Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# A performance evaluation of point pair features

Lilita Kiforenko[*,a], Bertram Drost[b], Federico Tombari[c], Norbert Krüger[a], Anders Glent Buch[a]

[a] The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
[b] MVTec Software GmbH, München, Germany
[c] Technische Universität München, Garching b. München, Germany

ARTICLE INFO

*Keywords:*
PPF
Point pair features
Object detection
Object recognition
Pose estimation
Feature description

*MSC:*
41A05
41A10
65D05
65D17

ABSTRACT

More than a decade ago, the point pair features (PPFs) were introduced, showing a great potential for 3D object detection and pose estimation under very different conditions. Many modifications have been made to the original PPF, in each case showing varying degrees of improvement for specific datasets. However, to the best of our knowledge, no comprehensive evaluation of these features has been made. In this work, we evaluate PPFs on a large set of 3D scenes. We not only compare PPFs to local point cloud descriptors, but also investigate the internal variations of PPFs (different types of relations between two points). Our comparison is made on 7 publicly available datasets, showing variations on a number of parameters, e.g. acquisition technique, the number of objects/scenes and the amount of occlusion and clutter. We evaluate feature performance both at a point-wise object-scene correspondence level and for overall object detection and pose estimation in a RANSAC pipeline. Additionally, we also present object detection and pose estimation results for the original, voting based, PPF algorithm. Our results show that in general PPF is the top performer, however, there are datasets, which have low resolution data, where local histogram features show a higher performance than PPFs. We also found that PPFs compared to most local histogram features degrade faster under disturbances such as occlusion and clutter, however, PPFs still remain more descriptive on an absolute scale. The main contribution of this paper is a detailed analysis of PPFs, which highlights under which conditions PPFs perform particularly well as well as its main weaknesses.

## 1. Introduction

Through the last three decades, many different 3D feature descriptors have been proposed. Usually, they are divided into two categories: global feature based methods (which describe the object using one global feature, e.g. Siddiqi et al., 1998 and Wahl et al., 2003) and local feature based methods (which describe the object using point neighbourhoods, e.g. Guo et al., 2014 and Wu et al., 2010). In the field of 3D object pose estimation, local feature descriptors have become more popular than global ones, since the local nature of such features makes the description tolerant to occlusions and clutter. Global descriptors are used primarily for object shape matching (object retrieval). The global descriptors represent the full object by some structure, e.g. skeletal graphs (Siddiqi et al., 1998) or a histogram over some relational features (Wahl et al., 2003). The global descriptors can be computationally expensive and require segmentation and full object shape, which makes them less stable under high occlusion. On the other hand, most local descriptors are by themselves computationally less expensive and more robust towards clutter and occlusion, but instead incur additional computation time in the following stages of matching and hypothesis verification. There have been proposals of combining both global and local descriptors. For example in Wu et al. (2010), a manifold harmonic analysis is used to design an isometry-invariant descriptor for 3D object shape comparison. Another type of a descriptor that captures local and global information is point pair features (PPFs) (Drost et al., 2010). This point relational descriptor has shown very successful object detection on different 3D datasets.

Local histogram based feature descriptors have been investigated in many works (Buch et al., 2016; Guo et al., 2016), mostly in order to design a better, more discriminative, faster and stable descriptor. Also, some work has been done in exhaustive evaluation of local descriptors, where popular descriptors have been evaluated on many datasets in order to conclude which ones are more robust, descriptive, scalable and efficient (Guo et al., 2016). In contrast to local descriptors, such evaluation has not been done yet for the PPF descriptor.

In this paper, we present a comprehensive analysis of the PPF descriptor and provide a comparison with several 3D local histogram feature descriptors (SHOT, ECSAD, FPFH, USC, SI). Here the 3D local

* Corresponding author.
*E-mail address:* lilita@kiforenko.eu (L. Kiforenko).

http://dx.doi.org/10.1016/j.cviu.2017.09.004
Received 21 November 2016; Received in revised form 12 June 2017; Accepted 17 September 2017

histogram feature descriptors serve as a baseline for the comparison to the PPF descriptor. Feature evaluations and comparisons are performed on 7 publicly available datasets, which contain large varieties of objects and scenes recorded by different sensors and for various purposes. One of the data sets (Sølund et al., 2016) provides 3204 scenes and 45 models and exceeds other data sets in terms of size and complexity. We start our investigations with a systematic evaluation of the relational space in the PPF descriptor (Section 5). We evaluate 19 different PPF variations and determine the best PPF for different data sources. Our results show that on average, the original PPF performs best, but there are special cases, e.g. where adding colour to the PPF descriptor boosts the performance.

After the initial PPF analysis, we perform a systematic comparison between the PPFs and 5 local histogram feature descriptors (Section 6). The results depend on the dataset, for example, if the data is recorded by a high precision laser scanner, then PPFs significantly outperform local histogram descriptors. The opposite occurs when the input data is less detailed and noisy, such as data from Kinect-like devices, where the highest performance is achieved by local histogram features.

Moreover, we investigate robustness towards noise, occlusion and clutter (Section 6.3). Our results show that performance of PPFs is decreasing significantly faster under increasing occlusion, clutter and noise compared to most local histogram descriptors.

One of the used datasets divides the objects into three different categories, which allows us to show feature performance for different object categories (Section 6.2). Our results also show that PPFs in comparison to local histogram features have a significant drop of initial precision (i.e. the inlier rate of the top ranked matches), which indicates that PPFs need to be used in more robust object detection algorithm.

After feature performance evaluations on a point-wise object-scene correspondence level, we present object detection and pose estimation results (Section 7). In order to be consistent, we compute object poses using the same RANSAC pose estimation pipeline for every used feature. We also provide object detection and pose estimation results for voting based pose clustering for PPFs as a reference.

Our object detection and pose estimation results show that PPFs clearly outperform local histogram features for two datasets, containing scenes with low quality and high levels of occlusion. On the other hand, if the scenes contain high-quality reconstructions with moderate occlusions and clutter, our results show that local features perform best. In one case, that is for a Kinect-based dataset, we get similar performances for local features and PPFs. Interestingly, this dataset does indeed differ from the others as it has low occlusions, but overall the quality of the point cloud data is poor.

In the end of the work, we provide a discussion of the achieved results (Section 8).

## 2. Related work

Object detection and pose estimation from 3D data have been a popular topic for many researchers, which lead to a development of a rich variety of feature descriptors. This section presents an overview of the popular feature descriptors, but mostly focusing on point pair features and their modifications.

Before the 3D sensors were available, a lot of work was focused on designing feature descriptors for images (2D data). The designed popular descriptor are still used to this data, for example, SIFT , SURF etc. Due to the descriptors good performance, some of them were extended to the 3D data. For example, 3D-SURF (Knopp et al., 2010) or SI-SIFT (Bayramoglu and Alatan, 2010).

The SURF or speeded up robust feature was initially inspired by SIFT and became a faster and more robust feature descriptor. The descriptor computes the Haar wavelet response of the feature point neighbourhood, it is scale and rotation invariant and can be also used as interest point detector.

The SI-SIFT feature descriptor integrates shape index with the SIFT 2D feature descriptor. It has shown a good performance for the data which is rotated, scaled and occluded. Shape index is a value, which describes the principal curvature of the 3D point of interest. For each 3D point, the shape index is computed and used to build a 2D image, which represents the depth discontinuities. SIFT features are computed using build 2D image as input.

Another often used feature descriptor is Normal Aligned Radial Feature (NARF) (Steder et al., 2010), which is not only a feature descriptor but also a robust interest point extractor. NARF extracts interest points, which have a stable normal and a significant change in depth.

Sun et al. (2009) proposed a Heat Kernel Signature (HKS) descriptor, which describes the heat distribution of the feature point neighbourhood. They showed that this descriptor is stable towards scale change and is isometric invariant. HKS capture both the local and global properties of the feature point.

PPFs were introduced for object recognition by Drost et al. (2010). Their point pair features use quite primitive relationships between any two points, such as distance and angle between normals. Together with a hash table and an efficient voting scheme, the method performs well in the case of occlusion, clutter and noise. The features were tested both on synthetic and real datasets. For a real dataset Drost et al. achieved a success rate of up to 97% for objects with occlusion levels less than 84%.

The method quickly became popular and many modifications have been proposed. Choi et al. (2012) proposed to use different types of relations for point pair features, for example, boundary to boundary relations or relations between two lines which are created by the edge points. Using these edge point relations decreases the number of features both for training and matching; consequently, it increases the detection speed. This modification shows in particular good performance for industrial (mostly planar) objects.

Kim and Medioni (2011) proposed to add a visibility context to the original PPF, creating a five-dimensional feature vector. They used three types of visibility - space, surface, invisible surface. Adding a visibility parameter improves the PPF matching. The approach was tested using a view-based object models on a data captured by RGB-D camera. The result shows clear outperformance of the original method on the same data.

Choi and Christensen (2012) described another modification of PPFs by adding a colour component to the traditional 4 dimensional point pair feature, creating CPPF - a 10 dimensional descriptor. The results showed good performance for 10 textured household objects in highly occluded and cluttered scenes.

Drost and Ilic (2012) computed PPFs for geometric edges (boundaries and silhouettes). In this case, the PPFs were computed slightly differently by the use of edge gradients. The evaluation was made on the ACCV3D dataset (Hinterstoisser et al., 2012), where the proposed method significantly improves the PPF descriptor in highly occluded scenes.

All the methods mentioned above were using very similar detection pipelines. Birdal and Ilic (2015) proposed another one, where a scene is first segmented and then PPFs are computed for each segment. The method was tested on the ACCV3D dataset (Hinterstoisser et al., 2012) with an average success rate of 88.77%, which is higher compared to original PPFs (81.18%).

The work by Tuzel et al. (2014) presents an approach for learning features. They showed that certain pairs do not have enough discriminative information (for example, pairs on the same planar area). Specifically, the features that were learned were the weights for the hash table bins and dummyTXdummy- weights for the object model points. The method was tested on two datasets. The results showed improvement in object detection for both of the tested datasets compared to the original PPF method.

The research into PPFs is still ongoing. Recently, a new method has been presented by Hinterstoisser et al. (2016), which proposes to use a different point sampling and pose voting approach. With their