EDVD – Enhanced Descriptor for Visual and Depth Data

Erickson R. Nascimento and William Robson Schwartz and Mario F. M. Campos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email:{erickson,william,mario}@dcc.ufmg.br

Abstract

Many problems in computer vision and robotics rely on automatically determining point correspondences from two images. Due to issues such as illumination variations, uncontrolled acquisition conditions and noise, this is a challenging problem. This work presents a method that combines visual and shape information to perform point correspondences which is invariant to rotation and scaling transformations in the image and geometry domains. Experimental results show that our approach is a robust and computationally efficient technique compared with classic descriptors in the literature.

1 Introduction

Matching image patches for building accurate 3D models of a scene or for recognizing objects are two examples of essential computer vision tasks that have to be solved accurately and efficiently. Most of the matching approaches rely on image features, therefore, descriptors that are capable to correctly and reliably establish the correspondence between pairs of points play a central role in computer vision.

For several years, textured images have been the choice input for computer vision algorithms since they provide a rich source of information. The Computer Vision literature presents numerous works that use different cues for correspondence based on texture. In virtually all of the approaches based on texture, feature descriptors are estimated from the two-dimensional images, and they seldom use other information such as the geometry of the scene. As a consequence, common issues concerning real scenes, such as variation in illumination and textureless objects, may dramatically decrease the performance of texture-based only descriptors.

With the increased availability of inexpensive real time range sensors, combined depth and visual data are becoming easily and readily obtainable. The fusion of visual (obtained from textured two-dimensional images), and shape (obtained from depth information) cues for object recognition has been increasingly attracting the attention of the computer vision community. Current RGB-D sensors have opened the way to obtain 3D information with unprecedented richness and speed. Therefore, due to the technological advances of RGB-D sensors and the use of large data sets, robust descriptors that efficiently use the available information are becoming critical for several tasks, such as object recognition.

Related Work In the last decade, a large number of techniques, each using different cues to obtain correspondence based on textural information has been reported in the computer vision literature. SIFT [6] and SURF [2] are, probably, the two most popular ones. More recently, new approaches that use feature descriptors based on local gradients such as [3, 8], have been introduced. Most of them bring forth descriptors that are invariant to rotation and/or scaling, and some even present reduced memory consumption and short processing time. However, those methods extract features only from two-dimension images. Therefore, they are more sensitive to variations in illumination and loose performance for images of textureless objects.

Depth information acquired by active 3D sensors is less sensitive to lighting conditions, since they typically cast structured IR lighting on the scene. These type of sensors are the core of current RGB-D sensors it has paved the way for the design of robust descriptors that wisely use the multiple sources of data. The MeshHOG [11] and CSHOT [10] are recent works that fuse visual and shape information. MeshHOG uses the texture information of 3D models as scalar functions defined over a 2D manifolds. CSHOT, on the other hand, creates the descriptor as a concatenation of two histograms: The first one is a histogram of the geometric features over

the spherical support around the keypoint, and the other is built from the sum of the absolute differences between the RGB triples of the each of its neighboring points.

Contributions In this paper, we present a novel and enhanced RGB-D descriptor, called EDVD, which efficiently combines visual and shape information to substantially improve discriminative power, enabling high matching performance. Unlike most current methodologies, our approach includes in its design, scale and rotation transforms in both image and geometrical domains. Experimental results show that our approach is robust and computationally efficient when compared to other well-known descriptors available in the literature.

2 Methodology

Let the pair (I, D) denote the output of a RGB-D system for which $I(\mathbf{x})$ and $D(\mathbf{x})$ provide color and depth information for a pixel $\mathbf{x}$, respectively. And $\mathcal{K}$ a list of detected keypoints. We provide the normal estimation for all $\mathbf{x}$ as a map N, where $N(\mathbf{x})$ is efficiently estimated by PCA over the surface defined by the depth map. Our descriptor is constructed for a small image patch, $\mathbf{p}$, centered at a keypoint $\mathbf{k} \in \mathcal{K}$. We use $p_i(\mathbf{x})$ and $p_n(\mathbf{x})$ to denote the pixel intensity and surface normal for a pixel $\mathbf{x} \in \mathbf{p}$, respectively.

In the first step (to exploit the visual information), we extract the visual features based on the direction of the gradient around a keypoint. The idea behind this step is similar to the one used by the Local Binary Patterns (LBP) [7]. Then, the gradient directions are computed using simple intensity difference tests, which have small memory consumption and modest processing time.

In the second step (to exploit the shape information), we build a rotation invariant representation based on the normals direction using an extended Gaussian image followed by the application of the Fourier transform. Finally, we concatenate both visual and shape vectors to create a robust descriptor that also improves discriminative power. This process is illustrated in Figure 1.

Patch Scale and Orientation Differently from image-based descriptors, which need to compute a pyramid and represent the scale of the keypoints using the scale-space, we directly use the depth information, d, from the RGB-D data to define the scale factor s:

$$s = \max\left(0.2, \frac{3.8 - 0.4\max(d_{\min}, d)}{3}\right), \quad (1)$$

which linearly scales the radius of a circular patch $\mathbf{p}$ from 9 to 24, and filters out depths with values less than $d_{\min}$ (in this work we used $d_{\min} = 2$ meters).

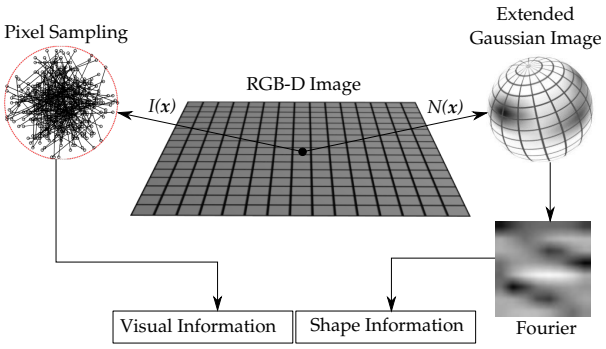

Figure 1. The proposed descriptor combines shape and visual information based on invariant measurement in both domains.

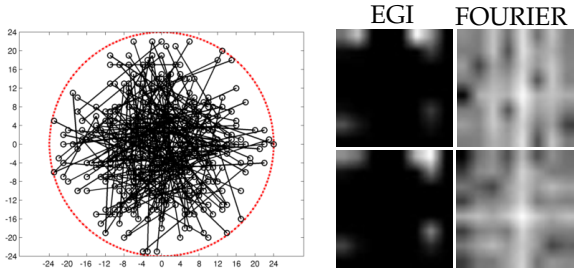

Figure 2. (a) Patch $\mathbf{p}$ with 48×48 pixels indicating 256 sampled pairs of pixel locations used to extract the visual information of the image; (b) Fourier transform highlighting differences.

For each keypoint, we compute the predominant orientation ω, with the fast orientation estimator presented in [2]. The orientation assignment for each keypoint is estimated by computing the Haar wavelet responses in both x and y directions. Unlike [2], that computes the radius of the circular neighborhood around the keypoint using the scale factor s at which the keypoint was detected, we use the keypoint depth information acquired from the RGB-D data. This value is used to scale the size of wavelets and to determine the standard deviation of the Gaussian used to weigh the wavelet.

Visual Information Given an image keypoint $\mathbf{k} \in \mathcal{K}$, assume an image patch $\mathbf{p}$ of size $S \times S$ (in this work we consider $18 \leq S \leq 48$) centered at $\mathbf{k}$. Figure 2(a) illustrates the patch where the set of pixel pairs $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbf{p}$ are indicated with line segments. We use a fixed pattern with locations given by an isotropic Gaussian distribu-

tion $\mathcal{N}(0, \frac{48^2}{25})$ for sampling pixel pairs, inspired by the work of [3]. However, differently from that work, we remove all pairs for which one or both points lay outside of the circle with radius equals to 24. We also smooth the patch with a Gaussian kernel with $\sigma = 2$ and a window with 9×9 pixels to decrease the sensitivity to noise and increase the stability in the pixel comparisons.

Let the fixed set of sampled pairs from $\mathbf{p}$ be $S = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, 256\}$. Before constructing the visual feature descriptor, the patch $\mathbf{p}$ is translated to the origin and then rotated and scaled by the transformation $\mathbf{T}_{\omega,s}$, which produces a set P, where

$$P = \{(\mathbf{T}_{\omega,s}(\mathbf{x}_i), \mathbf{T}_{\omega,s}(\mathbf{y}_i)) | (\mathbf{x}_i, \mathbf{y}_i) \in S\}. \quad (2)$$

This transformation normalizes the patch to allow comparisons between patches. Then, for each pair $(\mathbf{x}_i, \mathbf{y}_i) \in P$, we evaluate

$$f(\mathbf{x}_i, \mathbf{y}_i) = \begin{cases} 1 & \text{if } p_i(\mathbf{x}_i) < p_i(\mathbf{y}_i) \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where the comparison term captures gradient changes in the keypoint neighborhood. We group the results of eight tests and represent it as a floating point number. Therefore, we can use a vector $\mathbf{V}_v$ with 32 elements to store the results of all 256 comparisons.

Shape Information The second step of our methodology uses orientation histograms to capture the geometric characteristics of the patch $\mathbf{p}$ in the 3D domain. Since orientation histograms are approximations of Extended Gaussian Images (EGI), they constitute a powerful representation invariant to translational shift transformations.

Each normal $p_n(\mathbf{x})$ is represented in spherical coordinates (ϕ, θ). The coordinates ϕ and θ are discretized into 8 values each, and the number of normals falling inside each discretized orientation is accumulated. Figure 1 depicts the accumulation of normal directions in the sphere. Dark spots represent a large number of normals accumulated in that orientation.

Since rotations in the normal orientations become translations in the EGI domain, we apply the Fourier transform in the EGI to obtain a translation invariant Fourier spectrum. Finally, the Fourier spectrum is linearized and converted to a 64-dimension vector $\mathbf{V}_s$. In addition to the rotation invariance, the use of spectral information emphasizes differences among different descriptors (see Figure 2 (b)).

Final Descriptor Once the visual and shape information have been extracted, we concatenate the shape vector $\mathbf{V}_s$ and the visual vector $\mathbf{V}_v$, creating a 96-dimension vector which captures both visual and shape information.

Table 1. Average descriptor creation time (ms) and memory consumption (Kb).

Descriptor	Creation time (ms)	Memory (Kb)
EDVD	0.68	0.375
CSHOT	2.53	5.250

3 Experiments

We validate and compare our descriptor (EDVD) with classical descriptor approaches for two-dimensional images namely, SIFT [6] and SURF [2], with the geometric descriptor, spin-images [4], and the state-of-the-art in fusing both visual and shape information CSHOT [10].

We use the public[1] dataset presented in [9]. It contains several real world sequences of RGB-D data. We used three sequences in that dataset in our experiments: i) *freiburg2_xyz*, in which the Kinect is moving individually along the x/y/z axes; ii) *freiburg2_rpy* where the Kinect is sequentially rotated around the three axes and iii) the hand-held SLAM sequence *freiburg2_desk*.

Matching Performance In order to evaluate the performance of EDVD and to compare with other approaches, we use the criterion presented in [5] (we use Euclidean distance for SURF and SIFT and Correlation for spin-image and our descriptor to compare the keypoints signatures). The plots shown in Figure 3 depict the accuracy for each sequence. The results show that EDVD outperforms all other approaches, including the state-of-the-art CSHOT, for the three sequences, and it is 3.72 times faster than CSHOT using 14 times less memory space (see Table 1. Experiments were executed for 30K descriptors in an Intel Core i5 2.53GHz running only one core).

Rotation Invariance To evaluate our descriptor's robustness to rotation, we applied synthetic in-plane rotations and added Gaussian noise with standard deviation equal to 15. We then computed the keypoint descriptors using our methodology and SURF, and then performed a brute-force matching to find correspondences. The results are given in terms of percentage of inliers as a function of the rotation angle (Figure 4(a)). We also test the noise sensitivity of these algorithms. The results for the synthetic test for added noise with standard deviations of $15, 30, 45, 60$ and 75 are shown in Figure 4(b). It can be seen from the figure that our methodology presents smaller sensitivity to noise.

[1] https://cvpr.in.tum.de/data/datasets/rgbd-dataset

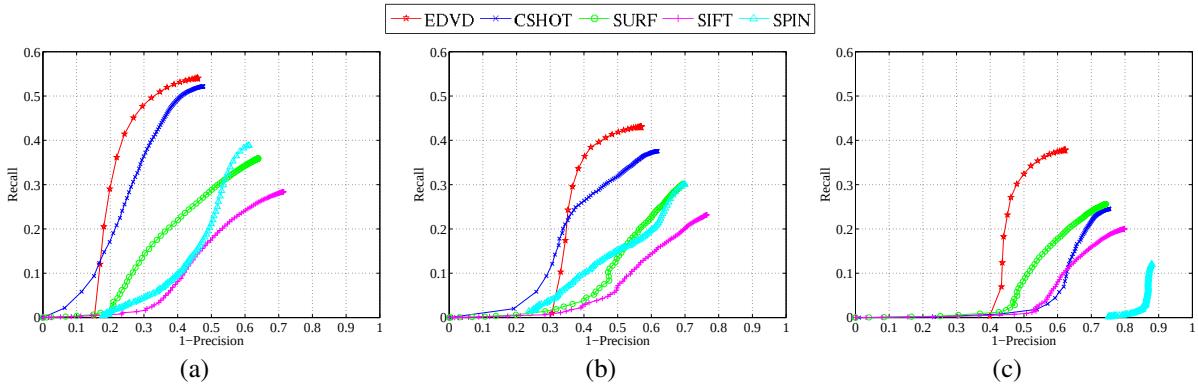

Figure 3. Precision-Recall curves for (a) freiburg2_xyz, (b) freiburg2_rpy and (c) freiburg2_desk. The keypoints were detected using STAR detector [1].

4 Conclusions

We proposed a new descriptor that takes into account both visual and shape information extracted from RGB-D data. Our methodology is robust to orientation and different illumination conditions and outperforms in processing time and memory consumption the standard descriptors in the literature. Our approach outperformed all the other descriptors, including the state of the art CSHOT descriptor, which also fuses both visual and shape information.

Acknowledgment The authors greatfully acknowledge the financial support of CNPq and CAPES.

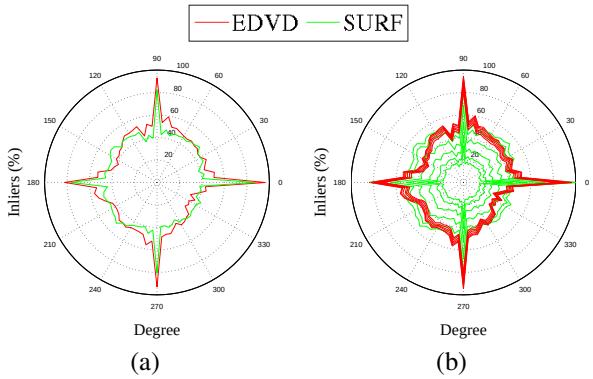

Figure 4. Percentage of inliers as a function of rotation angles. (a) EDVD and SURF matching performance under synthetic rotations with 15 of noise; (b) Matching sensitivity under 0, 15, 30, 45, 60 and 75 levels of noise.

References

[1] M. Agrawal, K. Konolige, and M. R. Blas. CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In *Proc. ECCV*, 2008.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *CVIU*, June 2008.

[3] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary Robust Independent Elementary Features. In *ECCV*, September 2010.

[4] A. E. Johnson and M. Hebert. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. *PAMI*, pages 433–449, 1999.

[5] Y. Ke and R. Sukthankar. PCA-SIFT: A More distinctive Representation for Local Image Descriptors. In *CVPR*, 2004.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, pages 91–110, 2004.

[7] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 1996.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: an efficient alternative to SIFT or SURF. In *ICCV*, Barcelona, November 2011.

[9] J. Sturm, S. Magnenat, N. Engelhard, F. Pomerleau, F. Colas, W. Burgard, D. Cremers, and R. Siegwart. Towards a benchmark for RGB-D SLAM evaluation. In *Proc. of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at RSS*, Los Angeles, USA, 2011.

[10] F. Tombari, S. Salti, and L. D. Stefano. A combined texture-shape descriptor for enhanced 3D feature matching. In *ICIP*, 2011.

[11] A. Zaharescu, E. Boyer, K. Varanasi, and R. P. Horaud. Surface Feature Detection and Description with Applications to Mesh Matching. In *CVPR*, 2009.