A COMBINED TEXTURE-SHAPE DESCRIPTOR FOR ENHANCED 3D FEATURE MATCHING

Federico Tombari, Samuele Salti, Luigi Di Stefano

CVLab - DEIS
University of Bologna, Italy
{ federico.tombari, samuele.salti, luigi.distefano } @unibo.it

ABSTRACT

Motivated by the increasing availability of 3D sensors capable of delivering both shape and texture information, this paper presents a novel descriptor for feature matching in 3D data enriched with texture. The proposed approach stems from the theory of a recently proposed descriptor for 3D data which relies on shape only, and represents its generalization to the case of multiple cues associated with a 3D mesh. The proposed descriptor, dubbed CSHOT, is demonstrated to notably improve the accuracy of feature matching in challenging object recognition scenarios characterized by the presence of clutter and occlusions.

Index Terms— 3D Descriptor, Surface Matching, 3D Features

1. INTRODUCTION

Automatic surface matching is attracting a growing interest in the research community, with applications found in areas such as shape retrieval, shape registration, object manipulation and grasping, robot localization and navigation. An important enabling factor for the development of this technology consists in the increasing availability of cheaper and more effective 3D sensors. Many of these sensors are able to acquire not only the 3D shape of the scene, but also its texture: this is the case, e.g. of stereo sensors, structure-from-motion systems, certain laser scanners as well as the recently proposed *Kinect* device by Microsoft.

In this paper we focus on solving the surface matching problem based on local features, i.e. by point-to-point correspondences obtained by matching local invariant descriptors of feature points. This approach has become the standard paradigm for tackling classical computer vision problems such as object recognition, automatic registration, image indexing, etc...

Several approaches have been proposed for 3D feature point detection and description [1–7]. A review of these methods is beyond the scope of this paper. However, we wish to highlight here that the majority of these proposals detect and describe a feature point by using shape data only. Recently, [4] has proposed the MeshDoG/HoG approach, whereby texture information can also be deployed.

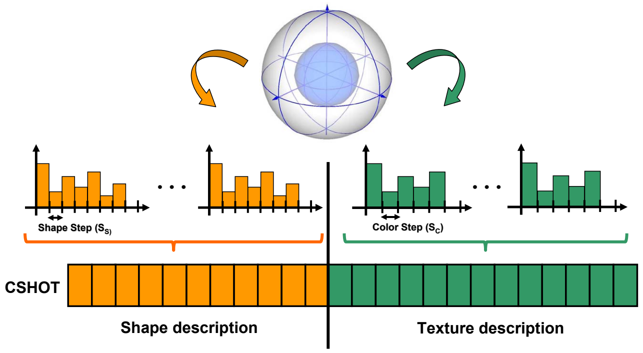

Fig. 1. The proposed descriptor merges together a signature of histograms of normal orientations and of texture-based measurements.

In this work we show that the design of the SHOT descriptor [7] can naturally be extended to incorporate texture (Sec. 2) and that such an extension allows for improved performances on publicly available datasets (Sec. 3). This results in a particularly interesting approach for carrying out surface matching tasks based on the output of modern 3D sensors capable of delivering both shape and texture.

2. COLOR SHOT: A COMBINED TEXTURE-SHAPE 3D DESCRIPTOR

We briefly summarize here the structure of the SHOT descriptor to make the paper self-contained. The reader is referred to [7] for details on the descriptor and a discussion on its properties. First of all, the descriptor relies on the definition of a repeatable local Reference Frame based on the Eigenvalue Decomposition of the scatter matrix of the neighborhood of a point. Given the local RF, an isotropic spherical grid is defined to encode spatially well localized information, i.e. to define a signature structure. For each sector of the grid an histogram of normals is defined and the overall descriptor re-

978-1-4577-1303-3/11/$26.00 ©2011 IEEE

sults from the juxtaposition of these histograms.

To generalize this design so as to include multiple cues, we denote here as $SH_{G,f}(P)$ the generic signature of histograms computed over the spherical support around feature point P. This signature of histograms relies upon two different entities: G, a vector-valued point-wise property of a vertex, and f, the metric used to compare two of such point-wise properties. To compute a histogram of the signature, f is applied over all pairs (G_P, G_Q), with Q representing a generic vertex belonging to the spherical support around feature point P. In the original SHOT formulation [7], G is the surface normal estimation, N, while $f(\cdot)$ is the dot product, denoted as $p(\cdot)$:

$$f\left(G_P, G_Q\right) = p\left(N_P, N_Q\right) = N_P \cdot N_Q \tag{1}$$

In the proposed generalization, m signatures of histograms relative to different *(property, metric)* pairs are computed on the spherical support and chained together in order to build the descriptor $D(P)$ for feature point P:

$$D(P) = \bigcup_{i=1}^{m} SH^{i}_{(G,f)}\left(P\right) \tag{2}$$

Although the formulation in (2) is general, we will hereinafter refer to the specific case of $m = 2$, so as to combine a signature of histograms of shape-related measurements together with a signature of texture-related measurements (Fig. 1). As for the former, we use the formulation of the original SHOT descriptor, i.e. as in (1). As for the latter, since we want here to embed texture information in the descriptor, we have to define a proper vector representing a point-wise property of the texture at each vertex and a suitable metric to compare two such texture-related properties. The overall descriptor, based on two signatures of histograms, will be dubbed hereinafter as Color-SHOT (CSHOT).

The most intuitive choice for a texture-based G vector is the RGB triplet of intensities associated to each vertex, referred to here as R. To properly compare RGB triplets, one option is to deploy the same metric as in SHOT, i.e. to use the dot product $p(R_P, R_Q)$. Alternatively, we have tested another possible metric based on the L_p norm between two triplets. In particular, we have implemented the operator based on the L_1 norm, referred to as $l(\cdot)$, which consists in the sum of the absolute differences between the triplets:

$$l\left(R_P, R_Q\right) = \sum_{i=1}^{3} \left|R_P(i) - R_Q(i)\right| \tag{3}$$

Moreover, we have investigated the possibility of using different color spaces rather than RGB. We have chosen the *CIELab* space given its well-known property of being more perceptually uniform than the RGB space [8]. Hence, as a different solution, vector G is represented by color triplets computed in this space, which will be referred to as C. Comparison between C triplets can be done using the metrics used for R triplets, i.e. the dot product $p(\cdot)$ or the L_1 norm $l(\cdot)$, leading to signatures of histograms relying, respectively, on $p\left(C_P, C_Q\right)$ and $l\left(C_P, C_Q\right)$.

In addition, we have investigated on the use of more specific metrics defined for the *CIELab* color space. In particular, we have deployed two metrics, known as *CIE94* and *CIE2000*, that were defined by the *CIE* Commission respectively in 1994 and 2000: their definitions is not reported here for lack of space and the reader is referred to [8] for additional details. These two metrics lead to two versions of operator $f(\cdot)$ which will be referred to, respectively, as $c_{94}(\cdot)$ and $c_{00}(\cdot)$. Hence, two additional signatures of histograms can be defined based on these two measures, denoted respectively as $c_{94}\left(C_P, C_Q\right)$ and $c_{00}\left(C_P, C_Q\right)$.

The CSHOT descriptor inherits SHOT parameters, i.e. the radius of the support and the number of bins in each histogram). However, given the different nature of the two signatures of histograms embedded in CSHOT, it is useful to allow for a different number of bins in the two histogram types. Thus, the CSHOT descriptor will have an additional parameter with respect to SHOT, indicating the number of bins in each texture histogram and referred to as Color Step (S_C, see Fig. 1).

3. EXPERIMENTAL RESULTS

The 6 different versions defined in Section 2 for the novel CSHOT descriptor are now evaluated in a typical 3D object recognition scenario where one or more objects have to be found in a scene with clutter and occlusions. The experimental evaluation is aimed at determining which version performs best in terms of both accuracy and efficiency. Furthermore, the best versions will be compared against the original SHOT descriptor as well as the MeshHoG descriptor, so as to evaluate the benefits brought in by the proposed approach.

In all experiments, features points are first extracted from a scene and an object, then they are described and matched based on the Euclidean distance between descriptors. As for the feature extraction stage, we rely on the same approach as in [7], i.e. features are first randomly extracted from the object, then the corresponding features are extracted from the scene[1] together with a set of additional features randomly extracted from clutter. All algorithms have been tested by keeping constant their parameters. In particular, all parameters that CSHOT shares with SHOT have been set the values originally proposed in [7]. Such values have been also used here for the tests concerning the SHOT descriptor. As for the additional parameter used by CSHOT (i.e., S_C), it has been tuned for each CSHOT version on a subset, made out of 3 scenes, of the *Spacetime Stereo* used in [7] and available on line[2]. This subset has been used to tune the radius and number of bins

[1] by means of available ground-truth information

[2] available at www.vision.deis.unibo.it/SHOT

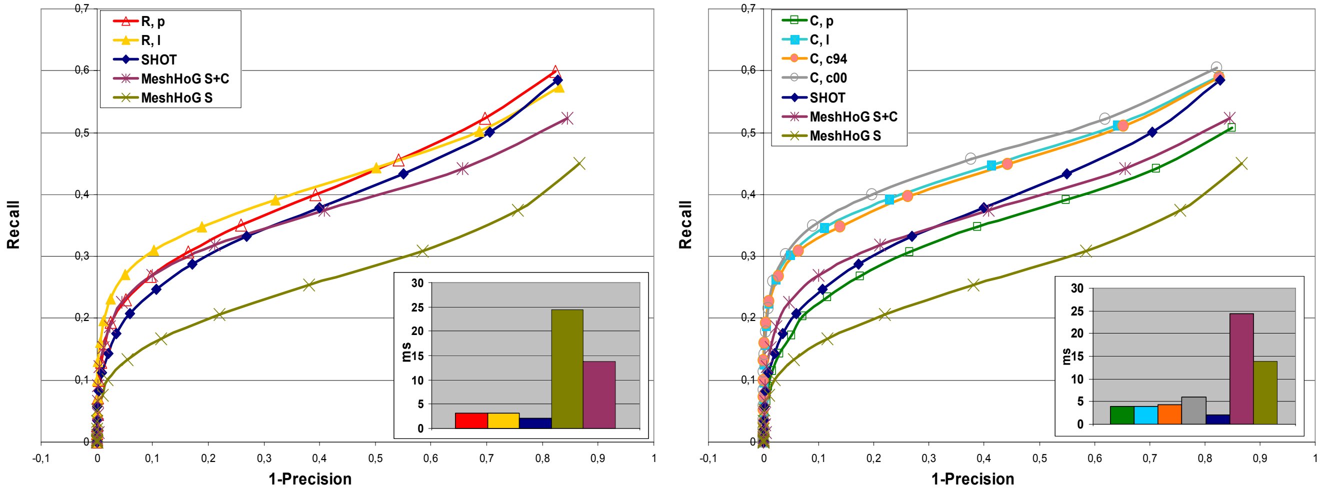

Fig. 2. Comparison in terms of accuracy (big chart) and efficiency (small chart) between CSHOTs with different measures in the *RGB* (left chart) and *CIELab* (right chart) color spaces on *Dataset 1*. SHOT and two variants of MeshHoG are also reported.

of the orientation histograms of MeshHoG, with the other parameters of the method kept as originally proposed in [4].

3.1. Comparison between color spaces and metrics

A first experimental evaluation has been carried out to identify the best CSHOT combinations for, respectively, the *RGB* and the *CIELab* color spaces. Results have been computed on a dataset composed of the 12 scenes not used for tuning of the *Spacetime Stereo* dataset proposed in [7]. This subset, hereinafter referred to as *Dataset 1*, includes scenes with clutter and occlusions of the objects to be recognized.

Figure 2 shows the comparison among the evaluated measures respectively in the RGB (left chart) and CIELab (right chart) color spaces. As for the former, the two *(property, metric)* pairs being compared are: (R, p) and (R, l). As for the latter, four pairs are compared, i.e. : (C, p), (C, l), (C, c_{94}), (C, c_{00}). Each comparison is carried out in terms of accuracy (big chart) and efficiency (small chart). As for the former, results are provided in terms of *Precision vs. Recall* curves computed on the output of the descriptor matching process carried out between the features extracted from the objects and those extracted from the scenes. Each object-scene pair of the dataset is then averaged to give out the final charts shown in the figure. As for efficiency, results are provided as the average amount of time (*ms*) needed to compute one correspondence between the scene and the object.

As for the *RGB* space, (R, l) proves to be more accurate than (R, p), and only slightly less efficient. As for the *CIELab* space, (C, l), (C, c_{94}) and (C, c_{00}) notably outperform (C, p), with (C, l) being slightly more accurate and more efficient than (C, c_{94}), and with (C, c_{00}) being by far the least efficient one. Hence, the two CSHOT versions that turn out more favorable in terms of the accuracy-efficiency trade-off are, respectively, (R, l) for the *RGB* space, and (C, l) for the *CIELab* space.

3.2. Comparison with SHOT and MeshHoG

We will now comment on the comparison between the two best CSHOT versions and the SHOT and MeshHoG descriptors, so as to assess the benefits brought in by the combined deployment of texture and shape in the proposed approach as well as to compare its overall performance with respect to state-of-the-art methods. We tested two versions of MeshHoG: one using only shape, as done by SHOT, and one deploying shape and texture. For shape-only MeshHoG, we used the mean curvature as feature. As reported in the experimental results section of [4] (Sec 6.1), the use of both shape and texture can be achieved by juxtaposing two MeshHoG descriptors, computed respectively using as feature the mean curvature and the color. Conversely to what reported in [4], on our dataset the shape-and-texture version of MeshHoG provides slightly better performance than the texture-only version: thus, it is the one included in our comparison. The two charts in Fig. 2 include the results yielded on *Dataset 1* by SHOT and the two considered variants of MeshHoG . In addition, Fig. 3 reports a further comparison performed among the same proposals on another dataset. This dataset, referred to here as *Dataset 2*, comprises 8 models and 16 scenes(2 models and 4 scenes of this dataset are shown on the left side of the Figure). *Dataset 2* differs from *Dataset 1* because the former includes objects having very similar shapes but different textures (i.e. different types of cans). Hence, it helps highlighting the importance of relying also on texture for the goal of 3D object recognition in cluttered scenes. Similarly

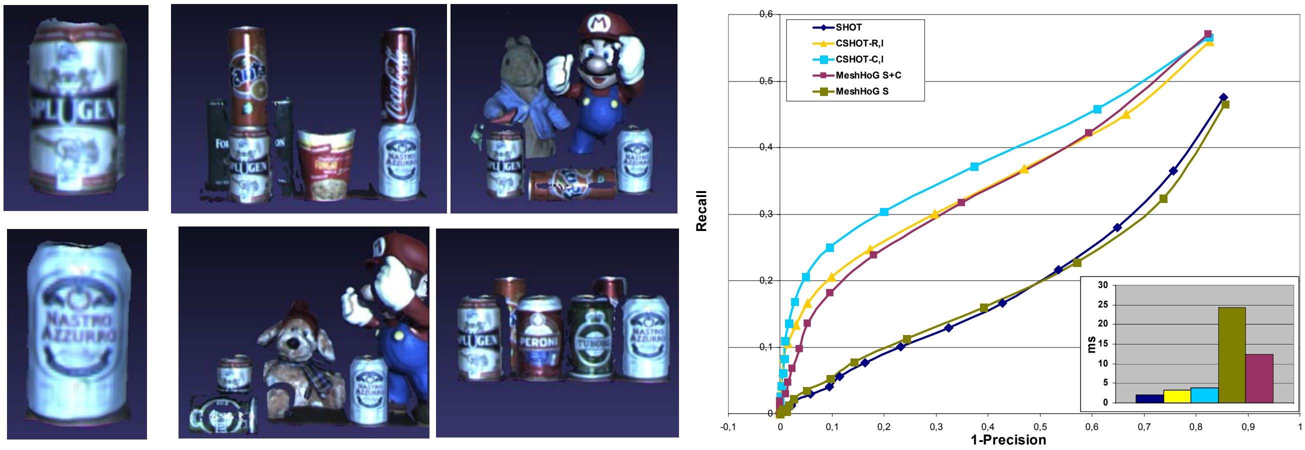

Fig. 3. Left: Two models and four scenes of *Dataset 2*. Right: Comparison in terms of accuracy (big chart) and efficiency (small chart) between the 2 best versions of CSHOT, SHOT and two variants of MeshHoG on *Dataset 2*.

to the previous experiment, results are given both in terms of accuracy (big chart) and efficiency (small chart).

Several observations can be made on these charts. First of all, on both dataset, the two best versions of CSHOT, i.e. (R, l) and (C, l) , notably outperform SHOT and the shape-only version of MeshHoG in terms of accuracy, with the gap in performance being more evident on *Dataset 2*, where the algorithms that only rely on shape fail since they do not hold enough discriminative power to cope with the traits of the dataset. The results on both datasets confirm the benefits of including texture information in the descriptor. Secondly, on both datasets the CSHOT descriptor based on (C, l) proves to be more effective than that relying on (R, l) as well as than the shape and texture version of MeshHoG, thus allowing for state-of-the-art performance on the considered datasets. Finally, as for efficiency, the CSHOT descriptor based on (C, l) is approximately twice as slow as SHOT and one order of magnitude faster than MeshHoG.

4. CONCLUSION

Starting from SHOT [7], a state-of-the-art descriptor for 3D features, we have presented a general formulation for multi-cue description of 3D data by signatures of histograms. We have then proposed a specific implementation of this formulation, CSHOT, that realizes a joint texture-shape 3D feature descriptor. CSHOT has been shown to improve the accuracy of SHOT and to obtain state-of-the-art performance on data comprising both shape and texture. By means of experimental evaluation, different combinations of metrics and color spaces have been tested: the L_1 norm in the *CIELab* color space turns out to be the most effective choices.

5. REFERENCES

[1] A. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3d scenes,” *PAMI*, vol. 21, no. 5, pp. 433–449, 1999.

[2] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik, “Recognizing objects in range data using regional point descriptors,” in *ECCV*, 2004, vol. 3, pp. 224–237.

[3] J. Novatnack and K. Nishino, “Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images,” in *ECCV*, 2008, pp. 440–453.

[4] Andrei Zaharescu, Edmond Boyer, Kiran Varanasi, and Radu P. Horaud, “Surface feature detection and description with applications to mesh matching,” in *Proc. CVPR*, June 2009.

[5] Ajmal S. Mian, Mohammed Bennamoun, and Robyn A. Owens, “On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes,” *IJCV*, vol. 89, no. 2-3, pp. 348–361, 2010.

[6] H. Chen and B. Bhanu, “3d free-form object recognition in range images using local surface patches,” *J. Patt. Recogn. Letters*, vol. 28, no. 10, pp. 1252–1262, 2007.

[7] F. Tombari, S. Salti, and L. Di Stefano, “Unique signatures of histograms for local surface description,” in *Proc. ECCV*, 2010.

[8] M.D. Fairchild, *Color Appearance Models*, John Wiley & Sons Ltd., 2005.