Multi-ToF sensor fusion for hand pose estimation

Preprint · September 2014

3 authors:

Thomas Kopinski
ENSTA ParisTech
23 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Alexander Gepperth
University of Applied Sciences Fulda
88 PUBLICATIONS **576** CITATIONS
SEE PROFILE

Uwe Handmann
Hochschule Ruhr West
96 PUBLICATIONS **688** CITATIONS
SEE PROFILE

Multi-ToF sensor fusion for hand pose estimation

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

1- University of Applied Sciences Bottrop - Schools of Informatics
Postfach 100755 - 45407 Mühlheim - Germany

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

No Institute Given

Abstract. We present a study on 3D based hand pose recognition using a new generation of low-cost ToF sensors. 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 means of a neural network and obtain, for certain descriptors and fusion strategies, a very satisfactory recognition performance.

1 Introduction

As "intelligent" 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-spred deployment in mass products. We will first discuss the related work relevant for our research and then go on to describe the sensors and the used database in Sec. 3. Subsequently, 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 sensors**, 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.

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 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 as resembling a certain *meaning* while remaining in place as opposed to the latter being in movement. Nevertheless to maintain a certain variability in the data, 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

Fig. 1: The current setup

Fig. 2: The hand pose database

2000 point clouds was recorded for each camera yielding a total dataset of 40.000 samples.

4 Descriptors

4.1 The VFH-descriptors

To analyze the point clouds the performance of the available algorithms from the Point Cloud Library (PCL) was evaluated. One generally distinguishes between local and global descriptors, the former describing the characteristic of a single point and the latter the cloud as a whole. The VFH [6] descriptor is a global descriptor partially based on the local FPFH [7] descriptor.

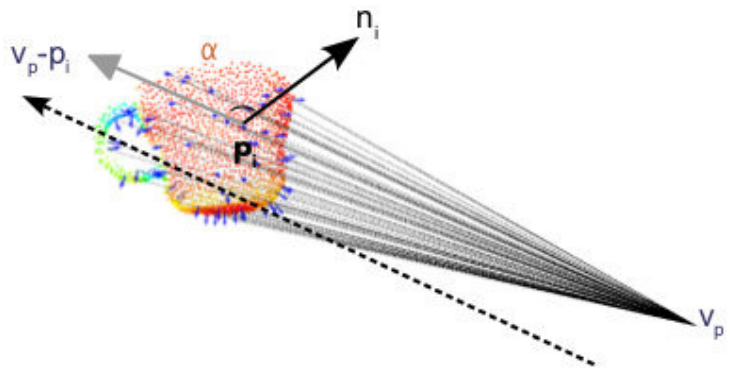

Fig. 3: The second component of the Viewpoint Feature Histogram (Image taken from www.pcl.com)

For our purposes we utilize the global VFH (Viewpoint Feature Histogram) descriptor which extends the idea of the FPFH by calculating the histogram for the centroid of the cloud with all the points set as neighbours and takes into

consideration the view angle between the origin of the source and each point's normal (cf. Fig.3). Here v_p denotes the origin of the viewpoint, p_i is the centroid of the point cloud and v_i - p_i is the vector between the centroid and the viewpoint origin. The vector n_i resembles the normal for each point in the cloud. Here the dashed arrow indicates the translation of the viewpoint origin to each point in the cloud which allows for the viewpoint invariance of the descriptor. This yields a scale-invariant descriptor while describing the point cloud as a whole. The remaining bins of the histogram consist of the SPFH (Simplified Point Feature Histogram) for the centroid of the cloud and the last component describes the 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 yielded results. A normal for a requested point in a point cloud can be estimated in a number of ways but in this case, since we work on unknown datasets, is estimated by the points in the surrounding environment. However as we have potentially highly noisy data, determining the right scale can have great influence on the results (cf. Fig.4). 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 radius r distorts the descriptor too much. The choice of the right search radius is application dependent, so together with advised radii from other applications and a couple of pre-tests we determined the influence of the normal radius to be in [0.03,0.09].

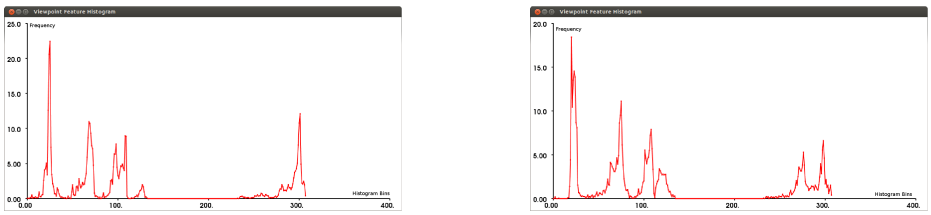

Fig. 4: Difference in histograms for the same point clouds with different normal radii of 5cm and 7cm.

4.2 The ESF-Descriptors

The ESF-Descriptor (Ensemble of Shape Function) [8] is another global descriptor which is however not geometric as it does not rely on the calculation of the normals. The resulting calculated descriptor consists of ten histograms each itself comprised of 64 bins. In an initial step 20000 points are sub-sampled from the cloud. Now in turn for each iteration sample three points randomly from the first step and calculate four measures. The *D2* measure checks whether the points on the line connecting two of the sampled points lie inside or outside the

surface or both. This 'convexity'-measure is binned into a histogram. The *D2 ratio* measures the ratio of these lines lying on the cloud or free. The *D3 ratio* calculates the square root of the area spanned by the sampled points and again checks for the position of the area relative to the cloud. Lastly the *A3* measure calculates the angles between the three points and creates the histogram in an analogous manner to the other measures. Although this descriptor does not rely on any normal information it represents the general shape of the data while retaining its global descriptive power.

5 Experiments

We use a multilayer perceptron, implemented using the freely available pybrain[9] library, to perform the final multi-class decision [1]. We will compare classification performance for the ESF descriptor (see Sec. 4.2) and the different parametrizations of the VFH descriptor (see Sec. 4.1). In all cases we will compare the *single-camera condition* to the *two-camera condition*, see Sec. 3, where in the two-camera condition the descriptors coming from each camera are concatenated. Networks contain a bias unit at each layer, training algorithm is "RProp-"[10], and network topology is N-10-10 (hidden layer sizes of up to 100 were tested without finding significant performance improvements), N indicating the size of the used descriptors depending on computation method and number of cameras. Activation functions are sigmoid for the hidden layer and sigmoid or softmax for the output layer, the latter being considered advantageous for non-ordinal multi-class categorization problems[11]. Each experiment is performed 10 times with different initial conditions and the best result is retained. Results for both experimental conditions are summarized in Tab. 1. They show a marked superiority of the VFH descriptor, and a slight parameter dependency in both experimental conditions. We also find that a sigmoid transfer function slightly outperforms a softmax one even though it is theoretically less appropriate. Training times are around 10min per single experiment, which outperforms an equivalent SVM-based (Support Vector Machine) implementation by a large margin as this would require 10 one-vs-all training runs with 20.000 examples which takes half day in total.

Descr. / Cond.	ESF	VFH						
		003	004	005	006	007	008	009
single-cam./softmax	25.33	18.98	19.09	20.07	26.54	29.8	30.44	30.17
two-cam./softmax	4.91	4.28	4.19	3.42	3.01	2.85	2.02	2.32
two-cam./sigmoid	1.56	4.28	4.19	3.42	3.01	2.85	2.02	2.32

Table 1

[1] The code and data for all experiments is available under `www.gepperth.net/alexander/downloads/2014_nn3D.tar.gz`

6 Discussion and outlook

Analyzing the results in the light of the key research questions formulated in Sec. **??**, we can state that, first of all, the VFH descriptor with a normal radius of about $5cm$ seems to be the most appropriate choice of describing hand poses. This is somewhat surprising as the used 3D data are rather noisy which could impair normal calculation, but apparently an increased radius can fix that. The ESF descriptor which does not use normals results in decent absolute performance as well although the errors are more than two times higher than those using VFH. Clearly, the use of a second sensor improves results tremendously, which is probably due to the resolution of viewpoint ambiguities so made possible. As for fusion strategies, a simple concatenation of feature vectors, followed by NN (Neural Network) classification (using sigmoid rather than softmax activation functions which would be appropriate from a theoretical point of view) is a very effective and computationally efficient strategy which, in addition, may be conveniently parallelized if required. In future work, we will aim to increase the number of recognized hand poses and include non-static hand gestures into our framework. Moreover our method allows for little to no need for spatial calibration of the sensors as opposed to other approaches where the cameras have to be aligned exactly for the designed algorithms.

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. 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.
7. 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.
8. 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.

9. T. Schaul, J. Bayer, D. Wierstra, Yi Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber. PyBrain. *Journal of Machine Learning Research*, 11:743–746, 2010.
10. S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall, 1999.
11. JS Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F Fogleman Soulie and J Herault, editors, *Neurocomputing: Algorithms, Architectures and Applications*. Springer, 1990.