for several years as the the driver can keep his hands close to the steering wheel while being able to focus on the surrounding environment. Pointing capabilities could be interesting to control content in the head-up displays. A good overview is given in [10].

Such scenarios demand robust data extraction techniques which is provided by the aforementioned ToF-sensor. Our approach shows that it is possible to achieve satisfactory results relying solely on depth data when detecting various hand poses. In merging information from a second depth sensor we are able to

boost our results significantly while always retaining the applicability under various lighting conditions - one of the greatest advantages of ToF-sensors compared to e.g. the frequently used Kinect sensor.

3 Database

The data was recorded using two ToF-Sensors (Figure 1 and 2) of type Camboard nano which provides depth images of resolution 165x120px with a frame rate of 90fps. The illumination wavelength is 850nm which makes the cameras applicable in various light conditions whilst maintaining robustness versus daylight interferences. Since the ToF-principle works by measuring the time the emitted light needs to travel from the sensor to an object and back pixel-wise the light is modulated by a frequency of 30MHz in order to be able to distinguish it from interferences. In a multi-sensor setup however this may lead to a distortion of measurements since both sensors have the same modulation frequency. To avoid such measurement errors, the data was recorded by taking alternating snapshots from each sensor. As can be seen in Figure 1 the cameras are mounted

Fig. 1: The current setup for 90°

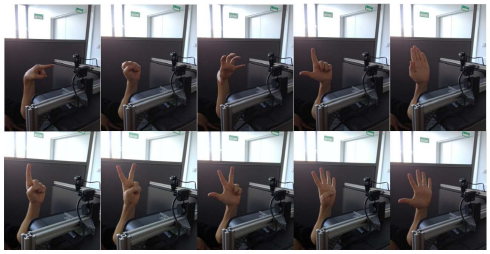

Fig. 2: The hand pose database

in a fixed position at a distance of approx 49.5cm and a perpendicular angle from the recorded object. This allows for a recording of the database such that the hand can be placed in an equal distance of about 35cm from each camera to the centroid of the resulting point cloud dataset and therefore each camera can also be calibrated to its needs. For the current experiments, focus has been put on the recognition of static hand gestures which are contrasted to dynamic hand gestures. Each set of poses was recorded with a variation of the hand posture in terms of translation and rotation of the hand and fingers. This results in an alphabet of ten hand poses: *point*, *fist*, *grip*, *L*, *stop* and counting from 1-5 (cf. Figure 2). For each pose, a set of 2000 point clouds was recorded for each camera. Since we recorded hand poses from four different persons independently, this yields a dataset of 160.000 samples. Additionally, we rotated one camera by 60° towards the other camera and recorded the same set now from an angle of 30° and compared the results to each other resulting in another dataset of

160.000 point clouds. The database is randomly split into two parts of equal size for training and evaluation purposes.

4 Point cloud descriptors

All used global descriptors were calculated using methods of the publicly available Point Cloud Library (PCL).

4.1 The ESF-Descriptor

The ESF-Descriptor (Ensemble of Shape Function) [11] is a global descriptor which does not rely on the calculation of the normals. First, 20000 points are sub-sampled from the input point cloud. Then, the algorithm repeatedly samples three points, from which four simple measures are calculated, which are discretized and used for histogram calculation.

4.2 The VFH-Descriptor

The VFH-Descriptor(Viewpoint Feature Histogram) [12] is a global descriptor partially based on the local FPFH (Fast Point Feature Histogram)[13] descriptor. It uses normal information, taking into consideration the view angle between the origin of the source and each point's normal. It furthermore includes the SPFH (Simplified Point Feature Histogram) for the centroid of the cloud, as well as a histogram of distances of the points in the cloud to the centroid. When calculating the VFHs for the various hand poses we have to take into consideration the influence of the normals on the results. In the described case the search parameter r guides the influence of the surrounding for the calculation of the normal. Choosing a small r can result in low descriptive power while a large r results in high computational load. We empirically chose a value of $r = 5cm$ and denote the resulting descriptor VFH5.

4.3 Neural network classification and fusion

With M cameras, N descriptors will be produced per frame (here: M=N) according to the methods described above. We use a multilayer perceptron (MLP) network[14] to implement the multi-class decision, which is either based on the the concatenation of all N descriptors ("early fusion"), or on each descriptor individually, with a subsequent combination of results ("late fusion"). The MLP training algorithm is "RProp"[14], with standard hyperparameters $\eta^+ = 1.2$, $\eta^- = 0.6$, $\Delta_0 = 0.1$, $\Delta_{\min} = 10^{-10}$ and $\Delta_{\max} = 5$. Network topology is NK-150-10 (hidden layers are fixed to 1[14], hidden layer sizes from 10-500 were tested), K indicating the method-dependent descriptor size, and N the number of cameras, here $N = 2$. Us usual, activation functions are sigmoid throughout the network. MLP classifiers have 10 output neurons (one per gesture class) with activities o_i. Thus, the final classification decision is obtained by taking the class of the

neuron with the highest output. However, we do not necessarily wish for every classification to be taken seriously, and we define several confidence measures $\text{conf}(\{o_i\})$ to this effect. Final decisions are thus taken in the following way:

$$\text{class} = \begin{cases} \text{argmax}_i o_i & \text{if conf}(\{o_i\}) > \theta_{\text{conf}} \\ \text{no decision} & \text{else} \end{cases}$$

We test three ad hoc confidence measures, which perform a mapping from $\mathbb{R}^{10} \rightarrow \mathbb{R}$: "confOfMax", "diffMeasure" and "varianceMeasure". Each of these measures is derived from the idea of approximating an entropy calculation, based on the information-theoretic idea that low entropy means high information content. The precise definitions are as follows:

$$\begin{aligned} \text{confOfMax}(\{o_i\}) &= \max\, o_i \\ \text{diffMeasure}(\{o_i\}) &= \max_i\, o_i - \max_i^2\, o_i \\ \text{varianceMeasure}(\{o_i\}) &= \frac{1}{N}\sum_i (o_i - E(\{o_i\}))^2 \end{aligned} \tag{1}$$

where $\max_i^2\, o_i$ indicates the second-strongest maximum over the neural outputs. For performing late fusion, that is, obtaining two independent classifications o_i^1, o_i^2 based on each camera's features, we simply calculate the arithmetic mean of both output vectors: $o_i^F = 0.5(o_i^1 + o_i^2)$. This intrinsically takes into account the variance in each response, as an output distribution strongly peaked on one class will dominate a flat (or less peaked) distribution. The resulting output distribution o_i^F can then be subjected to the decision rule of Eqn. (1).

5 Experiments

We implement a multilayer perceptron (MLP) as described in Sec. 4.3 using the freely available OpenCV library[15] and its C++ interface[1]. Each experiment is performed 10 times with different initial conditions for the MLP, and the best result is retained. In these experiments, we systematically evaluate the influence of different confidence measures("confOfMax","diffMeasure" or "varianceMeasure", see Sec. 4.3) on the fusion strategy ("add", see Sec. 4.3) while measuring the performance of the first camera, the second camera as well as an "early fusion" or a "late fusion" of the two cameras. In order to test the influence of different 3D descriptors, we perform an identical evaluation except that the VFH5 point cloud descriptors is replaced by ESF. Additionally, we perform the same evaluation on an analogous database using the VFH5 descriptor where the angle between ToF sensors is 90 deg. Results are evaluated by default according to whether one among the S strongest output neurons coincides with the true class of a point cloud ("S-peak measure"). Unless explicitly states, we use $S = 1$. Results are given in Fig. 3. Several important aspects may be perceived: first

[1] The code and data for all experiments is available under `www.gepperth.net/alexander/downloads/2014_icann.tar.gz`

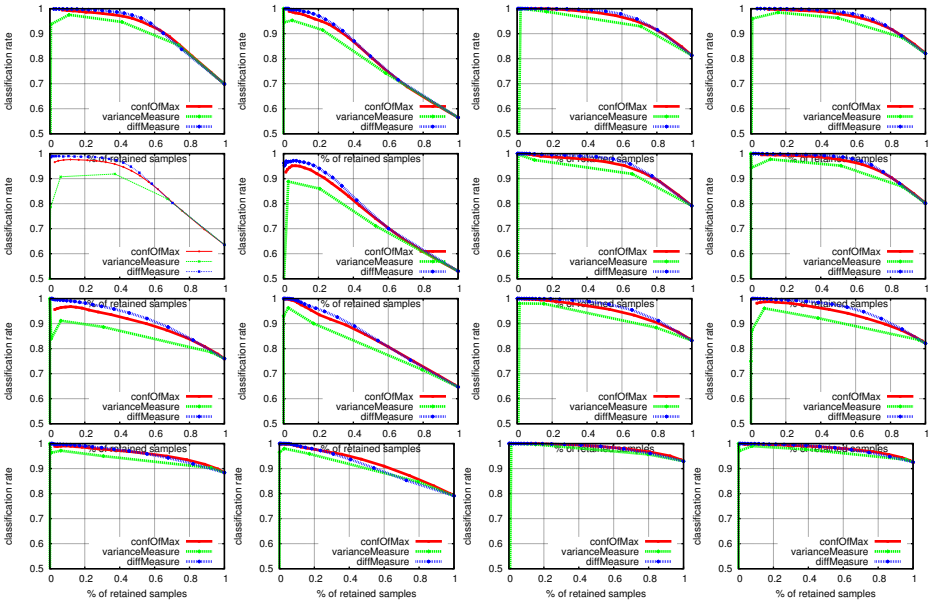

Fig. 3: Experimental results. First row: VFH5 descriptor, 30 degree between cameras. Second row: VFH5, 90 degrees between cameras. Third row: ESF descriptor, 30 degrees between cameras. Last row: Same as third row, only classification errors evaluated using the two-peak measure, see text. In all rows, the order of diagrams is, from left to right: 1,2) first/second sensor 3) late fusion 4) early fusion. Individual plots show the effects of varying confidence thresholds on classification accuracies for several possible online confidence measures. We do not show the method-dependent confidence thresholds but rather the acceptance rates which vary if thresholds are varied. At the far right of each diagram, we recover the classification performance obtained when not rejecting anything, naturally leading to reduced performance.

of all, fusion strongly improves results in comparison to any single sensor, w.r.t. to the efficiency of sample rejection but also in absolute terms when no samples are rejected, corresponding to the intersection of the graphs with the right boundary of the coordinate system. Secondly, early fusion has slightly superior performance than late fusion but the difference is marginal, potentially giving a preference to late fusion due to reduced computational complexity. Lastly, the different confidence measure are consistently ranked throughout all experiments, with the "diffMeasure" being the best-performing one, closely followed by "confOfMax". This is encouraging as especially confOfMax is computationally very lightweight, again favoring real-time execution. Thirdly, the angle between cameras does not seem to play a crucial role even though individual camera results differ considerably. Here, the beneficial aspects of fusion can be clearly demonstrated. And lastly, the ESF descriptor seems to perform slightly better than VFH5, which might lead us to prefer this descriptor as it is computation-

ally simpler and requires constant execution time regardless of point cloud size. An interesting observation is that the two-peak measure enormously improves classification rates in all conditions. This is very useful for an application, especially for temporal filtering, as the behaviour of the second-strongest output can obviously also provide valuable information about the true pose class.

Training times are around 10min per single experiment, which outperforms an equivalent SVM-based (Support Vector Machine) "one-versus-all" implementation by a large margin. Average execution times vary between 1-5 Hz depending of the use of the descriptor (ESF: 0.2s/0.2s for 30/90 deg. between cameras, VFH5: 0.4s/0.9s) whereas NN execution time is $< 0.005s$. On average the point clouds contain 1300-1600 points, depending on the angle between cameras and the distance of the recorded hand to each camera.

6 Discussion and outlook

Analyzing the results in the light of the key research questions formulated in Sec. 1, we can state that, first of all, fusion with data from a second ToF sensor improves results tremendously in all investigated conditions, camera setups and point cloud descriptors. Interestingly, late fusion performs globally just as well as early fusion, which is important as it has the potential to be much more computationally efficient. However, even when considering individual ToF sensors, the computation of confidence measures from output activity distributions is of tremendous impact as well. Confidence can be efficiently extracted at execution time (no need to see the class labels for this) and used to avoid classification decisions when they are likely to be incorrect anyway. We tested a number of information-theoretically motivated measures and luckily the most efficient measures seem to perform best. Concerning the influence of the used 3D descriptors: the ESF descriptor yields best performance with or without fusion. As this descriptor does not require normals computation and has approximately constant scaling behavior w.r.t. point cloud size, it is the most appropriate choice for real-time applications in the targeted automotive domain.

Summarizing, we have presented an adaptive data fusion approach for multiple ToF sensors adressing the generic task of 3D point cloud categorization in a multi-class setting. The fact of using a neural network for this purpose is of high advantage (besides very favorable database size scaling and multi-class issues) as the ensemble of normalized output confidences contains valuable information as well that can be efficiently exploited at runtime to improve results. Neural network learning furthermore removes the need for precise multi-sensor calibration as long as only categorization is targetted. Further work will include an implementation of this system in a true automotive setting, extensive performance evaluations, and possibly a fusion with a visual sensor as well.

References

1. S. Oprisescu, C. Rasche, and B. Su. Automatic static hand gesture recognition using tof cameras. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings*

of the 20th European, pages 2748–2751. IEEE, 2012.
2. E. Kollorz, J. Penne, J. Hornegger, and A. Barke. Gesture recognition with a time-of-flight camera. *International Journal of Intelligent Systems Technologies and Applications*, 5(3):334–343, 2008.
3. S. Soutschek, J. Penne, Jo. Hornegger, and J. Kornhuber. 3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.
4. Y. Wen, C. Hu, G. Yu, and C. Wang. A robust method of detecting hand gestures using depth sensors. In *Haptic Audio Visual Environments and Games (HAVE), 2012 IEEE International Workshop on*, pages 72–77. IEEE, 2012.
5. T. Kapuściński, M. Oszust, and M. Wysocki. Hand gesture recognition using time-of-flight camera and viewpoint feature histogram. In *Intelligent Systems in Technical and Medical Diagnostics*, pages 403–414. Springer, 2014.
6. Matthew Tang. Recognizing hand gestures with Microsoft's kinect. *Web Site: http://www.stanford.edu/ class/ee368/Project_11/ Reports/Tang_Hand_Gesture_Recognition. pdf*, 2011.
7. Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, pages 1–11, 2011.
8. Gilles Bailly, Robert Walter, Jörg Müller, Tongyan Ning, and Eric Lecolinet. Comparing free hand menu techniques for distant displays using linear, marking and finger-count menus. In *Human-Computer Interaction–INTERACT 2011*, pages 248–262. Springer, 2011.
9. Andrew D Wilson and Hrvoje Benko. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 273–282. ACM, 2010.
10. Carl A Pickering, Keith J Burnham, and Michael J Richardson. A research study of hand gesture recognition technologies and applications for human vehicle interaction. In *3rd Conf. on Automotive Electronics*. Citeseer, 2007.
11. W. Wohlkinger and M. Vincze. Ensemble of shape functions for 3d object classification. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pages 2987–2992. IEEE, 2011.
12. R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2155–2162. IEEE, 2010.
13. R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.
14. S Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall, 1999.
15. G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Incorporated, 2008.