Unique Shape Context for 3D Data Description

Federico Tombari
DEIS/ARCES
University of Bologna
Bologna, Italy
federico.tombari@unibo.it

Samuele Salti
DEIS/ARCES
University of Bologna
Bologna, Italy
samuele.salti@unibo.it

Luigi Di Stefano
DEIS/ARCES
University of Bologna
Bologna, Italy
luigi.distefano@unibo.it

ABSTRACT

The use of robust feature descriptors is now key for many 3D tasks such as 3D object recognition and surface alignment. Many descriptors have been proposed in literature which are based on a non-unique local Reference Frame and hence require the computation of multiple descriptions at each feature points. In this paper we show how to deploy a unique local Reference Frame to improve the accuracy and reduce the memory footprint of the well-known 3D Shape Context descriptor. We validate our proposal by means of an experimental analysis carried out on a large dataset of 3D scenes and addressing an object recognition scenario.

Categories and Subject Descriptors

I.4.8 [**Image processing and computer vision**]: Scene Analysis

General Terms

Theory

Keywords

3D Shape Context, 3D descriptor, surface matching, 3D object recognition

1. INTRODUCTION

Analysis of 3D data is a hot research topic in computer vision due to, one side, the continuous progresses in the field of 3D sensing technologies (laser scanner, stereo vision, Time-of-Flight cameras) and, on the other side, the demand for novel 3D applications calling for increasingly more reliable processing techniques. In this framework, one of the mostly investigated problems during the last decade has been 3D object recognition, that aims at detecting the presence of specific models in 3D data by means of surface analysis. One of the main challenges of 3D object recognition deals with the ability to recognize objects immersed in complex real-world environments, that is, in presence of a cluttered background and -possibly- with the object sought for only partially visible (i.e. *occluded*). Due to these nuisances, early approaches for 3D object recognition relying on global descriptions of the models (e.g. [15, 20]) did not prove effective enough in complex real-world settings. Indeed, a more successful and investigated approach makes use of local description of the 3D shapes through sets of distinctive features, that is performed by means of 3D feature detectors and descriptors [4, 5, 7, 9–11, 13, 14, 16, 17, 22]. With this approach, first salient and repeatable 3D features are extracted from both the model and the scene by means of a feature detector. Next, each feature is associated with a local description of its spatial neighborhood, which is usually invariant to rotation and scale. Finally, scene features are put in correspondence with the model features by matching their local descriptions.

The research activity concerning the field of 3D feature descriptors has been particularly active in the past few years. One of the most prominent proposals is *3D Shape Context* (3DSC) [7], which has shown state-of-the-art performance due to its descriptiveness and robustness to noise (e.g. outperforming the popular Spin Images technique [9]). One characteristic of 3DSC is that it requires the computation of multiple descriptions of each model feature point due to the lacking of the definition of a local Reference Frame (RF) to be associated with each feature point. Analogously to 3DSC, other state-of-the-art methods resort to the use of multiple descriptions to deal with the lacking of a full local 3D RF: this is the case of, e.g., [22] and [5]. More specifically, in [22] 4 different local RFs have to be taken into account for each feature point to deal with the intrinsic ambiguity on the sign of the local RF unit vectors. In other words, these 4 RFs are built up by unit vectors sharing the same directions but having opposite signs. As for [5], the local RF relies on a Repeatable Axis (the normal of the feature point), then a full local RF is defined by adding the direction and sign of the global maximum of the signed height of the 3D curve obtained by intersecting a sphere centered in the point with the surface. Since typically there are multiple maxima with the same height, each feature point is associated with multiple descriptions even in this proposal.

In this paper we show how *3D Shape Context* can benefit of the definition of an unique local 3D RF. In particular, the use of a local RF allows the method to avoid computing multiple descriptions at a given feature point. This directly results in a decreased memory footprint, hence a better scalability toward large 3D database, where the models to be stored could be on the order of thousands of points. Furthermore, as pointed out in Figure 1, we also aim at demonstrating how the proposed approach can yield higher accuracy in the matching stage, in particular by reducing the amount of wrong correspondences, thanks to the smaller set of model features to be put in correspondence with each scene (or query) feature. More specifically, we propose a novel descriptor that relies on the formulation of 3DSC but, differently from the original method, leverages on a recent study highlighting the importance of a repeat-

3DOR'10, October 25, 2010, Firenze, Italy.

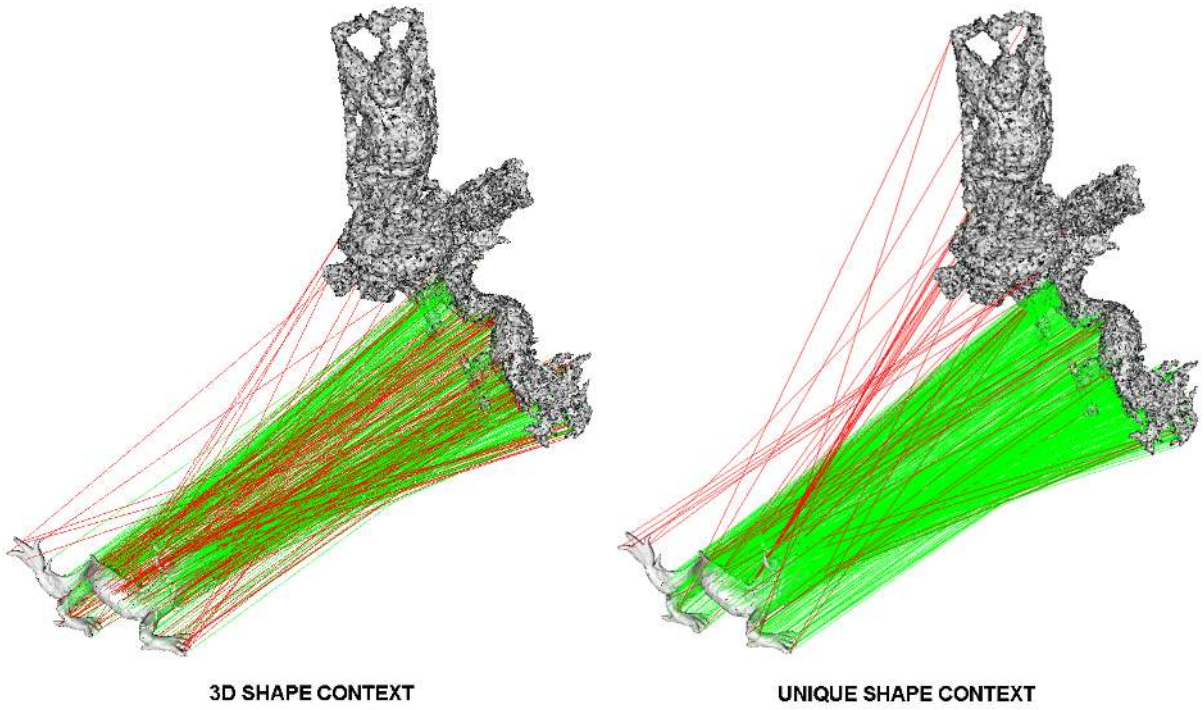

Figure 1: The proposed approach (right) increases the ability of 3D Shape Context (left) to find correct feature correspondences in cluttered and noisy point clouds (green lines: correct correspondences; red lines: wrong correspondences).

able and unambiguous local RF [18] in order to achieve a unique description at each feature point. We validate our proposal on a dataset composed of 60 scenes and by means of two different kinds of experiments: one addresses a typical feature matching for object recognition scenario, while the other one concerns object retrieval in a cluttered environment.

2. RELATED WORK

The 3DSC descriptor has been proposed in [7] and represents an extension to the 3D space of the 2D shape context introduced in [2]. It relies on a specific subdivision of the spherical volume around the feature point that needs to be described. In particular, as shown in Figure 2, a spherical grid is defined by means of subdivisions along the azimuth, elevation and radial dimensions. While the subdivision along the former 2 dimensions is equally spaced, that along the latter is logarithmically spaced. To account for generality, the number of subdivisions can be different along each dimension. Each bin of the resulting spherical grid accumulates a weighted sum of the surface points falling thereby, where each weight is inversely proportional to the local point density and the bin volume.

To achieve repeatability and to allow for comparison of 3DSC descriptors in different coordinate systems, the placement of the spherical grid around the feature point has to be unique, i.e. invariant to rotations and translations of the model and robust to the presence of noise. To this purpose, the north pole of the sphere is oriented with the estimation of the surface normal at the feature point. Nevertheless, this still leaves one degree of freedom for the reference vectors on the tangent plane. To solve this, a vector is randomly chosen on the plane and the L azimuth subdivisions of the descriptor sphere are created using it. This defines a set of L discrete rotations on the tangent plane, and each rotation is taken into account to build up a local 3D Reference Frame used to orientate

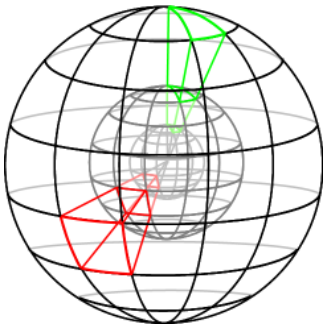

Figure 2: Spatial subdivision applied by the 3DSC descriptor over the spherical local support of a 3D feature.

the sphere grid. Hence, for each reference feature point (i.e. that referring to the model and computed offline) a total of L descriptors needs to be computed and stored to account for all the possible positions of the reference azimuth direction. Other than an increased memory footprint, this slows down the matching stage, where each scene feature has to be matched with all the possible L "variants" of each model feature. In addition, this overhead represents a disadvantage also in terms of matching accuracy, since a higher number of model feature descriptors can cause mismatches in terms of retrieved correspondences. In [7] an indexing scheme to speed up the matching stage is proposed, where the similarity measure is the Euclidean distance among a descriptors pair. Results shown in [7] demonstrate the higher descriptiveness and robustness of 3DSC compared to the popular Spin Images technique [9] in a typical object recognition scenario.

3. THE PROPOSED UNIQUE SHAPE CONTEXT DESCRIPTOR

As highlighted in the previous Section, 3DSC represents a state-of-the-art 3D feature descriptor technique. From our analysis, the main weak point of this technique is represented by the lacking of a repeatable local Reference Frame, with the aforementioned negative consequences in terms of memory footprint, efficiency and also accuracy. The main contribution of this work is the proposal of an improved Shape Context method which does not need to compute the descriptor over multiple rotations on different azimuth directions. For this aim, we propose to employ a unique, unambiguous local RF that can yield not only a repeatable normal axis, but also a unique pair of directions lying on the tangent plane. The main advantage of such an approach, that we dub *Unique Shape Context* (USC), is that it requires to compute, at each model feature, one single descriptor over the rotation indicated by the repeatable vectors over the tangent plane. This not only reduces the overall memory footprint, but also, and more importantly, notably simplifies the matching stage, since every scene feature needs to be matched with one single descriptor instance for every model feature. If standard matching approaches are employed, i.e. each scene feature is matched by exhaustively searching over all model features (*brute force*), the proposed approach allows speeding-up the matching stage given the highly reduced number of model features that each scene feature needs to be matched to. Instead, if efficient approximate indexing schemes are used to speed-up the matching stage (such as those based on Locality-Sensitive Hashing, as originally proposed in [7], or Kd-trees [1]) the main benefit attained by the use of USC in spite of 3DSC is the reduction of spurious correspondences that arise due to the higher number of possible matching candidates.

The definition of a unique, unambiguous local RF has been lacking in literature until recently [18]. Hence, in our approach, we employ the unique and unambiguous local RF proposed in [18], which is briefly reviewed here. Given a feature point p and a spherical neighborhood centered on p and defined by radius R, a weighted *covariance matrix* $\mathbf{M}$ of the points within the neighborhood is computed as:

$$\mathbf{M} = \frac{1}{Z} \sum_{i:d_i \leq R} (R - d_i)(\mathbf{p}_i - \mathbf{p})(\mathbf{p}_i - \mathbf{p})^T \quad (1)$$

where $d_i = ||\mathbf{p}_i - \mathbf{p}||_2$ and Z is a normalization factor computed as:

$$Z = \sum_{i:d_i \leq R} (R - d_i) \quad (2)$$

The weighting scheme embedded in (1) is performed so as to assign smaller weights to distant points, i.e. those more likely laying out of the object of interest. Then, a Total Least Squares (TLS) estimation of the 3 unit vectors of the local RF of p is performed by computing the EigenVector Decomposition (EVD) of $\mathbf{M}$. The TLS estimation of the normal direction is given by the eigenvector corresponding to the smallest eigenvalue of M. Unfortunately, as discussed in [3], due to the EVD ambiguity ,the sign of each unit vector obtained by such a procedure is itself a numerical accident, thus not repeatable. Hence, a further stage in the computation of the local RF proposed in [18] carries out a sign disambiguation step so as to yield a fully repeatable local RF. More specifically, the sign of each eigenvector is re-oriented so that it is coherent with the majority of the vectors it is representing. This is performed for the first and the third eigenvectors (i.e. those corresponding, respectively, to the biggest and smallest eigenvalues), while the sign of the last is obtained via the cross product due to the orthonormal constraint.

Hence, overall, for each feature point the USC descriptor is computed as follows. First, the aforementioned local RF is computed over a specific support around the feature point. Then, the spherical volume around the feature point is uniquely subdivided by means of a spherical grid oriented with the 3 repeatable directions yielded by the local RF. Each bin of the grid then accumulates a weighted sum of the surface points falling thereby analogously to the approach used by 3DSC. It is worth pointing out that, if the local support for the computation of the local RF is the same as that used to compute the descriptor (i.e. they have the same radius), the computation of the points falling within the two supports needs to be done only once, with benefits in terms of efficiency: in this specific case, in fact, the additional burden brought in by the local RF at description stage has an almost negligible overhead compared to the computational load of the 3DSC descriptor.

4. EXPERIMENTAL RESULTS

This Section aims at assessing the benefits brought in by the proposed USC descriptor by comparing it with the original 3DSC formulation in a typical feature matching scenario, which is the core stage of any object recognition process relying on 3D feature descriptors (Experiment 1). Additionally, we also propose qualitative results concerning a typical object retrieval experiment in a cluttered scene (Experiment 2).

In Experiment 1 we use two datasets composed of several scenes and models. More specifically, one dataset, referred to as *Stanford*, is composed of 6 models ("Armadillo", "Asian Dragon", "Thai Statue", "Bunny", "Happy Buddha", "Dragon") taken from the *Stanford 3D Scanning Repository* [1]. Given these models, 45 scenes are built up by randomly rotating and translating different subsets of the model set in order to create clutter[2]; then, similarly to [19], we add Gaussian random noise with increasing standard deviation, namely σ_1, σ_2 and σ_3 at respectively 10%, 20% and 30% of the average mesh resolution (computed on all models). A sample scene of the *Stanford* dataset is shown in Figure 3. The other dataset, referred to as *Lab*, consists of scenes and models acquired in our laboratory by means of a 3D sensing technique known as *Spacetime Stereo* [6], [21]. In particular, we compare 8 object models against 15 scenes characterized by clutter and occlusions, each scene containing two models. Some sample scenes of the *Lab* dataset are shown in Figure 4.

In this feature matching experiment we use the same feature detector for both algorithms: we randomly extract a set of feature points from each model, then we extract their corresponding points from the scene, so that performance of the descriptors is not affected by errors in the detection stage. More specifically, we extract 1000 feature points from each model in both datasets, while in the scenes we extract $n * 1000$ features per scene (n being the number of models in the scene) in the *Stanford* dataset, and 3000 features per scene in the *Lab* dataset.

Analogously, for what concerns the matching stage, we adopt the Euclidean distance as the matching measure for both algorithms. For each scene and model, we match each scene feature against all model features. We adopt a Kd-tree based indexing scheme which, for each given scene descriptor, finds its closest model descriptor as that yielding the minimum Euclidean distance (though with a different indexing algorithm, this same strategy is that originally proposed in [7]). According to the methodology for evaluation of 2D descriptors recommended in [12], we evaluate the accuracy of both methods in terms of *Precision-Recall* curves.

[1] `http://graphics.stanford.edu/data/3Dscanrep`

[2] 3 sets of 15 scenes each, containing respectively 3, 4 and 5 models

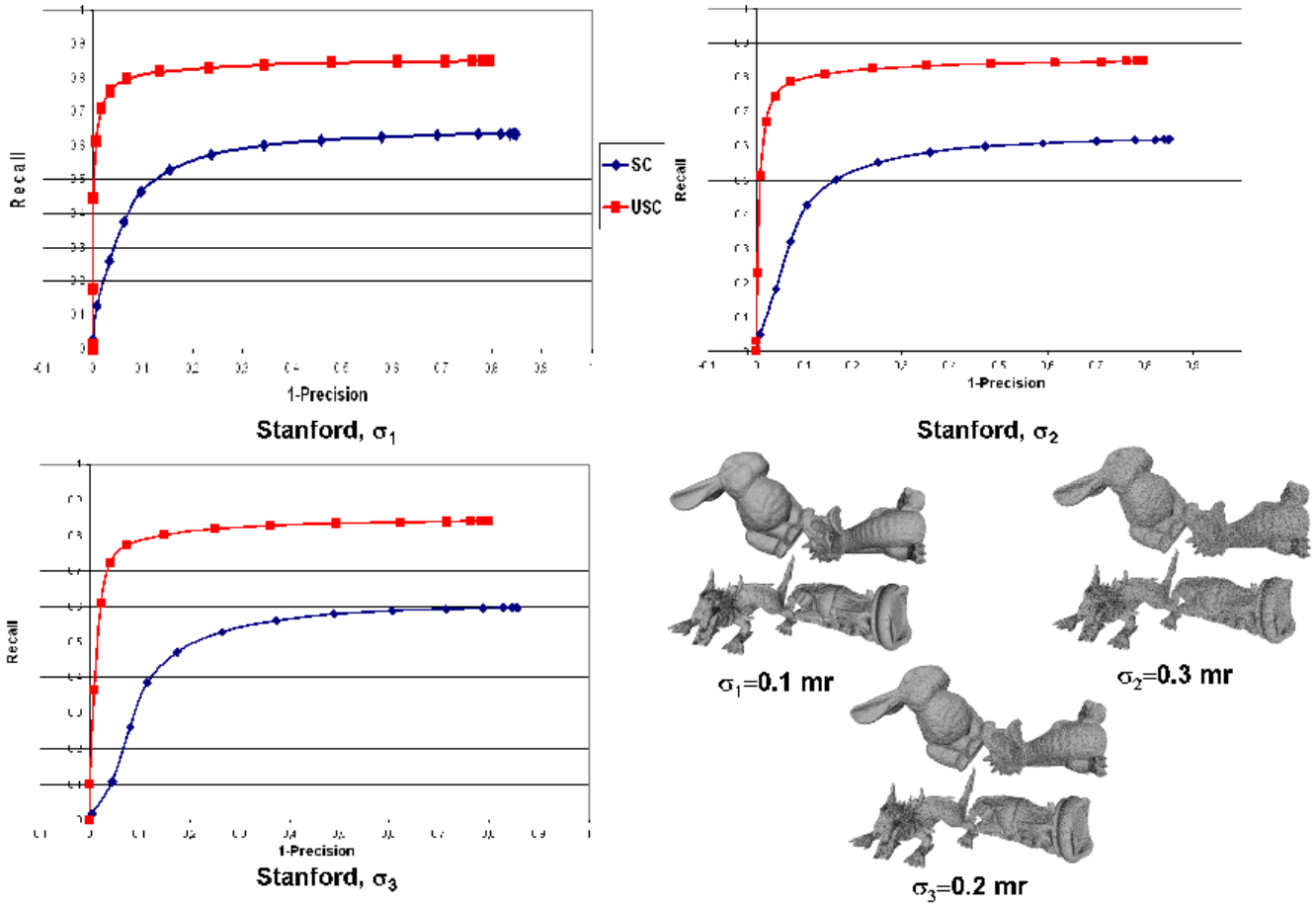

Figure 3: Experiment 1: *Precision-Recall* curves yielded by 3DSC and USC on the *Stanford* dataset at the 3 different noise levels Bottom-right: a scene from the dataset at the 3 different noise levels.

Parameter	3DSC	USC
Max Radius (r_{max})	20 m_r	20 m_r
Min Radius (r_{min})	0.1 r_{max}	0.1 r_{max}
RF Radius	//	20 m_r
Radial div. (J)	10	10
Elevation div. (K)	14	14
Azimuth div. (L)	16	14
Density Radius (δ)	2 m_r	2 m_r

Table 1: Tuned parameters for USC and 3DSC descriptors (m_r: mean mesh resolution of the models).

As for parameters, we have performed a tuning of both 3DSC and USC over a tuning scene corrupted with noise level σ_1 and built by rotating and translating three Stanford models ("Bunny", "Happy Buddha", "Dragon"). As for USC, we have constrained the support used to calculate the local RF to be the same as that used for description, to account for improved efficiency as previously explained. The parameter values obtained by the tuning procedure are then kept constant throughout all the experiments concerning the two datasets. Table 1 reports the tuned values for the parameters used by 3DSC and USC. For each parameter it shows, between brackets, the parameter names as originally defined in [7].

Figures 3 and 4 show the results yielded by USC and 3DSC and concerning Experiment 1. In particular, the 3 charts in Figure 3 are relative to the *Stanford* dataset, one for each of the 3 different noise levels added to each scene, while the chart in Figure 4 reports the result concerning the *Lab* dataset. As it can be clearly observed from the Figures, the novel USC approach yields substantial performance improvements with respect to original 3DSC proposal in the experiments concerning both datasets. More specifically, it yields improved results compared to 3DSC both in the case of data with synthetic noise as well as in the case of more realistic 3D data acquisition conditions. It is also worth pointing out that USC, similarly to 3DSC, shows small performance degradation (hence, good robustness) with respect to increasing noise.

In terms of efficiency, the use of the indexing scheme notably decreases the higher number of descriptors needed to be considered during the matching stage by 3DSC. On the other hand, the additional overhead required by USC is mainly represented by the computation of the local RF at description time, which tends to have a relatively small computational cost compared to the total load of the algorithm. Overall, the two effects tend to cancel out and our experiments showed practically equal measured execution times for USC and 3DSC. Nevertheless, in terms of memory requirements,

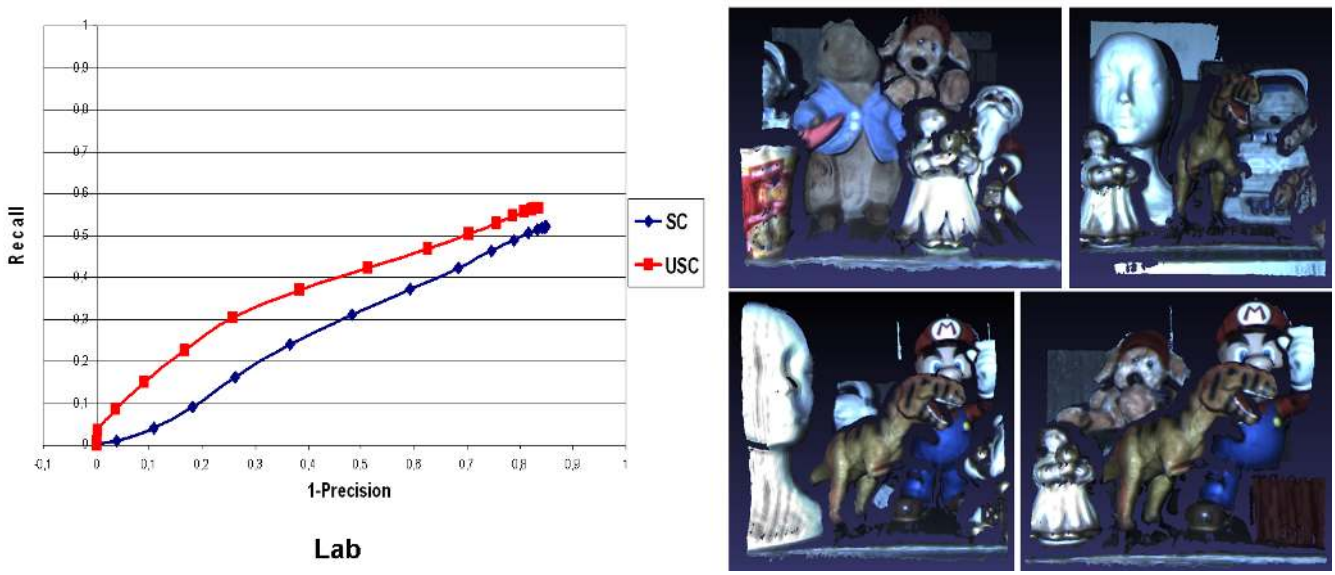

Figure 4: Experiment 1: left, *Precision-Recall* curves yielded by 3DSC and USC on the *Lab* dataset. Right, some scenes from the dataset.

and as previously mentioned, 3DSC has a memory footprint that is L times that of USC: in the experiments shown in this Section, as reported by Table 1, L was tuned to 16.

As for Experiment 2, we propose some qualitative results concerning an experiment of object retrieval from a cluttered scene. In particular, we have built up a database of 6 models, taken from the *Lab* dataset. We then aim at retrieving each object from a cluttered scene (also taken from the *Lab* dataset) that includes only 2 models out of the 6 in the database. The parameters of each descriptor are the same as those tuned in Experiment 1. After the matching stage, in order to compute a subset of reliable correspondences out of those yielded by the matcher, the Geometric Consistency approach is used, which is a typical object recognition algorithm [4,9]. More specifically, starting from a seed feature correspondence, correspondence grouping is carried out by iteratively aggregating those correspondences that satisfy geometric consistency constraints. If the number of the most numerous set of grouped correspondences is higher than a threshold, the object is found in the scene and a final stage based on Absolute Orientation [8] is performed in order to retrieve the pose of the object.

Figure 5 shows for both descriptors the 6 models and the scene used in the experiment, and reports with colored bounding boxes the objects from the database which were found by the algorithm, showing in the scene and with the same color the estimated object pose. As it can be seen, the USC descriptor yields an improved accuracy with regards to 3DSC, since it yields no false positives (one false positive is found by 3DSC). Also, the average pose estimation error for the two models present in the scene, computed in terms of RMSE, is $2.7 \cdot m_r$ (m_r being the mean mesh resolution of the models) for USC, against the $3.1 \cdot m_r$ yielded by 3DSC.

5. CONCLUSION

Inspired by a very recent study on the importance of the local RF for local 3D description, in this paper we have investigated on the deployment of a unique and unambiguous local 3D RF together with the well-known, state-of-the-art 3D Shape Context descriptor. Thanks to the use of a unique local RF, our proposal, dubbed Unique Shape Context, needs not to resort to multiple descriptions at each feature point. This significantly decrease memory occupancy and holds the potential to improve the accuracy of the feature matching stage. The latter benefit is vouched by our experimental evaluation, which shows how the proposed approach compares favorably with respect to the original 3D Shape Context method.

6. REFERENCES

[1] J. Beis and D. Lowe. Shape indexing using approximate nearest-neighbour search in high dimensional spaces. In *Proc. CVPR*, pages 1000–1006, 1997.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI*, 24(4):509–522, 2002.

[3] R. Bro, E. Acar, and T. Kolda. Resolving the sign ambiguity in the singular value decomposition. *J. Chemometrics*, 22:135–140, 2008.

[4] H. Chen and B. Bhanu. 3d free-form object recognition in range images using local surface patches. *Patt. Rec. Letters*, 28:1252–1262, 2007.

[5] C. S. Chua and R. Jarvis. Point signatures: A new representation for 3d object recognition. *IJCV*, 25(1):63–85, 1997.

[6] J. Davis, D. Nehab, R. Ramamoothi, and S. Rusinkiewicz. Spacetime stereo : A unifying framework for depth from triangulation. *PAMI*, 27(2):1615–1630, 2005.

[7] A. Frome, D. Huber, R. Kolluri, T. Bülow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *ECCV*, volume 3, pages 224–237, 2004.

[8] B. Horn. Closed-form solution of absolute orientation using unit quaternions. *J. Optical Society of America A*, 4(4):629–642, 1987.

[9] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *PAMI*, 21(5):433–449, 1999.

[10] A. Mian, M. Bennamoun, and R. Owens. A novel

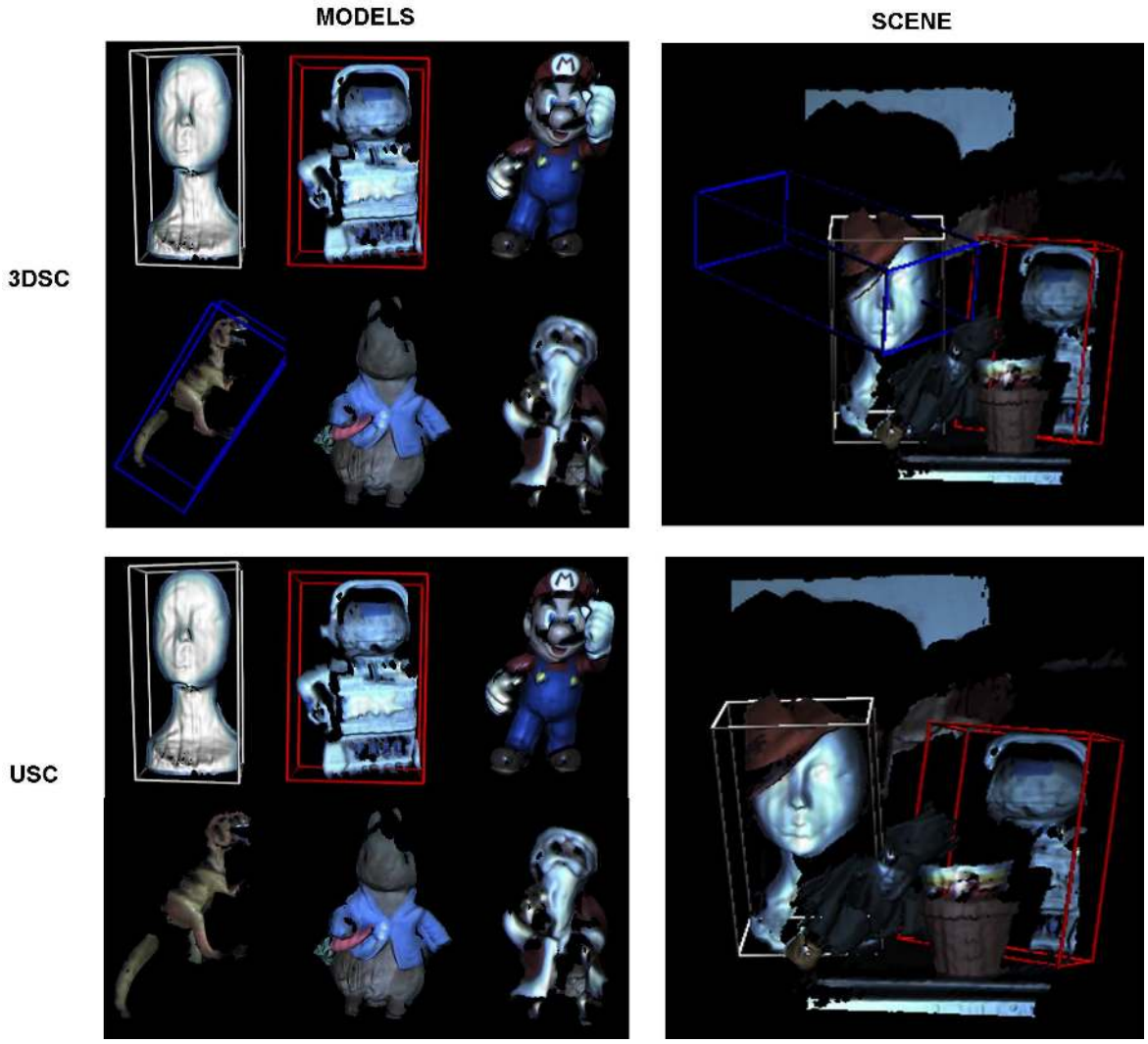

Figure 5: Experiment 2 (object retrieval in a cluttered scene) comparing 3DSC and USC. Retrieved models in the scene are marked with a colored bounding box, which also shows the estimated object pose. 3DSC yields a false positive, while USC correctly finds only the 2 models present in the scene.

representation and feature matching algorithm for automatic pairwise registration of range images. *IJCV*, 66(1):19–40, 2006.

[11] A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *Int. J. Computer Vision*, page to appear, 2009.

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[13] J. Novatnack and K. Nishino. Scale-dependent 3d geometric features. In *Proc. Int. Conf. on Computer Vision*, pages 1–8, 2007.

[14] 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.

[15] F. Solina and R. Bajcsy. Recovery of parametric models from range images: the case for superquadrics with global deformations. *PAMI*, 12(2):131–147, 1990.

[16] F. Stein and G. Medioni. Structural indexing: Efficient 3-d object recognition. *PAMI*, 14(2):125–145, 1992.

[17] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proc. Eurographics Symposium on Geometry Processing*, 2009.

[18] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *Proc. 11th Europ. Conf. on Computer Vision (ECCV 10)*, 2010.

[19] R. Unnikrishnan and M. Hebert. Multi-scale interest regions from unorganized point clouds. In *CVPR-WS: S3D*, 2008.

[20] K. Wu and M. Levine. Recovering parametrics geons from multiview range data. In *CVPR*, pages 159–166, 1994.

[21] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR*, 2003.

[22] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *ICCV-WS: 3dRR*, 2009.