2014 Second International Conference on 3D Vision

MCOV: a covariance descriptor for fusion of texture and shape features in 3D point clouds

Pol Cirujeda, Xavier Mateo, Yashin Dicente, Xavier Binefa
Department of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain
{*pol.cirujeda, javier.mateo, yashin.dicente, xavier.binefa*}*@upf.edu*

***Abstract*—In this paper we propose MCOV, a covariance-based descriptor for the fusion of shape and color information of 3D surfaces with associated texture aiming at a robust characterization and matching of areas in 3D point clouds. The proposed descriptor is based on the notion of covariance in order to create compact representations of the variations of texture and surface features in a radial neighbourhood, instead of using the absolute features themselves. Even if this representation is compact and low dimensional, it still offers discriminative power for complex scenes. The codification of feature variations in a close environment of a point provides invariance to rigid spatial transformations and robustness to changes in noise and scene resolution from a simple formulation perspective. Results on 3D points discrimination are validated by testing this approach performance on top of a selected database, corroborating the adequacy of our approach on the posed challenging conditions and outperforming other state-of-the-art 3D point descriptor methods. A qualitative test application on matching objects on scenes acquired with a common depth-sensor device is also provided.**

***Keywords*-3D shape descriptors; covariance feature fusion; 3D scene analysis**

I. Introduction

The description, detection and matching of points from complex scenes is a challenging task for many Computer Vision applications such as visual tracking, object modelling and recognition or scene reconstruction. Existing approaches make use of the available cues in the usual two channels of information: visual photometry such as color or texture, and shape and depth information from 3D sensors. State-of-the-art methods have provided successful outcomes in both areas separately. However, our goal is to provide a global method which can fuse information from both two worlds.

This paper focuses on the definition of the compact but at the same time descriptive capability of covariance matrices of feature variations. Encoding the correlating degrees between different texture and shape features together within a 3D point neighbourhood is more descriptive than using absolute features themselves, as in current histogram or keypoint-based approaches. This makes our fusion covariance descriptor adequate to avoid ambiguities in point matchings, and adds robustness to rigid spatial transformations, noise and resolution variations. The statistical nature of covariance also provides some added benefits to the descriptor: we provide an associated methodology for the analysis of salient points of the scene and for the estimation of the neighbourhood sampling radius of the descriptor. Last but not least, the MCOV covariance descriptor lays on a specific manifold topology, which makes that similar 3D scene points stay close in the descriptor space. Therefore, we propose that the comparison of descriptors for scene points matching can be performed by an adequate manifold metric.

II. Related work

Image processing applied to 3D data is currently an active topic in the computer vision literature. Advances in sensor technology have provided some affordable devices and acquisition techniques that make possible the capture of 3D information to a wider audience. This easy access to the capture technology has also produced an increase in the processing proposals for this kind of images during last years.

In the concrete context of 3D registration, a possible procedure is the use of the information obtained from a visible camera, previously calibrated with a range scanner, extracting information from the more well-known and deeply studied 2D image domain. The most usual method in order to match correspondences between two 2D images is undoubtedly the SIFT algorithm [1]. While SIFT is able to cope with small differences in the point of view, different methods which add partial 3D information to the SIFT algorithm are proposed in works as [2], [3], [4], estimating the surface normal at the 3D coordinate and performing a homography of the visible image as it would be seen from the front side of the keypoint.

On the other side there exist also descriptors which use exclusively the 3D information from the scene. Inside this category, Spin Images [5] (or related variants as [6]) are probably the most known method, encoding the neighbourhood of each 3D point in a 2D image. Other popular 3D descriptors are the point signatures [7], the 3D shape contexts [8], THRIFT [9] or, more recently, the Fast Point Feature Histograms [10].

978-1-4799-7000-1/14 $31.00 © 2014 IEEE
DOI 10.1109/3DV.2014.11

Finally, during last years some descriptors which encode simultaneously information from the 3D shape and the texture have been published in the literature. A good example is the MeshHOG descriptor [11], which performs a histogram of gradient of the neighborhood of a 3D point by using separately the texture information and the 3D curvature. In order to include both cues in the final descriptor, both representations can be directly concatenated. This same methodology is used from the authors of the CSHOT descriptor [12], which concatenates their SHOT descriptor [13] and the color information. Other contributions as Heat Kernel Signatures [14] follow a perspective which is closer to the one presented in this paper, and use manifold embedding procedures where photometric information is implicitly encoded as part of the coordinate projection parameters.

III. MCOV: Fusion of shape and texture information.

The method proposed in the present paper is focused on the combination of the visible and the 3D information in an implicit fusion and correlation analysis way, which is provided by means of the statistical concept of covariance. Covariance matrices in the Computer Vision domain where first used as descriptors by Tuzel *et al.* [15], [16] for the detection of objects and faces. They were proposed as a robust 2D color region estimator as the representation of variations amongst several color features, losing structural information and being robust to noisy inputs by construction, was shown to be more discriminative than encoding absolute features themselves. This framework has been extended to 3D surface description in few occasions: Fehr *et al.* [17] explore several shape measures which include the angular measures initially provided by Spin Images [5], or the normal vector directions at each point of the 3D scene, amongst others. Recently, Tabia *et al.* [18] have proposed a similar descriptor where features are based on direct point euclidean distances within the descriptor construction neighbourhood. Our descriptor approach is different in two aspects: on one hand, we integrate the addition of color features with shape measures for the fusion of texture information within the descriptor neighborhood. In a second place, for the definition of surfaces we propose a set of three angular features which are robust to rigid spatial transformations, noise and density variations due to their locally relative extraction, compared to the aforementioned approaches.

A. Feature fusion

The statistical notation of covariance is a measure of how several random variables change together and captures the intrinsic correlation between sampling distributions of the involved cues. In the context of a descriptor definition, the observed random variables are related to the set of observable features which can be extracted from points and their close localities in the scene, e.g. pixel color values, 3D coordinates, first or second order derivatives, etc. We define a feature selection function $\Phi(p, r)$ for a given 3D point p and its neighbourhood within spatial radius r in the scene:

$$\Phi(p, r) = \{\phi_{p_i}, \forall p_i \quad \text{s.t.} \quad |p - p_i| \leq r\} \tag{1}$$

where ϕ_{p_i} is the vector of random variables obtained at each one of the points p_i within the radial neighbourhood, and is defined as $\phi_{p_i} = (R_{p_i}, G_{p_i}, B_{p_i}, \alpha_{p_i}, \beta_{p_i}, \gamma_{p_i},)$.

This feature selection function includes the following observations that are computed relatively to the point for which the descriptor is being obtained: first of all, the visual information is taken into account in terms of R, G and B color space values. This values are enough for capturing texture and visual pattern information, but they could be easily changed to any other color space magnitudes, or obtained after a color invariance pre-processing stage -this is beyond the scope of the current approach and could be explored as future work.

α, β and γ values are angular measures which encode the surface information of the points within the descriptor center neighbourhood in the following way:

- α is the angle between the normal vector in p and the segment from p to p_i, and encodes the global concavity of the surface regarding the center of the descriptor.
- β is the angle between the same segment and the normal vector in p_i, and measures the local curvature at this point in the neighbourhood relative to the center p.
- γ is the angle between both normal vectors in p and p_i. Being a 3D angle, it helps encoding the local surface curvature in a non-ambiguous way.

Note that these angular measures are different from the ones proposed by Spin Images [5] or Fehr *et al.* [17]. In Figure 1 we show a schema of how these measures are obtained. All features are normalized in order to have an equivalent range both for angular and color measure.

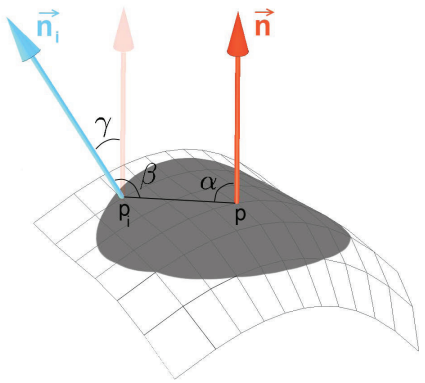

Figure 1. Schema of the used features for shape information encoding. For each p_i in the neighbourhood of p, α, β and γ are the rotational invariant angular measures.

Then, for a given point p of the scene the covariance descriptor for a radius r expressing the correlation of the defined cues can be obtained as:

$$C_r\left(\Phi(p,r)\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_{p_i}-\mu\right)\left(\phi_{p_i}-\mu\right)^T \qquad (2)$$

where μ is the vector mean of the set of vectors $\{\phi_{p_i}\}$ within the radial neighbourhood of N samples.

The resulting 6×6 matrix C_r will be a symmetric matrix where the diagonal entries will represent the variance of each one of the feature distributions, and the non-diagonal entries will represent their pairwise correlations. Figure 2 shows an example of a covariance descriptor. The abstract and compact notation of MCOV provides a representation which treats the observed features as samples of joint distributions: the structure information about the number of points and their ordering within the region it defines is lost during the construction of the descriptor. This is a desired advantage of the presented descriptor, as feature distributions will preserve their characterization even under changes of scale and rotation in data. Furthermore, this makes our descriptor robust to changes of resolution: according to central limit theorem, as long as a significant enough number of samples of the features distribution is used, this distribution will be correctly characterized within a certain confidence interval. Finally, noisy observations are also tolerated by the own nature of the covariance formulation, as outlier features are attenuated thanks to the mean subtraction during description computation. These characteristics of MCOV yield to a valuable discriminative performance boost in comparison to more rigid representations such as keypoint or histograms-based approaches.

B. Scene analysis from MCOV descriptors

Being covariance matrices, MCOV descriptors lay in the manifold of symmetric positive definite matrices. This spatial variety is of meaningful importance, as 3D regions sharing similar texture and shape characteristics will remain under close distances on the descriptor space. There exist several approaches for comparing symmetric positive definite matrices. Most of them are specifically focused on the retrieval of matrix similarities on close neighbourhoods [19], [20], fact which must consider prior knowledge about the different spatial clusters on the descriptor space. However, for a general descriptor comparison, we propose the use of the manifold metric defined by Förstner in [21]. This distance definition preserves the global geometric relationship of the descriptors as the involved generalized eigenvalues between two covariance matrices express the magnitude of their geodesic distance, respecting the curvature of the manifold:

$$\delta(C_r^1, C_r^2) = \sqrt{\sum_{i=1}^{6} \ln^2\lambda_i(C_r^1, C_r^2)} \qquad (3)$$

where $\lambda_i\left(C^1, C^2\right)$ is the set of generalized eigenvalues of C^1 and C^2 according to their dimensionality d ($d=6$ in our feature selection function).

In a second place, we propose a procedure for an accurate radius estimation according to the nature of the scene for which descriptors are being obtained. The sample mean is a good estimator of the population of a random variable distribution and its sampling size parameter, in order to lay within a confidence interval, is modelled by Chebyshev's inequality: $P\left(\mid \bar{X}-\mu \mid \geq \epsilon\right) \leq \sigma^2/\epsilon^2 n$, where μ and σ^2 are the mean and variance of the distribution we are considering; $\bar{X}$ is the sample mean according to the number of samples n we are observing; and ϵ is the threshold on data representation. Therefore we can generalize the following expression for an arbitrary feature distribution: $n \geq \sigma^2/\epsilon^2\left(1-p\right)$ where p is the desired confidence value. Usually we will use a threshold value $\epsilon = 0.1$ and a confidence interval of $p = 0.95$. This will provide a lower boundary of the needed number of samples n.

Relating this to our framework, we can obtain the sampling distributions of features along the whole scene whose, and then apply this boundary equation for each one of the feature distribution variances. This will define a set of 6 candidate sampling sizes, one for each cue. As this provides a lower boundary, we will keep the maximum value of all the candidate sizes, indicating the number of samples needed for assuring that descriptors encode correctly the scene feature distributions with a confidence of $p = 95\%$. We have found that this scene-dependant methodology provides accurate estimations for a discriminative behaviours of MCOV descriptor, as validated in our experiments. As an example, on scenes with different variations of shape and color and a density of 20000-30000 points, this estimation reflects the need of taking around 400-500 samples. This sampling size can be directly translated to a radius magnitude according to the density of the scene point cloud.

In a third place, covariance matrices can be also understood as salient point detectors. As defined after eq. (2), a covariance matrix C_r contains the variance of the observed features on its diagonal, and the covariance on the other entries. Computing the determinant of a covariance matrix is equivalent to obtaining the so-called *generalized variance*, which can be interpreted as a measure of the degree of homogeneity of each point in the scene [22]. As the used features have been previously normalized, there is no range variation which could interfere on this analysis. The points with higher determinant magnitudes on their associated descriptors can be identified as the points which belong to real interest areas, with inner significant variation

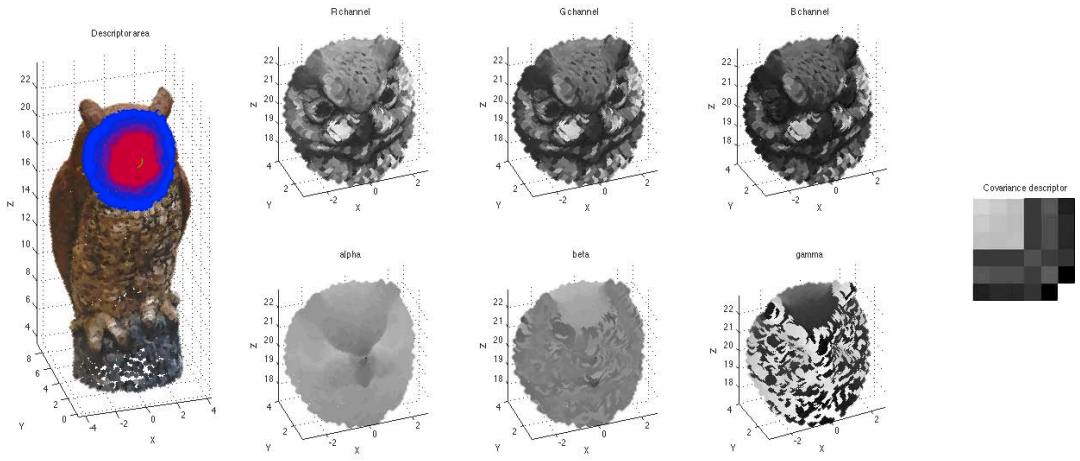

Figure 2. The left image shows the original 3D scene where a radial neighbourhood for computing the descriptor is coloured. The 6 central subfigures show the different used features, in terms of color (upper row) and shape description (bottom row). The resulting 6x6 covariance descriptor is represented in the right side.

in visual texture and 3D shape changes. It is worth to notice that these interest points are selected implicitly from a global point of view, combining both visual and shape saliency. Therefore, even in the case of an homogeneously coloured object like the one in Figure 3, keypoints are still obtained on significant parts such as eye holes or borders. Due to the nature of the used descriptor neighbourhood, relevant points tend to form small clusters, which could be further reduced with relevance sampling procedures like [23]. This property is commented for computational efficiency on big datasets and a further analysis is beyond the scope of this approach at its current stage, and left as future work.

Finally, as computing covariance descriptors does not involve any major operation, it is easy to extend them to a multi-scale framework by just adding several radius magnitudes for the neighbourhoods around the descriptor center point. Therefore, each point in the scene will receive not one, but a set of descriptors: $C_M(p) = \{C_r(\Phi(p,r)), \forall r \in \{r_1..r_s\}\}$ The idea behind using several neighbourhood radii is that discrimination performance can be improved if a point is supported by more than one descriptor, regarding a narrow to coarse set of surrounding areas. This can help to avoid repeatability problems and improve detection of points in edges or borders of scene objects. If needed, standard scale-space methodologies [24] can be used in order to determine the different radius factors in conjunction with the aforementioned radius estimation procedure.

IV. Experimental results

In order to compare the performance of the proposed method against other state-of-the-art approaches we provide a dataset combining 3D shape with visual information in 12 scenes which have been obtained using Autodesk 123D Catch [1] 3D modelling software. These models are stored as 3D meshes with photometric texture, where each vertex has a unique identifier in order to provide an unbiased groundtruth of labelled points. See Figure 4 for a visual representation of the 12 base models used. This dataset can be publicly accessed upon request by contacting the authors on the header of this paper. The contained objects have been particularly selected in order to include challenging handicaps as repeated areas, homogeneous surfaces and textures, and symmetries.

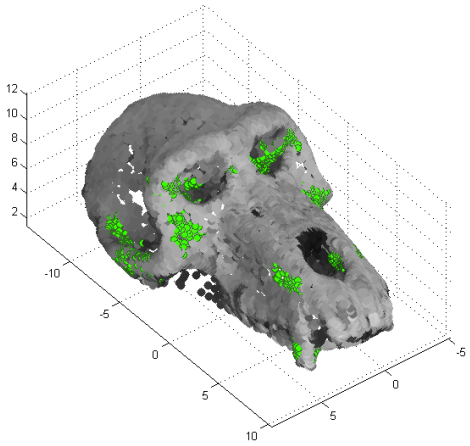

Figure 3. This figure shows the 1500 most significant points of a scene, according to the proposed *generalized variance* analysis. Even if the color information of the object is homogeneous, interest points have been detected on salient areas of the scene.

[1] http://www.123dapp.com/catch

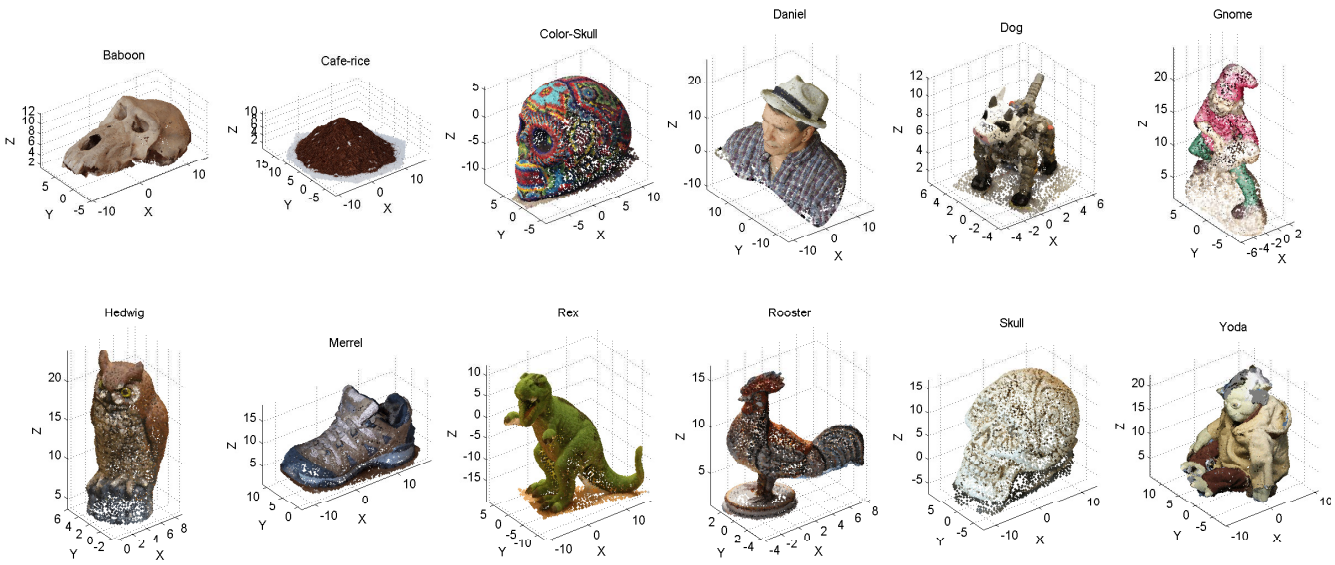

Figure 4. 3D plot of the 12 models included on the database. Full scenes are shown without added noise.

A. *Descriptor comparison against noise and resolution variations*

In order to test the descriptor performance, we will compare our MCOV Descriptor approach against the state-of-the-art methods MeshHOG [11], CSHOT [12] and Textured Spin Images [6] -which is a variation of the original Spin Images approach [5], still considered one of the classical 3D descriptors in the literature for successful matching of dense scenes. The compared descriptor approaches are used following the original implementation by their authors, and any needed parameter (radius, bin sizes) is set according to the recommendations of their original proposals -or to equivalent values regarding our approach in order to provide the most fair comparison as possible.

We have performed a cross-validation test, using 10 folds containing 10% of the number of points on each scene. Points are labelled in each model in the database, therefore we can compute the descriptor similarities regarding the same points on a variation of the model. The variation includes: *i)* an arbitrary rotation, *ii)* an arbitrary translation, and *iii)* an addition of noise to color and surface coordinates. Noise levels will follow different Gaussian noise distributions with standard deviations according to 2, 4, 6, 8 or 10% of each one of the data channels. The evaluation method consists of observing the amount of false and true positives, and false and true negatives, in terms of matching scene points by their according descriptor similarity measures. For our descriptor, we will use the metric defined in eq. (3). According to a *ratio* parameter we consider as true positives all those matches which are within the boundaries of *ratio* times the best similarity of this set of candidates. For each level of noise we move the *ratio* coefficient within a range of 1 to 5 and we obtain a set of ROC curves as exemplified in Figure 5. This is useful for comparing the behaviour of the different tested descriptors under all noise variations, for each one of the twelve available models. For a numerical comparison between these curves, their *Area Under the Curve (AUC)* measure can be obtained. This allows to numerically summarize the average performance of the four tested descriptors over all the models in our database, as seen in Table I.

	n002	n004	n006	n008	n010
MCOV	0.991	**0.976**	**0.961**	**0.953**	**0.917**
CSHOT	**0.992**	0.913	0.758	0.616	0.562
MeshHOG	0.963	0.819	0.704	0.607	0.577
TextSpinImg	0.750	0.614	0.615	0.564	0.533

Table I
AVERAGE AUC MEASURES FOR 12 MODELS, 100% VS 100% RESOLUTION, FOR 5 LEVELS OF NOISE.

According to the results, we can see how the proposed MCOV descriptor is more stable regarding the increases on the noise levels. Since other methods are working with local surface neighbourhoods and 3D coordinate histogram representations, they will quickly suffer this distortion on data, i.e. at bin discretisation. Textured Spin Images are clearly affected by color variance as the color sparsity is saturating the illuminant binning component of that descriptor. This was in fact identified as a possible drawback by their own authors in [6]. This puts into evidence the benefits of the statistical representation of the proposed MCOV descriptor, both for naturally attenuating noise effects and for fusing several cues of information in a flexible manner.

The same experiment has been also conducted by applying

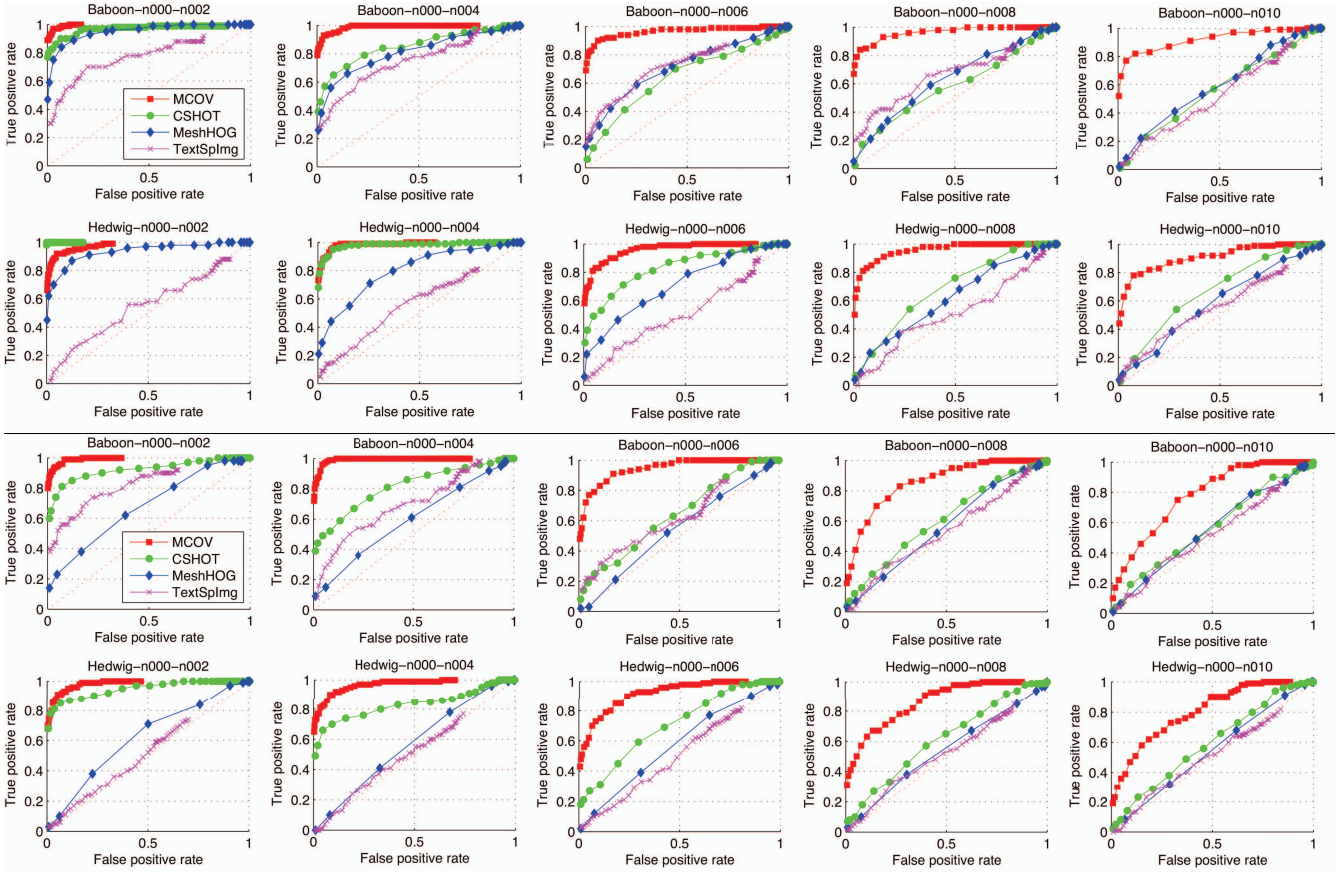

Figure 5. ROC curves for the performance comparison of the four tested approaches. Two upper rows depict the test on two different scenes for 100% vs. 100% resolution evaluation, while two bottom rows show the same scenes for 100% vs. 50% resolution evaluation. Each column shows the behaviour of the descriptors under different levels of additive noise over data (2, 4, 6, 8 and 10% of the standard deviation of color and surface coordinates).

	n002	n004	n006	n008	n010
MCOV	**0.984**	**0.967**	**0.924**	**0.871**	**0.812**
CSHOT	0.906	0.823	0.668	0.614	0.597
MeshHOG	0.616	0.597	0.522	0.517	0.521
TextSpinImg	0.662	0.613	0.563	0.534	0.520

Table II
AVERAGE AUC MEASURES FOR 12 MODELS, 50% VS 100% RESOLUTION, FOR 5 LEVELS OF NOISE.

a resolution variation over the models. The aim is to test the performance of descriptors when matching original models against a down-sampled variation to a 50% of their point cloud density. This down-sampling procedure is applied by randomly suppressing samples over the point clouds. Table II reflects the associated average AUC measures for these tests. The corresponding ROC curves to the *Baboon* and *Hedwig* models for an easier visualization of descriptor performance are plotted in the two bottom rows of Figure 5. Results suggest this is a more challenging experiment, as data is highly altered. Nevertheless, the statistical basis of our descriptor is valuable again in terms of resolution robustness: as long as a large enough number of samples is preserved, fact which we are assuring, covariance will still encode the underlying characteristics of feature distributions. In the other evaluated descriptors the changes on data resolution will incur on a bigger descent of their performance. A special consideration must be taken into account in the MeshHOG method, which requires mesh faces information in order to compute its descriptor. The applied resolution down-sampling implies the computation of an equivalent triangulation by using the edge collapse procedure [25]. This dependence has a drastic impact on MeshHOG performance as results depict.

In any case, from these experiments one can also conclude that even if a descriptor can be reliable at representing a given area, the presence of false positives could also be due to other unavoidable causes as the possibility of repetitions of visual patterns or surfaces in the scene. This puts into consideration the need of some sort of global mechanism which must be capable of finding these artifacts and filter out the non-positive matches according to global constraints such as geometric consistence. This can lead to future work in defining an accompanying method for taking into account

holistic scene aware observations.

B. Real-data matching qualitative evaluation

In this experiment we propose to test the descriptor in the context of scenes acquired with a Microsoft Kinect device. While the lack of a direct groundtruth information converts this set-up in a qualitative evaluation, it still justifies several benefits of our proposal: the usage of real data can validate our statement about the performance of the MCOV descriptor against noise and resolution changes (in this case, caused stochastically by the acquisition sensor). In a second place, we validate the application of our method under practical conditions like computational feasibility, or description of differently shaped objects -from planar to round. And finally, we provide an example of broadening the scope of our approach to areas such as scene understanding or object indexing.

We use objects and scenes from the publicly available RGB-D dataset presented in [26]. The goal is to perform a 3D object searching task: segmented objects available also on the dataset will be used as query instances to be found within the whole scenes, where they will be mixed with clutter elements and altered by changes on resolution, spatial transformations or incomplete views. Using a RANSAC standard implementation [27] we seek a spatial transformation between matches of the query instance and a set of geometrically coherent points in the whole scene. The spatially translated points from the query model regarding the whole scene will be considered as identifiers for the object segmentation points which will indicate the presence of the element in the scene. We have used 4 different cluttered scenes and 10 different query objects from the aforementioned dataset, with different shape and texture distributions. Qualitative results are shown in Figure 6.

This experiment has been conducted in an Intel Core i5 computer with 4Gb of RAM. As stated before, the implementation of the proposed approach does not pose major computational demands, and for the models in the database which have a density ranging from 80000 to 90000 points the whole descriptor calculation time takes around 50 seconds in a prototype, non-optimized implementation.

V. Conclusions

In this paper we have presented a novel descriptor specifically aimed at fusion of 3D shape and visual information under spatial transformations and changes in noise and scene resolution. The main benefit of the presented MCOV descriptor lays on the compact, yet discriminative representation present in encoding feature variations rather than rigidly represent features themselves. In a spatially close neighbourhood, this representation is robust to rigid spatial transformations and changes due to noise or resolution alterations. Experimental results have validated the discriminative capability of MCOV, which outperforms other state-of-the-art methods, specially in the case of noise over data or density variations. The analysis of several aspects of the descriptor also open the door to interesting future work, as a deep study of additional color features, geometric properties, or complementary constraint methods for globally scene understanding in object registration or recognition.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision, IJCV*, vol. 60, no. 2, pp. 91–110, 2004. 1

[2] E. R. Smith, B. J. King, C. V. Stewart, and R. J. Radke, "Registration of combined range and intensity scans: Initialization through verification," *Computer Vision and Image Understanding, CVIU*, vol. 110, no. 2, pp. 226 – 244, 2008. 1

[3] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys, "3d model matching with viewpoint-invariant patches (vip)," *Computer Vision and Pattern Recognition, CVPR*, pp. 1–8, 2008. 1

[4] G. T. Flitton, T. P. Breckon, and N. M. Bouallagu, "Object recognition using 3d sift in complex ct volumes." *BMVC*, pp. 1–12, 2010. 1

[5] A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999. 1, 2, 5

[6] N. Brusco, M. Andreetto, A. Giorgi, and G. M. Cortelazzo, "3d registration by textured spin-images," in *International Conference on 3D Digital Imaging and Modeling, 3DIM*, 2005, pp. 262–269. 1, 5

[7] C. Chua and R. Jarvis, "Point signatures: A new representation for 3d object recognition," *International Journal of Computer Vision, IJCV*, vol. 25, no. 1, pp. 63–85, 1997. 1

[8] A. Frome, D. Huber, R. Kolluri, T. Bülow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *European Conference on Computer Vision, ECCV*, 2004, vol. 3023, pp. 224–237. 1

[9] A. Flint, A. R. Dick, and A. Van Den Hengel, "Thrift: Local 3d structure recognition." vol. 7, 2007, pp. 182–188. 1

[10] R. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *International Conference on Robotics and Automation, ICRA*. IEEE, 2009, pp. 3212–3217. 1

[11] A. Zaharescu, E. Boyer, and R. Horaud, "Keypoints and local descriptors of scalar functions on 2d manifolds," *International Journal of Computer Vision, IJCV*, vol. 100, no. 1, pp. 78–98, 2012. 2, 5

[12] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3d feature matching," in *International Conf. on Image Processing*. IEEE, 2011, pp. 809–812. 2, 5

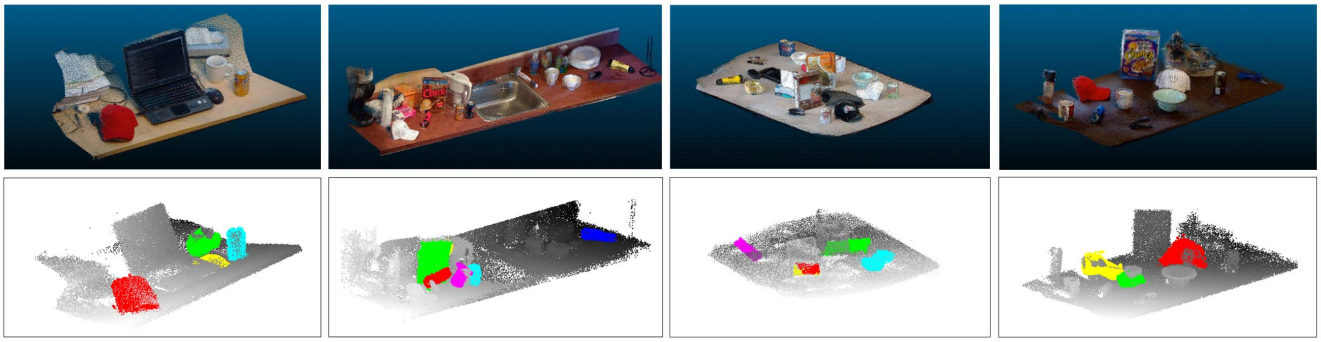

Figure 6. Results from the experimental setup performed on top of the RGB-D dataset. Top row shows four of the provided scenes. Bottom row depicts the results of our proposed query search method: for clarification we plot the depth map point cloud in grayscale and the found instances of different objects in solid colours.

[13] ——, "Unique signatures of histograms for local surface description," in *European Conf. on Computer Vision*, 2010, vol. 6313, pp. 356–369. 2

[14] A. Kovnatsky, M. M. Bronstein, A. M. Bronstein, and R. Kimmel, "Photometric heat kernel signatures," in *Scale Space and Variational Methods in Computer Vision*. Springer, 2012, pp. 616–627. 2

[15] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *European Conf. on Computer Vision*, pp. 589–600, 2006. 2

[16] ——, "Pedestrian detection via classification on riemannian manifolds," *Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008. 2

[17] D. Fehr *et al.*, "Compact covariance descriptors in 3d point clouds for object recognition," in *International Conference on Robotics and Automation, ICRA*, 2012, pp. 1793–1798. 2

[18] H. Tabia, H. Laga, D. Picard, and P.-H. Gosselin, "Covariance descriptors for 3d shape matching and retrieval," *CVPR*, 2014. 2

[19] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic resonance in medicine*, vol. 56, no. 2, pp. 411–421, 2006. 3

[20] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices," *IEEE Trans. on PAMI*, vol. 35, no. 9, pp. 2161–2174, 2013. 3

[21] W. Förstner and B. Moonen, "A metric for covariance matrices," *Quo vadis Geodesia*, pp. 113–128, 1999. 3

[22] S. S. Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, vol. 24, no. 3/4, pp. 471–494, 1932. 3

[23] A. Torsello, E. Rodolà, and A. Albarelli, "Sampling relevant points for surface registration," in *3DIMPVT*, 2011, pp. 290–295. 4

[24] T. Lindeberg, *Scale-space theory in Computer Vision*. Springer, 1993. 4

[25] D. P. Luebke, "A developer's survey of polygonal simplification algorithms," *Computer Graphics and Applications*, vol. 21, no. 3, pp. 24–35, 2001. 6

[26] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *International Conference on Robotics and Automation, ICRA*, 2011, pp. 1817–1824. 7

[27] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003. 7