Neurocomputing 120 (2013) 101–112

Contents lists available at ScienceDirect

Neurocomputing

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

Towards using covariance matrix pyramids as salient point descriptors in 3D point clouds

Moritz Kaiser [a,*], Xiao Xu [a], Bogdan Kwolek [a,b], Shamik Sural [a,c], Gerhard Rigoll [a]

[a] Institute for Human-Machine Communication, Technische Universität München, Munich, Germany
[b] Faculty of Electrical and Computer Engineering, Rzeszow University of Technology, Rzeszow, Poland
[c] School of Information Technology, Indian Institute of Technology, Kharagpur, India

ARTICLE INFO

Article history:
Received 18 December 2011
Received in revised form
25 May 2012
Accepted 9 June 2012
Available online 30 March 2013

Keywords:
Salient point descriptor
3D point clouds
Global optimization
Covariance matrix

ABSTRACT

In this work, a novel salient point descriptor for 3D point clouds, called Covariance Matrix Pyramids (CMPs), is presented. With CMPs it is possible to compare unstructured and unequal numbers of points which is an important characteristic when working with point clouds. Corresponding points from different scans are matched in a pyramidal approach combined with Particle Swarm Optimization. The flexibility of CMPs is demonstrated on the basis of several databases with objects, such as 3D faces, 3D apples, 3D kitchen scenes, 3D human–machine interaction gesture sequences, and 3D buildings all recorded with different 3D sensors. Quantitative results are given and compared with other state-of-the-art descriptors, whereby CMPs show promising performance.

1. Introduction

1.1. Motivation

In the computer vision domain conventional cameras which output one channel (gray) images or three channel (color) images are increasingly supplemented by information from novel sensors [1,2]. Especially 3D sensors are important to gather necessary information about the environment for all kinds of human–machine interaction applications. Examples for such sensors include the PointGrey Bumblebee XB3, the Velodyne LIDAR used in the DARPA Urban Challenge, the Siemens Structured-Light 3D Scanner, or Microsoft's Kinect sensor. The output of these devices is not conveniently structured as image but as 3D point cloud. The huge success of the NVidia/Google supported Point Cloud Library [3] and Microsoft's Kinect can be seen as indicator that in the near future point clouds will play an important role in the computer vision field and probably even replace conventional images for many applications. However, almost all salient point descriptors rely on dense gray or color images, and only little work has been done on matching points in point clouds. Therefore, we felt the need to present a new point descriptor that is able to cope with 3D point clouds. A possible application for such a descriptor would be automatic labeling of a database. The user could select salient points in, for example, one reference face and the other faces in a database are then automatically labeled.

1.2. Related work

There exists a considerable number of salient point descriptors. Among the most prominent ones are KLT [4], SIFT [5], PCA-SIFT [6], and SURF [7]. In [8], a comparison among state-of-the-art point descriptors is given, in which the SIFT descriptor performs best. Also for tracking, accurate optical flow methods exist, such as [9–12]. The SURF descriptor has been further refined in [13], where the FAIR-SURF descriptor has been proposed. In [14], the authors present a scale invariant method for image matching which applies weighted voting on a 3D affinity matrix.

Covariance matrices have been used in [15,16], where both approaches are applied to conventional images. In [17], the authors propose a similar approach, called Sigma Set, which is computationally less demanding. In [18], Pang et al. applied Gabor-based covariance matrices for face recognition. This approach has been further refined in [19], where the Kernel Gabor Region Covariance Matrix has been presented and also applied for face recognition tasks. In [20], the authors explore smart possibilities to extract features from co-occurrence histograms of oriented gradients (CoHOGs) for person detection. However, all these methods rely on conventional images. Thus, they are not suited for 3D point clouds. In [21], the authors propose an interesting approach where

* Corresponding author. Tel.: +49 89 289 28547; fax: +49 89 289 28535.
E-mail addresses: moritz.kaiser@mytum.de, moritz.kaiser@tum.de (M. Kaiser), xiao.xu@tum.de (X. Xu), bkwolek@prz.edu.pl (B. Kwolek), shamik@sit.iitkgp.ernet.in (S. Sural), gerhard.rigoll@tum.de (G. Rigoll).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.06.058

SIFT features are adapted for 2.5D range data with image structure and without texture.

There have also been contributions with methods that work directly with point clouds. Frome et al. presented 3D shape contexts and harmonic shape contexts to classify whole shapes without using texture [22]. In [23], the authors introduced a technique for the registration of 3D point clouds and Brostow et al. presented a work on semantic segmentation based on 3D point clouds in [24]. Another promising approach is spin images [25]. Note that the point matching strategy is brute force. Furthermore, spin-images are quite restrictive, i.e., they are designed to match points from exactly the same object, while matching, for example, facial feature points of two different individuals might fail. Rusu et al. [26] presented the Persistent Point Feature Histograms (PFH) for 3D point clouds that are also already available in Willow Garage's Point Cloud Library [3].

1.3. Overview

In this work, covariance matrix pyramids (CMPs), that have been presented in [16], are used for point clouds. Since images and point clouds are structurally different, the method substantially changed in order to work for point clouds. The result is a new, highly flexible salient point descriptor that works directly on 3D point clouds. The method is summarized as follows:

- A list of potential features for the description of the salient point's neighborhood is presented. With a training set, adequate features are selected via Sequential Forward Selection (SFS) with discrete weights (Section 2).
- Features are summarized by a covariance matrix. Employing a covariance matrix as salient point descriptor is practical for matching salient points. In contrast to many previously proposed descriptors (SIFT, SURF, local optical flow, etc.), it provides a convenient way to fuse conventional features (red, green, blue) with non-conventional features (depth, infrared, etc.). Spatial distribution is captured by the covariance between x, y, or z-coordinates of the points and their other features. Furthermore, covariance matrices are, to a certain extent, robust against noise and illumination offset, because both are filtered out by an average filter during covariance computation (Section 3).
- Corresponding points from different scans are matched. To allow for larger displacements covariance matrices are used in pyramids, motivating the name *covariance matrix pyramid*. Particle Swarm Optimization (PSO) is employed to find the best match at each pyramid level (Section 4).

Five application scenarios are given in Section 5. In the first two experiments, salient points in 3D faces are matched. Two publicly available databases with handlabeled landmarks have been employed. With these landmarks as ground truth quantitative results can be given and it can also be shown that PSO reduces computation time while not affecting matching accuracy.

Further, salient points in 3D apples, gesture sequences, kitchen scenes, and buildings are matched. The matching accuracy is compared to another point descriptor for 3D point clouds and two other point descriptors that rely on 2D images. All experiments demonstrate promising performance of CMPs. In Section 6, the work is concluded and future scope is outlined.

2. Adequate features

2.1. Output from sensors

We assume that sensors output an unstructured 3D point cloud. Examples for these sensors include the PointGrey Bumblebee XB3, the Velodyne LIDAR used in the DARPA Urban Challenge, the Siemens Structured-Light 3D Scanner, Inspeck Mega Capturor II 3D, Di3D Dynamic Imaging System, or Microsoft's Kinect sensor. Each point has spatial attributes (x,y,z) and color attributes (r,g,b). If one of the points is selected as salient point, information about this point and its neighborhood must be extracted for its representation. For this purpose, features are extracted, as explained in the next section.

2.2. Feature extraction

For a salient point a set of features is computed. We propose a list of potential features (depicted in Fig. 2 for a face of the Bosphorus database [27]) of which the best features can be selected automatically if a training set is available. Spatial information (x,y,z) can be directly taken. Hue H, saturation S, and value V are computed from each point's rgb-values.

The surface normal $\boldsymbol{n}_i$ for point i, which is depicted in Fig. 1, is computed as follows. The point cloud is triangulated with Delaunay triangulation. The surface normal $\boldsymbol{n}_t$ at the triangle centroid is computed. For the triangle t $(\boldsymbol{p}_1,\boldsymbol{p}_2,\boldsymbol{p}_3)$ the surface normal is

$$\boldsymbol{n}_t = \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = (\boldsymbol{p}_2 - \boldsymbol{p}_1) \times (\boldsymbol{p}_3 - \boldsymbol{p}_1). \tag{1}$$

The surface normal $\boldsymbol{n}_i$ of point i is then the average of all surface normals of the triangles of which point i is a vertex:

$$\boldsymbol{n}_i = \frac{1}{\sum_t \omega_t} \sum_t \omega_t \cdot \boldsymbol{n}_t, \tag{2}$$

where ω_t is a weight that depends on the distance between the centroid of triangle t and point i and $\sum_t \omega_t = 1$.

There is no straightforward way to compute the intensity gradient for point clouds, as for conventional images, so an alternative measure is considered. The *intensity normals* $\boldsymbol{g}_i$ are computed similar to the surface normal, except that the third component of the triangle point is the intensity instead of z: $\boldsymbol{p}_j = (x_j, y_j, I_j)^T$.

A further feature is the intensity entropy. To compute the entropy, all points in the neighborhood of point i are taken. We set the neighborhood size to 2% of the object height. A histogram of the intensity values of all points in the neighborhood is created. With this histogram a numerical probability p_g can be assigned to each gray value $g \in (0,255)$. The intensity entropy is then

$$H(I) = \sum_{g=0}^{255} p_g \cdot \log p_g. \tag{3}$$

We also perform several operations on these features that are inspired by a mean filter, a mean of absolute values filter, and a Laplace filter for conventional images. These three operations are applied to all three components of the surface normal (n_x, n_y, n_z)

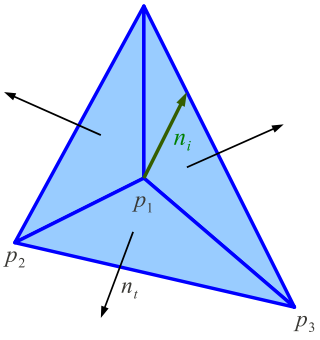

Fig. 1. The surface normal $\boldsymbol{n}_i$ of point i is the average of the surface normals of adjacent triangles.