Unique Signatures of Histograms for Local Surface Description

Federico Tombari, Samuele Salti, and Luigi Di Stefano

CVLab - DEIS, University of Bologna,
Viale Risorgimento, 2 - 40135 Bologna, Italy
{federico.tombari,samuele.salti,luigi.distefano}@unibo.it
http://www.vision.deis.unibo.it

Abstract. This paper deals with local 3D descriptors for surface matching. First, we categorize existing methods into two classes: *Signatures* and *Histograms*. Then, by discussion and experiments alike, we point out the key issues of uniqueness and repeatability of the local reference frame. Based on these observations, we formulate a novel comprehensive proposal for surface representation, which encompasses a new unique and repeatable local reference frame as well as a new 3D descriptor. The latter lays at the intersection between Signatures and Histograms, so as to possibly achieve a better balance between descriptiveness and robustness. Experiments on publicly available datasets as well as on range scans obtained with *Spacetime Stereo* provide a thorough validation of our proposal.

1 Introduction and Previous Work

The ability of computing similarities between 3D surfaces, sometimes referred to as *surface matching* [1], is a key for computer vision tasks such as 3D object recognition and surface alignment. These tasks find numerous applications in fields such as robotics, automation, biometric systems, reverse engineering, search in 3D object databases [1] [2] [3].

There has been strong research interest in surface matching since the 1980's. Early works were based on fitting 3D data with global parametric surfaces such as *geons* [4] or *superquadrics* [5]. For the last 15 years though, the most popular trend for surface matching exploits a compact local representation of the input data, known as *descriptor*, and shares basic motivations with the successful approaches for matching 2D images that rely on local invariant features. Local correspondences established by matching 3D descriptors (Fig. 1) can then be used to solve higher level tasks such as 3D object recognition. This approach allows for dealing effectively with issues such as occlusion, clutter and changes of viewpoint. As a result, a variety of proposals for 3D descriptors can be found in recent literature.

In Table 1 we propose a categorization of the main proposals in the field. As shown in the second column, we divide proposals for 3D descriptors into two main categories, namely *Signature* and *Histogram*. The first category, that includes earliest works on the subject, describes the 3D surface neighborhood of a given point (hereinafter *support*) by defining an invariant local Reference Frame (RF) and encoding, according to the local coordinates, one or more geometric measurements computed individually on each

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part III, LNCS 6313, pp. 356–369, 2010.

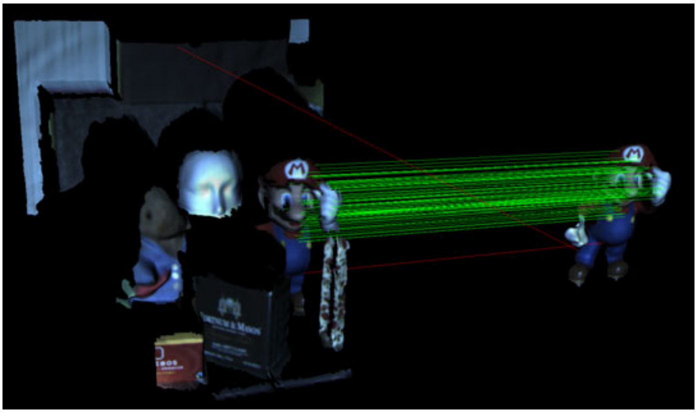


Fig. 1. Example of matching local descriptors in an 3D object recognition scenario. Green lines identify correct matches, whereas red ones represent wrong correspondences.

point of a subset of the support. On the other hand, Histogram-based methods describe the support by accumulating local geometrical or topological measurements (e.g. point counts, mesh triangle areas) into histograms according to a specific quantized domain (e.g. point coordinates, curvatures) which requires the definition of either a Reference Axis (RA) or a local RF. In broad terms, signatures are potentially highly descriptive thanks to the use of spatially well localized information, whereas histograms trade-off descriptive power for robustness by compressing geometric structure into bins.

As far as Signature-based methods are concerned, one of the first proposals is *Structural Indexing* [6], which builds up a representation based on either a *3D curve* or a *Splash* depending on the characteristics of the 3D support. The former encodes the angles between consecutive segments of the polygonal approximation of edges (corresponding to depth or orientation discontinuities) on the surface. The latter encodes as a 3D curve the local distribution of surface orientations along a geodesic circle centered on the point. In *Point Signatures* [7] the signature is given by the signed height of the 3D curve obtained by intersecting a sphere centered in the point with the surface. *3D Point Fingerprint* [8] encodes the normal angle variations and the contour radius variations along different geodesic circles projected on the tangent plane. Recently, *Exponential Mapping* [9] proposed a descriptor that encodes the components of the normals within the support by deploying a 2D parametrization of the local surface.

As for Histogram-based methods, those relying on the definition of just a RA are typically based on the feature point normal. For example, *Spin Images* [1], arguably the most popular method for 3D mesh description, computes 2D histograms of points falling within a cylindrical volume by means of a plane that ”spins” around the normal. Within the same subclass, *Local Surface Patches* [10] computes histograms of normals and *shape indexes* [11] of the points belonging to the support. As for methods relying on the definition of a full local RF, *3D Shape Context* [12] modifies the basic idea of Spin Images by accumulating 3D histograms of points within a sphere centered at the feature point. *Intrinsic Shape Signatures* [13] proposed an improvement of [12] based on a different partitioning of the 3D local volume as well as on a different definition of the local RF. Finally, Mian et al. [2] accumulate 3D histograms (*Tensors*) of mesh triangle areas within a cubic support.

As pointed out in Tab. 1, all proposals rely on the definition of a local RF or, at least, a repeatable RA. However, we believe that the importance of the choice of the local reference for a 3D descriptor is underrated in literature, with efforts mainly focused on the development of discriminative descriptors. As a consequence, approaches for the choice of the local reference are ambiguous, or not unique, or too sensitive to noise and also lack specific experimental validation. Instead, as we will show in the remainder of the paper, the repeatability of the local RF (or, analogously, of the RA) is mandatory to achieve effective local surface description.

Table 1. Taxonomy of 3D descriptors

Method	Category	Local RF	
		Unique	Unambig.
StInd [6]	Signature	No	Yes
PS [7]	Signature	No	Yes
3DPF [8]	Signature	No	Yes
EM [9]	Signature	Yes	No
SI [1]	Histogram	RA	
LSP [10]	Histogram	RA	
3DSC [12]	Histogram	No	Yes
ISS [13]	Histogram	Yes	No
Tensor [2]	Histogram	No	Yes
SHOT	**Both**	**Yes**	**Yes**

Therefore, the first contribution of this paper is a specific study upon local RFs. We carry out an analysis of repeatability and robustness on proposed local RFs, and provide experiments that demonstrate the strong impact of the choice of the RF on the performance of a 3D descriptor (Sec. 2). Given the impact of such a choice, the second contribution of this paper is a robust local RF that, unlike all other proposals, is unique and unambiguous(Sec. 3).

As for the descriptor, based on the nature of existing approaches highlighted by the proposed categorization, it is our belief that an effective and robust solution to the problem of 3D shape description can be found as a proper combination of *Signatures* and *Histograms*. Hence, the third contribution of the paper is a novel 3D descriptor aware of the proposed categorization (Sec. 4). Its design, inspired by an analysis of the successful choices performed in the related field of 2D descriptors, has been explicitly conceived to achieve computational efficiency, descriptive power and robustness. Finally, we provide a thorough experimental validation of our proposals (Sec. 5). We compare them to three state-of-the-art methods in surface matching experiments run on publicly available datasets as well as on range scans acquired in our lab.

2 On the Traits and Importance of the Local RF

The definition of a local RF, invariant to translations and rotations and robust to noise and clutter, has been the preferred option to endow a 3D descriptor with invariance to the same sources of variations, similarly to the way rotation and/or scale invariance is injected into 2D descriptors. On the other hand, the definition of such an invariant frame is challenging. Furthermore, although almost every new proposal for local shape description is equipped with its own local RF, experimental validation has always been focused on the results obtained by the joint used of an RF and a descriptor, whilst the impact of the selected local RF on the descriptor performance has not been investigated in literature.

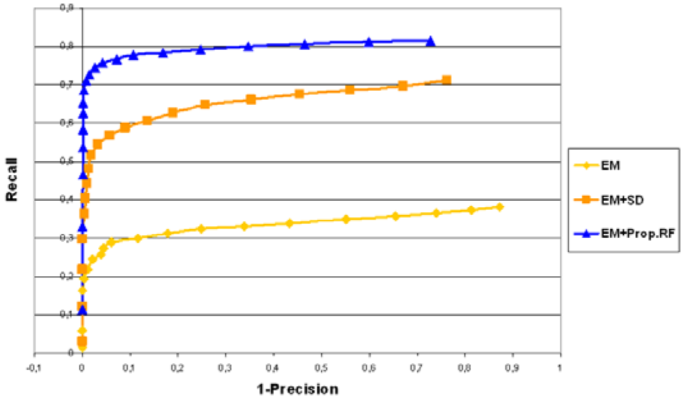


Fig. 2. Impact of the local RF on a descriptor performance

In Table 1 we have reported for each proposal the properties of uniqueness and unambiguity of their local RF. As highlighted in the third column, the majority of proposals are based on RFs that are not *unique* [6] [7] [8] [12] [2], i.e. to obtain an invariant description they require multiple descriptors to be computed at each feature point. This is usually handled by describing a ”model” point using multiple descriptors, each based on a different local RFs, and a ”scene” point with just one of them. This approach causes additional ambiguity to the correspondence problem since it shifts the intrinsic non-uniqueness of the local RF to the matching stage, thus increasing potential mismatches, computational requirements and sometimes also memory footprint. Another disadvantage brought in by the use of multiple local RFs is that the proposed matching stage is so tailored on the descriptor that it prevents the use of off-the-shelf efficient solutions for matching and indexing, that in principle could be advantageously performed orthogonally with respect to the description. This may result in a severe loss of computational efficiency.

In addition to multiple RFs, another limit of current proposals consists in the intrinsic ambiguity of the sign of the local RF axes. For example, in [9] and [13], normals and principal curvature directions are used. The main problem with this choice is that principal directions are not vectors, i.e. their sign is not defined. From a practical point of view, principal directions are computed using Singular Value Decomposition (SVD) or Eigenvalue Decomposition (EVD) of the covariance matrix of the point coordinates within the support. Of course, the output of the algorithm is a vector with a sign. Nevertheless, this sign is simply a numerical accident and, thus, is not repeatable on different (e.g. rotated) instances of the same mesh, even though the same SVD/EVD algorithm is used, as clearly discussed in [14]. Therefore, such an approach to the definition of the local RF is inherently ambiguous and thus not repeatable. [13] resorts to multiple RFs to overcome this limitation, while [9] does not deal with it explicitly.

To highlight the impact of the local RF on a descriptor performance, we show in Fig. 2 the performance of the EM descriptor [9] with different local RFs. Results are reported as *Recall vs 1-Precision* curves (see Sec. 5 for a discussion about this choice and for the settings used in all our experiments). The ambiguous RF used in [9] leads to unsatisfactory performances (yellow curve). Using exactly the same settings and exactly the same descriptor, we can boost performances simply by deploying the Sign Disambiguation technique recently proposed in [14]. Furthermore, using the more robust and more repeatable local RF that we propose in next section we can obtain another significant improvement (e.g. at recall 0.7 precision raises from 0.308 to 0.994) without changing the descriptive power of the descriptor.

3 Disambiguated EVD for a Repeatable RF

As shown by Table 1, none of current local RF proposals is at the same time unique and unambiguous. To fill this gap we have designed and extensively tested a variety of novel unique and unambiguous local RFs. We present here the method that turned out to be the most robust in our thorough experimental evaluation. It builds on a well known technique presented in [15] and [16], where the problem of normal estimation in presence of noise is specifically addressed. A Total Least Squares (TLS) estimation of the normal direction is obtained in [15] and [16] by EVD of the covariance matrix $\mathbf{M}$ of the $k-$nearest neighbors p_i of the point, defined by

$$\mathbf{M} = \frac{1}{k}\sum_{i=0}^{k}(\mathbf{p}_i - \hat{\mathbf{p}})(\mathbf{p}_i - \hat{\mathbf{p}})^T,\ \hat{\mathbf{p}} = \frac{1}{k}\sum_{i=0}^{k}\mathbf{p}_i\ . \tag{1}$$

In particular, the TLS estimation of the normal direction is given by the eigenvector corresponding to the smallest eigenvalue of M. Finally, they perform the sign disambiguation of the normals *globally* by means of sign consistency, i.e. propagating the sign from a seed chosen heuristically.

While this has proven to be a robust and effective technique for surface reconstruction of a single object, it cannot work for local surface description since in the latter case signs must be repeatable across any possible object pose as well as in scenes with multiple objects, so that a *local* rather than global sign disambiguation method is mandatory. Moreover, Hoppe's sign disambiguation concerns the normal only, hence it leaves ambiguous the signs of the remaining two axes.

In our proposal, we start by modifying (1) so as to assign distant points smaller weights, in order to increase repeatability in presence of clutter. Then, to improve robustness, all points laying within the spherical support (of radius R) which are used to compute the descriptor are used also to calculate $\mathbf{M}$. For the sake of efficiency, we also neglect the centroid computation, replacing it with the feature point $\mathbf{p}$. Therefore, we compute M as a weighted linear combination,

$$\mathbf{M} = \frac{1}{\sum_{i:d_i \le R}(R - d_i)}\sum_{i:d_i \le R}(R - d_i)(\mathbf{p}_i - \mathbf{p})(\mathbf{p}_i - \mathbf{p})^T \tag{2}$$

where $d_i = ||\mathbf{p}_i - \mathbf{p}||_2$. Our experimental evaluation indicates that the eigenvectors of M define repeatable, orthogonal directions in presence of noise and clutter. It is worth pointing out that, compared to [15] and [16], in our proposal the third eigenvector no longer represents the TLS estimation of the normal direction and sometimes it notably differs from it. However, this does not affect performance, since in the case of local surface description what matters is a highly repeatable and robust triplet of orthogonal directions, and not its geometrical or topological meaning.

Hence, eigenvectors of (2) represent a good starting point, but they need to be disambiguated to yield a repeatable local RF. The problem of sign disambiguation for EVD and SVD has been recently addressed in [14]. Their proposal basically reorients the sign of each singular or eigenvector so that its sign is coherent with the majority of the vectors it is representing. We determine the sign on the local x and z axes according to this principle. In the following we refer to the three eigenvectors in decreasing eigenvalue order as the $\mathbf{x}^+$, $\mathbf{y}^+$ and $\mathbf{z}^+$ axis, respectively. With $\mathbf{x}^-$, $\mathbf{y}^-$ and $\mathbf{z}^-$, we denote instead the opposite vectors. Hence, the final disambiguated x axis is defined as

$$S_x^+ \doteq \left\{ i : d_i \leq R \ \wedge \ (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^+ \geq 0 \right\} \tag{3}$$

$$S_x^- \doteq \left\{ i : d_i \leq R \ \wedge \ (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^- > 0 \right\} \tag{4}$$

$$\mathbf{x} = \begin{cases} \mathbf{x}^+, & |S_x^+| \geq |S_x^-| \\ \mathbf{x}^-, & \text{otherwise} \end{cases} \tag{5}$$

The same procedure is used to disambiguate the z axis. Finally, the y axis is obtained as $\mathbf{z} \times \mathbf{x}$.

We compare the repeatability of our proposal against two representative RFs: that of PS and that of EM, respectively a not-unique solution and an ambiguous one. To prevent these shortcomings from invalidating the comparison we consider only the global maximum of the height [7] for PS and we add the sign disambiguation of [14] to EM (EM+SD), thereby obtaining two unique and unambiguous RFs. We also consider the original EM approach to show the effectiveness of sign disambiguation. Using again the settings detailed in Sec. 5, in Fig. 3 we plot, for 5 increasing noise levels, the mean cosine between corresponding axes of the local RFs computed on two instances of the same mesh, i.e. the original one and a rotated and noisy instance. On one hand, ambiguity is clearly the most serious nuisance, as the low performances of the original EM proposal demonstrate. On the other hand, the use of a higher number of points to compute the local RF (i.e. the whole surface contained in the spherical support, as done by EM, instead of the 3D curve resulting by the intersection of the spherical support with the surface, as done by PS) yields better robustness, as shown by the relative drop of EM with respect to PS when noise increases. The disambiguation introduced in EM+SD dramatically enhances repeatability. However, both EM and EM+SD subordinate computation of the directions on the tangent plane to the normal estimation (i.e. , the repeatable directions they compute are then projected onto the tangent plane to create an orthogonal basis). This choice sums noise on the normal to the noise inevitably affecting the other directions, thereby leading to increased sensitivity of the estimation of the axes on the tangent plane and finally to poor repeatability. Our proposal, instead, estimates all axes simultaneously and turns out to be the most effective, thanks to the combination of its noise and clutter-aware definition, the effectiveness of the proposed disambiguation and the inherent uniqueness deriving from its theoretical formulation.

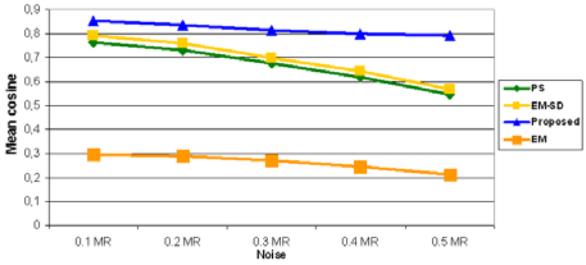


Fig. 3. Comparison between local RFs

4 Description by Signatures of Histograms

In Sec. 1 we have classified 3D descriptors as based on either histograms or signatures. We have designed our proposal following this intuition and aiming at a local representation that is efficient, descriptive, robust to noise and clutter as well as to point density variation. The point density issue is specific to the 3D scenario, where the same 3D volume of the real world may be represented with different amounts of vertexes in its mesh approximation, e.g. due to the use of different 3D sensors (stereo, ToF cameras, LIDARs, etc...) or different acquisition distances.

Beside our taxonomy, another source of inspiration has been the related field of 2D feature descriptors, which has reached a remarkable maturity during the last years. By analyzing SIFT [17], arguably the most successful and widespread proposal among 2D descriptors, we have singled out what we believe are among the major reasons behind its effectiveness. First of all, the use of histograms is spread throughout the algorithm, from the definition of the local orientation to the descriptor itself, this accounting for its robustness. Since a single global histogram computed on the whole patch would be not descriptive enough, SIFT relies on a set of local histograms, that are computed on specific subsets of pixels defined by a regular grid superimposed on the patch. The use of this coarse geometric information creates what we identify as a signature-like structure. Moreover, the elements of these local histograms are based on first order derivatives describing the signal of interest, i.e. intensity gradients. Although it has been argued that building a descriptor based on differential entities may result in poor robustness to noise [7], they hold high descriptive power, as the effectiveness of SIFT clearly demonstrates. Therefore, we believe they can provide a more effective solution for a descriptor than point coordinates [1] [12]. Yet, to account for robustness to noise, differential entities have to be filtered, and not deployed directly, e.g. as done in [9].

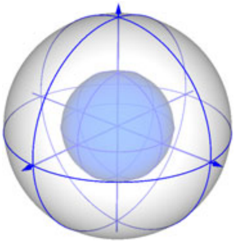


Fig. 4. Signature structure for SHOT

Based on these considerations, we propose a 3D descriptor that encodes histograms of basic first-order differential entities (i.e. the normals of the points within the support), which are more representative of the local structure of the surface compared to plain 3D coordinates. The use of histograms brings in the filtering effect required to achieve robustness to noise. Having defined an unique and robust 3D local RF (see Sec. 3), it is possible to enhance the discriminative power of the descriptor by introducing geometric information concerning the location of the points within the support, thereby mimicking a signature. This is done by first computing a set of local histograms over the 3D volumes defined by a 3D grid superimposed on the support and then grouping together all local histograms to form the actual descriptor. Hence, our descriptor lays at the intersection between Histograms and Signatures: we dub it Signature of Histograms of OrienTations (SHOT).

For each of the local histograms, we accumulate point counts into bins according to a function of the angle, θ_i, between the normal at each point within the corresponding part of the grid, $\mathbf{n}_{v_i}$, and the normal at the feature point, $\mathbf{n}_u$. This function is $cos\theta_i$, the reason being twofold: it can be computed fast, since $cos\theta_i = \mathbf{n}_u \cdot \mathbf{n}_{v_i}$; an equally spaced binning on $cos\theta_i$ is equivalent to a spatially varying binning on θ_i, whereby a coarser binning is created for directions close to the reference normal direction and a finer one for orthogonal directions. In this way, small differences in orthogonal directions to the normal, i.e. presumably the most informative ones, cause a point to be accumulated in different bins leading to different histograms. Moreover, in presence of quasi-planar regions (i.e. not very descriptive ones) this choice limits histogram differences due to noise by concentrating counts in a fewer number of bins.

As for the structure of the signature, we use an isotropic spherical grid that encompasses partitions along the radial, azimuth and elevation axes, as sketched in Fig. 4. Since each volume of the grid encodes a very descriptive entity represented by the local histogram, we can use a coarse partitioning of the spatial grid and hence a small cardinality of the descriptor. In particular, our experimentations indicate that 32 is a proper number of spatial bins, resulting from 8 azimuth divisions, 2 elevation divisions and 2 radial divisions (though, for clarity, only 4 azimuth divisions are shown in Fig. 4).

Since our descriptor is based upon local histograms, it is important to avoid boundary effects, as pointed out e.g. in [1] [17]. Furthermore, due to the spatial subdivision of the support, boundary effects might arise also in presence of perturbations of the local RF. Therefore, for each point being accumulated into a specific local histogram bin, we perform quadrilinear interpolation with its neighbors, i.e. the neighboring bins in the local histogram and the bins having the same index in the local histograms corresponding to the neighboring volumes of the grid. In particular, each count is multiplied by a weight of $1 - d$ for each dimension. As for the local histogram, d is the distance of the current entry from the central value of the bin. As for elevation and azimuth, d is the angular distance of the entry from the central value of the volume. Along the radial dimension, d is the Euclidean distance of the entry from the central value of the volume.

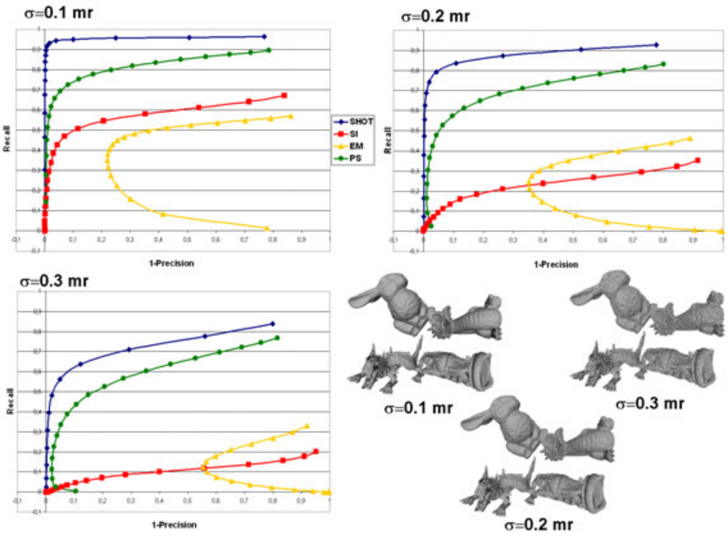


Fig. 5. Exp. 1: Precision-Recall curves on Stanford dataset and a scene at the 3 noise levels

Along each dimension, d is measured in units of the histogram or grid spacing, i.e. it is normalized by the distance between two neighbor bins or volumes.

To achieve robustness to variations of the point density, we normalize the whole descriptor to sum up to 1. This is preferable to the solution proposed in [12], i.e. normalizing each bin with the inverse of the point density and bin volume. In fact, while [12] implicitly assumes that the sampling density may vary independently in every bin, and thus discards as not informative the differences in point density among bins, we assume global (or at least regional) variations of the density and keep the local differences as a source of discriminative information.

5 Experimental Results

In this section we provide experimental validation of our proposals, i.e. the unique local RF together with the SHOT descriptor. To this purpose, we carry out a quantitative comparison against three state-of-the-art approaches in a typical surface matching scenario, where correspondences have to be established between a set of features extracted from a scene and those extracted from a number of models. The considered approaches are: *Spin Images* (SI), as representative of Histogram-based methods due to its vast popularity in the addressed scenario; *Exponential Mapping* (EM) and *Point Signatures* (PS) as representatives of Signature-based methods, the former since it is a very recent approach, the latter given its importance in literature. All methods were implemented in C++ and are made publicly available together with the datasets (www.vision.deis.unibo.it/SHOT).

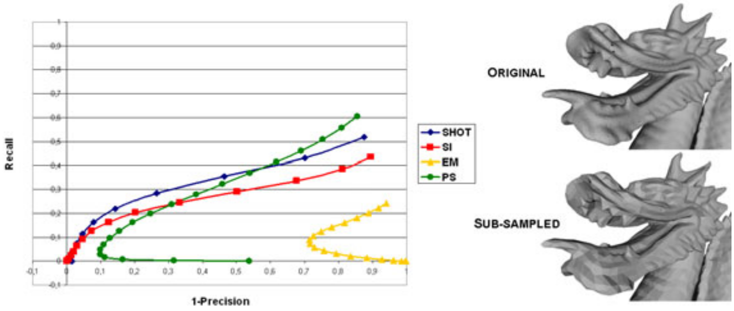


Fig. 6. Exp. 2: Precision-Recall curves on subsampled dataset and a detail from one scene

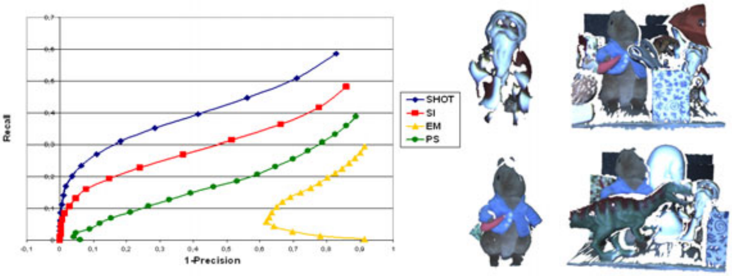
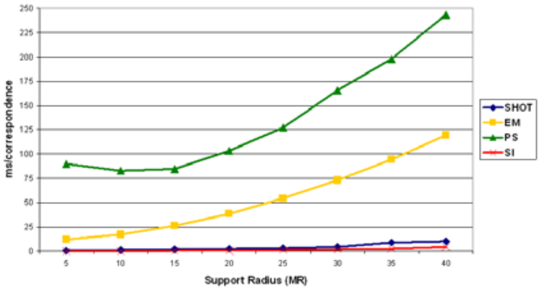
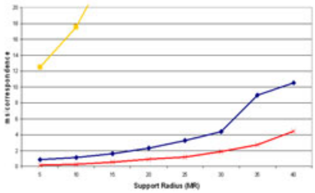


Fig. 7. Exp. 3: Results on Spacetime Stereo dataset and two models (middle) and scenes (right)

For a fair comparison, we use the same feature detector for all algorithms: in particular, 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 of the detector. Analogously, for what concerns the matching stage, we adopt the same matching measure for all algorithms, i.e. , as proposed in [1], the Euclidean distance. We could also have evaluated the synergistic effect of description and matching for those methods that explicitly include a proposal for the latter, e.g. the tolerance band for PS. In turn, we did experiments on the whole dataset with the original EM and PS matching schemes, obtaining slightly worse performance for both. This, and the attempt to be as fair as possible, leaned us to use the same measure for all algorithms. However, we did not discard the characteristics of the descriptors that required a specific treatment during matching: in particular, since EM is a sparse descriptor, we compute the Euclidean distance only on the overlapping subset of EM descriptor pairs, as proposed by the authors; and for PS we use the matching scheme proposed by the authors to disambiguate its not-unique local RF [7]. For each scene and model, we match each scene feature against all model features and we compute the ratio between the nearest neighbor and the second best (as in [17]): if the ratio is below a threshold a correspondence is established between the scene feature and its closest model feature.





	Time (s)	Radius (mr)	Length
SHOT	4.8	15	320
SI	5.6	30	100
EM	52.6	10	2700
PS	248.8	10	90

Fig. 8. Charts: ms/correspondence vs. support radius (in the smaller chart the time axis is zoomed in for better comparison between SI and SHOT). Table: measured execution times (in Experiment 1) and tuned parameter values. Radius values are reported in mesh resolution units. As for SI, the support radius is the product of the bin size by the number of bins in each side of the spin image

According to the methodology for evaluation of 2D descriptors recommended in [18], we provide results in terms of *Recall* versus *1-Precision* curves. This choice is preferable compared to ROC curves (i.e. *True Positive Rate* versus *False Positive rate*) when comparing descriptors or detectors due to the ambiguity in calculating the *False Positive Rate* [19]. We present three different experiments. Experiment 1 deals with 6 models (”Armadillo”, ”Asian Dragon”, ”Thai Statue”, ”Bunny”, ”Happy Buddha”, ”Dragon”) taken from the *Stanford 3D Scanning Repository* [1]. We build up 45 scenes by randomly rotating and translating different subsets of the model set so to create clutter[2]; then, similarly to [20], 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). In Experiment 2 we consider the same models and scenes as in Experiment 1, add noise (i.e. σ_1) and resample the 3D meshes down to $1/8$ of their original point density. For a fair comparison in this experiment, our implementation of SI -used throughout all the evaluation- normalizes each descriptor to the unit vector to make it more robust to density variations [3]. Finally, in Experiment 3 the dataset consists of scenes and models acquired in our lab by means of a 3D sensing technique known as *Spacetime Stereo* [21], [22]. In particular, we compare 8 object models against 15 scenes characterized by clutter and occlusions, each scene containing two models. Fig. 7 shows two scenes together with the models appearing in them. In each of the three experiments, 1000 feature points were extracted from each model. As for the scenes, in Exp. 1 and 2 we extract $n * 1000$ features per scene (n being the number of models in the scene) whereas in Exp. 3 we extract 3000 features per scene.

Throughout all the three experiments we used the same values for the parameters of considered methods. In particular, we tuned the two parameters of each descriptor (*support radius* and *length of the descriptor*) based on a tuning scene corrupted

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

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

with noise level σ_1 and built rotating and translating three Stanford models (”Bunny”, ”Happy Buddha”, ”Dragon”). The values resulting from the tuning process are reported in the last two columns of the Table in Fig. 8. It is worth noting that our tuning yielded comparable values of the support radius among the various methods, and that, for SI and PS, the resulting parameter values are coherent, as far as the order of magnitude is concerned, with those originally proposed by their authors (no indication about EM parameters is given in [9]). Yet, we used the finely tuned values instead of those originally proposed by the authors since the former yield higher performance in these experiments.

Results for the three Experiments are reported in Figure 5, 6 and 7, respectively. Experiment 1 focuses on robustness to noise. Given the reported results, it is clear that SHOT performs better than the other methods at all different noise levels on the Stanford dataset. We can observe that, comparing the two Signature methods, PS exhibits a higher robustness than EM. We address this mainly to the higher robustness of its local RF, as shown in Fig. 3. As for SI, it appears to be highly susceptible to noise, its performance notably deteriorating as the noise level increases. This is due to the fact that this descriptor is highly sensitive to small variations in the normal estimation (i.e. SI Reference Axis), that here we compute as proposed in [1]. This is also consistent with the results reported in [12]. As for Experiment 2, it is clear that the point density variation is the most challenging nuisance among those accounted for, causing a severe performance loss of all methods. SHOT, PS and SI obtain comparable performance, nevertheless for high values of precision, that are typical working points for real applications, SHOT obtains the highest levels of Recall. Experiment 3 shows that under real working conditions SHOT outperforms the other methods. It is worth noting that this experiment is especially focused on the descriptiveness of evaluated approaches, since the much smoother shapes of the objects surfaces compared to those of the Stanford models make the former harder to discriminate. Hence, results demonstrate the higher descriptiveness embedded in SHOT with respect to the other proposals.

In addition, we have compared the methods in terms of their computational efficiency and memory requirements. Since, as discussed in Sec. 2, descriptors based on multiple RFs, like PS, can not deploy efficient indexing to speed-up the matching stage, we use a full search strategy for all methods. Results are reported in Fig. 8. The two charts in the Figure, showing the number of milliseconds per correspondence needed by the various methods using different support sizes, demonstrate the notable differences in computational efficiency between the algorithms. In particular, SI and SHOT run one order of magnitude faster than EM and almost two orders of magnitude faster than PS, with SI turning out consistently slightly faster than SHOT at each support size. As for EM, efficiency is mainly affected by the re-parametrization of the support needed to describe each feature point and to the large memory footprint (see next). With regards to PS, as discussed in Sec. (2) the use of multiple local RFs dramatically slows down the matching stage. These results are confirmed by the Table in the Figure (first column), which reports the measured times required to match the scene to the models in Experiment 1 (i.e. 3000 scene features and 3000 models features) using the tuned parameter values. Here, the larger support needed by SI allows SHOT to run slightly faster. As for memory requirements, the reported descriptor length (third column) highlights the much higher memory footprint required by EM compared to other methods.

Finally, as a practical application in a challenging and active research area, we demonstrate the use of SHOT correspondences to perform fully automatic 3D Reconstruction from Spacetime Stereo data. We merge 18 views covering a 360° field of view of one of the smooth objects used in Experiment 3. We follow a 2 steps procedure: 1) we obtain a coarse registration by estimating the 3D transformations between every pair of views and retaining only those maximizing the global area of overlap; 2) we use the coarse registration as initial guess for a final global registration carried out using a standard external tool (*Scanalyze*). In the first step, correspondences among views are established by computing and matching SHOT descriptors on 1000 randomly selected feature points. 3D transformations are estimated by applying a well known Absolute Orientation algorithm [23] on such correspondences and filtering outliers with RANSAC. Maximization of the area of overlap is achieved through the Maximum Spanning Tree approach described in [9]. As shown in Fig. 9, without any assumptions about the initial poses, SHOT correspondences allows for attaining a coarse alignment which is an accurate enough initial guess to successfully reconstruct the 3D shape of the object without any manual intervention. To the best of our knowledge, fully automatic 3D reconstruction from multiple Spacetime Stereo views has not been demonstrated yet.

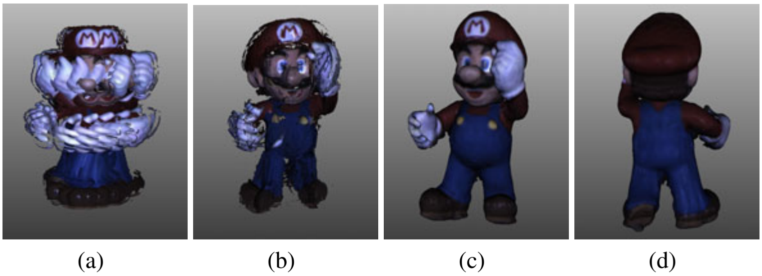

(a) (b) (c) (d)

Fig. 9. 3D Reconstruction from Spacetime Stereo views: (a) initial set of views (b) coarse registration (c) global registration frontal view (d) global registration rear view

6 Conclusion and Future Work

Overall, our proposals compare favorably with the considered methods. The results validate the proposed categorization as well as the intuition that the synergy between the design of a repeatable local RF and the embedding of an hybrid signature/histogram nature into SHOT allows for achieving at the same time state-of-the-art robustness and descriptiveness. Remarkably, our proposal delivers such notable performances with high computational efficiency. As for future work, we plan to investigate on how to improve robustness to point density variations. Comparing our proposal with other relevant methods and on a larger dataset is another main direction of our research.

Acknowledgments. The authors would like to deeply thank Alioscia Petrelli and Alessandro Franchi, of the Computer Vision Lab of the University of Bologna, who helped implementing all the proposals used in our experiments.

References

1. Johnson, A., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. PAMI 21, 433–449 (1999)
2. Mian, A., Bennamoun, M., Owens, R.: A novel representation and feature matching algorithm for automatic pairwise registration of range images. IJCV 66, 19–40 (2006)
3. Conde, C., Rodríguez-Aragón, L.J., Cabello, E.: Automatic 3D face feature points extraction with spin images. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4142, pp. 317–328. Springer, Heidelberg (2006)
4. Wu, K., Levine, M.: Recovering parametrics geons from multiview range data. In: CVPR, pp. 159–166 (1994)
5. Solina, F., Bajcsy, R.: Recovery of parametric models from range images: the case for superquadrics with global deformations. PAMI 12, 131–147 (1990)
6. Stein, F., Medioni, G.: Structural indexing: Efficient 3-D object recognition. PAMI 14, 125–145 (1992)
7. Chua, C.S., Jarvis, R.: Point signatures: A new representation for 3D object recognition. IJCV 25, 63–85 (1997)
8. Sun, Y., Abidi, M.A.: Surface matching by 3D point's fingerprint. ICCV 2, 263–269 (2001)
9. Novatnack, J., Nishino, K.: Scale-dependent/invariant local 3D shape descriptors for fully automatic registration of multiple sets of range images. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 440–453. Springer, Heidelberg (2008)
10. Chen, H., Bhanu, B.: 3D free-form object recognition in range images using local surface patches. Patt. Rec. Letters 28, 1252–1262 (2007)
11. Koenderink, J., Doorn, A.: Surface shape and curvature scales. Image Vision Computing 8, 557–565 (1992)
12. Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J.: Recognizing objects in range data using regional point descriptors. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 224–237. Springer, Heidelberg (2004)
13. Zhong, Y.: Intrinsic shape signatures: A shape descriptor for 3D object recognition. In: ICCV-WS: 3DRR (2009)
14. Bro, R., Acar, E., Kolda, T.: Resolving the sign ambiguity in the singular value decomposition. J. Chemometrics 22, 135–140 (2008)
15. Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W.: Surface reconstruction from unorganized points. In: SIGGRAPH, pp. 71–78 (1992)
16. Mitra, N.J., Nguyen, A., Guibas, L.: Estimating surface normals in noisy point cloud data. Int. J. of Computational Geometry and Applications 14, 261–276 (2004)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110 (2004)
18. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI 27, 1615–1630 (2005)
19. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: CVPR (2004)
20. Unnikrishnan, R., Hebert, M.: Multi-scale interest regions from unorganized point clouds. In: CVPR-WS: S3D (2008)
21. Davis, J., Nehab, D., Ramamoothi, R., Rusinkiewicz, S.: Spacetime stereo: A unifying framework for depth from triangulation. PAMI 27, 1615–1630 (2005)
22. Zhang, L., Curless, B., Seitz, S.: Spacetime stereo: Shape recovery for dynamic scenes. In: CVPR (2003)
23. Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. J. of the Optical Society of America A 4, 629–642 (1987)