Spherical Blurred Shape Model for 3-D Object and Pose Recognition: Quantitative Analysis and HCI Applications in Smart Environments

Oscar Lopes, Miguel Reyes, Sergio Escalera, and Jordi Gonzàlez

***Abstract*—The use of depth maps is of increasing interest after the advent of cheap multisensor devices based on structured light, such as Kinect. In this context, there is a strong need of powerful 3-D shape descriptors able to generate rich object representations. Although several 3-D descriptors have been already proposed in the literature, the research of discriminative and computationally efficient descriptors is still an open issue. In this paper, we propose a novel point cloud descriptor called spherical blurred shape model (SBSM) that successfully encodes the structure density and local variabilities of an object based on shape voxel distances and a neighborhood propagation strategy. The proposed SBSM is proven to be rotation and scale invariant, robust to noise and occlusions, highly discriminative for multiple categories of complex objects like the human hand, and computationally efficient since the SBSM complexity is linear to the number of object voxels. Experimental evaluation in public depth multiclass object data, 3-D facial expressions data, and a novel hand poses data sets show significant performance improvements in relation to state-of-the-art approaches. Moreover, the effectiveness of the proposal is also proved for object spotting in 3-D scenes and for real-time automatic hand pose recognition in human computer interaction scenarios.**

***Index Terms*—Depth image analysis, human computer interaction (HCI), image descriptors, object and pose recognition, smart environments.**

I. Introduction

COMPUTER vision research on 3-D point cloud analysis has recently received a lot of attention because of the availability of cheap multisensor devices based on structured light, such as Kinect. This RGB-Depth camera is compact and portable, so it can be easily installed in any environment to understand 3-D scenes. This way there are multiple applications which can benefit from the analysis of 3-D objects in scenes [1], [17], [31], [33], [37]. However, recognition of 3-D objects is still a challenging problem; in addition to the typical issues tackled by 2-D object recognition approaches (such as robustness to noise and occlusions, discriminate power, and computational complexity), the captured sequences are usually sampled at discrete points, so the finer details of the 3-D object are usually lost.

Under these assumptions, there exists a strong interest for designing new 3-D object descriptors [3], [18], [25], [34]. We next revisit the literature by dividing the existing approaches into those descriptors based on pure 3-D geometric properties and those extended from already existing 2-D object descriptors.

Describing 3-D geometric information has been proven to be useful when classifying everyday objects like cans, glasses or doors, and for 3-D scene analysis. For example, some approaches take into account the set of normals of the surface defined by a given point and its neighbors [2], [23]. As an example, the SHOT descriptor proposed in [35] defines a surface representation based on point normals. It is based on counting the points that fall into bins according to a function of the angle between the normal at each point within the corresponding part of the grid and the normal at the feature point. However, in this case, the descriptor is local and usually requires a previous keypoint detection step, which complicates its adaptation to recognize nonrigid shapes. The use of normals are useful to recognize 3-D objects since they encode the implicit surface that neighboring points define, although they depend on the density of the underlying points and the smoothness of 3-D object surfaces to give accurate results. Also, spherical harmonics [10] have been used to design 3-D descriptors invariant to rotation [19] or have been considered directly as features [29]. Conformal factors have also been considered [6], measuring the relative curvature of a vertex given the total curvature. The result can be viewed as a vector which is not only invariant to rigid body transformations, but also to changes in the pose. The point feature histogram (PFH) local descriptor proposed in [28] is used to recognize points conforming planes, cylinders, and other geometric primitives. As an extension, the fast PFH (FPFH) descriptor [26] is based on codifying angle relations among 3-D points. FPFH optimizes the PFH computation to make it usable in real-time 3-D registration applications. The

Manuscript received July 30, 2013; revised November 28, 2013 and February 14, 2014; accepted February 17, 2014. This work was supported in part by the European project DiCoMa under Grant TSI-020400-2011-55 and in part by the Spanish projects under Grant TIN2009-14501-C02-02 and Grant TIN2012-39051. This paper was recommended by Associate Editor X. Li.

O. Lopes is with the University of Utrecht, Utrecht 3512 JE, The Netherlands (e-mail: oscar.pino.lopes@gmail.com).

M. Reyes is with the University of Barcelona, Barcelona 08007, Spain (e-mail: mreyes@gmail.com).

S. Escalera is with the department of Matemàtica Aplicada i Anàlisis, Facultat de Matemàtiques, Universitat de Barcelona, Barcelona 08007, Spain (e-mail: sergio@maia.ub.es).

J. Gonzàlez is with the department of Computer Science, Universitat Autònoma de Barcelona, Barcelona 08193, Spain (e-mail: poal@cvc.uab.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2307121

viewpoint feature histogram (VFH) [27] combines an extended version of FPFH with statistics between the viewpoint and the surface normals on the 3-D object. Recently, Wohlkinger and Vincze [36] have presented an ensemble of shape functions (ESF) approach to describe 3-D objects, which benefits from several combinations of histograms for codifying 3-D object relations of angles, areas, and distances among points.

Unfortunately, the point clouds captured from Kinect-like devices usually contain holes, since data are sampled at discrete points. Consequently, in all the aforementioned approaches (which rely on an accurate computation of 3-D geometric primitives) their performance is usually downgraded. Alternatively, some recent 3-D regional descriptors have been defined as an extension of classical derivative-based 2-D features, such as HOG, SIFT, and SURF [20], [24]. For example, as a generalization of the 2-D shape context descriptor presented in [5], Frome *et al*. [16] propose a 3-D shape context descriptor which is compared with a classical spin-image representation and a novel harmonic shape context (HSC) descriptor for 3-D car model classification. Despite the excellent results reported, these methods require the computation of a large number of shape points relations.

In this paper, we propose a novel 3-D object descriptor, called spherical blurred shape model (SBSM). SBSM is inspired in the blurred shape model (BSM) descriptor presented in [15] and [14]. The novel SBSM descriptor codifies the object structure density and local variabilities in the 3-D space. Similar to the Zoning descriptor and 3-D shape histograms of [4], SBSM bases on a linear computation of spatial relation of shape points to 3-D bin centroid, but including a propagation *blurring* degree to define a compact and discriminative 3-D object descriptor. In this sense, Zoning and the descriptor in [4] can be seen as an instance of the proposed descriptor when the defined blurring degree is null. As it is reported in the results, when increasing the blurring factor the overall classification rate of the system is improved. In addition to provide a 3-D generalization of BSM, SBSM introduces the following enhancements; 1) a 3-D spherical grid which partitions the 3-D space into 3-D shape bins, 2) a 3-D Gaussian-based weight propagation schema controlling the blurring level based on shape voxel distances, and 3) a quaternion-based rotation strategy based on sphere axis densities to define a 3-D rotation invariant descriptor. As a result, the proposed SBSM is a global descriptor that encodes the shape of an object, being rotation and scale invariant, computationally efficient, and highly discriminative.

We evaluate the descriptor on public and novel 3-D object and hand poses data set, showing significant performance improvements in comparison to the state-of-the-art approaches. We also test the descriptor in front of deformation in depth coordinates and noise point removal. As a result, we can show that the SBSM copes with the noise and occlusions typically present in the point clouds acquired by range scanner sensors. Additionally, we show four real applications where we apply the descriptor. In the first, we perform object spotting in public 3-D scenes. The last three applications correspond to human computer interaction (HCI) scenarios. In the second, we present a real-time fully-automatic HCI system for medical image volume navigation, segmenting human hands, and classifying multiple hand poses using the proposed SBSM descriptor. In the third, the same approach is applied in a multicamera setup to perform intelligent retail. Finally, we present a prototype for intelligent navigation through a repository of books in a living lab which may represent the library of the future. Although pointing recognition and gaze tracking in multicamera setups for HCI has been previously addressed in [12] and [38], here, we show how complex multicamera setups can be avoided and recognition for interaction can be improved within different HCI application contexts using the proposed approach.[1]

The rest of the paper is organized as follows. Section II presents the SBSM descriptor. SBSM is evaluated and compared to the state-of-the-art approaches in Section III. Section IV presents four real applications that uses the proposed descriptor, including 3-D object spotting in real scenes and different HCI scenarios. Finally, Section V concludes the paper.

II. Method

In this section, we present the novel SBSM to describe 3-D objects.

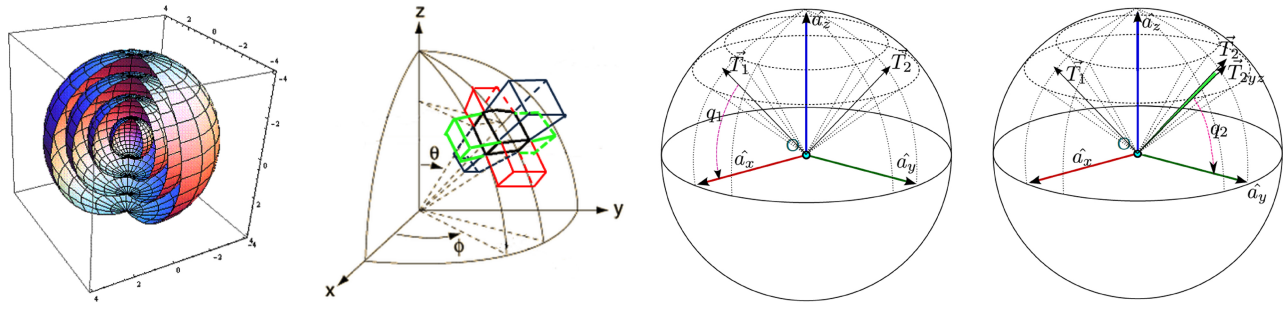
A. Spherical Blurred Shape Model

The SBSM is inspired in 3-D grid approaches and in the discriminative power of SIFT and HOG descriptors to codify object information based on the distribution of object gradients and orientations. However, instead of performing computation of 3-D object derivatives, SBSM just requires the computation of object shape voxel distances between neighbors in order to codify the object structure density and local variabilities in the 3-D space. As a result, SBSM is a computationally efficient descriptor, with a complexity linear to the number of object voxels $O(|P|)$, with an upper bound of $27 \cdot |P|$ simple operations for a point cloud P of $|P|$ shape points (defined based on a 26-connectivity of regions in the 3-D space of bins).

As in the case of 2-D and 3-D object descriptors, an initial grid is fitted to contain the region of interest to describe. In our case, to describe 3-D regions, a spherical grid containing a set of 3-D bins is defined, which contain the set of voxels P of the point cloud to be described. Our description methodology computes for each voxel P contained in the grid a set of voxel-bin spatial relations that are included in a global region descriptor. Next, we describe in detail each step of the description procedure.

In the first step, a discrete spherical grid partitions the 3-D space in a set of bins, as shown in Fig. 1(a). Let $P = \{p_i | p_i \in \mathbb{R}^3\}$, C, N_L, N_θ, N_ϕ, R, and σ define the set of voxels of the point cloud, point cloud centroid, number of layers, number of angular divisions for θ, number of angular divisions for ϕ, radius length, and sigma value for the gaussian distance

[1]We include as supplemental material the SBSM descriptor code, the novel ASL 3-D hand poses data set, and a demonstration video with the descriptor running real-time in different HCI scenarios.



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

Fig. 1. Illustration of SBSM descriptor computation. (a) Sphere bins. (b) Example of neighbor bins. (c) and (d) Example of the estimation of two main quaternion to rotate feature vector in the 3-D space.

metric, respectively. Some of these parameters are illustrated in Fig. 1(b). Then, $d_R = R/N_L$, $d_\theta = 2\pi/N_\theta$, and $d_\phi = 2\pi/N_\phi$ are computed as the distance between consecutive layers and the degrees in θ and ϕ polar coordinates between consecutive sectors, respectively. Using this division, we next define the set B of sphere bins $b_{\{i,j,k\}}$ as follows:

$$\begin{aligned} B = \{&b_{\{0,0,0\}}, ..., b_{\{i,j,k\}}, ..., b_{\{N_L-1,N_\theta-1,N_\phi-1\}}\} \\ &\forall i \in \{0, 1, ..., N_L - 1\} \forall j \in \{0, 1, ..., N_\theta - 1\}, \\ &\forall k \in \{0, 1, ..., N_\phi - 1\} \end{aligned} \tag{1}$$

where bin $b_{\{i,j,k\}}$ is the 3-D bin defined as the cartesian product of intervals $[i \cdot d_R, (i+1) \cdot d_R)$, $[j \cdot d_\theta, (j+1) \cdot d_\theta)$, and $[k \cdot d_\phi, (k+1) \cdot d_\phi)$ in relation to the center of the spherical grid, θ, and ϕ, respectively. This way B defines a partition in $\mathbb{R}^3$ of the object of interest. Then, the centroid coordinates for all each bin $b^*_{\{i,j,k\}} \in B^*$ are computed as follows:

$$\begin{aligned} b^*_{\{i,j,k\}} = &\left(i \cdot d_R + \frac{d_R}{2}, j \cdot d_\theta + \frac{d_\theta}{2}, k \cdot d_\phi + \frac{d_\phi}{2} \right) \\ &\forall i \in \{0, 1, ..., N_L - 1\}, \forall j \in \{0, 1, ..., N_\theta - 1\} \\ &\forall k \in \{0, 1, ..., N_\phi - 1\}. \end{aligned} \tag{2}$$

An example of some 3-D bin neighbors of the spherical descriptor is shown in Fig. 1(b). Once the 3-D spatial bins are defined, the SBSM feature vector is initialized as

$$W_i = 0, \forall i \in \{1, 2, ..., N_L \cdot N_\theta \cdot N_\phi\}. \tag{3}$$

Subsequently, for each voxel in the point cloud $p_z \in \{P | P \subset B\}$, the distance of that voxel to its neighbor bins is estimated based on a Gaussian distance metric, and the normalized weights are added to the corresponding descriptor bin locations. For this task, let $b_z = b_{\{i,j,k\}} | p_z \subset b_{\{i,j,k\}}$ be the bin containing voxel p_z. First, the lists containing bin weights and index bins for p_z are initialized to $W^* = \{0\}$ and $I^* = \{\{i, j, k\}\}$, respectively. Then, the iterative procedure updates W^* and I^* for each $b_{\{i,j,k\}} \in N(b_z)$, where $N(b_z)$ is the set of neighbors bins of b_z in a 27-neighborhood for inner sphere bins and 18-neighborhood for external sphere surface bins (including the reference bin). So the list of weights is updated as

$$W^* = W^* \cup \left\{ e^{-\frac{||p_z - b^*_{\{i,j,k\}}||}{R \cdot \sigma}} \right\} \tag{4}$$

and the list of indexes as

$$I^* = I^* \cup \{\{i, j, k\}\}. \tag{5}$$

As a result, the normalized weights for p_z are added to its corresponding positions of W as follows:

$$W_{I^*_i} = W_{I^*_i} + \frac{W^*_i}{\sum_{j=1}^{|W^*|} W^*_j}, \forall i \in \{1, 2, .., |I^*|\}. \tag{6}$$

In this way, a new shape point of the point cloud will include a weight to its belonging bin centroid and neighbor centroid based on a Gaussian function of the distance and a blurring level defined by σ. This values defines the degree of influence of each neighbor bin for each point cloud voxel. Note that when σ parameter is set to zero, the descriptor is equivalent to a dense sampling of the point cloud as in the classical state-of-the-art Zoning descriptor, but defined in the 3-D space [14]. It is important to remark that the voxels that are not contained within the spherical grid bins are not considered in the descriptor computation. On the other hand, the voxels that intersect with the spherical surface are c onsidered as inner voxels and thus, considered in the descriptor estimation. Fig. 2 shows an example of an hypothetical sphere slice for $\phi = k$ and the analysis of a point cloud voxel to update the SBSM descriptor.

Once the procedure is repeated for all points $p_z \in P$, the final feature vector W is normalized as follows:

$$W_i = \frac{W_i}{\sum_{j=1}^{N_L \cdot N_\theta \cdot N_\phi} W_j}, \forall i \in \{1, 2, .., N_L \cdot N_\theta \cdot N_\phi\}. \tag{7}$$

Given that all the voxels within the point cloud where the descriptor is computed contribute with the same cost and that the final vector is normalized, it becomes scale invariant. Thus, if different instances of a 3-D object category are fitted with the spherical descriptor, even with different sizes, all the descriptors are comparable and can be trained with the same classifier.

B. 3-D Rotation Invariant SBSM

Once SBSM is computed based on the predefined number of layers, bin orientations, and σ value for the Gaussian function, the descriptor is able to encode the local density and

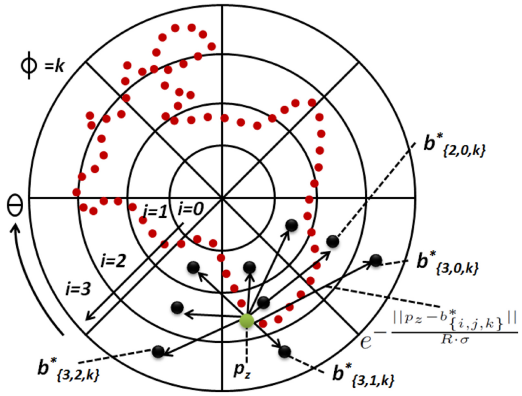


Fig. 2. Example of point cloud voxel for an hypothetical sphere slice for $\phi = k$. Voxels of the point cloud visible on that slice are shown as red dots. An example of a voxel estimation p_z is shown in green. For this point, neighbor bins centroids are shown as black dots. For each of these relations (note that in the 3-D space a total of 27 relations will be computed), equation 4 is computed, and the estimated value is added to descriptor position corresponding to its corresponding bin.

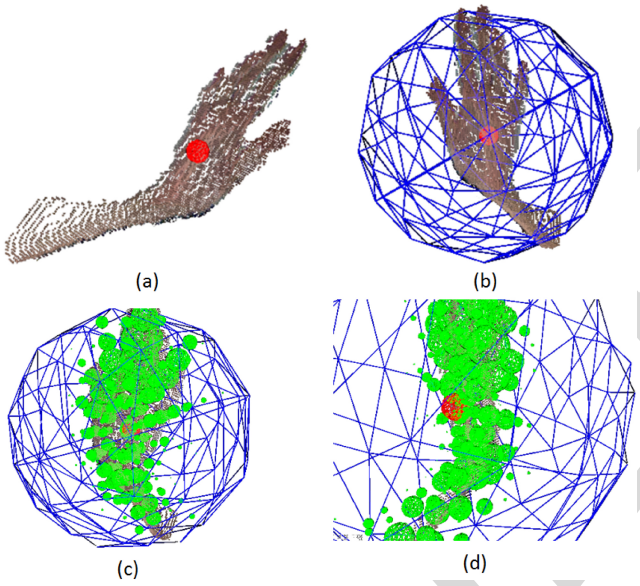


Fig. 3. (a) Initial hand point cloud and computed center. (b) Sphere including a point cloud corresponding to a 3-D hand pose. (c) Same sphere where SBSM descriptor has been computed. The density of the green dots represents the centroid bin values, and the whole descriptor has been rotated based on the quaternion codified by two main descriptor axis densities. (d) Alternative view of the computed SBSM descriptor.

spatial relations of 3-D shape points for a particular granularity degree. Rotation invariance is achieved by considering the main spherical axis densities to compute the main vector orientations in quaternion coordinates as a reference axis that rotates feature vector bins. As a result, this feature vector reordering step makes the descriptor rotation invariant for similar 3-D objects.

The use of unit quaternion instead of rotation matrices provides a fast computation for rotation invariance, and at the same time, it is simpler to enforce that quaternions have unit magnitude than constrain rotation matrices to be orthogonal [32]. The procedure is detailed next.

First, we compute the density of the descriptor for each axis defined by the angles θ and ϕ as follows:

$$f(\theta, \phi) = \sum_{r=1}^{N_L} W_{\{r,\theta,\phi\}} \tag{8}$$

and the two maximum axis densities are found

$$\overrightarrow{T}_1 = \arg\max_{\theta,\phi} f(\theta, \phi),\ \overrightarrow{T}_2 = \arg\max_{\theta,\phi \setminus \overrightarrow{T}_1} f(\theta, \phi). \tag{9}$$

Back to Cartesian coordinates, we compute the component of $\overrightarrow{T}_2$ vertical to $\overrightarrow{T}_1$ by projecting $\overrightarrow{T}_2$ onto the plane perpendicular to $\overrightarrow{T}_1$ as follows:

$$\overrightarrow{T}_{2yz} = \overrightarrow{T}_2 - \overrightarrow{T}_1^T \overrightarrow{T}_2 \overrightarrow{T}_1. \tag{10}$$

Subsequently, we compute the rotation that aligns the axis $\overrightarrow{T}_1$, $\overrightarrow{T}_2$ with $\hat{a}_x$ and $\hat{a}_y$, respectively. Where $\hat{a}_x = [100]^T$ and $\hat{a}_y = [010]^T$. The rotation quaternion q can be computed as the combination of two quaternions q_1 and q_2, so that $q = q_2 q_1$, where q_1 rotates $\overrightarrow{T}_1$ to $\hat{a}_x$ and q_2 aligns $\overrightarrow{T}_2$ with $\hat{a}_y$ [Fig. 1(d)].

Finally, the values of the bin locations are rotated based on the quaternion q, such that each bin $b^{*r}_{i,j,k} \in B^*$ is computed as

$$b^{*r}_{i,j,k} = q b^*_{i,j,k} q^* \tag{11}$$

using the Hamilton product, where q^* is the conjugate of the quaternion q. Abusing of notation, b^* and b^{*r} also denote the corresponding pure quaternion to each bin. So, we take advantage of this rotation order to obtain the rotation invariant feature vector $W^r_{\{i,j,k\}} = W_{\{i,j,k\}}$.

An example of the two main quaternions for an hypothetical spherical toy problem is shown in Fig. 1(c) and (d), respectively. In Fig. 3(a), a real example of a 3-D hand pose is shown. Fig. 3(b) shows the centered point cloud within the 3-D correlogram containing SBSM bins. The result after computing the SBSM descriptor from hand point cloud and performing rotation invariance is shown in Fig. 3(c) and (d) for two different points of view.

III. Quantitative Analysis of SBSM

In order to present the results, we first describe the training data and settings of the experiments.

A. Data Sets

We test our methodology on three data sets; a public 3-D object category data set, a new 3-D hand pose data set, and a public subject identification data set.

1) *RGB-D Object Data Set:* The RGB-D Object data set is a large collection of 300 common household objects [21]. All these objects are organized into 51 categories arranged using WordNet hypernym-hyponym relationships (similar to ImageNet). This data set was recorded using a Kinect style 3-D camera that recorded a set of synchronized and aligned 640×480 RGB-D images at 30 Hz. Each object was placed on a turntable, and the video sequences were captured for a single, full rotation. For each object, three video sequences

Fig. 4. RGB-Depth object data set category samples [21].

Fig. 5. American sign language data set categories.

were recorded with the camera mounted at different heights so that the object could be viewed from different angles with respect to the horizon. Example of segmented objects are shown in Fig. 4.

2) *American Signal Language Data Set:* We have recorded a novel 3-D hand poses data set based on the American sign language vocabulary. The data set is composed of 23 categories with around 47K instances of both hands. The Kinect device was used to extract the hands and their point clouds data using standard segmentation and detection algorithms. Also, both hands were captured not only considering a frontal view but also including variabilities in terms of scale, hand orientation, and finger joint articulations. This way, the complexity and variability of the overall data set was enriched. Examples of hand categories are shown in Fig. 5.[2]

3) *Bosphorus 3-D Face Expression Data Set:* The Bosphorus 3-D faces data set [30] contains several users performing nine natural facial expressions. From the Bosphorus face dataset it was considered a subset of 21 individuals, including 1039 samples. For each individual, the original structure of the dataset was kept, and all the corresponding nine facial expressions were considered. Each sample of the original Bosphorus dataset is composed by roughly 45000 points. For performance issues, it was performed a down-sampling using a Voxel grid filter, obtaining samples comprising approximately 9 000 points (some examples of the data set and the computed point clouds are shown in Fig. 6).

[2]Data set included as supplemental material (∼1Gb).

B. Settings and Evaluation Metrics

A multiclass classifier is trained using the proposed SBSM descriptor. Specifically, feature vectors are trained in a one-versus-one SVM classifier using a RBF kernel, and optimizing the parameters C and γ by means of cross-validation using LibSVM [11]. SBSM descriptor size was experimentally set to $N_L = 8, N_\theta = 8, N_\phi = 8$ for all the experiments, with a total descriptor length of 512. [3] We compare the SBSM descriptor with different state-of-the-art methods on different experiments [7], [8], [21], [22]. We also include in the comparative the VFH [27] and ESF [36] descriptors by also training the feature vectors with one-versus-one SVM classifier using a RBF kernel and optimizing the parameters as in the case of SBSM. VFH and ESF have been selected since they are recent, representative, and robust well-known descriptors for shape estimation and codification of normal vectors distribution.

We validate the object classification experiments by means of recognition rate applying stratified ten-fold cross-validation and estimating the confidence interval with a two-tailed t-test. We also test and compare the descriptors against depth distortions and noisy data to compute their statistical significance based on Friedman and Nemenyi statistics [13].

C. Experiments

We next present the multiclass 3-D object categorization performance using SBSM in the RGB-D object, sign language, and 3-D facial expression data sets. Once we demonstrate the performance of the proposed descriptor, we test it against different 3-D object distortions.

1) *Analysis of Classification Performance:* For the RGB-D object data set, we use the turntable data for both training and evaluation, thus classifying 51 different 3-D object categories using depth information only. For the object recognition experiments on cropped images, we apply a leave-one-out strategy as described in [21]. For comparison with the state-of-the-art, we compare our SBSM performance with the previous results provided on the same data set using the same data partitions for evaluation [7], [8], [21], [22] and ESF [36] and VFH [27] descriptors, as shown in Table I.

Subsequently, we show the importance of the weight propagation strategy in the SBSM descriptor by setting $\sigma = 1$ and $\sigma = 0$. These two values define the presence or absence of the propagation step, respectively. [4] Based on the mean data set samples volume radius length, we set $R = 0.15$. Results reported in Table I show that the SBSM descriptor clearly outperforms previous state-of-the-art results on this data set. In particular, it is concluded that using neighbor propagation, the performance improves by more than 16% the best result reported in [8] for this data set. This experiment shows that when a neighboring measure of the shape point is taken into account to update neighbor bins, the local variations of shape objects are better learnt by the classifier. Consequently, the intraclass variability is reduced without the need of increasing the computational complexity of the descriptor.

[3]The SBSM descriptor code is included as supplemental material.

[4]We also tested for different values of σ and experimentally found $\sigma = 1$ to obtain the best results.

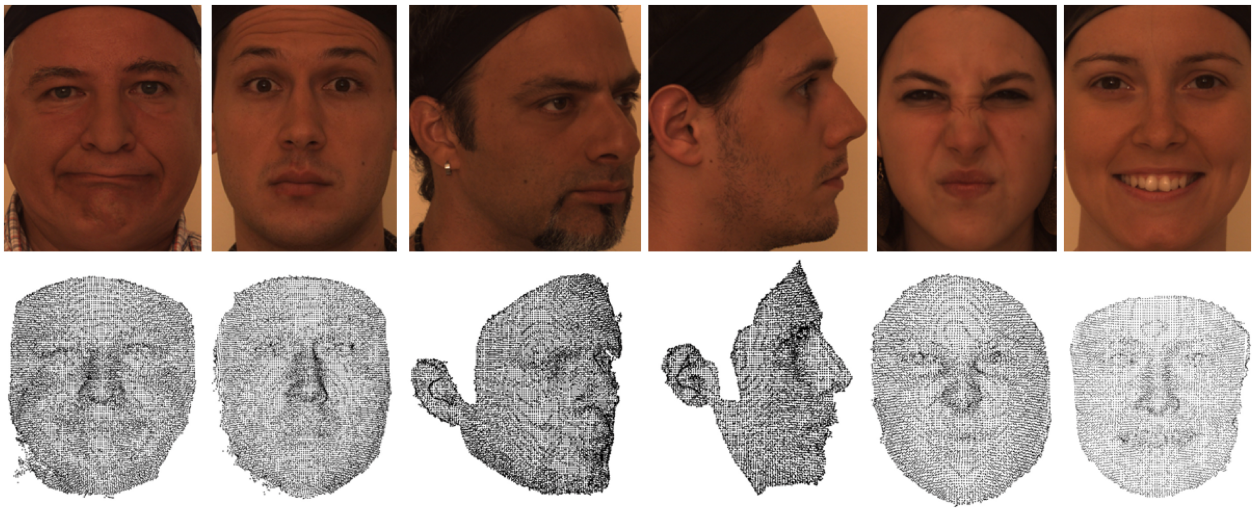


Fig. 6. Bosphorus 3-D face expression data set. Top: RGB samples. Bottom: corresponding point cloud samples.

TABLE I
CLASSIFICATION PERFORMANCE ON THE RGB-DEPTH DATA SET [21]

Method	Recognition rate
SIFT + Texton + Color + Spin [21]	64.7%
Sparse Distance Learning [22]	70.2%
VFH + SVM	77.5%
RGB-D Kernel Descriptors [7]	80.3%
Hierarchical Matching Pursuit [8]	81.2%
ESF + SVM	84.9%
SBSM, $\sigma = 0$ + SVM	96.7%
SBSM, $\sigma = 1$ + SVM	**97.9%**

TABLE II
CLASSIFICATION PERFORMANCE AND CONFIDENCE INTERVAL OF THE DIFFERENT DESCRIPTORS ON THE NOVEL AMERICAN SIGN LANGUAGE DATA SET

Method	Recognition rate
VFH + SVM	88.7%±0.2
ESF + SVM	92.3%±0.2
SBSM, $\sigma = 0$ + SVM	97.1%±0.6
SBSM, $\sigma = 1$ + SVM	**99.3%±0.4**

We also show the performance of the VFH, ESF, and SBSM on the novel ASL data set. The final performance is obtained by applying a stratified ten-fold cross-validation and testing the confidence interval with a two-tailed t-test. In this case, the spherical grid size is fitted to the minimum spherical grid size containing all the voxels for each data sample. Results are shown in Table II. One can see how our descriptor obtains better performances than using VFH or ESF, and that the best performance is achieved when weight propagation is taken into account.

Finally, we also compare the different descriptors on the public Bosphorus 3-D face expression data set. In this experiment, we perform user recognition from the set of 21 users taking into account the nine different facial expressions as well as the different head poses present in the data set. Applying stratified 5-fold cross-validation, the recognition rate results for ESF, VFH, and SBSM are shown in Table III. One can see that the proposed descriptor performs substantially better than ESF and VFH counterparts. In addition, it is also shown that considering the blurring degree to be 1 the performance of the SBSM descriptor is improved. Looking at the confidence intervals of Tables II and III, one can also see that SBSM variance in the recognition rate is smaller in comparison to the rest of methods, despite VFH and ESF which are kept small for ASL data set.

TABLE III
CLASSIFICATION PERFORMANCE AND CONFIDENCE INTERVAL OF THE DIFFERENT DESCRIPTORS ON THE PUBLIC BOSPHORUS 3-D FACE EXPRESSION DATA SET

Method	Recognition rate
VFH + SVM	69.7%±1.2
ESF + SVM	66.3%±2.1
SBSM, $\sigma = 0$ + SVM	73.1%±1.1
SBSM, $\sigma = 1$ + SVM	**77.3%±1.0**

2) *Robustness to Noise and Deformations:* In this section, we demonstrate the robustness of the SBSM when describing and classifying 3-D objects that suffer from noisy captures and deformations due to different ambient conditions or deviations captured by the sensor, as well as partial occlusions. To achieve this goal, we designed two different settings. In the first one, we analyze the robustness of the descriptor when objects suffer from deviations in the depth dimension in a range from 0 up to 20 mm in both directions of the z-axis [Fig. 7(a)–(c)]. This distortion simulates non-accurate reading errors of the sensors because of distance precision and ambient conditions. In the second test we progressively remove shape points from the point cloud from 0 up to 50% of the voxels for each object sample [Fig. 7(d)–(f)]. This distortion simulated local occlusions and reading errors that may produce the removal of some voxel points of the region of interest. Thus, in the depth distortion, the resulting point cloud has the same number of available voxels, though they are distorted in the z-axis, meanwhile in the removing distortion, the resulting point cloud contains less voxel points based on the distortion percentage.

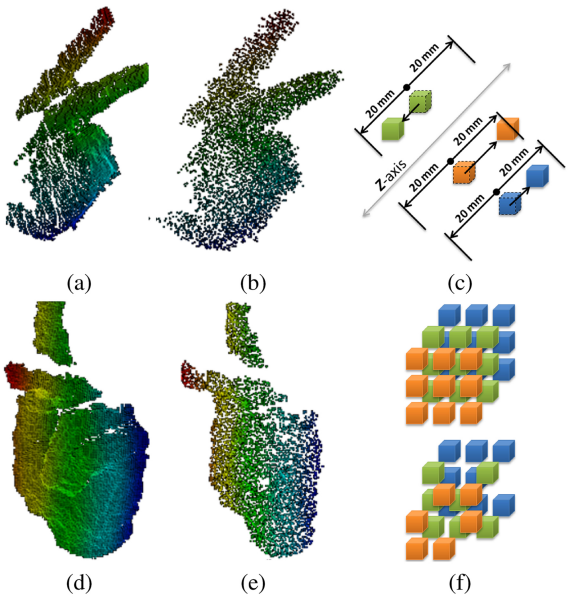


Fig. 7. (a) Input point cloud for a hand pose instance. (b) Example of distortion in the depth axis for (a). (c) For this distortion each voxel is randomly displaced in the z-axis with a maximum distortion of 20 mm in both directions of the axis. (d) Input point cloud for a hand pose instance. (e) Example of cloud removal distortion for (d). (f) For this distortion, each voxel of the original point cloud (top) is removed based on a probability value defined by the distortion (bottom).

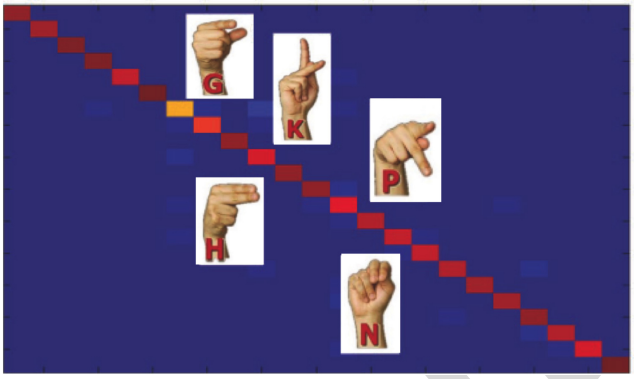


Fig. 8. Mean confusion matrix of the ASL data set using the SBSM descriptor $\sigma = 1$. The five most confused categories are displayed.

We fixed this range of distortions to be representative of the maximum distortion that we can find on recorded samples in real scenarios under different conditions and errors produced by different types of depth sensors.

In order to perform these analysis, we selected those five categories from the novel 3-D human poses data set that achieved the highest confusion in the previous section. The chosen hand categories are displayed in the confusion matrix of Fig. 8. In this image the confusion matrix is the mean computed for SBSM on that particular data set.

We show the recognition rate results when applying distortion in the depth axis in Fig. 9: for each degree of distortion, the mean recognition rate and confidence interval for 10 runs of ESF, VFH, and SBSM $\sigma = 1$ descriptors are computed. At each iteration, the percentage of distortion is randomly computed for each object voxel within different depth ranges

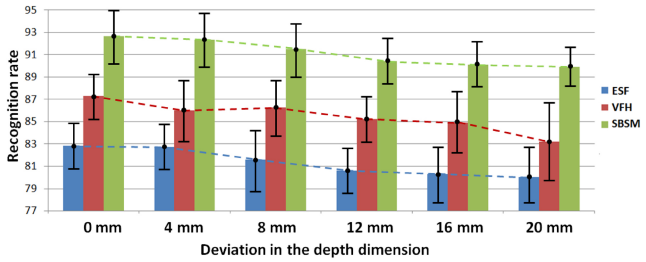


Fig. 9. Classification performance of different classification strategies under different degrees of distortion in the depth axis on the five selected categories in the ASL data set.

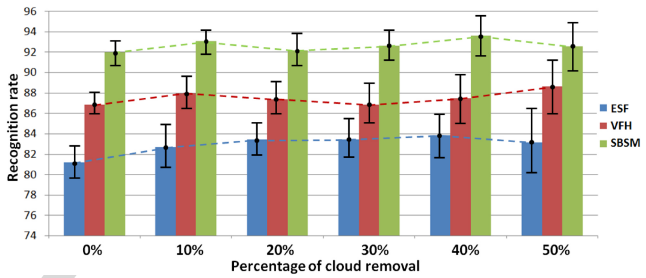


Fig. 10. Classification performance of different classification strategies under different degrees of cloud removal on the five selected categories in the ASL data set.

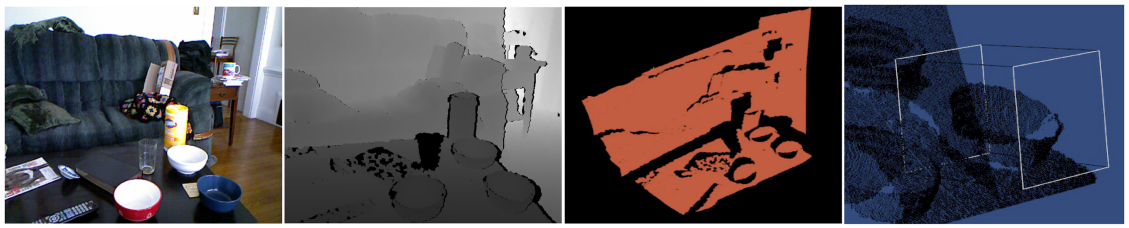
TABLE IV
MEAN RANK FOR THE COMPARED DESCRIPTORS CONSIDERING ALL THE EXPERIMENTS

	VFH	ESF	**SBSM**
Mean rank	2.13	2.86	**1.00**

in millimeters. The maximum value was set to 20 mm since it is hard to obtain higher deformations produced by the sensor. As expected, the recognition rate for all three descriptors decrease w.r.t. the depth distortion. One can see that for all the different tests of this experiments and descriptors, SBSM still obtains the best performance and VFH suffers the worst decrease in recognition (around 4%).

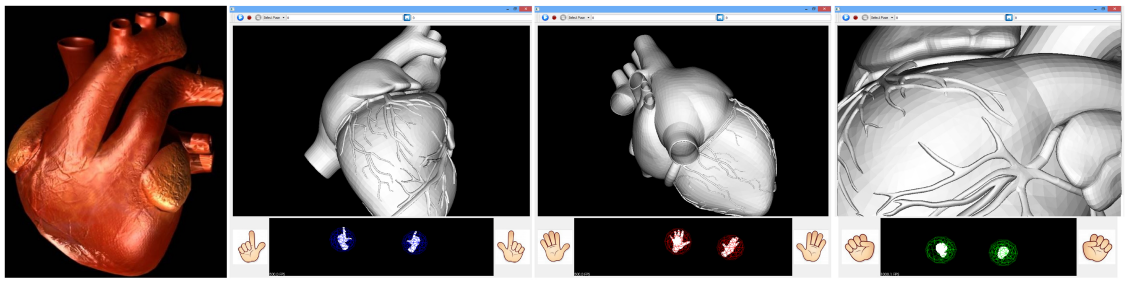
In Fig. 10, we show the results when applying cloud removal. For each percentage of removed voxels, the mean recognition rate and confidence internal are shown. At each iteration, a percentage of distortion is randomly generated, and different voxels are removed at each time satisfying the percentage of information to be removed. One can see that the general performance ranking in recognition is SBSM, VFH and finally ESF descriptor. Moreover, independently of the percentage of removed number of shape points, the recognition rate for the three methods is maintained in a small range of performance.

3) *Statistical Significance:* In order to compare the performances computed by the different experiments considered, Table IV shows the mean rank for each descriptor considering 15 different experiments (three data sets and 6×2 distortion experiments). The rankings are obtained by estimating each particular ranking r_i^j for each data set and experiment i and each descriptor strategy j, and computing the mean ranking R for each configuration as $R_j = \frac{1}{j}\sum_i r_i^j$, where J is the total number of tests.



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

Fig. 11. Object spotting in 3-D scenes. (a) Example of RGB image of a multimodal Berkeley data set. (b) Depth image of the same scene. (c) Computed point cloud from the scene. (d) Bowl spotting using SBSM (first positive 3-D object prediction is shown based on minimum Euclidean distance).



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

Fig. 12. (a) Original 3-D Heart volume of http://thefree3dmodels.com. Automatic interaction with the volume with (b) translation, (c) rotation and (d) zoom manipulation.

In order to reject the *null hypothesis*, i.e., measured ranks may not differ from the mean rank and these may be also affected by randomness in the results, we use the Friedman test [13]. The Friedman statistic value is computed as follows:

$$X_F^2 = \frac{12J}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)^2}{4}\right]. \quad (12)$$

In our case, since $K = 3$ descriptors are compared, $X_F^2 = 25.74$. This value is undesirable conservative, so Iman and Davenport proposed a corrected statistic instead

$$F_F = \frac{(J-1)X_F^2}{J(K-1) - X_F^2}. \quad (13)$$

Applying this correction, we obtain $F_F = 84.76$. With K=3 methods and J=15 experiments, F_F is distributed according to the F distribution with $(K-1) = 2$ and $(K-1)\cdot(J-1)$=28 degrees of freedom. The critical value of $F(2, 28)$ for 0.05 is 3.34. As the value of $F_F = 84.76$ is clearly higher than 3.34, we can reject the null hypothesis.

Once we have checked for the nonrandomness of the results, we perform a post ad-hoc test to check if one of the configurations could be statistically singled out. For this purpose, we use the Nemenyi test, in which the Nemenyi statistic is obtained as follows:

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6J}}. \quad (14)$$

In our case with $K = 3$ description strategies to compare and $J = 15$ tests, the critical value for a 95% of confidence ($q_\alpha = 2.35$) is $CD = 0.85$. As the ranking of the proposed SBSM approach does not intersect with any rank for that value of the CD, we can state that for the reported experiments our results are statistically significant with respect to VFH and ESF results. In the case of VFH and ESF descriptors, since their rank intersect with the CD value we can not state that there exists statistical differences between both strategies.

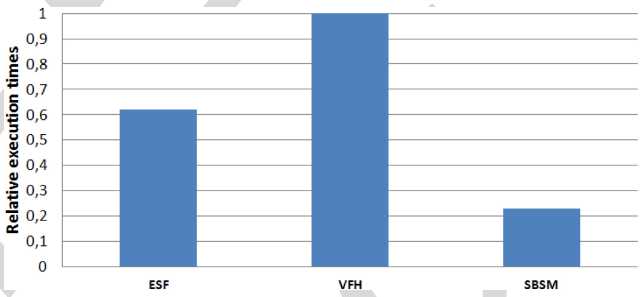


Fig. 13. Mean relative execution time in the range [0, ..., 1] among ESF, VFH, and SBSM descriptors. The relative execution time value is computed in proportion to the slowest method (set to value 1).

Finally, in order to compare the execution times of the considered ESF, VHF, and SBSM descriptors, Fig. 13 shows their mean relative execution time considering all the performed experiments. In order to compare the description complexity, the relative execution times only considers the description step, without taking into account the learning strategy. One can see that the proposed SBSM descriptor does not only obtain the best recognition rates, but also is more efficiently computed than the methods in the comparative. In particular, SBSM is more than four times faster in comparison to VFH and more than two times faster in comparison to ESF.

IV. Qualitative Analysis of SBSM

In this section, we present four real applications that use the proposed SBSM descriptor, object spotting in 3-D scenes and three fully-functional applications for a real-time HCI: 3-D

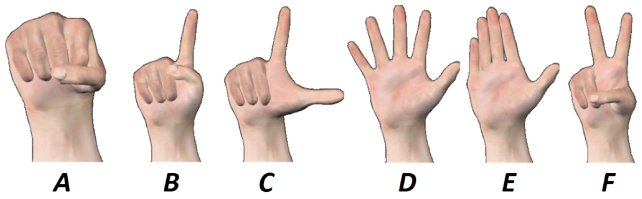


Fig. 14. HCI hand poses data set categories.

medical volume navigation, intelligent retail, and living labs; the library of the future. For this task, an additional novel set of hand poses for the HCI navigation applications was designed.

A. Object Spotting in 3-D Scenes

After showing the discriminative power and robustness of our proposed descriptor, we next illustrate the generality of SBSM when applied in real scenarios. To achieve this end, we consider a single scene obtained in the public 3-D Berkeley data set.[5] The public RGB and corresponding depth map for the selected 3-D scene are shown in Fig. 11(a) and (b), respectively. Using these data, we computed the point cloud scene shown in Fig. 11(c). In this particular case, we selected one *bowl* object to describe it and perform object spotting within the whole scene. We manually selected one bowl from a training image, computed its SBSM descriptor, and performed sliding window search over the three dimensions of the point cloud shown Fig. 11(c) for different scale hypotheses of the target object. SBSM descriptor size was set to $N_L = 8$, $N_\theta = 8$, $N_\phi = 8$, and $\sigma = 1$, with a total descriptor length of 512. The radius of the sphere is set in the range 5 to 20 cm, with an increment of 1 cm, and two voxel displacement increments in the three axis among iterations of the sliding windows approach. The first matched region of interest based on the best score obtained by the minimum Euclidean distance among the computed descriptors in the scene is shown in Fig. 11(d). Note the accurate fitting of the captured 3-D bowl object in the test 3-D scene. Moreover, the system spends 13 s in a conventional 2.7 GHz 2Core 4 Gb RAM computer to run this experiment for the tested scene. Given the performed exhaustive search and that we ran the experiment iteratively without any kind of parallelism, this experiment show the generality of the descriptor to be applied for scene analysis purposes.

B. HCI Application for Medical Navigation

Given the high discriminative power and fast computation of the proposed descriptor in comparison to the state-of-the-art approaches, we also designed different fully-functional applications to take benefit from it. The first application is an automatic HCI system for medical volume navigation, which is able to detect user, hands, poses, and gestures, manipulating a medical volume of interest.

The application was developed based upon the MS Kinect SDK to capture the RGB-D data stream from the depth sensor. The depth maps were converted into world coordinates projecting the registered 3-D image by means of pin-hole model and intrinsic camera parameters. Then, point cloud library (PCL) is used to work with the point cloud, and VTK framework is used for medical image visualization purposes. The hand detection algorithm takes advantage of the Body Pose Skeleton to find the hand wrist joints. Using this information, a fixed radius of interest of 15 cm centered in each joint is defined to segment hand point clouds. The usage of the skeleton also enables us to define a heuristic to discard false positive hand poses in the cases that the hands are near the body, or below the waist line. In a later step, the detected hand point clouds within the sphere are used to refine sphere center by computing center of mass and relocating sphere center, and points are classified using the proposed SBSM methodology. A reduced data set of 18K samples for the six classes shown in Fig. 14 were recorded and trained with SBSM and SVM for this purpose.

The detected pose label is then combined with the hand movement (3-D object displacements in real coordinates), and used as input of a hidden Markov model to obtain a hand gesture classification. The system is able to recognize and control zoom, rotation, and translation operations. In the user interface, the user can visualize its detected hand point clouds in real-time, having a feed-back with the displayed prototypes of both recognized hand poses. Also, the user can visualize the volumetric medical model and the real-time interactions caused by the hand gestures. The overall application enables a powerful and automatic volumetric medical model interaction and visualization. Fig. 12 shows some examples of detected poses, gestures, interactions, and visualizations on a public 3-D heart volume.[6]

C. HCI Application for Intelligent Retail

Based on the same procedure for hand pose recognition than in the medical volume navigation approach, we designed a fully-functional application for intelligent retail. In this scenario, hands are tracked in a multidevice setup and two main poses (open and close hand) are recognized. The user can automatically interact with a set of products in an interface designed with the Unity engine[7] so that information about the product and manipulation by 3-D rotation based on hand trajectories on two main object axis can be performed. In this scenario, given that we work with bigger displays and only one Kinect may not cover most part of the pointcloud related to one hand, we included an extra Kinects in order to recover voxels of the same hand from near complementary views and reconstruct a new hand with more voxel information. Thanks to the fusion of two views, more information about the tracked hand is available, and thus we make the descriptor more discriminative to classify multiple hand poses useful for interaction purposes.

In order to achieve registration of two Kinect cameras, we have mounted two cameras in a rigid setup with a baseline of 1.30 m and an angle of 18 degrees. In this way we can acquire depth images and convert each pixel to real-world coordinates by applying the intrinsic parameters of the cameras. In order to complete the registration, we need the

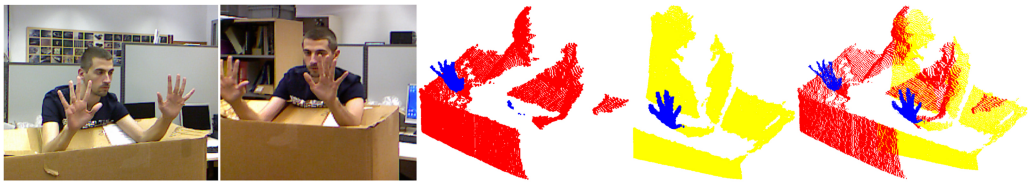
[5]http://kinectdata.com/

[6]See the supplemental material video for a system demonstration.

[7]http://unity3-D.com/

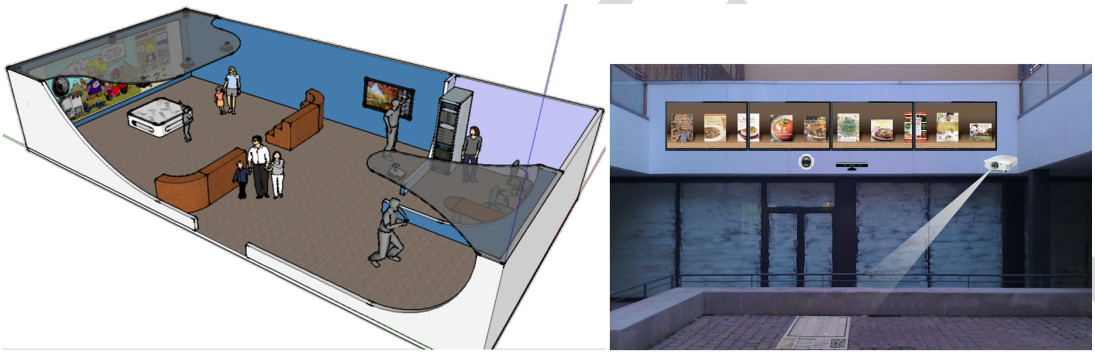
(a) (b) (c)

Fig. 15. HCI for retail. (a) Designed 3-D retail scenario using Unity engine. (b) User interaction with the scenario. (c) User manipulation by 3-D rotation.



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

Fig. 16. (e) Hand reconstruction using (a) and (b) two Kinect views (RGB) and (c) and (d) point clouds.



(a) (b)

Fig. 17. (a) Different HCI applications designed to implement within the Living Lab project. (b) Real scenario simulating the setup of the implementation of the prototype.

extrinsic parameters between both depth cameras, represented by a rotation matrix and a translation vector that allows us to convert $Kinect_2$ depth points to $Kinect_1$ depth coordinate system. Intrinsic and extrinsic parameters can be obtained by doing a stereo calibration. In our case, we use the *Camera Calibration Toolbox* [9], that gives us a rotation vector related through the *rodrigues* formula and a translation vector. For this procedure we only need few images from both depth cameras looking at the same planar checker-board pattern. Once the calibration is performed we can related $Kinect_1$ and $Kinect_2$ depth points as follows:

$$\mathcal{P}_{reference} = \mathcal{R}\mathcal{P} + \mathcal{T}$$

being $Kinect_1$ be the camera reference, $\mathcal{P}_{reference} = (\mathcal{X}_{ref}, \mathcal{Y}_{ref}, \mathcal{Z}_{ref})$ a point in the camera reference coordinate system, $\mathcal{R}$ the rotation matrix obtained in calibration step, $\mathcal{P}$ a point of the $Kinect_2$, and $\mathcal{T} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ the translation vector obtained in calibration step.

Fig. 16 shows an example of a hand reconstruction by two different views. From the 3-D reconstructed hand, SBSM is computed, and the automatic interactive process is performed. The designed 3-D scenario and real-time use case scenarios are shown in Fig. 15. In this case we found that increasing the point of view of the hand because of the use of two cameras allowed for a more natural movements of the user while keeping the recognition rate of the system.[8]

D. HCI Application in Living Labs

For this last scenario, we designed a first prototype for intelligent interaction in a virtual repository of books within a Smart Environment or Libing Lab corresponding to a new generation of libraries (VL^3 [6]Volpelleres Library Living Labs' project). For this scenario, a one Kinect camera setup was designed, and the same recognition procedure than in the retail scenario was performed. The 3-D scenario was designed with the Unity engine. The real environment to implement the prototype is shown in Fig. 17, which is located in the region of Volpelleres in Barcelona. In this prototype, the user is able to navigate through a catalogue of books and read the selected ones. The designed 3-D scenario and real-time use case scenarios are shown in Fig. 18.

[8]See the supplemental material video for a system demonstration.

(a) (b) (c)

Fig. 18. HCI Living lab prototype use case scenarios. (a) Designed 3-D Unity library environments. (b) Book appears on the scene and the user can open the hand and move through different books. (c) When close hand pose is performed, the book is selected and opened for reading.

Finally, it is important to remark that the Kinect acquired a variable frame rate between the range 20–30 FPS. Our procedure works real time for all the presented HCI applications, and thus, we are able to process (segment, describe, and classify segmented point clouds from the region of interest) for all the frames acquired by the Kinect device.

V. Conclusion

We presented the SBSM descriptor. SBSM is computationally efficient and highly discriminative. The computed descriptor codifies the spatial relations among object voxels and spherical bins given a granularity degree and a blurring factor defined by a Gaussian-based weight propagation function. The descriptor is rotation invariant by relocating descriptor bins using the two main quaternion based on the two major 3-D descriptor axis densities. In our experimental evaluation, we found that our methodology outperformed state-of-the-art results up to 16.7% on public depth multiclass object recognition data, also obtaining significant performance improvements in other two data sets; a novel ASL multiclass data set and a public 3-D face expression data set. Moreover, we tested the descriptor against different 3-D data distortions, obtaining high recognition rates and significant performance improvements in relation to standard approaches.

We also tested the descriptors in four real scenarios: object spotting in 3-D scenes, within a probabilistic gesture recognition pipeline for real-time HCI in medical volume navigation scenarios, HCI for intelligent retail in a multicamera setup, and within a prototype for catalogue navigation in a living lab library, showing the high discriminative power, efficiency and generality of the proposed descriptor to be applied in new generation of HCI applications and smart environments.

Acknowledgment

The authors would like to thank M. Pousa, C. Antens, J. Abella, A. Borras, M. Angel Blanco, J. Mas, R. Alcaide, A. Sabate, J. R. Jiménez, and D. Karatzas, from the Computer Vision Center, for their support during the development of the HCI applications.

References

[1] J. Abella *et al.*, "Multi-modal descriptors for multi-class hand pose recognition in human computer interaction systems," in *Proc. Chalearn Multi-Modal Gesture Recognit. Workshop ICMI*, 2013, pp. 503–508.

[2] M. Alexa and A. Adamson, "On normals and projection operators for surfaces defined by point sets," in *Proc. Eurographics Symp. Point-Based Graph.*, 2004, pp. 149–155.

[3] L. Alexandre, "3-D descriptors for object and category recognition: A comparative evaluation," in *Proc. Int. Conf. IROS*, 2012.

[4] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, "3-D shape histograms for similarity search and classification in spatial databases," in *Proc. 6th Int. Symp. Advances Spatial Databases*, 1999, pp. 207–226.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[6] M. Ben-Chen and C. Gotsman, "Characterizing shape using conformal factors," in *Proc. Eurographics Conf. 3DOR*, 2008, pp. 1–8.

[7] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. Int. Conf. IROS*, 2011.

[8] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Proc. ISER*, Jun. 2012, pp. 387–402.

[9] J.-Y. Bouguet. (2013). "Camera calibration toolbox for MATLAB." [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/.

[10] G. Burel and H. Henoco, "Determination of the orientation of 3-D objects using spherical harmonics," *Graph Models Image Process*, vol. 57, no. 5, pp. 400–408, 1995.

[11] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, 2011.

[12] C. Colombo, A. D. Bimbo, and A. Valli, "Visual capture and understanding of hand pointing actions in a 3-D environment," *IEEE Trans. Syst., Man, Cybern. B*, vol. 33, no. 4, pp. 677–686, Aug. 2003.

[13] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learning Res.*, vol. 7, pp. 1–30, Jan. 2006.

[14] S. Escalera, A. Fornes, O. Pujol, J. Llados, and P. Radeva, "Circular blurred shape model for multiclass symbol recognition," *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 41, no. 2, pp. 497–506, Apr. 2011.

[15] S. Escalera, A. Fornes, O. Pujol, P. Radeva, G. Sanchez, and J. Llados, "Blurred shape model for binary and grey-level symbol recognition," *Pattern Recognit. Letters*, vol. 30, no. 15, pp. 1424–1433, 2009.

[16] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *Proc. ECCV*, 2004.

[17] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. CVPR*, 2010, pp. 755–762.

[18] Point Cloud Library. (2013) [Online]. Available: http://docs.pointclouds.org/trunk/a02944.html.

[19] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3-D shape descriptors," in *Proc. SIGGRAPH Symp. Geometry Process.*, 2003, pp. 156–164.

[20] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. V. Gool, "Hough transform and 3-D SURF for robust three dimensional classification," in *Proc. ECCV*, 2010.

[21] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. Int. Conf. Robot. Autom.*, 2011.

[22] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *Proc. Int. Conf. Robot. Autom.*, 2011.

[23] N. Mitra, A. Nguyen, and L. Guibas, "Estimating surface normals in noisy point cloud data," *Int. J. Computational Geometry Applicat.*, vol. 14, nos. 4–5, pp. 261–276, 2004.

[24] C. Redondo-Cabrera, R. J. Lopez-Sastre, F. Acevedo-Rodriguez, and S. Maldonado-Bascon, "SURFing the point clouds: Selective 3-D spatial pyramids for category-level object recognition," in *Proc. IEEE CVPR*, 2012, pp. 3458–3465.
[25] M. Ruggeri, G. Patane, M. Spagnuolo, and D. Saupe, "Spectral-driven isometry-invariant matching of 3-D shapes," *IJCV*, vol. 89, pp. 248–265, 2010.
[26] R. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3-D registration," in *Proc. Int. Conf. Robot. Autom.*, 2009, pp. 1848–1853.
[27] R. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3-D recognition and pose using the viewpoint feature histogram," in *Proc. Int. Conf. IROS*, 2010, pp. 2155–2162.
[28] R. Rusu, Z. Marton, N. Blodow, and M. Beetz, "Learning informative point classes for the acquisition of object model maps," in *Proc. Control Autom. Robot. Vision*, 2008, pp. 643–650.
[29] D. Saupe and D. V. Vrani, "3-D model retrieval with spherical harmonics and moments," in *Proc. DAGM*, vol. 2191, 2001, pp. 392–397.
[30] A. Savran, B. Sankur, and M. Bilge, "Regression-based intensity estimation of facial action units," *Image Vision Comput.*, vol. 30, no. 10, pp. 774–784, 2012.
[31] J. Shen, D. Wang, and X. Li, "Depth-aware image seam carving," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1453–1461, Oct. 2013.
[32] K. Shoemake, "Animating rotation with quaternion curves," in *Proc. Comput. Graph. Interactive Techniques*, pp. 245–254, 1985.
[33] H. P. H. Shum, E. S. L. Ho, Y. Jiang, and S. Takagi, "Real-time posture reconstruction for Microsoft Kinect," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 43, no. 5, pp. 1357–1369, Oct. 2013.
[34] B. Steder, R. Rusu, K. Konolige, and W. Burgard, "NARF: 3-D range image features for object recognition," in *Proc. Int. Conf. IROS*, 2010.
[35] F. Tombari, S. Salti, and L. D. Stefano, "Unique signatures of histograms for local surface description," in *Proc. ECCV*, 2010, pp. 356–369.
[36] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3-D object classification," in *Proc. IEEE ROBIO*, 2011, pp. 2987–2992.
[37] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *Proc. ICCV*, 2013.
[38] P. Zhang, Z. Wang, S. Zheng, and X. Gu, "A design and research of eye gaze tracking system based on stereovision," *Emerging Intell. Comput. Technol. Applicat., Lecture Notes Comput. Sci.*, vol. 5754, no. 4, pp. 278–286, 2009.

Oscar Lopes received the master's degree in Computer Vision and in Multimedia Technologies from Universitat Autònoma de Barcelona (UAB), Barcelona, Spain. He is currently pursuing Ph.D. degree in the Information and Computing Sciences at the University of Utrecht, Utrecht, The Netherlands.

His current research interests include pattern recognition, machine learning, and mobile computing, to leverage the creation of rich HCI systems based on gestural semantics.

Mr. Lopes collaborated in the design of a hand pose and gesture recognition system for which he was awarded 3rd place at the ICPR held in Tsukuba, Japan, in 2012.

Miguel Reyes received the bachelor's degree in computer science at Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2010, and the Master's degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2011. He is currently pursuing the Ph.D. degree in math in the area of computer science and artificial intelligence with the University of Barcelona, Barcelona.

His current research interests include pattern recognition, signal processing and visual object recognition, and their application to health care systems.

Mr. Reyes is a member of the Human Pose Recovery and Behavior Analysis group and the Computer Vision Center.

Sergio Escalera received the Ph.D. degree on multiclass visual categorization systems at Computer Vision Center, Universitat Autònoma de Barcelona (UAB), Barcelona, Spain.

He leads the Human Pose Recovery and Behavior Analysis Group. He is currently a Lecturer with the Department of Applied Mathematics and Analysis, Universitat de Barcelona, Barcelona, Spain. He is the Editor-in-Chief of American Journal of Intelligent Systems and the Advisor and Director of the ChaLearn Challenges in Machine Learning. He is also a part time Professor at Universitat Oberta de Catalunya, Barcelona, Spain. His current research interests include statistical pattern recognition, visual object recognition, and HCI systems, with special interest in human pose recovery and behavior analysis.

Dr. Escalera received the 2008 Best Thesis Award on Computer Science at UAB. He is a member of the Computer Vision Center at Campus UAB.

Jordi Gonzàlez received the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2004.

He is currently an Associate Professor in computer science with the Department de Ciències de la Computació at UAB. He is also a Research Fellow at the Computer Vision Center, Barcelona. His current research interests include cover pattern recognition and machine learning techniques for the computational interpretation of human behaviors in image sequences and video hermeneutics.

Dr. Gonzàlez has coorganized Special Issues in IJPRAI journals in 2009, CVIU journals in 2012, and MVA journals in 2013. He is a member of the Editorial Board of CVIU and IET-CVI journals.

Spherical Blurred Shape Model for 3-D Object and Pose Recognition: Quantitative Analysis and HCI Applications in Smart Environments

Oscar Lopes, Miguel Reyes, Sergio Escalera, and Jordi Gonzàlez

***Abstract*—The use of depth maps is of increasing interest after the advent of cheap multisensor devices based on structured light, such as Kinect. In this context, there is a strong need of powerful 3-D shape descriptors able to generate rich object representations. Although several 3-D descriptors have been already proposed in the literature, the research of discriminative and computationally efficient descriptors is still an open issue. In this paper, we propose a novel point cloud descriptor called spherical blurred shape model (SBSM) that successfully encodes the structure density and local variabilities of an object based on shape voxel distances and a neighborhood propagation strategy. The proposed SBSM is proven to be rotation and scale invariant, robust to noise and occlusions, highly discriminative for multiple categories of complex objects like the human hand, and computationally efficient since the SBSM complexity is linear to the number of object voxels. Experimental evaluation in public depth multiclass object data, 3-D facial expressions data, and a novel hand poses data sets show significant performance improvements in relation to state-of-the-art approaches. Moreover, the effectiveness of the proposal is also proved for object spotting in 3-D scenes and for real-time automatic hand pose recognition in human computer interaction scenarios.**

***Index Terms*—Depth image analysis, human computer interaction (HCI), image descriptors, object and pose recognition, smart environments.**

I. Introduction

COMPUTER vision research on 3-D point cloud analysis has recently received a lot of attention because of the availability of cheap multisensor devices based on structured light, such as Kinect. This RGB-Depth camera is compact and portable, so it can be easily installed in any environment to understand 3-D scenes. This way there are multiple applications which can benefit from the analysis of 3-D objects in scenes [1], [17], [31], [33], [37]. However, recognition of 3-D objects is still a challenging problem; in addition to the typical issues tackled by 2-D object recognition approaches (such as robustness to noise and occlusions, discriminate power, and computational complexity), the captured sequences are usually sampled at discrete points, so the finer details of the 3-D object are usually lost.

Under these assumptions, there exists a strong interest for designing new 3-D object descriptors [3], [18], [25], [34]. We next revisit the literature by dividing the existing approaches into those descriptors based on pure 3-D geometric properties and those extended from already existing 2-D object descriptors.

Describing 3-D geometric information has been proven to be useful when classifying everyday objects like cans, glasses or doors, and for 3-D scene analysis. For example, some approaches take into account the set of normals of the surface defined by a given point and its neighbors [2], [23]. As an example, the SHOT descriptor proposed in [35] defines a surface representation based on point normals. It is based on counting the points that fall into bins according to a function of the angle between the normal at each point within the corresponding part of the grid and the normal at the feature point. However, in this case, the descriptor is local and usually requires a previous keypoint detection step, which complicates its adaptation to recognize nonrigid shapes. The use of normals are useful to recognize 3-D objects since they encode the implicit surface that neighboring points define, although they depend on the density of the underlying points and the smoothness of 3-D object surfaces to give accurate results. Also, spherical harmonics [10] have been used to design 3-D descriptors invariant to rotation [19] or have been considered directly as features [29]. Conformal factors have also been considered [6], measuring the relative curvature of a vertex given the total curvature. The result can be viewed as a vector which is not only invariant to rigid body transformations, but also to changes in the pose. The point feature histogram (PFH) local descriptor proposed in [28] is used to recognize points conforming planes, cylinders, and other geometric primitives. As an extension, the fast PFH (FPFH) descriptor [26] is based on codifying angle relations among 3-D points. FPFH optimizes the PFH computation to make it usable in real-time 3-D registration applications. The

Manuscript received July 30, 2013; revised November 28, 2013 and February 14, 2014; accepted February 17, 2014. This work was supported in part by the European project DiCoMa under Grant TSI-020400-2011-55 and in part by the Spanish projects under Grant TIN2009-14501-C02-02 and Grant TIN2012-39051. This paper was recommended by Associate Editor X. Li.

O. Lopes is with the University of Utrecht, Utrecht 3512 JE, The Netherlands (e-mail: oscar.pino.lopes@gmail.com).

M. Reyes is with the University of Barcelona, Barcelona 08007, Spain (e-mail: mreyes@gmail.com).

S. Escalera is with the department of Matemàtica Aplicada i Anàlisis, Facultat de Matemàtiques, Universitat de Barcelona, Barcelona 08007, Spain (e-mail: sergio@maia.ub.es).

J. Gonzàlez is with the department of Computer Science, Universitat Autònoma de Barcelona, Barcelona 08193, Spain (e-mail: poal@cvc.uab.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2307121

viewpoint feature histogram (VFH) [27] combines an extended version of FPFH with statistics between the viewpoint and the surface normals on the 3-D object. Recently, Wohlkinger and Vincze [36] have presented an ensemble of shape functions (ESF) approach to describe 3-D objects, which benefits from several combinations of histograms for codifying 3-D object relations of angles, areas, and distances among points.

Unfortunately, the point clouds captured from Kinect-like devices usually contain holes, since data are sampled at discrete points. Consequently, in all the aforementioned approaches (which rely on an accurate computation of 3-D geometric primitives) their performance is usually downgraded. Alternatively, some recent 3-D regional descriptors have been defined as an extension of classical derivative-based 2-D features, such as HOG, SIFT, and SURF [20], [24]. For example, as a generalization of the 2-D shape context descriptor presented in [5], Frome *et al*. [16] propose a 3-D shape context descriptor which is compared with a classical spin-image representation and a novel harmonic shape context (HSC) descriptor for 3-D car model classification. Despite the excellent results reported, these methods require the computation of a large number of shape points relations.

In this paper, we propose a novel 3-D object descriptor, called spherical blurred shape model (SBSM). SBSM is inspired in the blurred shape model (BSM) descriptor presented in [15] and [14]. The novel SBSM descriptor codifies the object structure density and local variabilities in the 3-D space. Similar to the Zoning descriptor and 3-D shape histograms of [4], SBSM bases on a linear computation of spatial relation of shape points to 3-D bin centroid, but including a propagation *blurring* degree to define a compact and discriminative 3-D object descriptor. In this sense, Zoning and the descriptor in [4] can be seen as an instance of the proposed descriptor when the defined blurring degree is null. As it is reported in the results, when increasing the blurring factor the overall classification rate of the system is improved. In addition to provide a 3-D generalization of BSM, SBSM introduces the following enhancements; 1) a 3-D spherical grid which partitions the 3-D space into 3-D shape bins, 2) a 3-D Gaussian-based weight propagation schema controlling the blurring level based on shape voxel distances, and 3) a quaternion-based rotation strategy based on sphere axis densities to define a 3-D rotation invariant descriptor. As a result, the proposed SBSM is a global descriptor that encodes the shape of an object, being rotation and scale invariant, computationally efficient, and highly discriminative.

We evaluate the descriptor on public and novel 3-D object and hand poses data set, showing significant performance improvements in comparison to the state-of-the-art approaches. We also test the descriptor in front of deformation in depth coordinates and noise point removal. As a result, we can show that the SBSM copes with the noise and occlusions typically present in the point clouds acquired by range scanner sensors. Additionally, we show four real applications where we apply the descriptor. In the first, we perform object spotting in public 3-D scenes. The last three applications correspond to human computer interaction (HCI) scenarios. In the second, we present a real-time fully-automatic HCI system for medical image volume navigation, segmenting human hands, and classifying multiple hand poses using the proposed SBSM descriptor. In the third, the same approach is applied in a multicamera setup to perform intelligent retail. Finally, we present a prototype for intelligent navigation through a repository of books in a living lab which may represent the library of the future. Although pointing recognition and gaze tracking in multicamera setups for HCI has been previously addressed in [12] and [38], here, we show how complex multicamera setups can be avoided and recognition for interaction can be improved within different HCI application contexts using the proposed approach.[1]

The rest of the paper is organized as follows. Section II presents the SBSM descriptor. SBSM is evaluated and compared to the state-of-the-art approaches in Section III. Section IV presents four real applications that uses the proposed descriptor, including 3-D object spotting in real scenes and different HCI scenarios. Finally, Section V concludes the paper.

II. Method

In this section, we present the novel SBSM to describe 3-D objects.

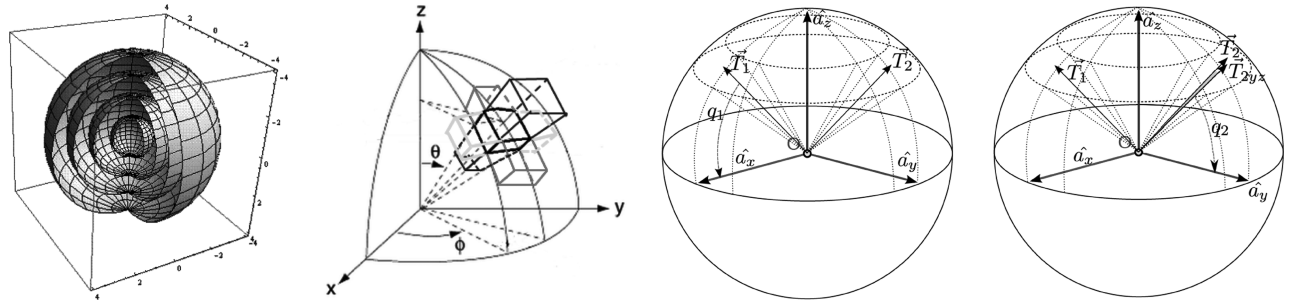
A. Spherical Blurred Shape Model

The SBSM is inspired in 3-D grid approaches and in the discriminative power of SIFT and HOG descriptors to codify object information based on the distribution of object gradients and orientations. However, instead of performing computation of 3-D object derivatives, SBSM just requires the computation of object shape voxel distances between neighbors in order to codify the object structure density and local variabilities in the 3-D space. As a result, SBSM is a computationally efficient descriptor, with a complexity linear to the number of object voxels $O(|P|)$, with an upper bound of $27 \cdot |P|$ simple operations for a point cloud P of $|P|$ shape points (defined based on a 26-connectivity of regions in the 3-D space of bins).

As in the case of 2-D and 3-D object descriptors, an initial grid is fitted to contain the region of interest to describe. In our case, to describe 3-D regions, a spherical grid containing a set of 3-D bins is defined, which contain the set of voxels P of the point cloud to be described. Our description methodology computes for each voxel P contained in the grid a set of voxel-bin spatial relations that are included in a global region descriptor. Next, we describe in detail each step of the description procedure.

In the first step, a discrete spherical grid partitions the 3-D space in a set of bins, as shown in Fig. 1(a). Let $P = \{p_i | p_i \in \mathbb{R}^3\}$, C, N_L, N_θ, N_ϕ, R, and σ define the set of voxels of the point cloud, point cloud centroid, number of layers, number of angular divisions for θ, number of angular divisions for ϕ, radius length, and sigma value for the gaussian distance

[1]We include as supplemental material the SBSM descriptor code, the novel ASL 3-D hand poses data set, and a demonstration video with the descriptor running real-time in different HCI scenarios.



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

Fig. 1. Illustration of SBSM descriptor computation. (a) Sphere bins. (b) Example of neighbor bins. (c) and (d) Example of the estimation of two main quaternion to rotate feature vector in the 3-D space.

metric, respectively. Some of these parameters are illustrated in Fig. 1(b). Then, $d_R = R/N_L$, $d_\theta = 2\pi/N_\theta$, and $d_\phi = 2\pi/N_\phi$ are computed as the distance between consecutive layers and the degrees in θ and ϕ polar coordinates between consecutive sectors, respectively. Using this division, we next define the set B of sphere bins $b_{\{i,j,k\}}$ as follows:

$$\begin{aligned} B = \{&b_{\{0,0,0\}}, ..., b_{\{i,j,k\}}, ..., b_{\{N_L-1,N_\theta-1,N_\phi-1\}}\} \\ &\forall i \in \{0, 1, ..., N_L - 1\} \forall j \in \{0, 1, ..., N_\theta - 1\}, \\ &\forall k \in \{0, 1, ..., N_\phi - 1\} \end{aligned} \tag{1}$$

where bin $b_{\{i,j,k\}}$ is the 3-D bin defined as the cartesian product of intervals $[i \cdot d_R, (i+1) \cdot d_R)$, $[j \cdot d_\theta, (j+1) \cdot d_\theta)$, and $[k \cdot d_\phi, (k+1) \cdot d_\phi)$ in relation to the center of the spherical grid, θ, and ϕ, respectively. This way B defines a partition in $\mathbb{R}^3$ of the object of interest. Then, the centroid coordinates for all each bin $b^*_{\{i,j,k\}} \in B^*$ are computed as follows:

$$\begin{aligned} b^*_{\{i,j,k\}} = &\left(i \cdot d_R + \frac{d_R}{2}, j \cdot d_\theta + \frac{d_\theta}{2}, k \cdot d_\phi + \frac{d_\phi}{2} \right) \\ &\forall i \in \{0, 1, ..., N_L - 1\}, \forall j \in \{0, 1, ..., N_\theta - 1\} \\ &\forall k \in \{0, 1, ..., N_\phi - 1\}. \end{aligned} \tag{2}$$

An example of some 3-D bin neighbors of the spherical descriptor is shown in Fig. 1(b). Once the 3-D spatial bins are defined, the SBSM feature vector is initialized as

$$W_i = 0, \forall i \in \{1, 2, ..., N_L \cdot N_\theta \cdot N_\phi\}. \tag{3}$$

Subsequently, for each voxel in the point cloud $p_z \in \{P | P \subset B\}$, the distance of that voxel to its neighbor bins is estimated based on a Gaussian distance metric, and the normalized weights are added to the corresponding descriptor bin locations. For this task, let $b_z = b_{\{i,j,k\}} | p_z \subset b_{\{i,j,k\}}$ be the bin containing voxel p_z. First, the lists containing bin weights and index bins for p_z are initialized to $W^* = \{0\}$ and $I^* = \{\{i, j, k\}\}$, respectively. Then, the iterative procedure updates W^* and I^* for each $b_{\{i,j,k\}} \in N(b_z)$, where $N(b_z)$ is the set of neighbors bins of b_z in a 27-neighborhood for inner sphere bins and 18-neighborhood for external sphere surface bins (including the reference bin). So the list of weights is updated as

$$W^* = W^* \cup \left\{ e^{-\frac{||p_z - b^*_{\{i,j,k\}}||}{R \cdot \sigma}} \right\} \tag{4}$$

and the list of indexes as

$$I^* = I^* \cup \{\{i, j, k\}\}. \tag{5}$$

As a result, the normalized weights for p_z are added to its corresponding positions of W as follows:

$$W_{I^*_i} = W_{I^*_i} + \frac{W^*_i}{\sum_{j=1}^{|W^*|} W^*_j}, \forall i \in \{1, 2, .., |I^*|\}. \tag{6}$$

In this way, a new shape point of the point cloud will include a weight to its belonging bin centroid and neighbor centroid based on a Gaussian function of the distance and a blurring level defined by σ. This values defines the degree of influence of each neighbor bin for each point cloud voxel. Note that when σ parameter is set to zero, the descriptor is equivalent to a dense sampling of the point cloud as in the classical state-of-the-art Zoning descriptor, but defined in the 3-D space [14]. It is important to remark that the voxels that are not contained within the spherical grid bins are not considered in the descriptor computation. On the other hand, the voxels that intersect with the spherical surface are c onsidered as inner voxels and thus, considered in the descriptor estimation. Fig. 2 shows an example of an hypothetical sphere slice for $\phi = k$ and the analysis of a point cloud voxel to update the SBSM descriptor.

Once the procedure is repeated for all points $p_z \in P$, the final feature vector W is normalized as follows:

$$W_i = \frac{W_i}{\sum_{j=1}^{N_L \cdot N_\theta \cdot N_\phi} W_j}, \forall i \in \{1, 2, .., N_L \cdot N_\theta \cdot N_\phi\}. \tag{7}$$

Given that all the voxels within the point cloud where the descriptor is computed contribute with the same cost and that the final vector is normalized, it becomes scale invariant. Thus, if different instances of a 3-D object category are fitted with the spherical descriptor, even with different sizes, all the descriptors are comparable and can be trained with the same classifier.

B. 3-D Rotation Invariant SBSM

Once SBSM is computed based on the predefined number of layers, bin orientations, and σ value for the Gaussian function, the descriptor is able to encode the local density and

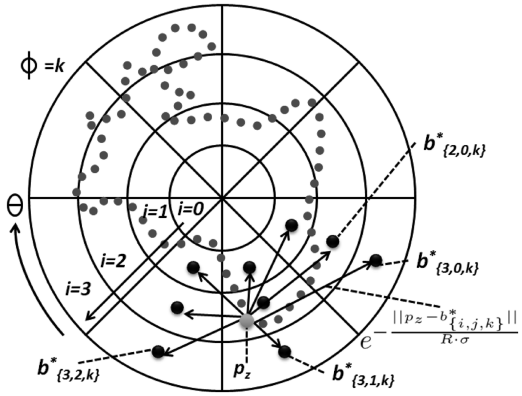


Fig. 2. Example of point cloud voxel for an hypothetical sphere slice for $\phi = k$. Voxels of the point cloud visible on that slice are shown as red dots. An example of a voxel estimation p_z is shown in green. For this point, neighbor bins centroids are shown as black dots. For each of these relations (note that in the 3-D space a total of 27 relations will be computed), equation 4 is computed, and the estimated value is added to descriptor position corresponding to its corresponding bin.

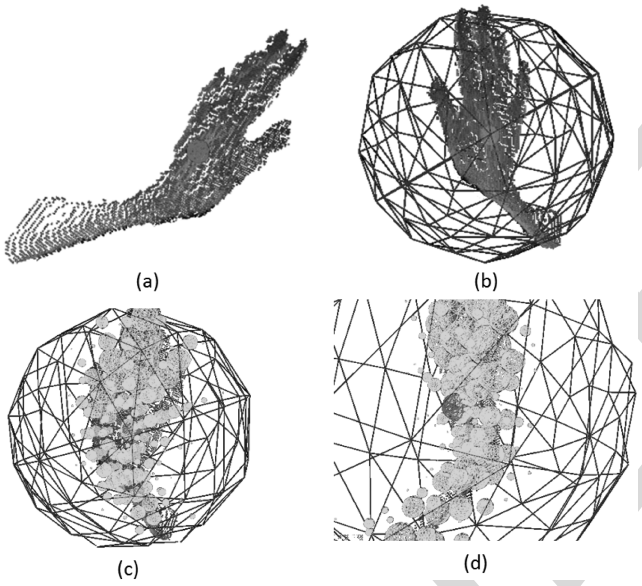


Fig. 3. (a) Initial hand point cloud and computed center. (b) Sphere including a point cloud corresponding to a 3-D hand pose. (c) Same sphere where SBSM descriptor has been computed. The density of the green dots represents the centroid bin values, and the whole descriptor has been rotated based on the quaternion codified by two main descriptor axis densities. (d) Alternative view of the computed SBSM descriptor.

spatial relations of 3-D shape points for a particular granularity degree. Rotation invariance is achieved by considering the main spherical axis densities to compute the main vector orientations in quaternion coordinates as a reference axis that rotates feature vector bins. As a result, this feature vector reordering step makes the descriptor rotation invariant for similar 3-D objects.

The use of unit quaternion instead of rotation matrices provides a fast computation for rotation invariance, and at the same time, it is simpler to enforce that quaternions have unit magnitude than constrain rotation matrices to be orthogonal [32]. The procedure is detailed next.

First, we compute the density of the descriptor for each axis defined by the angles θ and ϕ as follows:

$$f(\theta, \phi) = \sum_{r=1}^{N_L} W_{\{r,\theta,\phi\}} \tag{8}$$

and the two maximum axis densities are found

$$\overrightarrow{T}_1 = \arg\max_{\theta,\phi} f(\theta, \phi),\ \overrightarrow{T}_2 = \arg\max_{\theta,\phi \setminus \overrightarrow{T}_1} f(\theta, \phi). \tag{9}$$

Back to Cartesian coordinates, we compute the component of $\overrightarrow{T}_2$ vertical to $\overrightarrow{T}_1$ by projecting $\overrightarrow{T}_2$ onto the plane perpendicular to $\overrightarrow{T}_1$ as follows:

$$\overrightarrow{T}_{2yz} = \overrightarrow{T}_2 - \overrightarrow{T}_1^T \overrightarrow{T}_2 \overrightarrow{T}_1. \tag{10}$$

Subsequently, we compute the rotation that aligns the axis $\overrightarrow{T}_1$, $\overrightarrow{T}_2$ with $\hat{a}_x$ and $\hat{a}_y$, respectively. Where $\hat{a}_x = [100]^T$ and $\hat{a}_y = [010]^T$. The rotation quaternion q can be computed as the combination of two quaternions q_1 and q_2, so that $q = q_2 q_1$, where q_1 rotates $\overrightarrow{T}_1$ to $\hat{a}_x$ and q_2 aligns $\overrightarrow{T}_2$ with $\hat{a}_y$ [Fig. 1(d)].

Finally, the values of the bin locations are rotated based on the quaternion q, such that each bin $b^{*r}_{i,j,k} \in B^*$ is computed as

$$b^{*r}_{i,j,k} = q b^*_{i,j,k} q^* \tag{11}$$

using the Hamilton product, where q^* is the conjugate of the quaternion q. Abusing of notation, b^* and b^{*r} also denote the corresponding pure quaternion to each bin. So, we take advantage of this rotation order to obtain the rotation invariant feature vector $W^r_{\{i,j,k\}} = W_{\{i,j,k\}}$.

An example of the two main quaternions for an hypothetical spherical toy problem is shown in Fig. 1(c) and (d), respectively. In Fig. 3(a), a real example of a 3-D hand pose is shown. Fig. 3(b) shows the centered point cloud within the 3-D correlogram containing SBSM bins. The result after computing the SBSM descriptor from hand point cloud and performing rotation invariance is shown in Fig. 3(c) and (d) for two different points of view.

III. Quantitative Analysis of SBSM

In order to present the results, we first describe the training data and settings of the experiments.

A. Data Sets

We test our methodology on three data sets; a public 3-D object category data set, a new 3-D hand pose data set, and a public subject identification data set.

1) *RGB-D Object Data Set:* The RGB-D Object data set is a large collection of 300 common household objects [21]. All these objects are organized into 51 categories arranged using WordNet hypernym-hyponym relationships (similar to ImageNet). This data set was recorded using a Kinect style 3-D camera that recorded a set of synchronized and aligned 640×480 RGB-D images at 30 Hz. Each object was placed on a turntable, and the video sequences were captured for a single, full rotation. For each object, three video sequences

Fig. 4. RGB-Depth object data set category samples [21].

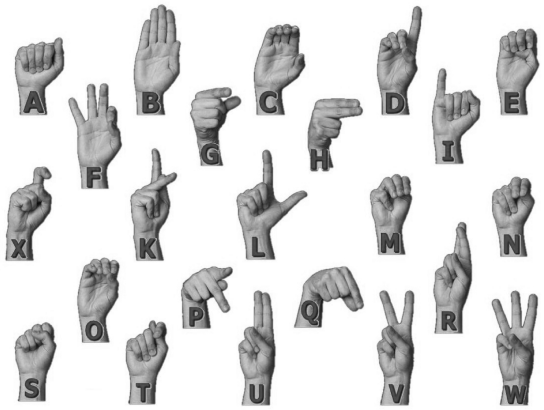


Fig. 5. American sign language data set categories.

were recorded with the camera mounted at different heights so that the object could be viewed from different angles with respect to the horizon. Example of segmented objects are shown in Fig. 4.

2) *American Signal Language Data Set:* We have recorded a novel 3-D hand poses data set based on the American sign language vocabulary. The data set is composed of 23 categories with around 47K instances of both hands. The Kinect device was used to extract the hands and their point clouds data using standard segmentation and detection algorithms. Also, both hands were captured not only considering a frontal view but also including variabilities in terms of scale, hand orientation, and finger joint articulations. This way, the complexity and variability of the overall data set was enriched. Examples of hand categories are shown in Fig. 5.[2]

3) *Bosphorus 3-D Face Expression Data Set:* The Bosphorus 3-D faces data set [30] contains several users performing nine natural facial expressions. From the Bosphorus face dataset it was considered a subset of 21 individuals, including 1039 samples. For each individual, the original structure of the dataset was kept, and all the corresponding nine facial expressions were considered. Each sample of the original Bosphorus dataset is composed by roughly 45000 points. For performance issues, it was performed a down-sampling using a Voxel grid filter, obtaining samples comprising approximately 9 000 points (some examples of the data set and the computed point clouds are shown in Fig. 6).

[2]Data set included as supplemental material (∼1Gb).

B. Settings and Evaluation Metrics

A multiclass classifier is trained using the proposed SBSM descriptor. Specifically, feature vectors are trained in a one-versus-one SVM classifier using a RBF kernel, and optimizing the parameters C and γ by means of cross-validation using LibSVM [11]. SBSM descriptor size was experimentally set to $N_L = 8, N_\theta = 8, N_\phi = 8$ for all the experiments, with a total descriptor length of 512. [3] We compare the SBSM descriptor with different state-of-the-art methods on different experiments [7], [8], [21], [22]. We also include in the comparative the VFH [27] and ESF [36] descriptors by also training the feature vectors with one-versus-one SVM classifier using a RBF kernel and optimizing the parameters as in the case of SBSM. VFH and ESF have been selected since they are recent, representative, and robust well-known descriptors for shape estimation and codification of normal vectors distribution.

We validate the object classification experiments by means of recognition rate applying stratified ten-fold cross-validation and estimating the confidence interval with a two-tailed t-test. We also test and compare the descriptors against depth distortions and noisy data to compute their statistical significance based on Friedman and Nemenyi statistics [13].

C. Experiments

We next present the multiclass 3-D object categorization performance using SBSM in the RGB-D object, sign language, and 3-D facial expression data sets. Once we demonstrate the performance of the proposed descriptor, we test it against different 3-D object distortions.

1) *Analysis of Classification Performance:* For the RGB-D object data set, we use the turntable data for both training and evaluation, thus classifying 51 different 3-D object categories using depth information only. For the object recognition experiments on cropped images, we apply a leave-one-out strategy as described in [21]. For comparison with the state-of-the-art, we compare our SBSM performance with the previous results provided on the same data set using the same data partitions for evaluation [7], [8], [21], [22] and ESF [36] and VFH [27] descriptors, as shown in Table I.

Subsequently, we show the importance of the weight propagation strategy in the SBSM descriptor by setting $\sigma = 1$ and $\sigma = 0$. These two values define the presence or absence of the propagation step, respectively. [4] Based on the mean data set samples volume radius length, we set $R = 0.15$. Results reported in Table I show that the SBSM descriptor clearly outperforms previous state-of-the-art results on this data set. In particular, it is concluded that using neighbor propagation, the performance improves by more than 16% the best result reported in [8] for this data set. This experiment shows that when a neighboring measure of the shape point is taken into account to update neighbor bins, the local variations of shape objects are better learnt by the classifier. Consequently, the intraclass variability is reduced without the need of increasing the computational complexity of the descriptor.

[3]The SBSM descriptor code is included as supplemental material.

[4]We also tested for different values of σ and experimentally found $\sigma = 1$ to obtain the best results.

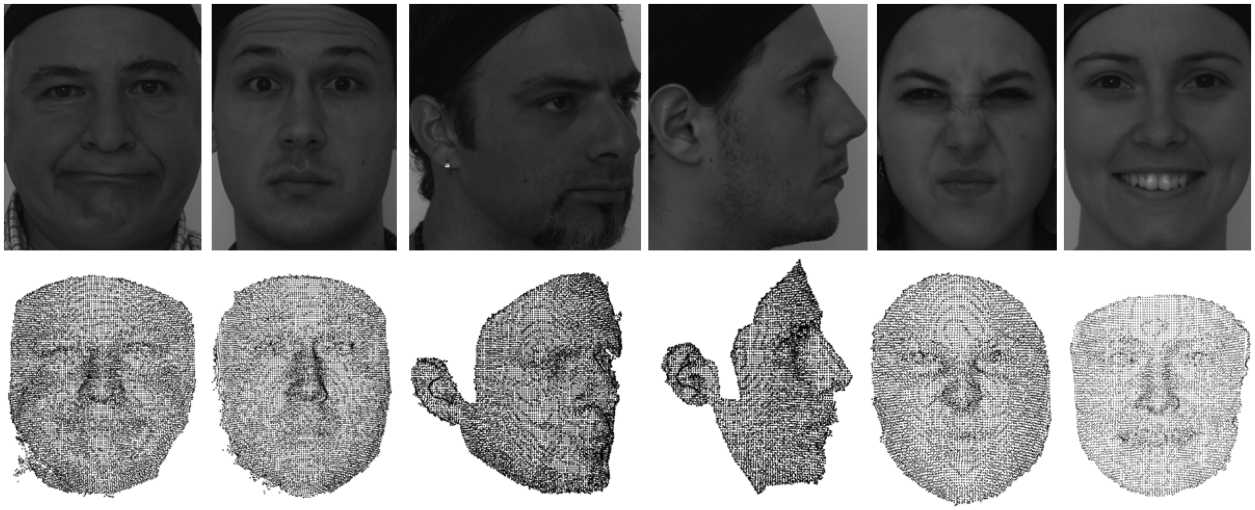


Fig. 6. Bosphorus 3-D face expression data set. Top: RGB samples. Bottom: corresponding point cloud samples.

TABLE I
CLASSIFICATION PERFORMANCE ON THE RGB-DEPTH DATA SET [21]

Method	Recognition rate
SIFT + Texton + Color + Spin [21]	64.7%
Sparse Distance Learning [22]	70.2%
VFH + SVM	77.5%
RGB-D Kernel Descriptors [7]	80.3%
Hierarchical Matching Pursuit [8]	81.2%
ESF + SVM	84.9%
SBSM, $\sigma = 0$ + SVM	96.7%
SBSM, $\sigma = 1$ + SVM	**97.9%**

TABLE II
CLASSIFICATION PERFORMANCE AND CONFIDENCE INTERVAL OF THE DIFFERENT DESCRIPTORS ON THE NOVEL AMERICAN SIGN LANGUAGE DATA SET

Method	Recognition rate
VFH + SVM	88.7%±0.2
ESF + SVM	92.3%±0.2
SBSM, $\sigma = 0$ + SVM	97.1%±0.6
SBSM, $\sigma = 1$ + SVM	**99.3%±0.4**

We also show the performance of the VFH, ESF, and SBSM on the novel ASL data set. The final performance is obtained by applying a stratified ten-fold cross-validation and testing the confidence interval with a two-tailed t-test. In this case, the spherical grid size is fitted to the minimum spherical grid size containing all the voxels for each data sample. Results are shown in Table II. One can see how our descriptor obtains better performances than using VFH or ESF, and that the best performance is achieved when weight propagation is taken into account.

Finally, we also compare the different descriptors on the public Bosphorus 3-D face expression data set. In this experiment, we perform user recognition from the set of 21 users taking into account the nine different facial expressions as well as the different head poses present in the data set. Applying stratified 5-fold cross-validation, the recognition rate results for ESF, VFH, and SBSM are shown in Table III. One can see that the proposed descriptor performs substantially better than ESF and VFH counterparts. In addition, it is also shown that considering the blurring degree to be 1 the performance of the SBSM descriptor is improved. Looking at the confidence intervals of Tables II and III, one can also see that SBSM variance in the recognition rate is smaller in comparison to the rest of methods, despite VFH and ESF which are kept small for ASL data set.

TABLE III
CLASSIFICATION PERFORMANCE AND CONFIDENCE INTERVAL OF THE DIFFERENT DESCRIPTORS ON THE PUBLIC BOSPHORUS 3-D FACE EXPRESSION DATA SET

Method	Recognition rate
VFH + SVM	69.7%±1.2
ESF + SVM	66.3%±2.1
SBSM, $\sigma = 0$ + SVM	73.1%±1.1
SBSM, $\sigma = 1$ + SVM	**77.3%±1.0**

2) *Robustness to Noise and Deformations:* In this section, we demonstrate the robustness of the SBSM when describing and classifying 3-D objects that suffer from noisy captures and deformations due to different ambient conditions or deviations captured by the sensor, as well as partial occlusions. To achieve this goal, we designed two different settings. In the first one, we analyze the robustness of the descriptor when objects suffer from deviations in the depth dimension in a range from 0 up to 20 mm in both directions of the z-axis [Fig. 7(a)–(c)]. This distortion simulates non-accurate reading errors of the sensors because of distance precision and ambient conditions. In the second test we progressively remove shape points from the point cloud from 0 up to 50% of the voxels for each object sample [Fig. 7(d)–(f)]. This distortion simulated local occlusions and reading errors that may produce the removal of some voxel points of the region of interest. Thus, in the depth distortion, the resulting point cloud has the same number of available voxels, though they are distorted in the z-axis, meanwhile in the removing distortion, the resulting point cloud contains less voxel points based on the distortion percentage.

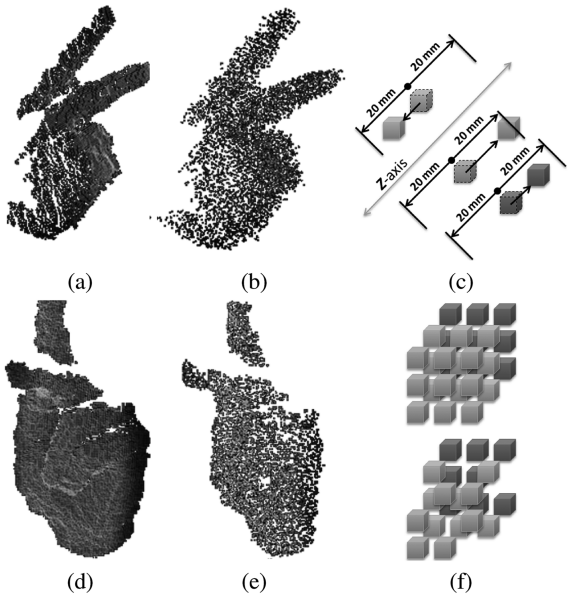


Fig. 7. (a) Input point cloud for a hand pose instance. (b) Example of distortion in the depth axis for (a). (c) For this distortion each voxel is randomly displaced in the z-axis with a maximum distortion of 20 mm in both directions of the axis. (d) Input point cloud for a hand pose instance. (e) Example of cloud removal distortion for (d). (f) For this distortion, each voxel of the original point cloud (top) is removed based on a probability value defined by the distortion (bottom).

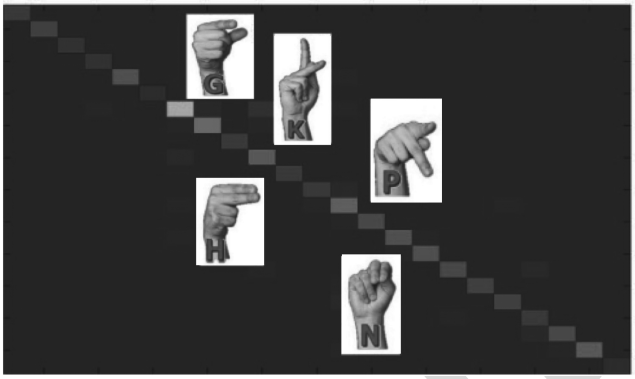


Fig. 8. Mean confusion matrix of the ASL data set using the SBSM descriptor $\sigma = 1$. The five most confused categories are displayed.

We fixed this range of distortions to be representative of the maximum distortion that we can find on recorded samples in real scenarios under different conditions and errors produced by different types of depth sensors.

In order to perform these analysis, we selected those five categories from the novel 3-D human poses data set that achieved the highest confusion in the previous section. The chosen hand categories are displayed in the confusion matrix of Fig. 8. In this image the confusion matrix is the mean computed for SBSM on that particular data set.

We show the recognition rate results when applying distortion in the depth axis in Fig. 9: for each degree of distortion, the mean recognition rate and confidence interval for 10 runs of ESF, VFH, and SBSM $\sigma = 1$ descriptors are computed. At each iteration, the percentage of distortion is randomly computed for each object voxel within different depth ranges

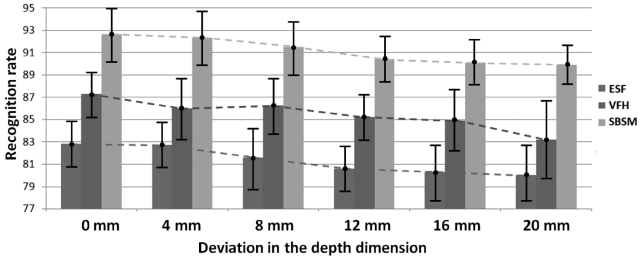


Fig. 9. Classification performance of different classification strategies under different degrees of distortion in the depth axis on the five selected categories in the ASL data set.

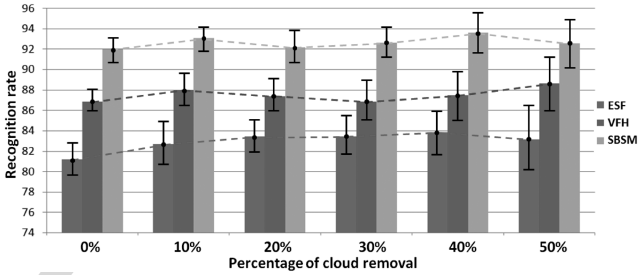


Fig. 10. Classification performance of different classification strategies under different degrees of cloud removal on the five selected categories in the ASL data set.

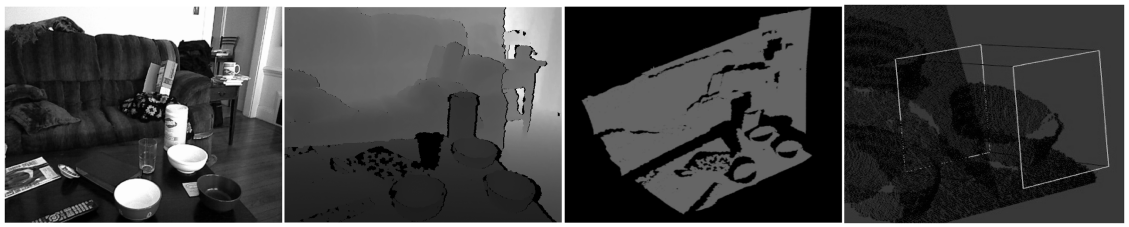
TABLE IV
MEAN RANK FOR THE COMPARED DESCRIPTORS CONSIDERING ALL THE EXPERIMENTS

	VFH	ESF	**SBSM**
Mean rank	2.13	2.86	**1.00**

in millimeters. The maximum value was set to 20 mm since it is hard to obtain higher deformations produced by the sensor. As expected, the recognition rate for all three descriptors decrease w.r.t. the depth distortion. One can see that for all the different tests of this experiments and descriptors, SBSM still obtains the best performance and VFH suffers the worst decrease in recognition (around 4%).

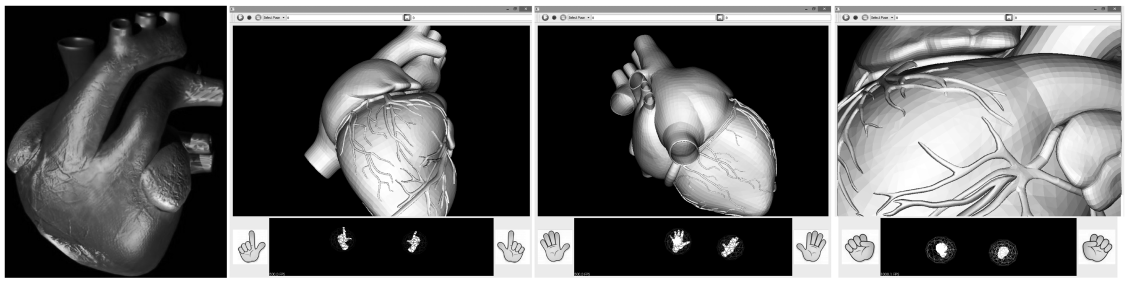
In Fig. 10, we show the results when applying cloud removal. For each percentage of removed voxels, the mean recognition rate and confidence internal are shown. At each iteration, a percentage of distortion is randomly generated, and different voxels are removed at each time satisfying the percentage of information to be removed. One can see that the general performance ranking in recognition is SBSM, VFH and finally ESF descriptor. Moreover, independently of the percentage of removed number of shape points, the recognition rate for the three methods is maintained in a small range of performance.

3) *Statistical Significance:* In order to compare the performances computed by the different experiments considered, Table IV shows the mean rank for each descriptor considering 15 different experiments (three data sets and 6×2 distortion experiments). The rankings are obtained by estimating each particular ranking r_i^j for each data set and experiment i and each descriptor strategy j, and computing the mean ranking R for each configuration as $R_j = \frac{1}{j}\sum_i r_i^j$, where J is the total number of tests.



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

Fig. 11. Object spotting in 3-D scenes. (a) Example of RGB image of a multimodal Berkeley data set. (b) Depth image of the same scene. (c) Computed point cloud from the scene. (d) Bowl spotting using SBSM (first positive 3-D object prediction is shown based on minimum Euclidean distance).



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

Fig. 12. (a) Original 3-D Heart volume of http://thefree3dmodels.com. Automatic interaction with the volume with (b) translation, (c) rotation and (d) zoom manipulation.

In order to reject the *null hypothesis*, i.e., measured ranks may not differ from the mean rank and these may be also affected by randomness in the results, we use the Friedman test [13]. The Friedman statistic value is computed as follows:

$$X_F^2 = \frac{12J}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)^2}{4}\right]. \quad (12)$$

In our case, since $K = 3$ descriptors are compared, $X_F^2 = 25.74$. This value is undesirable conservative, so Iman and Davenport proposed a corrected statistic instead

$$F_F = \frac{(J-1)X_F^2}{J(K-1) - X_F^2}. \quad (13)$$

Applying this correction, we obtain $F_F = 84.76$. With K=3 methods and J=15 experiments, F_F is distributed according to the F distribution with $(K-1) = 2$ and $(K-1)\cdot(J-1)$=28 degrees of freedom. The critical value of $F(2, 28)$ for 0.05 is 3.34. As the value of $F_F = 84.76$ is clearly higher than 3.34, we can reject the null hypothesis.

Once we have checked for the nonrandomness of the results, we perform a post ad-hoc test to check if one of the configurations could be statistically singled out. For this purpose, we use the Nemenyi test, in which the Nemenyi statistic is obtained as follows:

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6J}}. \quad (14)$$

In our case with $K = 3$ description strategies to compare and $J = 15$ tests, the critical value for a 95% of confidence ($q_\alpha = 2.35$) is $CD = 0.85$. As the ranking of the proposed SBSM approach does not intersect with any rank for that value of the CD, we can state that for the reported experiments our results are statistically significant with respect to VFH and ESF results. In the case of VFH and ESF descriptors, since their rank intersect with the CD value we can not state that there exists statistical differences between both strategies.

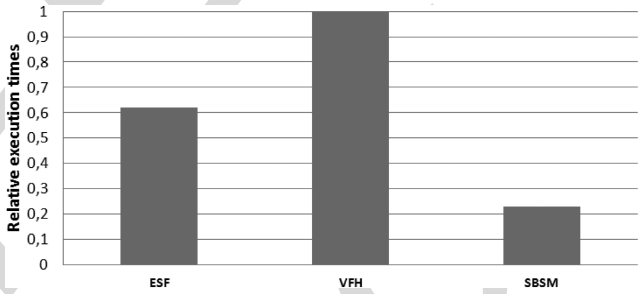


Fig. 13. Mean relative execution time in the range [0, ..., 1] among ESF, VFH, and SBSM descriptors. The relative execution time value is computed in proportion to the slowest method (set to value 1).

Finally, in order to compare the execution times of the considered ESF, VHF, and SBSM descriptors, Fig. 13 shows their mean relative execution time considering all the performed experiments. In order to compare the description complexity, the relative execution times only considers the description step, without taking into account the learning strategy. One can see that the proposed SBSM descriptor does not only obtain the best recognition rates, but also is more efficiently computed than the methods in the comparative. In particular, SBSM is more than four times faster in comparison to VFH and more than two times faster in comparison to ESF.

IV. Qualitative Analysis of SBSM

In this section, we present four real applications that use the proposed SBSM descriptor, object spotting in 3-D scenes and three fully-functional applications for a real-time HCI: 3-D

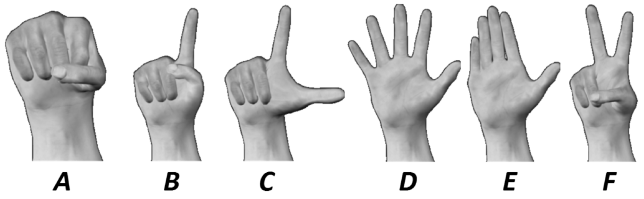


Fig. 14. HCI hand poses data set categories.

medical volume navigation, intelligent retail, and living labs; the library of the future. For this task, an additional novel set of hand poses for the HCI navigation applications was designed.

A. Object Spotting in 3-D Scenes

After showing the discriminative power and robustness of our proposed descriptor, we next illustrate the generality of SBSM when applied in real scenarios. To achieve this end, we consider a single scene obtained in the public 3-D Berkeley data set.[5] The public RGB and corresponding depth map for the selected 3-D scene are shown in Fig. 11(a) and (b), respectively. Using these data, we computed the point cloud scene shown in Fig. 11(c). In this particular case, we selected one *bowl* object to describe it and perform object spotting within the whole scene. We manually selected one bowl from a training image, computed its SBSM descriptor, and performed sliding window search over the three dimensions of the point cloud shown Fig. 11(c) for different scale hypotheses of the target object. SBSM descriptor size was set to $N_L = 8$, $N_\theta = 8$, $N_\phi = 8$, and $\sigma = 1$, with a total descriptor length of 512. The radius of the sphere is set in the range 5 to 20 cm, with an increment of 1 cm, and two voxel displacement increments in the three axis among iterations of the sliding windows approach. The first matched region of interest based on the best score obtained by the minimum Euclidean distance among the computed descriptors in the scene is shown in Fig. 11(d). Note the accurate fitting of the captured 3-D bowl object in the test 3-D scene. Moreover, the system spends 13 s in a conventional 2.7 GHz 2Core 4 Gb RAM computer to run this experiment for the tested scene. Given the performed exhaustive search and that we ran the experiment iteratively without any kind of parallelism, this experiment show the generality of the descriptor to be applied for scene analysis purposes.

B. HCI Application for Medical Navigation

Given the high discriminative power and fast computation of the proposed descriptor in comparison to the state-of-the-art approaches, we also designed different fully-functional applications to take benefit from it. The first application is an automatic HCI system for medical volume navigation, which is able to detect user, hands, poses, and gestures, manipulating a medical volume of interest.

The application was developed based upon the MS Kinect SDK to capture the RGB-D data stream from the depth sensor. The depth maps were converted into world coordinates projecting the registered 3-D image by means of pin-hole model and intrinsic camera parameters. Then, point cloud library (PCL) is used to work with the point cloud, and VTK framework is used for medical image visualization purposes. The hand detection algorithm takes advantage of the Body Pose Skeleton to find the hand wrist joints. Using this information, a fixed radius of interest of 15 cm centered in each joint is defined to segment hand point clouds. The usage of the skeleton also enables us to define a heuristic to discard false positive hand poses in the cases that the hands are near the body, or below the waist line. In a later step, the detected hand point clouds within the sphere are used to refine sphere center by computing center of mass and relocating sphere center, and points are classified using the proposed SBSM methodology. A reduced data set of 18K samples for the six classes shown in Fig. 14 were recorded and trained with SBSM and SVM for this purpose.

The detected pose label is then combined with the hand movement (3-D object displacements in real coordinates), and used as input of a hidden Markov model to obtain a hand gesture classification. The system is able to recognize and control zoom, rotation, and translation operations. In the user interface, the user can visualize its detected hand point clouds in real-time, having a feed-back with the displayed prototypes of both recognized hand poses. Also, the user can visualize the volumetric medical model and the real-time interactions caused by the hand gestures. The overall application enables a powerful and automatic volumetric medical model interaction and visualization. Fig. 12 shows some examples of detected poses, gestures, interactions, and visualizations on a public 3-D heart volume.[6]

C. HCI Application for Intelligent Retail

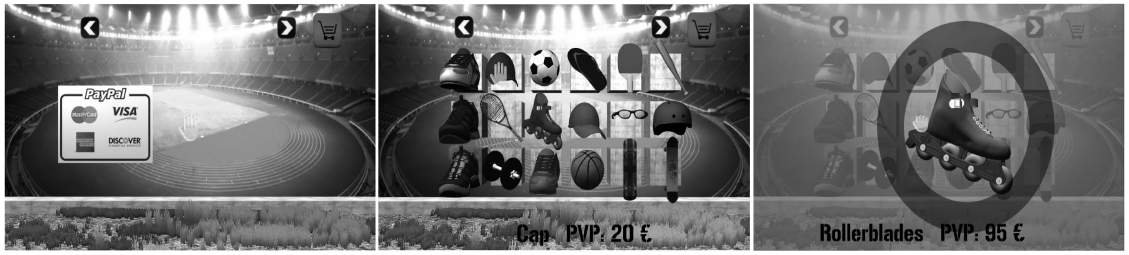
Based on the same procedure for hand pose recognition than in the medical volume navigation approach, we designed a fully-functional application for intelligent retail. In this scenario, hands are tracked in a multidevice setup and two main poses (open and close hand) are recognized. The user can automatically interact with a set of products in an interface designed with the Unity engine[7] so that information about the product and manipulation by 3-D rotation based on hand trajectories on two main object axis can be performed. In this scenario, given that we work with bigger displays and only one Kinect may not cover most part of the pointcloud related to one hand, we included an extra Kinects in order to recover voxels of the same hand from near complementary views and reconstruct a new hand with more voxel information. Thanks to the fusion of two views, more information about the tracked hand is available, and thus we make the descriptor more discriminative to classify multiple hand poses useful for interaction purposes.

In order to achieve registration of two Kinect cameras, we have mounted two cameras in a rigid setup with a baseline of 1.30 m and an angle of 18 degrees. In this way we can acquire depth images and convert each pixel to real-world coordinates by applying the intrinsic parameters of the cameras. In order to complete the registration, we need the

[5]http://kinectdata.com/

[6]See the supplemental material video for a system demonstration.

[7]http://unity3-D.com/

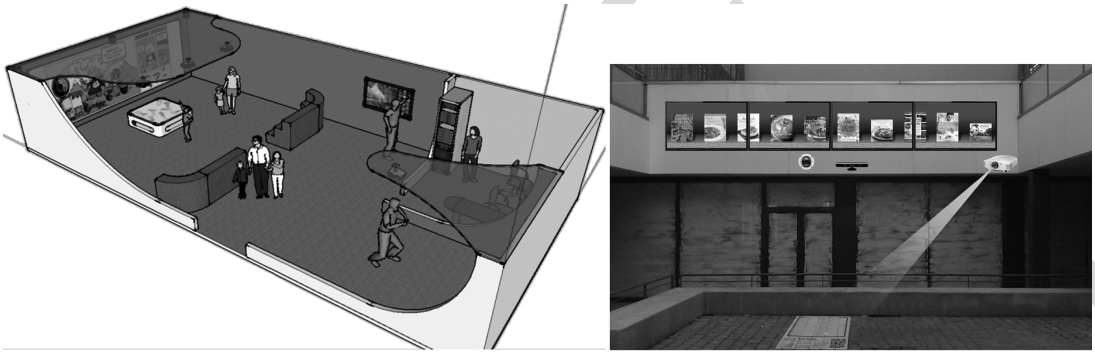


(a) (b) (c)

Fig. 15. HCI for retail. (a) Designed 3-D retail scenario using Unity engine. (b) User interaction with the scenario. (c) User manipulation by 3-D rotation.

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

Fig. 16. (e) Hand reconstruction using (a) and (b) two Kinect views (RGB) and (c) and (d) point clouds.



(a) (b)

Fig. 17. (a) Different HCI applications designed to implement within the Living Lab project. (b) Real scenario simulating the setup of the implementation of the prototype.

extrinsic parameters between both depth cameras, represented by a rotation matrix and a translation vector that allows us to convert $Kinect_2$ depth points to $Kinect_1$ depth coordinate system. Intrinsic and extrinsic parameters can be obtained by doing a stereo calibration. In our case, we use the *Camera Calibration Toolbox* [9], that gives us a rotation vector related through the *rodrigues* formula and a translation vector. For this procedure we only need few images from both depth cameras looking at the same planar checker-board pattern. Once the calibration is performed we can related $Kinect_1$ and $Kinect_2$ depth points as follows:

$$\mathcal{P}_{reference} = \mathcal{R}\mathcal{P} + \mathcal{T}$$

being $Kinect_1$ be the camera reference, $\mathcal{P}_{reference} = (\mathcal{X}_{ref}, \mathcal{Y}_{ref}, \mathcal{Z}_{ref})$ a point in the camera reference coordinate system, $\mathcal{R}$ the rotation matrix obtained in calibration step, $\mathcal{P}$ a point of the $Kinect_2$, and $\mathcal{T} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ the translation vector obtained in calibration step.

Fig. 16 shows an example of a hand reconstruction by two different views. From the 3-D reconstructed hand, SBSM is computed, and the automatic interactive process is performed. The designed 3-D scenario and real-time use case scenarios are shown in Fig. 15. In this case we found that increasing the point of view of the hand because of the use of two cameras allowed for a more natural movements of the user while keeping the recognition rate of the system.[8]

D. HCI Application in Living Labs

For this last scenario, we designed a first prototype for intelligent interaction in a virtual repository of books within a Smart Environment or Libing Lab corresponding to a new generation of libraries (VL^3 [6]Volpelleres Library Living Labs' project). For this scenario, a one Kinect camera setup was designed, and the same recognition procedure than in the retail scenario was performed. The 3-D scenario was designed with the Unity engine. The real environment to implement the prototype is shown in Fig. 17, which is located in the region of Volpelleres in Barcelona. In this prototype, the user is able to navigate through a catalogue of books and read the selected ones. The designed 3-D scenario and real-time use case scenarios are shown in Fig. 18.

[8]See the supplemental material video for a system demonstration.

(a) (b) (c)

Fig. 18. HCI Living lab prototype use case scenarios. (a) Designed 3-D Unity library environments. (b) Book appears on the scene and the user can open the hand and move through different books. (c) When close hand pose is performed, the book is selected and opened for reading.

Finally, it is important to remark that the Kinect acquired a variable frame rate between the range 20–30 FPS. Our procedure works real time for all the presented HCI applications, and thus, we are able to process (segment, describe, and classify segmented point clouds from the region of interest) for all the frames acquired by the Kinect device.

V. Conclusion

We presented the SBSM descriptor. SBSM is computationally efficient and highly discriminative. The computed descriptor codifies the spatial relations among object voxels and spherical bins given a granularity degree and a blurring factor defined by a Gaussian-based weight propagation function. The descriptor is rotation invariant by relocating descriptor bins using the two main quaternion based on the two major 3-D descriptor axis densities. In our experimental evaluation, we found that our methodology outperformed state-of-the-art results up to 16.7% on public depth multiclass object recognition data, also obtaining significant performance improvements in other two data sets; a novel ASL multiclass data set and a public 3-D face expression data set. Moreover, we tested the descriptor against different 3-D data distortions, obtaining high recognition rates and significant performance improvements in relation to standard approaches.

We also tested the descriptors in four real scenarios: object spotting in 3-D scenes, within a probabilistic gesture recognition pipeline for real-time HCI in medical volume navigation scenarios, HCI for intelligent retail in a multicamera setup, and within a prototype for catalogue navigation in a living lab library, showing the high discriminative power, efficiency and generality of the proposed descriptor to be applied in new generation of HCI applications and smart environments.

Acknowledgment

The authors would like to thank M. Pousa, C. Antens, J. Abella, A. Borras, M. Angel Blanco, J. Mas, R. Alcaide, A. Sabate, J. R. Jiménez, and D. Karatzas, from the Computer Vision Center, for their support during the development of the HCI applications.

References

[1] J. Abella *et al.*, "Multi-modal descriptors for multi-class hand pose recognition in human computer interaction systems," in *Proc. Chalearn Multi-Modal Gesture Recognit. Workshop ICMI*, 2013, pp. 503–508.

[2] M. Alexa and A. Adamson, "On normals and projection operators for surfaces defined by point sets," in *Proc. Eurographics Symp. Point-Based Graph.*, 2004, pp. 149–155.

[3] L. Alexandre, "3-D descriptors for object and category recognition: A comparative evaluation," in *Proc. Int. Conf. IROS*, 2012.

[4] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, "3-D shape histograms for similarity search and classification in spatial databases," in *Proc. 6th Int. Symp. Advances Spatial Databases*, 1999, pp. 207–226.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[6] M. Ben-Chen and C. Gotsman, "Characterizing shape using conformal factors," in *Proc. Eurographics Conf. 3DOR*, 2008, pp. 1–8.

[7] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. Int. Conf. IROS*, 2011.

[8] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Proc. ISER*, Jun. 2012, pp. 387–402.

[9] J.-Y. Bouguet. (2013). "Camera calibration toolbox for MATLAB." [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/.

[10] G. Burel and H. Henoco, "Determination of the orientation of 3-D objects using spherical harmonics," *Graph Models Image Process*, vol. 57, no. 5, pp. 400–408, 1995.

[11] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, 2011.

[12] C. Colombo, A. D. Bimbo, and A. Valli, "Visual capture and understanding of hand pointing actions in a 3-D environment," *IEEE Trans. Syst., Man, Cybern. B*, vol. 33, no. 4, pp. 677–686, Aug. 2003.

[13] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learning Res.*, vol. 7, pp. 1–30, Jan. 2006.

[14] S. Escalera, A. Fornes, O. Pujol, J. Llados, and P. Radeva, "Circular blurred shape model for multiclass symbol recognition," *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 41, no. 2, pp. 497–506, Apr. 2011.

[15] S. Escalera, A. Fornes, O. Pujol, P. Radeva, G. Sanchez, and J. Llados, "Blurred shape model for binary and grey-level symbol recognition," *Pattern Recognit. Letters*, vol. 30, no. 15, pp. 1424–1433, 2009.

[16] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *Proc. ECCV*, 2004.

[17] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. CVPR*, 2010, pp. 755–762.

[18] Point Cloud Library. (2013) [Online]. Available: http://docs.pointclouds.org/trunk/a02944.html.

[19] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3-D shape descriptors," in *Proc. SIGGRAPH Symp. Geometry Process.*, 2003, pp. 156–164.

[20] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. V. Gool, "Hough transform and 3-D SURF for robust three dimensional classification," in *Proc. ECCV*, 2010.

[21] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. Int. Conf. Robot. Autom.*, 2011.

[22] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *Proc. Int. Conf. Robot. Autom.*, 2011.

[23] N. Mitra, A. Nguyen, and L. Guibas, "Estimating surface normals in noisy point cloud data," *Int. J. Computational Geometry Applicat.*, vol. 14, nos. 4–5, pp. 261–276, 2004.

[24] C. Redondo-Cabrera, R. J. Lopez-Sastre, F. Acevedo-Rodriguez, and S. Maldonado-Bascon, "SURFing the point clouds: Selective 3-D spatial pyramids for category-level object recognition," in *Proc. IEEE CVPR*, 2012, pp. 3458–3465.
[25] M. Ruggeri, G. Patane, M. Spagnuolo, and D. Saupe, "Spectral-driven isometry-invariant matching of 3-D shapes," *IJCV*, vol. 89, pp. 248–265, 2010.
[26] R. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3-D registration," in *Proc. Int. Conf. Robot. Autom.*, 2009, pp. 1848–1853.
[27] R. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3-D recognition and pose using the viewpoint feature histogram," in *Proc. Int. Conf. IROS*, 2010, pp. 2155–2162.
[28] R. Rusu, Z. Marton, N. Blodow, and M. Beetz, "Learning informative point classes for the acquisition of object model maps," in *Proc. Control Autom. Robot. Vision*, 2008, pp. 643–650.
[29] D. Saupe and D. V. Vrani, "3-D model retrieval with spherical harmonics and moments," in *Proc. DAGM*, vol. 2191, 2001, pp. 392–397.
[30] A. Savran, B. Sankur, and M. Bilge, "Regression-based intensity estimation of facial action units," *Image Vision Comput.*, vol. 30, no. 10, pp. 774–784, 2012.
[31] J. Shen, D. Wang, and X. Li, "Depth-aware image seam carving," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1453–1461, Oct. 2013.
[32] K. Shoemake, "Animating rotation with quaternion curves," in *Proc. Comput. Graph. Interactive Techniques*, pp. 245–254, 1985.
[33] H. P. H. Shum, E. S. L. Ho, Y. Jiang, and S. Takagi, "Real-time posture reconstruction for Microsoft Kinect," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 43, no. 5, pp. 1357–1369, Oct. 2013.
[34] B. Steder, R. Rusu, K. Konolige, and W. Burgard, "NARF: 3-D range image features for object recognition," in *Proc. Int. Conf. IROS*, 2010.
[35] F. Tombari, S. Salti, and L. D. Stefano, "Unique signatures of histograms for local surface description," in *Proc. ECCV*, 2010, pp. 356–369.
[36] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3-D object classification," in *Proc. IEEE ROBIO*, 2011, pp. 2987–2992.
[37] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *Proc. ICCV*, 2013.
[38] P. Zhang, Z. Wang, S. Zheng, and X. Gu, "A design and research of eye gaze tracking system based on stereovision," *Emerging Intell. Comput. Technol. Applicat., Lecture Notes Comput. Sci.*, vol. 5754, no. 4, pp. 278–286, 2009.

Oscar Lopes received the master's degree in Computer Vision and in Multimedia Technologies from Universitat Autònoma de Barcelona (UAB), Barcelona, Spain. He is currently pursuing Ph.D. degree in the Information and Computing Sciences at the University of Utrecht, Utrecht, The Netherlands.

His current research interests include pattern recognition, machine learning, and mobile computing, to leverage the creation of rich HCI systems based on gestural semantics.

Mr. Lopes collaborated in the design of a hand pose and gesture recognition system for which he was awarded 3rd place at the ICPR held in Tsukuba, Japan, in 2012.

Miguel Reyes received the bachelor's degree in computer science at Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2010, and the Master's degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2011. He is currently pursuing the Ph.D. degree in math in the area of computer science and artificial intelligence with the University of Barcelona, Barcelona.

His current research interests include pattern recognition, signal processing and visual object recognition, and their application to health care systems.

Mr. Reyes is a member of the Human Pose Recovery and Behavior Analysis group and the Computer Vision Center.

Sergio Escalera received the Ph.D. degree on multiclass visual categorization systems at Computer Vision Center, Universitat Autònoma de Barcelona (UAB), Barcelona, Spain.

He leads the Human Pose Recovery and Behavior Analysis Group. He is currently a Lecturer with the Department of Applied Mathematics and Analysis, Universitat de Barcelona, Barcelona, Spain. He is the Editor-in-Chief of American Journal of Intelligent Systems and the Advisor and Director of the ChaLearn Challenges in Machine Learning. He is also a part time Professor at Universitat Oberta de Catalunya, Barcelona, Spain. His current research interests include statistical pattern recognition, visual object recognition, and HCI systems, with special interest in human pose recovery and behavior analysis.

Dr. Escalera received the 2008 Best Thesis Award on Computer Science at UAB. He is a member of the Computer Vision Center at Campus UAB.

Jordi Gonzàlez received the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2004.

He is currently an Associate Professor in computer science with the Department de Ciències de la Computació at UAB. He is also a Research Fellow at the Computer Vision Center, Barcelona. His current research interests include cover pattern recognition and machine learning techniques for the computational interpretation of human behaviors in image sequences and video hermeneutics.

Dr. Gonzàlez has coorganized Special Issues in IJPRAI journals in 2009, CVIU journals in 2012, and MVA journals in 2013. He is a member of the Editorial Board of CVIU and IET-CVI journals.