Int J Comput Vis
DOI 10.1007/s11263-015-0820-2

A 3D Scene Registration Method via Covariance Descriptors and an Evolutionary Stable Strategy Game Theory Solver

Fusing Photometric and Shape-Based Features

Pol Cirujeda[1] · **Yashin Dicente Cid**[1] · **Xavier Mateo**[1] · **Xavier Binefa**[1]

Received: 26 April 2014 / Accepted: 20 March 2015

Abstract In this paper we provide an integrated approach for matching patterns in scenes combining 3D and visual information. For local definition of points we propose a descriptor based on the notion of covariance of features for fusion of shape and color information of 3D surfaces, so-called multi-scale covariance descriptor (MCOV). The intrinsic properties of this descriptor are many: it is invariant to spatial rigid transformations, and robust to noise and resolution changes; it can also be used for characteristic point detection; and lies on top of a manifold topology which allows the use of analytical metric properties. This descriptor is complemented with a game theoretic approach for solving the matching correspondences under global geometric constraints. This layer offers a comprehensive understanding of the scene and avoids possible mismatches due to repeated areas or symmetries—which would be impossibly identified by the detector solely at a local level. Our solution is able to accurately match different views of a scene even under spatial transformations, high noise levels and with small overlap between views, outperforming state-of-the-art approaches. Results are validated by comparing MCOV against other state-of-the-art 3D point descriptor methods, and matching complex 3D and color scenes under several challenging conditions.

Keywords 3D scene registration · Covariance descriptor · Evolutionary game theory · Feature fusion

Communicated by S. Soatto.

✉ Pol Cirujeda
pol.cirujeda@upf.edu

Yashin Dicente Cid
yashin.dicente@upf.edu

Xavier Mateo
javier.mateo@upf.edu

Xavier Binefa
xavier.binefa@upf.edu

[1] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Tànger 122-140, 08018 Barcelona, Spain

1 Introduction

The description, detection and matching of points from different complex scenes is a challenging task for many Computer Vision applications such as tracking, object modelling and recognition or scene reconstruction. Existing approaches make use of all the available cues in the usual two channels of information: visual photometry such as color or textures, and shape and depth information from 3D sensors. While state-of-the-art methods have given successful outcomes in both areas, as reviewed in the following section, we are strongly encouraged to find a global method which can fuse information from both two worlds, and provide a descriptive unit which is able to encode surface definition and its correlated texture or pattern information together, as seen in Fig. 1. This must be supported with a global matching procedure so the scene can be observed from an overall perspective in order to cope with challenging conditions as local repetitive patterns or symmetries. Our aim is to avoid ambiguities which can be reduced at all levels: both locally if a shape is defined in conjunction with its associated visual cues, and globally with a holistic match refinement procedure.

This paper provides a threefold contribution: first, we develop the formulation of a covariance-based descriptor which is able to gather shape and visual information together within a radial 3D area. Thanks to its fundamentals, this descriptor is robust to noise changes, rigid spatial transfor-

Published online: 03 April 2015

Fig. 1 Example of a coherent visual and shape aware descriptor for matching in a 3D scene. While the visual appearance of the ball is similar on its real appearance and the paper printed representation, the matches should be correctly considered only on the true 3D points, since shape information should also encoded on the used descriptor for matching

mations or even resolution variations; and because of its low computational cost it can be extended to a multi-scale context for better discrimination performance. In a second place, we review an intrinsic property of the descriptor thanks to which it offers a procedure for keypoint extraction. Therefore, salient points in the scene (in terms of major color and shape variation areas) can be detected at the same stage where descriptors are being obtained. And finally, we provide a Game Theory based solution method which integrates local descriptor similarity with global 3D geometric consistency for a possible registration of several scene views. This method efficiently looks for the global minimal reconstruction error, taking into account all the available descriptor matches and avoiding local minima which other methods could find due to symmetries or local repetitions.

The remainder of this paper is organized as follows: in Sect. 2 we review the state-of-the-art approaches and related works. Sections 3 and 4 introduce our contribution in two separate sections: the covariance descriptor and scene analysis framework itself, and the Game Theory based solution approach for scene reconstruction. Section 5 presents and discusses the results, before concluding in Sect. 6.

2 Related Work

3D scene registration is currently an active topic in the computer vision literature, recently compelled by advances in the sensors technology which have provided some affordable devices and acquisition techniques. This has eased the capture of 3D information to the mass public and also produced an increase in the processing proposals for this kind of images during the last years.

This topic has been however studied since some time ago from several perspectives. One of the first proposals, which is still currently considered as one of the main methods in the 3D registration area, is the Iterative Closest Point (Besl and McKay 1992) proposed by Besl and MacKay. Their method estimates the registration between two 3D point clouds, performing an iterative process in order to minimize the mean square distance between two sets. The main problem of this algorithm is the need of a good initialization if we desire that the iterative process converges to a global minimum and not to a local minimum. In order to achieve this initial approximation, the typical procedure consists in the establishment of some correspondences between specific points of the two 3D points clouds. Once these correspondences have been established, both subsets of points can be registered by solving the classical problem of absolute orientation (Horn 1987).

Other state-of-the-art approaches for point set registration commonly use iterative algorithms such as RANSAC (Fischler and Bolles 1981) or its variants (Chum et al. 2004; Chum and Matas 2005, 2008) which allow the integration of geometric consistency as a measure for minimizing the correspondence error. This is basically an heuristic for comparing how a set of points fits within some geometric constraints: projections to a coordinate system, error measurements regarding a rigid transformation, or relative distances amongst connected point sets. Despite the existence of other possibilities, RANSAC is undoubtedly the predominant algorithm for the geometrically consistence situation in 3D registration, thanks to its good results and its standardized implementations. Authors like Johnson, in his Spin Images approach (Johnson 1997), also propose his own geometric consistency algorithm based in the same conceptualization of the Spin Image, but he also finally refers to RANSAC as the appropriate technique for more problematic cases. In fact, RANSAC has also been used as the basis for well-known methods of 3D scene registration which do not even use the correspondences information and only rely on the iteratively search of the RANSAC algorithm, as shown in Chen et al. (1999).

In any case, it is obvious that any registration process must rely on a previous search for correspondent points, which must take into account the similarity amongst these candidate matches. During last years, this selection of candidate correspondences has been achieved by using descriptors

which encode exclusively the 3D information from the scene points and can provide similitude measures amongst points. Inside this category, Spin Images (Johnson and Hebert 1999) is probably the most known method, representing the neighbourhood of each 3D point into a 2D image and later comparing it against other Spin Images by a simple correlation factor. Other popular 3D descriptors are the point signatures (Chua and Jarvis 1997), the 3D shape contexts (Frome et al. 2004), THRIFT (Flint et al. 2007) or, more recently, the Fast Point Feature Histograms (Rusu et al. 2009).

However, thanks to the availability of 3D scanners which can also capture texture information, some descriptors which encode simultaneously information from the 3D shape and the color have been recently published in the literature. Textured Spin Images (Brusco et al. 2005) are a good example of this trend. Novel approaches include the MeshHOG descriptor (Zaharescu et al. 2009), which performs a histogram of gradient of a 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 also used from the authors of the CSHOT descriptor (Tombari et al. 2011), which concatenates their SHOT descriptor (Tombari et al. 2010) and the color information. Other contributions as Kovnatsky et al. (2012) follow a more geometric perspective, where manifold embedding procedures are used and photometric information is implicitly encoded as part of the coordinate projection parameters.

Once the different correspondences have been established by comparing the descriptors, this first set of matches can be filtered by the aforementioned iterative registration methods, in order to discard the correspondences which can be incorrect or not accurate enough. A current challenge at this stage is posed by effects like symmetries or repetitive patterns in the scene: a correspondence between two 3D points could seem correct if we look individually, but incorrect in a more global context. A global algorithm taking into account all the previously found correspondences must be defined, allowing to obtain at the end of the process a subset of the initial group of correspondences which are geometrically consistent between them keeping local similitude as well.

3 Covariance Framework for Scene Analysis

Covariance matrices in the computer vision context arose as a way for relating several image feature statistics inside a region of interest, keeping a lower dimensionality space. This usage of covariance magnitudes as descriptive units was first introduced by Tuzel et al. (2006) for the detection of objects and faces. With promisingly good results, the approach was extended to more complex frameworks for pedestrian or objects detection from visual cues (Tuzel et al. 2008, 2007; Yao et al. 2008). Regarding 3D surface description (Fehr et al. 2012) is the only approach, to our knowledge, which explores different combinations of features obtained from range images related under a covariance analysis framework. Taking this as preliminary work, we extend it in order to deal with 3D point cloud scenes, making use also of correlated color information and with a formulation which is proven to be invariant to rotations, viewpoint, noise and density variations (Fig. 1).

3.1 Covariance Matrix Descriptor for Fusion of Shape and Visual Information

From a statistics point of view, covariance can be understood as a measure of how several variables change together. Within the context of the descriptor definition, the set of random variables must correspond to a set of observable features which can be extracted from points in the scene, e.g. pixel color values, depth magnitudes, 3D coordinates, etc. Therefore, the first step for the computation of the descriptor is the definition of a feature selection function $\Phi(p, r)$ for a given 3D point p and its neighbourhood within radius r in the scene:

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

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:

$$\phi_{p_i} = \left(R_{p_i}, G_{p_i}, B_{p_i}, \alpha_{p_i}, \beta_{p_i}, \gamma_{p_i}, \right) \tag{2}$$

This feature selection function includes the following observations that are robust to spatial transformations, as they 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. 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.

In Fig. 2 we show an example of how these measures are obtained. As these selected features are relative measures

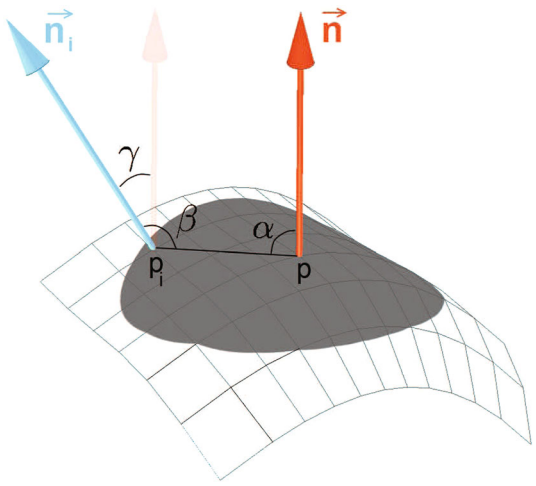


Fig. 2 Scheme of the used features for shape information encoding. For each p_i in the neighbourhood of p, α, β and γ are the rotational invariant angular measures

in terms of shape description, their usage in the covariance descriptor formulation guarantees a rotation and view invariance, which is a desired behaviour in descriptor performance. RGB space color values also lose structural information and become observations of a sampling distribution within the covariance descriptor formulation, therefore they will become invariant to rigid transformations in the scene. Even if in a more formal sense an intermediate colour invariant projection must be performed for minimizing the impact of illumination variations and offering a true robustness to view changes, we consider this is beyond the scope of our approach at its current stage and will be included as a future extension—thanks to the ease of our presented descriptor for including new features. In any case, for small descriptor localities, RGB color space values have demonstrated to be significant enough. Finally, variables are normalized in order to have an equivalent range both for angular and color measure.

Then, for a given point p of the scene the covariance descriptor can be obtained as:

$$C_r\left(\Phi(p,r)\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_{p_i}-\mu\right)\left(\phi_{p_i}-\mu\right)^T \qquad (3)$$

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.

A covariance descriptor can be seen as a high level and abstract representation which treats the observed features as samples of joint distributions, and looses all the spatial notion (information about the number of points and their ordering) within the region. This compactness provides a combination of flexibility—feature distributions will contribute to the descriptor still preserving their inner characteristics even under changes of scale and rotation in data—and robustness—according to central limit theorem, as long as a significant enough number of samples is used the data within a certain range within the features distribution will be correctly represented. In addition, these two facts yield a valuable performance boost in comparison to other descriptors based on more rigid representations such as histograms. Figure 3 shows an example of a covariance descriptor.

3.2 Manifold Topology of the Descriptor

A remarkable consideration about the proposed descriptor is its geometrical topology. Covariance matrices, being symmetric positive definite matrices, do not lay on a Euclidean space, but on a Riemannian manifold. Indeed, covariance descriptors form the $d \times d$ dimensional space of symmetric positive definite matrices, where d is the number of used features ($d = 6$ in our descriptor approach), and the main concern is that this descriptor space is meaningful for scene definition purposes as it abstractly represents a geometrical location of shape and texture distributions within a scene point area. This assertion can be visualized by a proof of concept as shown in Fig. 4. In an instance of a scene we have computed descriptors at different areas from different nature in shape and colour. Once the manifold distances amongst the set of descriptors have been computed, we have applied a Multidimensional Scaling embedding onto a 2D coordinate space in order to graphically represent the consistency of our descriptor space. The plot demonstrates how different points coming from different areas in the scene are located in the descriptor space.

The most important implication of this manifold descriptor space is that it provides a formal way of comparing descriptors, while other approaches are forced to use distance approximations as histogram differences or correlations. While there exist different approaches in the literature based in local Euclidean approximations (Cherian et al. 2011; Arsigny et al. 2006) with an efficiency compromise in mind, we have opted for the use of the geodesic distance proposed by Förstner in Förstner and Moonen (1999). This is adequate to our context as no prior knowledge might be available between two arbitrary descriptor points in a context of scene description and matching, therefore a local Euclidean approximation might not be accurate in most of the cases. Furthermore, with such a low dimensionality in our descriptor definition, the computational expense regarding the accuracy gain of a manifold aware distance is permissible.

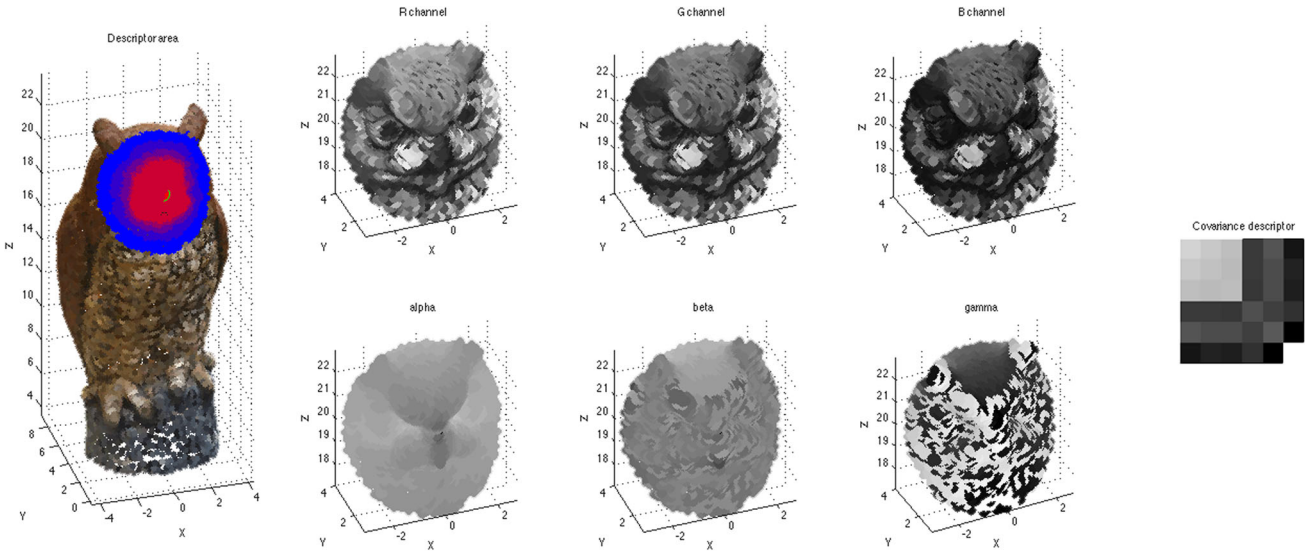


Fig. 3 Example of a scene view where a multi-scale covariance descriptor is extracted on the face of an owl model. The left image shows the original 3D scene where the overlap gradient of colors from *red* to *blue* depicts 5 different scales used for obtaining a multi-scale descriptor. The 6 central subfigures show the different used features, in terms of color (*upper row*) and shape description (*bottom row*). Finally, on the right, a single scale 6×6 covariance descriptor is graphically represented (Color figure online)

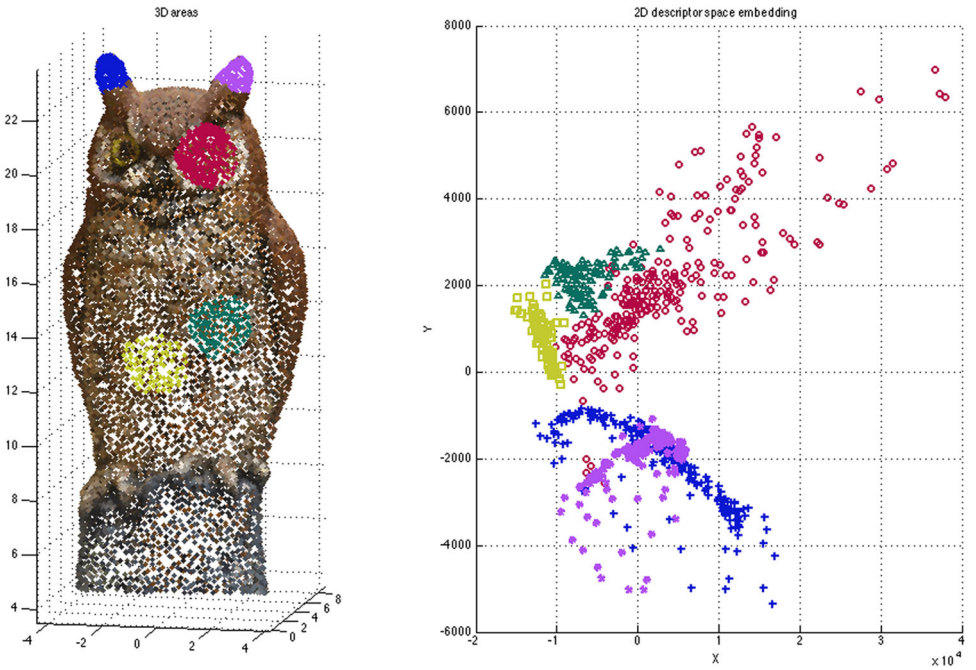


Fig. 4 Set of scene areas where descriptors are obtained and their embedding to a two-dimensional space. Scene areas include two ears (marked in *blue* and *purple*) which are similar and therefore overlapped on the descriptor space plot. As these areas suffer changes in shape, its clusters are visually disperse. The *red marked points*, belonging to an eye area with changes in both colour and shape, appear separated from other clusters and also disperse due to this intra-area variations. Finally, yellow and green points belong to different homogeneous body areas, therefore they appear close in the 2D descriptor space (with a slight location variation due to slight differences in texture tone), and with a certain cluster compactness (due to their similar shape) (Color figure online)

Therefore, in order to measure the similarity of two arbitrary descriptors, the metric for computing distances between two covariance matrices C_r^1 and C_r^2, is defined as follows:

$$\delta(C_r^1, C_r^2) = \sqrt{\sum_{i=1}^{6} \ln^2 \lambda_i(C_r^1, C_r^2)} \tag{4}$$

where $\lambda_i\left(C_r^1, C_r^2\right)$ is the set of generalized eigenvalues of C_r^1 and C_r^2, whose magnitude express the geodesic distance between the compared points, preserving its curvature along the manifold.

3.3 Multi-scale Covariance Descriptor, MCOV

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:

$$C_M(p) = \{C_r\left(\Phi(p,r)\right), \forall r \in \{r_1..r_s\}\} \tag{5}$$

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, as depicted in the most left sub-image in Fig. 3. Then, we are intentionally seeking matches of points which are locally similar, but also related in a more global area. This can help to avoid repeatability problems and improve detection of points in edges or borders of scene objects. The radius estimation procedure, which is the only parameter the proposed method needs, is self-contained in our approach and commented below in Sect. 3.4.

Finally, in the multi-scale descriptor framework, it is easy to extend the metric defined in Eq. (4) in the following way:

$$\delta_M(C_M^1, C_M^2) = \sum_{i=r_1..r_s} \delta(C_i^1, C_i^2) - \max_{j=r_1..r_s}\left[\delta(C_j^1, C_j^2)\right] \tag{6}$$

where C_i^1 and C_i^2 are the covariance descriptors belonging to each one of the $i = r_1...r_s$ radius scales, at each one of both scenes respectively. The formulation behind Eq. (6) takes into account the similarities of all scales except the one providing a lower similarity j, which is ignored because it might contain a major dissimilarity at a given scale—due to a possible dissimilarity on a border, an occlusion, or other artifacts.

3.4 Covariance Descriptor Properties for Scene Pre-analysis

Covariance matrices as descriptors have still other desirable outcomes thanks to their mathematical underlying fundamentals which allow several previous scene analyses. One of them is that they can be also used as keypoint detectors in a direct way. As defined after Eq. (3), 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 (Wilks 1932). As the used features have been previously normalized, there is no range variation which could interfere on this analysis. Starting with an arbitrarily big radius parameter at a single scale (empirically we determined this as the magnitude corresponding to the 5 % of the scene coordinates volume range) one can compute the covariance descriptor matrices for all the points in the scene, and observe all their determinants. Then, the ones with higher values can be interpreted as the points which belong to real interest areas, with inner significant variation 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 Fig. 5, keypoints are still obtained on significant parts such as eye holes or borders.

Due to the nature of the proposed descriptor radial neighbourhood, relevant points might tend to form small clusters as samples could be shared for closer points, therefore producing similar descriptors. This can be reduced with relevance sampling procedures like the one proposed in Torsello et al. (2011). In our approach, we naturally relate this to the aforementioned concept *generalized variance* and also exploit it as the associated relevance of a point in the scene. We want to explore all possible saliency clusters and isolate those points with major relevance in a similar formulation as the one introduced in Torsello et al. (2011): for each one of the previously obtained keypoints kp at each scene S, we will compute its *relevance region* R_{kp} as:

$$R_{kp} = \left\{q \in S \,|\, \hat{\sigma}_{kp} - \hat{\sigma}_q > T, \forall q \text{ s.t. } |kp - q| \leq r\right\} \tag{7}$$

where $\hat{\sigma}$ is the *generalized variance* of each point, r is a radius parameter and T is a threshold parameter which we empirically set to 0.7 times the maximum *generalized variance* found in the points of the scene. Finally, a measure of distinctiveness can be assigned to each one of the relevance regions R_{kp}:

$$f(p) = \|R_{kp}\|^{-k} \tag{8}$$

where $\|R_{kp}\|$ is the 2-norm of the points in R_{kp} and k is an equalization parameter in order to change the relative weight of really distinctive points (the larger its value, the more distinctiveness of points in a small patch is emphasized). We empirically set k value to 1. We can finally keep the points with maximal values according to the distinctiveness features and observe how these belong to local isolated points within the original saliency clusters as depicted in the right image in Fig. 5.

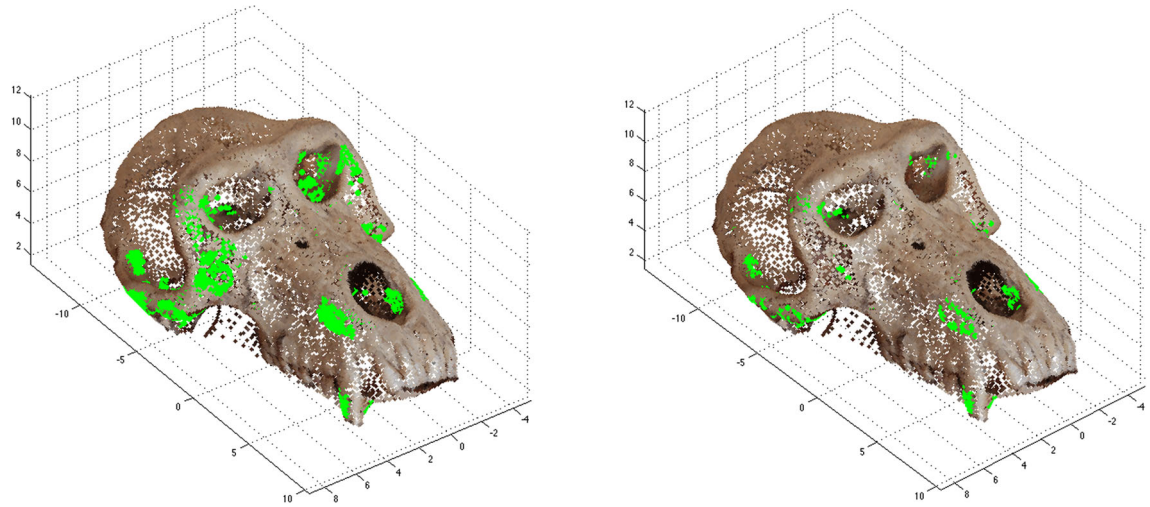


Fig. 5 Visual example of keypoint analysis by generalized variance. The *left sub-figure* shows the 1500 most significant points of the scene, marked by sorting their covariance descriptor determinants (generalized variances) in descendant order. Even if the color information of the object is homogeneous, interest points have been detected on salient areas of the scene. The computational cost of such task is minimal. The *right image* shows the set of points after the relevance sampling procedure, which in this case isolates 488 salient points with a major degree of sparsity regarding the previous saliency clusters. This can help reducing further registration errors (Color figure online)

We can also precede this saliency analysis by a point suppression stage thanks to the analysis of the rank of the covariance matrix descriptors. If different feature observations within the neighbourhood of a given point are correlated, which is not desirable, the rank of the descriptor matrices will be lower than the number of used feature dimensions. This straight criterion allows discarding uninteresting points where the covariance descriptor does not capture any significant differentiation between features.

Finally, we can integrate an estimation procedure of a more narrow radius value, which can take into account the nature of the scene in order to fit its probabilistic definition of points with a more accurate area sensitivity. From statistics theory we know that 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 with the following expression: $P\left(\mid \bar{X} - \mu \mid \geq \epsilon\right) \leq \sigma^2 / \epsilon^2 n$, 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. As an example, if we want to infer the number of samples such that data will lay within 0.1 units the original distribution, with a confidence of the 95 %, this can be expressed as $P\left(\mid \bar{X} - \mu \mid < 0.1\right) \geq 0.95$. This is equivalent to $P\left(\mid \bar{X} - \mu \mid \geq 0.1\right) \leq 1 - 0.95$, therefore we can relate it to Chebyshev's inequality and generalize the following expression for an arbitrary feature distribution:

$$n \geq \frac{\sigma^2}{\epsilon^2 \left(1 - p\right)} \tag{9}$$

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 an upper bound on the minimum needed number of samples n necessary to limit the sampling mean confidence error to a given value. Relating this to our framework, while calculating the covariance descriptors for the proposed scene pre-analysis we are already observing each one of the six used feature distributions at each scene point neighbourhood, and this is being kept encoded at the diagonal values of the set of descriptors. Therefore, we can apply the boundary equation defined in (9) for each one of the feature variances, defining a set of 6 candidate sampling sizes. As this provides a lower boundary, we will keep the maximum value of the candidate sizes. While this is a scene-dependant, quite flexible methodology, it allows for an adaptive method in the case there are areas with specific high variation. Its formulation is coherent along the points of the whole scene and provides specific descriptor constructions, rather than using a static radius parameter for the whole scene. Usual analyses for scenes with average homogeneity of shape and color (as most of the ones depicted in Fig. 8, whose point clouds have densities ranging from 20000 to 30000 points) reflect the need of taking around 400-500 samples within the radial neighbourhood. This sampling size can be translated to a radius magnitude according to the density of the scene point cloud. For the multiscale approach, we propose the usage of 5 different scales, with radius scales $s = \{1, 1.1, 1.3, 1.6, 2\}$ times the single-scale descriptor radius. This scaling distribution focuses the attention on narrow neighborhoods, while coarse areas are still present for disambiguation.

We finally want to reinforce the idea that the proposed descriptor methodology is not only suitable for the core task of 3D scene point definition, but also integrates a set of possibilities on the statistical analysis of data, as gathered up on this section, which provides an added value to the framework.

4 Globally Aware Scene Registration by an Evolutionary Game Theory Approach

The descriptor introduced so far has proven to be discriminative enough for a local recognition of a point in different views of a scene. The associated salient point detector, which pre-selects a set of relevant points according to what has been explained on previous section, is also helpful on this high descriptiveness level. Nevertheless, if the scene contains unavoidably similar points due to facts as repetitive patterns of an object or symmetries, this could pose a bigger challenge for the descriptor which could only be addressed with the help of scene-wise knowledge taking into account these particularities. This focuses our interest on the proposal of a descriptor matching methodology which does not only takes into account relationships between local point similarities, but also encodes a set of global restrictions in order to avoid possible ambiguities in the whole scene. Our proposed framework will perform a rejection of all the points which do not fit into this set of scene-wise constraints, leaving only a selection of correctly considered point matches. We will propose a global heuristic for match evaluation based in the so called geometric consistency.

As previously stated in Sect. 2, state-of-the-art approaches in scene registration are commonly based on top of iterative algorithms such as RANSAC or its later variants (Fischler and Bolles 1981; Chum et al. 2004; Chum and Matas 2005, 2008). However, there are several aspects in these algorithms which do not suit our purposes. The most important one is the presence of a possibly high number of outlier matches which could have an impact in the performance of the method due to the number of needed iterations, or even worst, the achievement of a convergent solution which is not a correct registration. In the context of registering complex scenes with a huge number of descriptor matches to be discarded, the impact of high amounts of outlier candidates might become computationally intractable as well. Other drawback considerations include the need of input parameters which must be tuned in order to obtain a valid solution, and the need of a high number of random evaluations of possible combinations producing, in the specific case of registration, an unoptimized performance of the method and probably different solutions for two different executions limited in time.

We introduce a registration method for the context of matching 3D scene views which entails a significant conceptual innovation regarding the aforementioned methods. In a first place, the change of paradigm implies not to compute implicitly the spatial transformation between scenes at each iteration (and to temporarily evaluate it), but to perform a rejection of all those matches which do not satisfy a set of constraints. The spatial transformation will be computed in a final instance, as the result of a limited set of leftover candidate points. In a second place, the presented method will allow to enclose a formulation expressing the adequacy of each match within the final solution: we propose to combine the geometric consistency together with the point descriptor likelihood (which, as a reminder, is well defined by the metric in Eq. 6). Therefore the set of constraints will be joining both local and global information. And, last but not least, the proposed method will be independent of the presence of outlier candidate matches, and theoretically guaranteed to converge asymptotically towards the solution: at each iteration, it will discard one candidate match. Therefore, the maximum number of trials will depend on the number of match samples, which is a clear advantage in a high presence of noise or outliers regarding an approach like RANSAC (this will be commented hereafter in an experimental set-up in Sect. 5).

We propose to translate the global scene matching problem to a Game Theoretic field using the so called Evolutionary Stable Strategy—ESS—solver introduced in Albarelli et al. (2009). This approach presents a framework where a set of abstract candidates of a system are successively discarded in order to obtain the best remaining combination of them according to a defined heuristic function. While the ESS method defined in Albarelli et al. (2009) is a standard methodology for game-theoretic problem solving, Albarelli et al. have used this approach in a more scene registration focused application context in Albarelli et al. (2010) and Rodolà et al. (2013). In these cases, they use the Game Theoretic solver for refining matches that have already been filtered by a standard descriptor pairing algorithm MeshHOG descriptor and associated MeshDOG keypoint locator (Zaharescu et al. 2009). Therefore, their heuristic functions for game definition are limited to simple spatial constraints for a final discard of those matches. The underlying idea is to consider all the pairwise matches of two compared scenes, and calculate a penalty value associated to the cost of hypothetically choosing each one of these correspondences as part of the final registration solution. In the Game Theoretic framework these values are named payoffs, and can be computed at once for a set of correspondences resulting from a descriptor matching stage. In an analogy to a game, these payoffs will be the set of "rules", each scene view will be a player and each match candidate a game turn. Therefore, the best play of the game (the best registration between scene views) will take place by the best set of turns for both players—that is, the best set of matches at each scene incurring on the best global cost. The aforementioned set of payoffs is codified in a matrix

notation where all the possible pair choices are being taken into
account.

In the following subsection we propose the definition of a payoff term which is able to integrate both global scene geometric structure constraints and descriptor similarities. We believe that a Game Theoretic based solver is powerful enough for taking into consideration similarity constraints of point descriptors, and scene-wise geometrical structure restrictions for relevant challenges as symmetries or repeated areas, in a single game definition. This is different to other current approaches which depend on a previous descriptor extraction and correspondence stage and use the Game Theory framework as a gathering of rules in order to reject match candidates which do not fit some local surface characteristics.

4.1 Modelling the Game

The main complexity of the Game Theoretic matching framework lies on the payoff matrix building step, which must take into account all the possible pairwise affections between candidate matches of both scenes. We emphasize that each match payoff must encode all the information that must be assigned to a pair of points, both in a positive or a negative way: in this sense, we will put together the cost related to local likelihood, geometric consistence, and relative distance of points. This latter term helps in a better discard of undesired matches: the ideal keeping of point candidates is that set which is sparse enough with local similarities of point descriptors. This can be seen as finding the most spaced-out sub-graph common to both scenes graphs, where the vertices are similar enough (see Fig. 6).

We propose to use a Game Theory based solution as a holistic way of grouping all the information available both in terms of descriptor similarities and scene-whise geometric prior knowledge. With these conditions, we propose the building of the payoff matrix as follows. Let A and B be the scenes we want to register. Let $\{a_m\}$ be the set of points in A and $\{b_n\}$ the set of points in B. Then we have a set of k candidate pairs $\{(a_i, b_j)\}$ which have been preselected according to the best covariance descriptor likelihoods between scenes. For each pair, we can evaluate its game payoff regarding any other pair of matches, exhaustively. Therefore, a matrix C of game payoffs, of size $k \times k$, is defined for all combinations of pairs $\{(a_i, b_j), (a_k, b_l)\}$, and it will take into account all the relationships in the scenes with the corresponding incidence over the global registration error:

$$c_{(a_i,b_j)(a_k,b_l)} = P_{\text{desc}} \cdot P_{\text{geom}} \tag{10}$$

where P_{desc} is the payoff related to the covariance descriptor similarity and P_{geom} is the payoff related to the geometric consistency. In more detail, both values are defined as follows:

$$P_{\text{desc}} = f(a_i, b_j) \cdot f(a_k, b_l) \tag{11}$$

$$f(a, b) = e^{-\delta_{\text{M}}(C_M(a), C_M(b))} \tag{12}$$

where δ_{M} is the multi-scale Förstner distance between the covariance descriptors of a and b. With this term we are taking into consideration a normalized payoff value associated to the local similarity of points.

Regarding geometric consistency, we define:

$$P_{\text{geom}} = \frac{\min(d(a_i, a_k), d(b_j, b_l))}{\max(d(a_i, a_k), d(b_j, b_l))} \cdot g(a_i, a_k, b_j, b_l) \tag{13}$$

$$g(a_1, a_2, b_1, b_2) = e^{-|\text{d}(\text{a}_1,\text{a}_2)-\text{d}(\text{b}_1,\text{b}_2)|} \tag{14}$$

where $d(x, y)$ is the Euclidean distance between 3D points x and y. As we are working with 3D information, we are able to ensure that the euclidean distance between points of the object is the same from every point of view.

The first *min/max* term in this geometric payoff was originally used in Albarelli et al. (2010) and Rodolà et al. (2013), and penalizes elements which are closer in the scene as they would incur in more error if they where selected as a wrong part of the registration solution. Note that these approaches are based on a previous keypoint detection and matching stage, therefore a single spatial constraint such as this one is enough for the rejection of erroneous pair candidates, provided the keypoints are correctly matched on controlled scenes. As we are focusing our approach to the registration of complex, textured scenes with still many point candidates at this point, we add a second term in this payoff value defined in Eq. 14. This adds a normalized coefficient which indicates the structure similarity between points in both scenes. All these constraints, together with the aforementioned P_{desc} term in Eq. 12, define a single game which is capable of

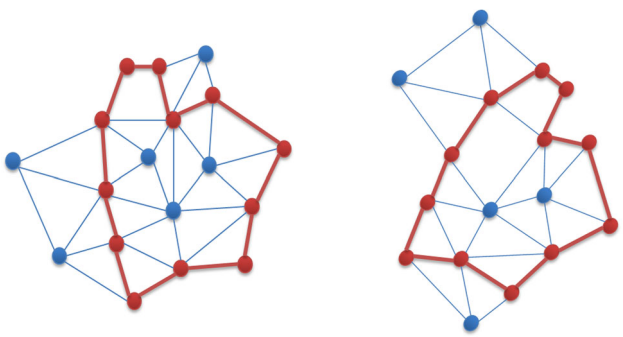


Fig. 6 Schema of how a common sub-graph must be selected by the game theory solution. On the *left:* the graph obtained by the cloud points from the first scene; on the *right;* the graph obtained by the cloud points of the second scene. Marked in *red* there is the most suitable common sub-graph found within both graphs (Color figure online)

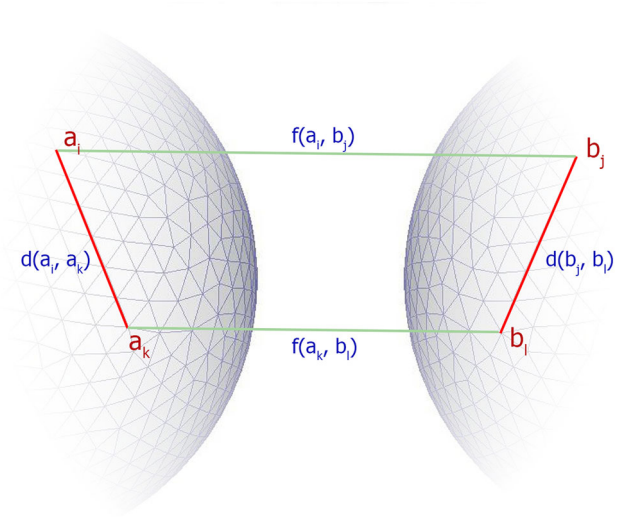


Fig. 7 Scheme of the elements involved in payoff calculations. $f(a, b)$ expresses the descriptor likelihood between a pair of matches. $d(a_1, a_2)$ evaluates the geometric consistency on the match candidates within the pair of matches which is being evaluated

selecting registered point pairs taking into account both texture and shape information. See Fig. 7 for a clarification of the elements involved in such calculations.

A visual representation of a general payoff matrix remains as follows:

$$C = \begin{array}{c} \begin{array}{ccccccc} & \cdots & a_i & \cdots & a_k & \cdots & \\ & \cdots & b_j & \cdots & b_l & \cdots & \end{array} \\ \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 0 & \cdots & c_{(a_k b_l)(a_i b_j)} & \cdots & a_i \;\; b_j \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & c_{(a_i b_j)(a_k b_l)} & \cdots & 0 & \cdots & a_k \;\; b_l \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix} \end{array} \tag{15}$$

4.2 Playing the Game

Finally, we want to find a stable solution to the game represented by the payoffs modelled in C, which are in fact the implicit restrictions of the candidate matches of the scene registration. According to Albarelli et al. (2009), the evolutionary stable solution of the game is the so-called support vector x whose response to the game is maxima: $x^T C x \geq y^T C y \, \forall y \in \Delta$, where $\Delta = \left\{ x \in \mathbb{R}^k : \sum_{i=1}^{k} x_i = 1 \text{ and } x_i \geq 0 \right\}$, the space of all vectors which are solutions to the game.

The support vector x can be found via the *Evolutionary Stable Strategy solver* algorithm proposed in Albarelli et al. (2009). If $x_i > 0$, then the match belonging to column or row i in C is marked as a positive correspondence. The own values x_i can be considered as normalized weights expressing the confidence associated to each correspondence. In Albarelli et al. (2009) some other valuable details are examined: in a first instance, it is shown that if we want a mixed solution (that is, more than one element in x satisfies $x_i > 0$), then we need that $c_{ii} = 0$ and $c_{ij} \geq 0 \;\; \forall i \neq j$. In a second place, the algorithm is proven to converge to a unique and global solution which takes into account all the payoff values associated to all the possible matches between scenes. And finally, this convergence is guaranteed in an asymptotic way and with a linear time complexity per iteration.

Therefore the solution found in the support vector indicates the indices of the subsets of correspondences in both scenes which can be used to find the spatial transformation needed in order to change the coordinates of one scene into the other, by solving the problem of absolute orientation (Horn 1987) for registration. We summarise all the involved steps of our approach in Algorithm 1.

Algorithm 1: Overview of the proposed registration procedure

input : Two scene views in the form of 3xN point clouds
output: Rigid transformation (rotation R and translation T) between scene views

Stage 1: *scene pre-analysis*
- Compute pre-descriptors at view 1 and view 2.
- Perform generalized variance analysis.
- Prune salient areas by relevance sampling.
- Estimate *covRadius* by confidence intervals.

Out: *keypoints kpSc1, kpSc2, covRadius*

Stage 2: *descriptors obtention*
for $kpSc1 \leftarrow 1$ ***to*** $\#kpSc1$ **do**
    - compute each MCOV(kpSc1)
end
for $kpSc2 \leftarrow 1$ ***to*** $\#kpSc2$ **do**
    - compute each MCOV(kpSc2)
end
- compute distance matrix between descriptors

Out: *distMatrix size* $\#kpSc1 \times \#kpSc2$

Stage 3: *get candidate correspondences*
for $r \leftarrow 1$ ***to*** *#rows distMatrix* **do**
    - get best correspondence according to MCOV distances (*inclusive/exclusive ratio criterion*)
end

Out: *set of n candidate pair matches*

Stage 4: *registration via Evolutionary Game Theory*
- build payoff matrix C
- apply Evolutionary Stable Strategy Solver (Albarelli et al. 2009)

Out: *set of m final matches*

- Compute $[R, T]$, rigid transformation from set of matches (absolute orientation)

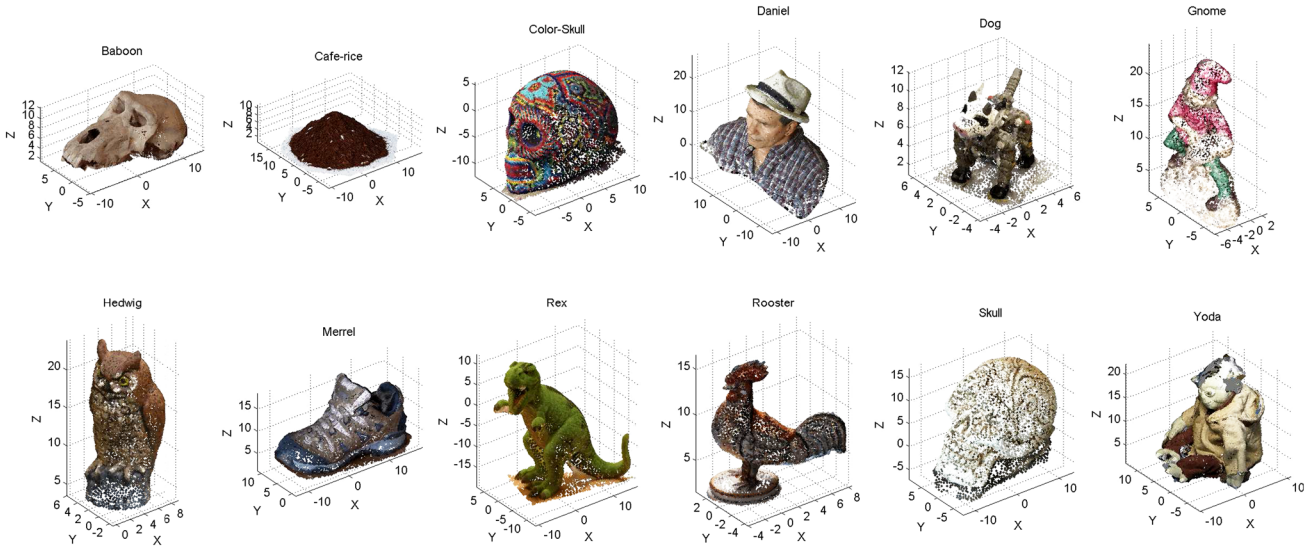


Fig. 8 3D plot of the 12 models included on our database. Full scenes are shown without added noise

5 Experimental Results

We are validating our proposed descriptor approach on top of a dataset combining 3D shape with visual information. This dataset contains 12 scenes which have been obtained using Autodesk 123D Catch [1] 3D modelling software. Our dataset combines scenes of originally acquired objects and others available on the 123D Catch website under a Creative Commons license. These models are stored as 3D meshes with photometric texture, where each vertex has a unique identifier for experimental ground-truth purposes. See Fig. 8 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 for testing the performance of our method: repeated areas, homogeneous surfaces and textures, and symmetries.

5.1 Descriptor Comparison

In order to test the descriptor performance, we will compare our MCOV Covariance Descriptor approach against the state-of-the-art methods MeshHOG (Zaharescu et al. 2012) and CSHOT (Tombari et al. 2011). In addition, we will also evaluate the performance of the Textured Spin Images approach (Brusco et al. 2005) which, even if it dates back to 2005, is a variation of the original Spin Images approach (Johnson and Hebert 1999). This is still considered one of the classical 3D descriptors in the literature for successful matching of dense scenes and we want to include its results in our comparison as a base line of a method which set up a standard in 3D scene matching. The compared descriptor approaches are used following the original implementation by their authors, and any needed parameter (radius, bin size) 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.

5.1.1 Performance Over Noise Variations

In this experimental evaluation we are testing the descriptor tolerance to noise variations. Each model in our database is affected by a variation including (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 distributions with standard deviations according to 2, 4, 6, 8 or 10 % of each one of the data channels. Therefore, for each model we have performed a cross validation procedure including 10 folds of 100 randomly selected points along the surface of the scene. For each one of the evaluated points, we have computed its descriptor likelihood against the same set of points on the variation of the model. The evaluation method consists on observing the amount of false and true positives, and false and true negatives averaged along the cross validation test, in terms of matching scene points by their according descriptor likelihood measures. For our descriptor, we will use the metric defined in Eq. (6). According to a *ratio* parameter, we present two criteria for evaluation:

- The so-called *exclusive ratio*, considers a match as a true positive if and only if the descriptor likelihood between the match points is *ratio* times better than the second

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

best match candidate likelihood. This criterion variant is inspired in current approaches as SIFT (Lowe 2004) and has the particularity of being more restrictive on finding true positive matches, reducing also the apparition of false positives. Due to its behaviour, this selection is suitable for the evaluation of the descriptor performance itself.

- In the so-called *inclusive ratio*, we consider as true positives all those matches which are within the boundaries of *ratio* times the best likelihood of this set of candidates. In this case the rate of true positive candidates is increased, but this has the expense of increasing the risk of appearance of false positives. This criterion is suitable for a whole registration procedure as a point is associated to many matches as long as they are similar within a range of likelihood measurements, at the expense of requiring a rejection method afterwards in order to deal with elements external to the descriptor itself, as pattern repetitions in the scene or symmetries.

Both matching criteria are presented and evaluated in the experiments as they might be of different adequacy regarding the application context of the descriptor: as we stated before, the main difference between both methods is the amount of tolerated false positives they allow. Assuming a descriptor is reliable at representing a given area, the presence of false positives is not a drawback by itself; it is just a side effect due to the possibility of repetitions of visual patterns or surfaces in the scene. Therefore, the *inclusive ratio* criterion is more flexible and is allowing this fact to happen. In some applications, such as object detection or scene registration, this can be a desired feature, but it puts into consideration the needing of some sort of global mechanism which must be capable of finding repetitions, symmetries, etc. and filter out the non-positive matches according to global error minimizing constraints such as geometric consistence, which is why our descriptor proposal is paired with a scene-wise Game Theoretic solver definition. Nevertheless, both criteria are complementary, with a common point when both exclusive or inclusive ratio parameter is set to 1. In this case, both criteria are conceptually the same one.

The results of the experiment are presented as follows: for each level of noise we move the *ratio* coefficient within a range of 1–5 and we obtain a set of ROC curves as exemplified in Figs. 9 and 10 for *exclusive* and *inclusive* ratio methods respectively. This is useful for comparing the behaviour of the different tested descriptors under all noise variations, for each one of the twelve available models. As it can be observed in the separate figures, due to aforementioned complementarity the ROC curve plots belonging to inclusive criterion are the continuation of the exclusive criterion ones (please note the later ones are zoomed in in order to offer a better visualization). We agree that in a more formal sense, the plots should be presented continuously, and with a more extense *ratio* parameter space exploration in order to offer normalized coordinate axis between 0 and 1 values. However, the current figures intend to offer more detail in order to interpret the results: using the current disposition, the figures clearly display that when the *ratio* parameter is set to a defined higher value of 5, some of the descriptors have a false positive rate of 1, while other ones still maintain this value in a lower level.

Table 1 Average AUC measures for 12 models, *exclusive ratio* evaluation, 100 versus 100 % resolution, for 5 levels of noise

	n002	n004	n006	n008	n010
MCOV	0.896	**0.868**	**0.781**	**0.758**	**0.710**
CSHOT	**0.911**	0.799	0.602	0.534	0.511
MeshHOG	0.745	0.703	0.613	0.528	0.506
Text. SpinImages	0.619	0.544	0.540	0.523	0.503

Bold values indicate the best performance in each case

Table 2 Average AUC measures for 12 models, *inclusive ratio* evaluation, 100 versus 100 % resolution, for 5 levels of noise

	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
Text. SpinImages	0.750	0.614	0.615	0.564	0.533

Bold values indicate the best performance in each case

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 Tables 1 and 2 for exclusive and inclusive ratio criteria, respectively.

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. On the contrary, the MCOV descriptor offers a more flexible representation since it considers 3D points as samples of a distribution and, by construction, subtracts the mean of this samples distribution: therefore, in case of noise, it will be naturally attenuated.

Thanks to this experimental set-up, we can also extract a conclusion about fusion of color together with shape information, as the three database models selected to be represented column-wise in Figs. 9 and 10 provide three different challenging scenarios in terms of color homogeneity, repetitive patterns or great color variation, respectively. In this sense, one can see how classical histogram representations, as in the basis for MeshHOG or Textured Spin Images, are clearly affected by color variance. The usage of the *Hedwig* model

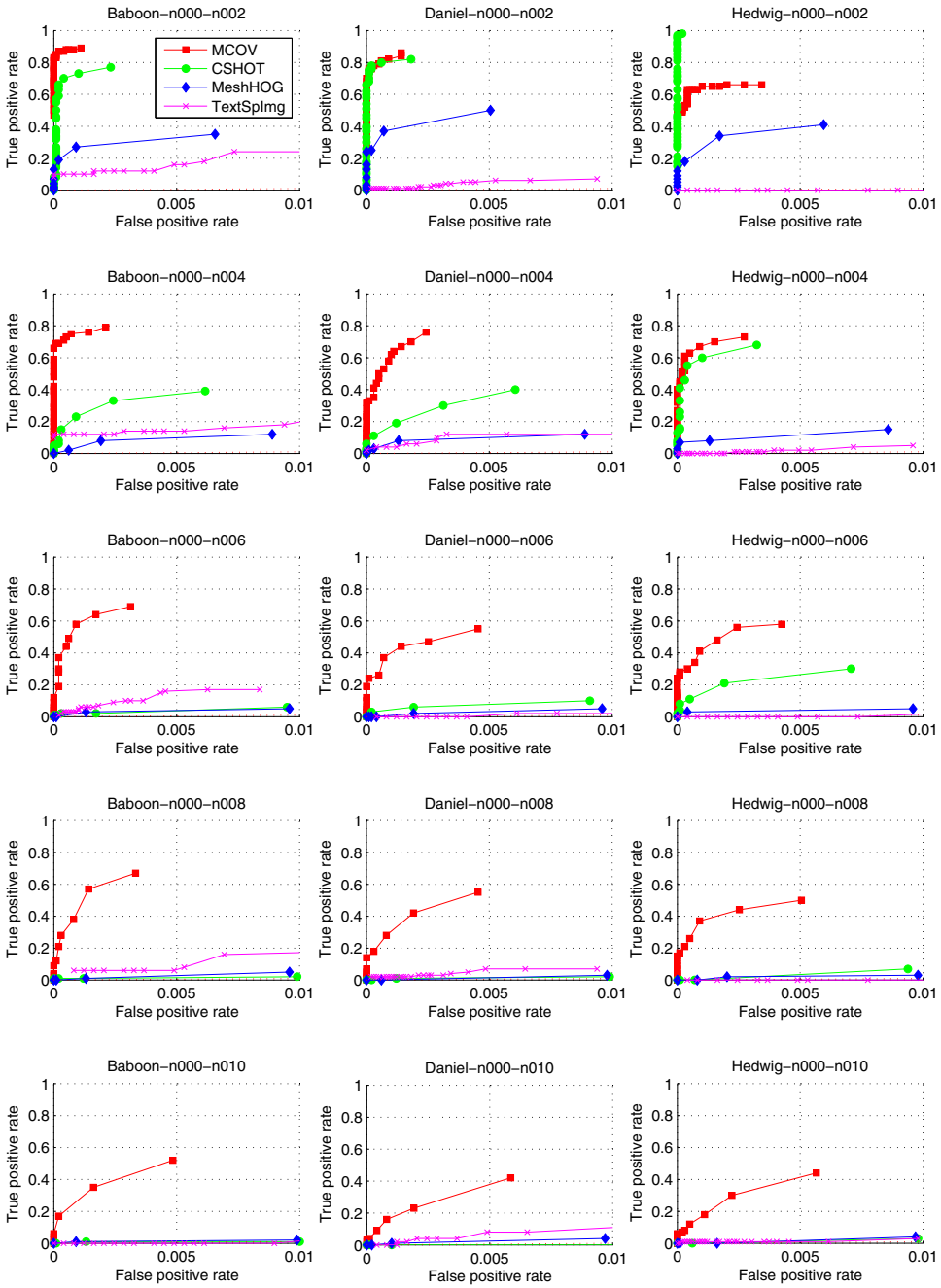


Fig. 9 ROC curves for comparison of several 3D descriptors, using the *exclusive ratio* criterion. Each *column* depicts a test on a different model of the database. Each *row* shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10 % of the standard deviation of color and surface coordinates) (Color figure online)

is a clear challenge for the Textured Spin Images approach 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 Brusco et al. (2005). Other more flexible approaches as CSHOT or our statistical-based Covariance Descriptor offer more robustness in its representation until bigger amounts of applied noise.

5.1.2 Performance Against Resolution Changes

A very similar experiment to the one presented on the previous section is also conducted by applying a high 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.

Again, by moving the *ratio* coefficient within a range of 1–5, we can obtain a set of ROC curves for the 12 tested models under the same 5 noise variations as in the previous experiment. Tables 3 and 4 reflect the associated average AUC measures for exclusive and inclusive ratio criteria, respectively. As in the previous experiment, we also attach the ROC curves corresponding to the *Baboon*, *Daniel*

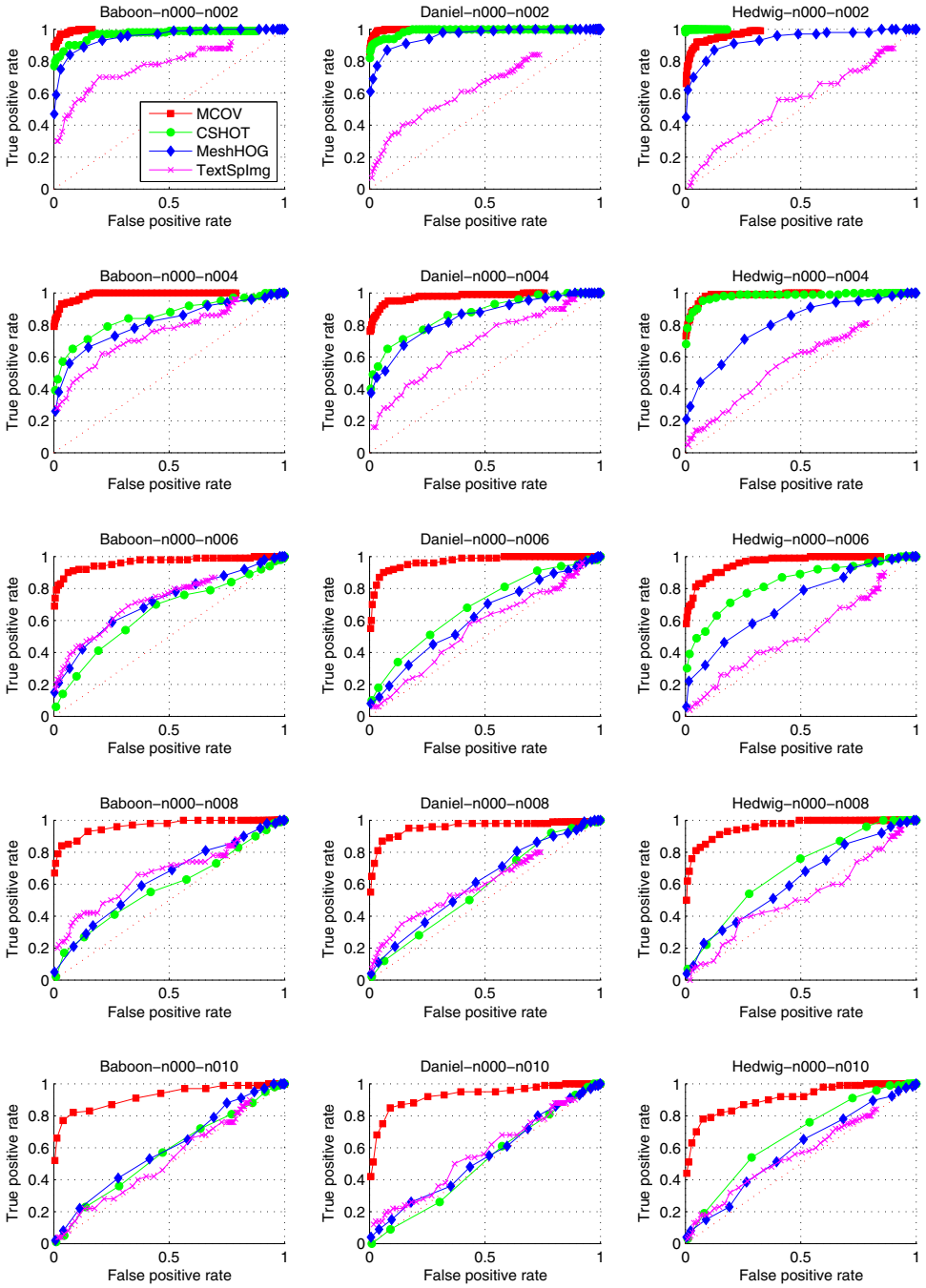


Fig. 10 ROC curves for comparison of several 3D descriptors, using the *inclusive ratio* criterion. Each *column* depicts a test on a different model of the database. Each *row* shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 1 % of the standard deviation of color and surface coordinates) (Color figure online)

Table 3 Average AUC measures for 12 models, *exclusive ratio* evaluation, 50 versus 100 % resolution, for 5 levels of noise

	n002	n004	n006	n008	n010
MCOV	**0.874**	**0.813**	**0.732**	**0.657**	**0.599**
CSHOT	0.772	0.651	0.572	0.515	0.510
MeshHOG	0.561	0.547	0.523	0.521	0.511
Text. SpinImages	0.572	0.522	0.527	0.498	0.498

Bold values indicate the best performance in each case

Table 4 Average AUC measures for 12 models, *inclusive ratio* evaluation, 50 versus 100 % resolution, for 5 levels of noise

	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
Text. SpinImages	0.662	0.613	0.563	0.534	0.520

Bold values indicate the best performance in each case

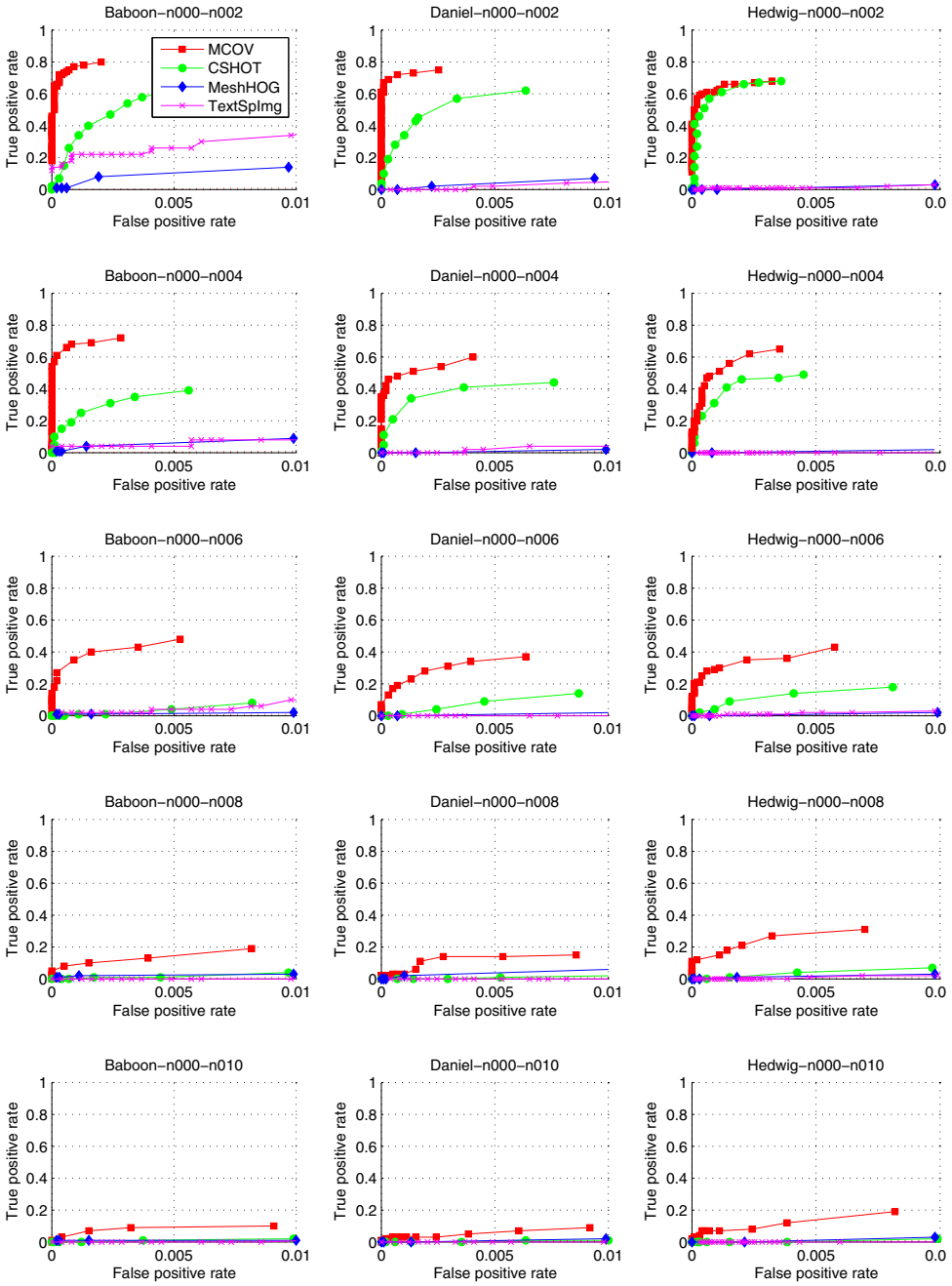


Fig. 11 ROC curves for comparison of several 3D descriptors, using the *exclusive ratio* criterion and reducing the resolution of the second scene to the 50 %. Each *column* depicts a test on a different model of the database. Each *row* shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10 % of the standard deviation of color and surface coordinates) (Color figure online)

and *Hedwig* models for an easier visualization of descriptor performances. These are plotted in Figs. 11 and 12, again taking into account the proposed *exclusive* and *inclusive* ratio matching criteria.

As we can see, both numerical and ROC curve 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. An special consideration must be taken into account in the MeshHOG method, which requires 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 (Luebke 2001). This has a drastic impact on its performance as ROC curves and AUC values suggest.

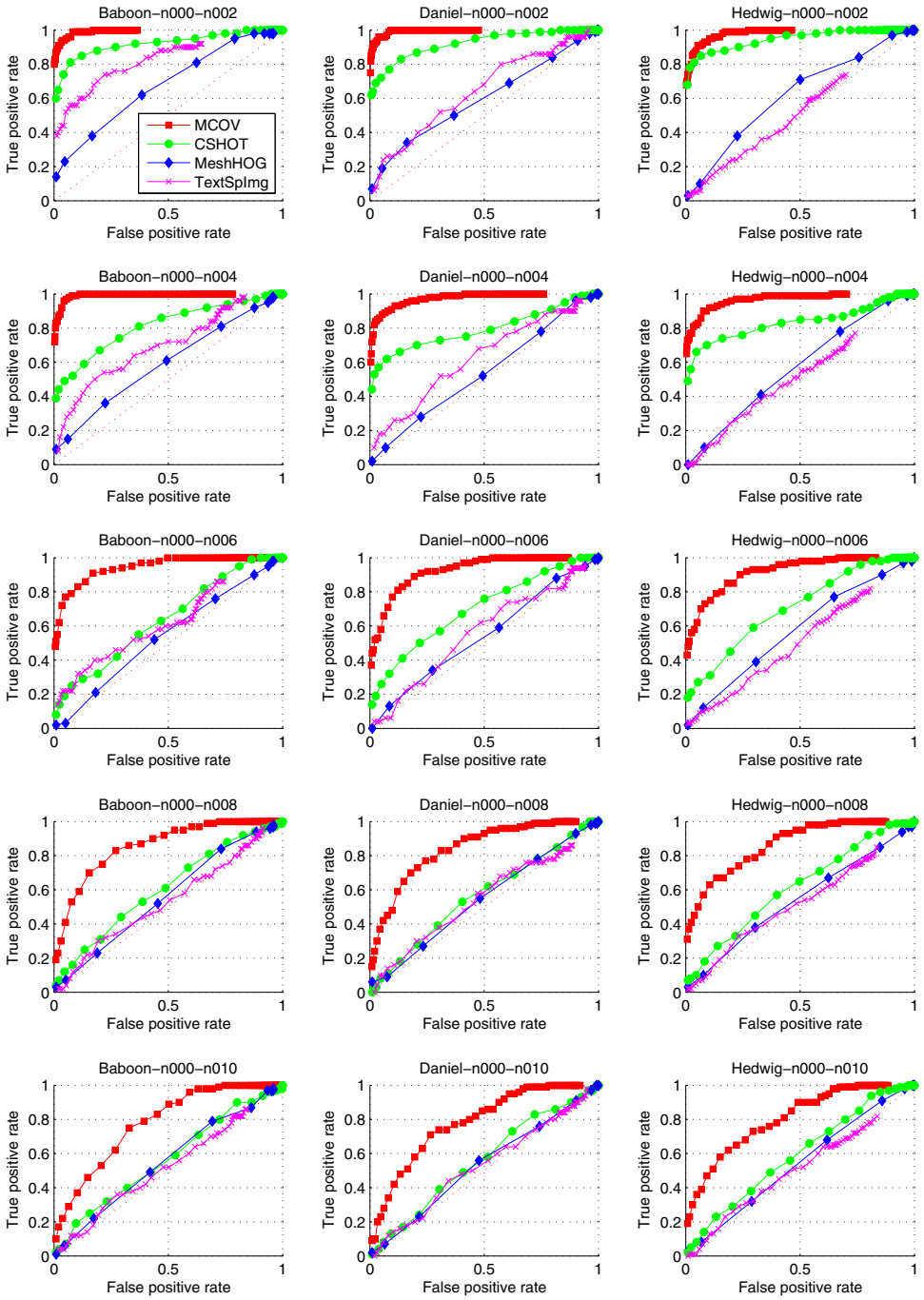


Fig. 12 ROC curves for comparison of several 3D descriptors, using the *inclusive ratio* criterion and reducing the resolution of the second scene to the 50 %. Each *column* depicts a test on a different model of the database. Each *row* shows the behaviour of the descriptor under different levels of additive noise over data (2, 4, 6, 8 and 10 % of the standard deviation of color and surface coordinates) (Color figure online)

5.1.3 Exclusive/Inclusive Ratio Matching Evaluation

In Fig. 13 we present a complementary qualitative result for visually observing the different impact of the aforementioned matching criteria on the descriptive performance of our approach. The assignation of different *ratio* values, as well as the performed criterion, affects on the number of established matches. The equivalent case between two methods takes place when *ratio* parameter is set to 1.

While the *exclusive ratio* criterion is a usual procedure found in other approaches, its application is of limited feasibility in the context of registration of arbitrarily repetitive scenes. As several challenging conditions must occur, it is better to intentionally allow a certain flexibility on point matches, in order to keep all the locally similar areas of the scene. Later on, the parts with local similarities will be filtered by the Game Theory geometric consistence methodology. The conclusion we extract from this experimental

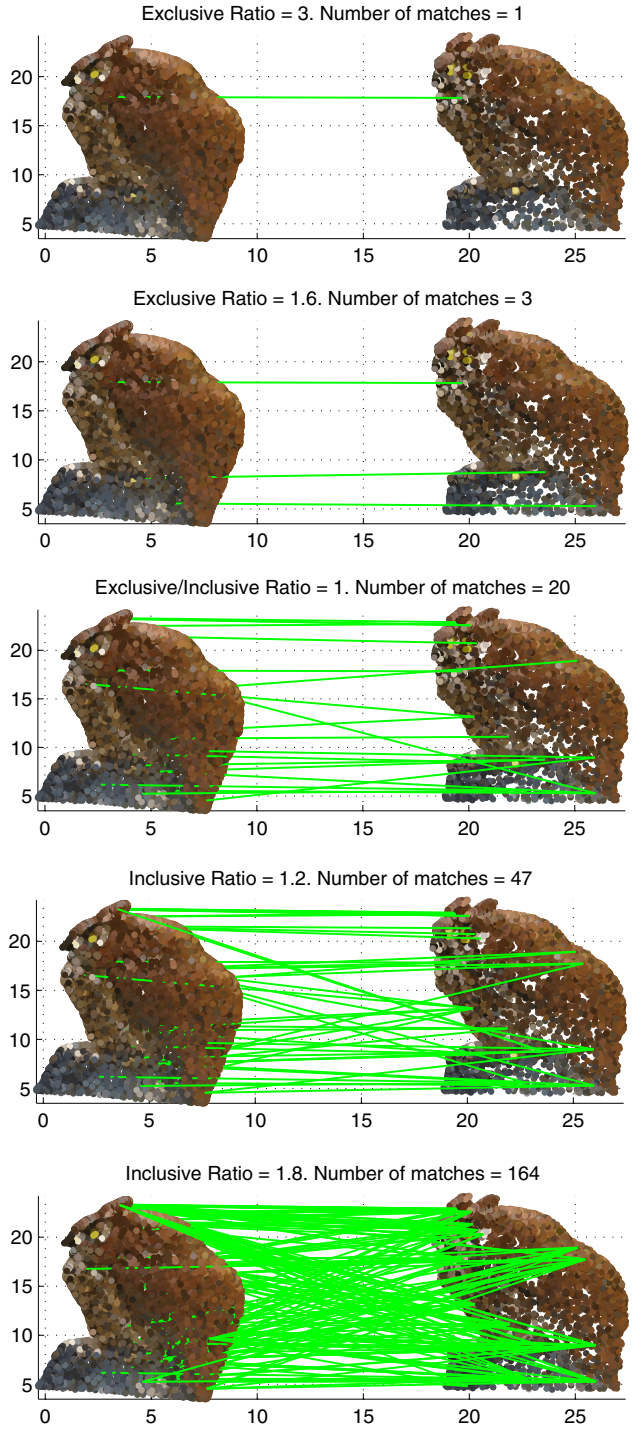


Fig. 13 Effects of the different matching criteria over the matches of *Hedwig* model, which is considered specially challenging due to homogeneous pattern areas. The test is performed using a variation of the second scene under 50 % resolution and 2 % noise. The number of keypoints has been limited to 20, as can be seen in the simulation conducted when *ratio* parameter is set to 1

set-up is that the most feasible criterion in order to perform this match candidate selection is the *inclusive ratio* criterion.

5.2 Game Theory Evolutionary Stable Strategy Solver Validation

The procedure of our Game Theoretic approach consists in the successive removal of error inducing correspondences, from the set of initial descriptor matches, until the algorithm converges to a limited set of point correspondences. These must be consistent in terms of local descriptor likelihood as well as scene-wise geometric consistency. Since the main reason for adopting this methodology in order to discard the incorrect scene correspondences is its faster convergence to a global solution, and its computational feasibility over scenes with huge amounts of outlier correspondences, we want to provide an experimental set-up which validates the performance of the proposed methodology under different deliberate amounts of outlier descriptor correspondences. For this reason the following procedure is carried out for different models and levels of noise:

- A fixed number of 500 correspondences is established between two instances of a scene. This amount of candidate matches will be intentionally corrupted with increasing levels of outlier correspondences, from 5 to 95 % regarding the total amount of candidate pairs. We consider a corruption of a candidate match as an alteration of the coordinates from the two scene views which should have been matched according to a local similarity of the descriptor likelihoods. Thus, an outlier correspondence would pose a challenge to the registration approach in the sense that it will not fulfil the geometric consistency constraints.
- Along the different iterations of the Evolutionary Stable Strategy solver algorithm (Albarelli et al. 2009), successive incoherent correspondences will be removed according to the consecutive game payoff evaluations. As we know which are the manually altered correspondences, we can evaluate the evolution of the ratio of inlier candidates regarding the remaining set. Ideally, we must validate how the tendency to keep the correct candidate pairs is monotonically increasing.
- Starting from the initial set of 500 correspondences until a minimum of 3 finally selected matches (which is the minimum amount of correspondences needed in order to estimate a rigid spatial transformation for scene registration), we can display the averaged results of this experimental set-up in the 12 models of our database as depicted in Fig. 14. We show different simulations starting with different outlier percentages within the 500 correspondences. Our intention is to visualize the evolution of the inlier ratio, which should converge to a remaining set of correspondences where this ratio is 100 %. This convergence must be reached in a monotonic increasing way as defined in Albarelli et al. (2009). Although there exist unusual cases where the inlier ratio evolution temporarily

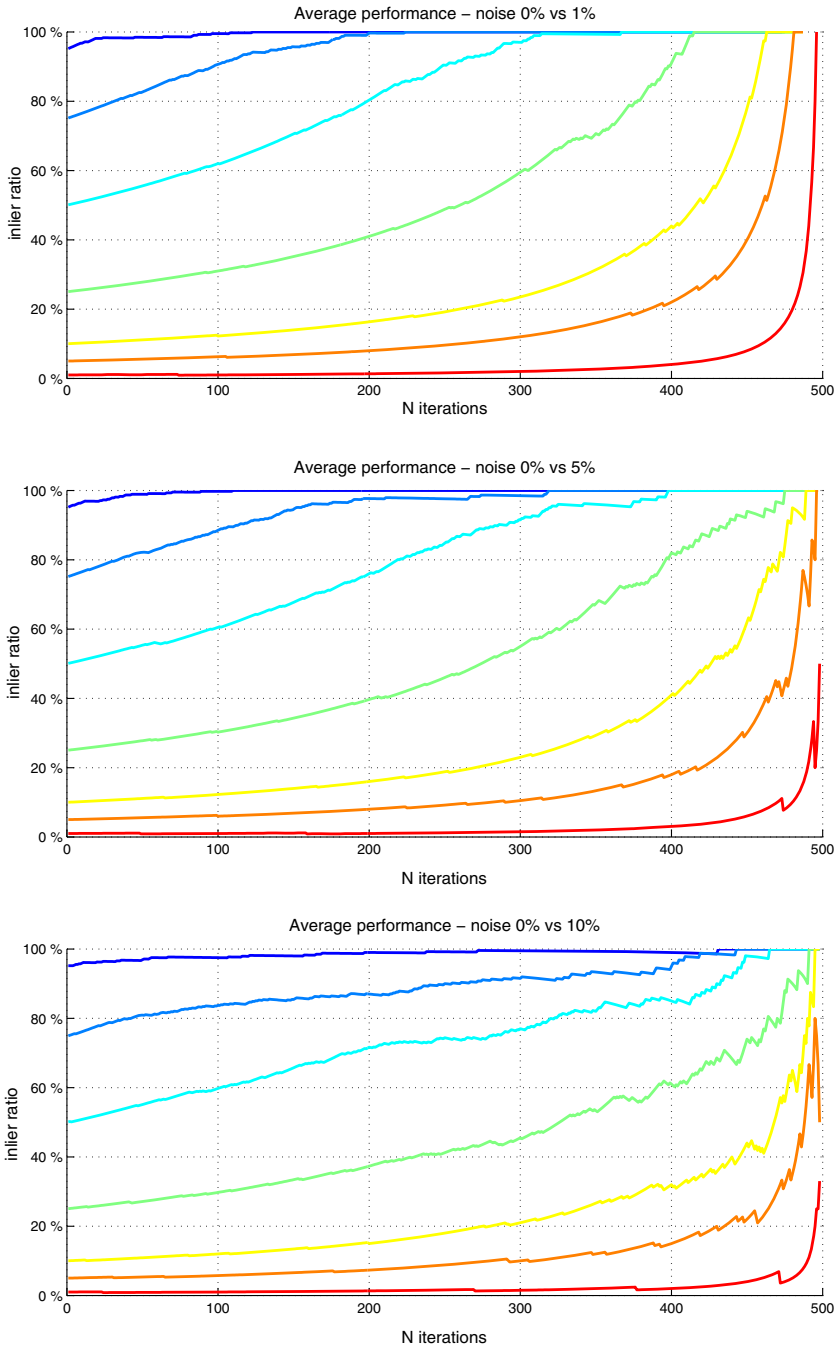


Fig. 14 Evolutionary Game Theory approach performance on the removal of correspondences outliers. Each *row* depicts the average performance on the proposed set-up for all 12 models in the database, for different levels of noise: 1, 5 and 10 % the standard deviation of each feature. Each plot displays the evolution for 7 different initial situations: 95, 75, 50, 25, 10, 5 and 1 % of inliers: along the different iterations of the Evolutionary Game Theory approach the percentage of correct correspondences is evaluated

decreases (indicating the *sacrifice* of a correct correspondence at a given iteration), the overall crescent evolution shown in the figures certifies the correct performance of our approach even in cases where we set up a very low presence of initial correct matches.

An ideal comparative experiment must contrast the performance of our Evolutionary Game Theory-based approach against any commonly used iterative RANSAC-based approach. But the difference on paradigms complicates the establishment of any comparative criterion: while the approach presented in this paper is based in a constraint-based rejection methodology, and we can evaluate the inlier ratio evolution of the remaining set of match candidates at each iteration; any RANSAC-based approach is usually based in an iterative hypothesis evaluation until a minimal error is found. At each iteration, a sub-sampling of point candidates takes place: this means that the ratio of inlier candidates will evolve in an erratic way during iterations. Furthermore, the number of iterations will vary for different

executions of the method. And as each iteration hypothesis is a spatial transformation itself, its evaluation will also depend on an error threshold parameter which will directly affect the inlier discrimination. Finally, for this same reason, noise on data will also affect the system. This will not happen on a Game Theory approach as this method will implicitly select the elements in the payoff matrix with less noise, and the solution will be consistent no matter how many different executions are done.

For all these reasons, we finally compare both approaches via the number of iterations needed, which can be a good justification baseline. The proposed Evolutionary Game Theory method, being a rejection based approach, will not surpass the number of initial candidates—500 in our current set-up—and this upper bound will be constant regardless the initial inlier ratio. On the other side, in an iterative approach the number of needed iterations can not be exactly established a priori: in this sense, Hartley and Zisserman (2003) provides a theoretic approximation for an estimation of RANSAC number of needed iterations N in order to solve a registration with a given amount of outlier elements:

$$N = \frac{\log(1-p)}{\log\left(1-(1-\epsilon)^s\right)} \tag{16}$$

where s is the sampling size which is taken at each iteration, ϵ is the probability that any match pair is an outlier, and p is the probability that the sampling set is free from outliers (usually set to 0.99). In a registration context the subsampling size for estimating a rigid transformation is set as $s = 4$. This is taken as the baseline of a common and standardized outlier removal procedure and can provide estimations like the following ones:

inlier %	75	50	25	10	5	1
N	8	34	292	**4603**	**36839**	**4605168**

For the experimental set-up proposed here, a 21 % of inliers would suppose the threshold value for which it is advantageous to choose the Evolutionary Game Theory method. As a conclusion we can see the validity of the proposed method in the context of real scenes registration, where a limited set of initial candidate matches can be provided by a descriptor matching, but the nature of the scene itself can still provide a presence of repeated areas or symmetries, resulting in outlier match candidates. The presented approach can provide an efficient solution in a robust way, regardless of challenging conditions such as a high presence of outliers or noise on data.

5.3 Global Matching Evaluation

For testing the overall performance of the descriptor in conjunction with the correspondence selection stage, we have designed an exhaustive scene registration test where each one of the twelve models has been split in halves of different common overlap (from 10 to 70 % of the surface in common). A random spatial transformation (arbitrary rotation and translation) is introduced to one of the halves. In addition, each model is tested under different levels of noise, from 0 to 10 % the standard deviation of color and surface coordinate values. For each scene, the experiment is conducted 5 different times so different halves and noise applications are considered. This leaves a total of 4620 registration executions, which have been evaluated as follows.

In order to consider a registration as correct, we evaluate its registration error measure by looking at the average Euclidean distance of groundtruth points in the common overlap surface. This is done after applying the found rotation and translation which undoes the arbitrarily applied rigid transformation. In the case of executions with applied noise, the system is solved using the modified data but the performance is evaluated on the equivalent un-noised scenes in order to be coherent on performance comparison. Object spatial coordinates are normalized so they fit within the boundaries of a prism of unitary volume, therefore the error measure is also normalized and is finally expressed as a percent ratio regarding the overall scene range. This way, results between different size scenes can be coherently compared.

We have chosen an error acceptance threshold of 0.02, which would mean that objects of one cubic meter of volume should have an average error of 2 centimeters. By establishing this threshold we can represent the execution of all registrations by a histogram of how many of them are considered as correct, for each experiment conditions of noise and overlap. See Fig. 15 for such results representation. By watching this histogram, one can conclude which is the minimum overlap between scenes for which our approach is valid at a 20 % of common surface our method is able to perform most of the scene registrations in a correct way.

In Fig. 16 we can see the distribution of error magnitudes for the aforementioned correct scene registrations. As expected, the most challenging conditions are those where the system is tested with a smaller overlap and a higher noise. Nevertheless, by watching the value distributions on these figures, we can conclude that our approach is more sensitive to the minimum overlap need rather than to the noise tolerance, which is coped by the descriptor performance—as tested in Experiment 4.1. This is easily arguable, as our system needs a minimum of heterogeneous observable areas in order to find symmetric or repeated areas.

This experiment has been conducted in an Intel Core i5 computer with 4Gb of RAM. As stated before, the imple-

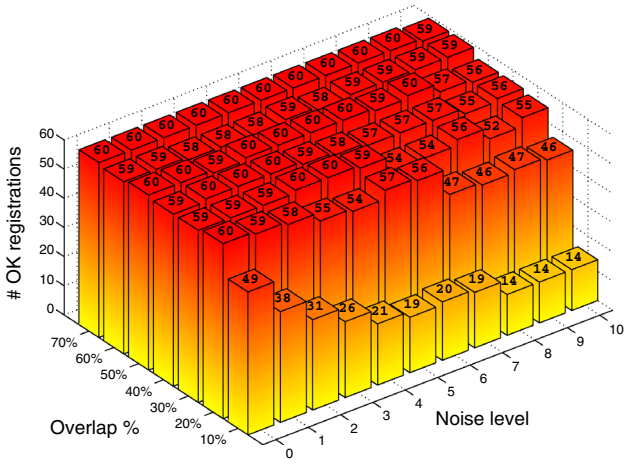


Fig. 15 Histogram of correct registrations (for an error threshold of 0.02). As we can see, the performance of our approach is rather homogeneous on most of the experimental conditions, even with *low* overlap between scenes and high levels of noise applied to data

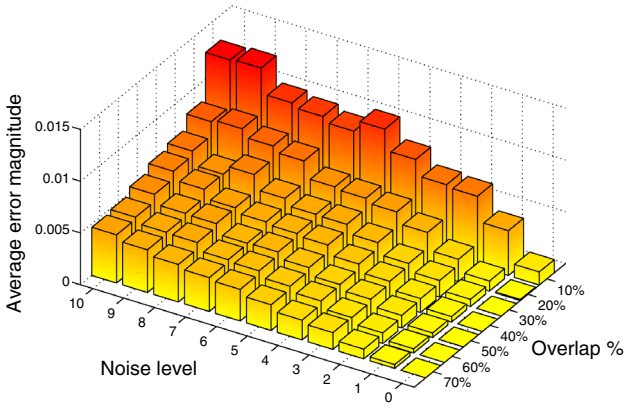


Fig. 16 Average error distribution of those registrations considered as correct. As one could expect, major errors occur on the cases of higher noise levels and less overlap

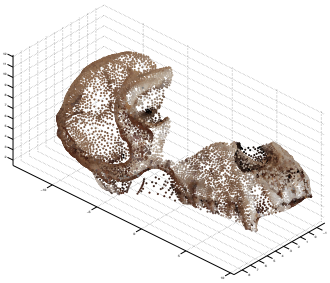
mentation of the proposed approach does not pose major computational demands, and for the models in our database which have a density ranging from 20,000 to 30,000 points the whole registration execution time takes around 140 s in a prototype, non-optimized implementation. From this time, the scene analysis stage takes an average time of 17 s; the descriptor candidate matching and payoff matrix construction an average time of 90 s; and the Evolutionary Game Theory solver algorithm an average time of 30 s. Figure 17 shows some steps involved in the registration procedure, as well as a qualitative result on one concrete model of the database.

5.4 Real-Data Matching Qualitative Valuation

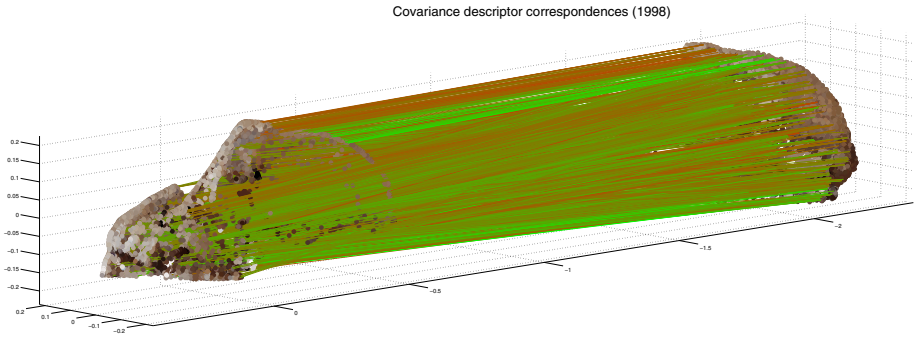
In this last experiment, we propose to test our complete approach in the context of scenes acquired with a Microsoft Kinect device , which suffer from sensor noise and artefacts along the capture of the different views. Therefore, the registration of such scenes has the drawback of not allowing a direct quantitative evaluation, as there does not exist any direct groundtruth information of correspondences between different views, and this converts this experimental set-up in a mere qualitative evaluation. Nevertheless, there are still several benefits in this framework which can help extracting conclusions about our proposal: in a first place, the usage of real data can validate our statement about the performance of the Covariance Descriptor against noise and resolution changes (in this case, caused stochastically by the acquisition sensor). In a second place, we will 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 will provide an example of broadening the scope of our approach to other areas such as scene understanding or object indexing under challenging conditions: the query objects belong to different views and can suffer small shape variations or lack of detail in certain parts. The main power of our Game-Theoretic approach is that it will act as a best subset selector of points present both in the object query and the scene, enabling a robust searching procedure.

This experiment is performed on top of the publicly available RGB-D dataset presented in Lai et al. (2011), which contains 300 objects organized into 51 categories as well as 22 different complete scenes. The main interest of using this database is that included objects and scenes suffer unstructured noise due to the own acquisition sensor, and segmented objects may have been acquired at different resolutions and static conditions with respect to the scenes. The goal is to perform a 3D object searching task: segmented objects will be used as query instances to be found within the whole scenes, where these instances will be mixed with clutter elements and altered by changes on resolution, spatial transformations or incomplete views. In this context, as we know that there will be at most one instance of each object in the scene, we justify the usage of the *exclusive ratio* criterion for initial match candidates. The rest of the methodology presented before remains unaltered, but instead of putting two scene views of a similar size into correspondence, we seek a spatial registration for matching a smaller object inside the scene. The spatially translated points from the query model regarding the whole scene will be considered as object identifying points, therefore inferring the presence of the element in the scene.

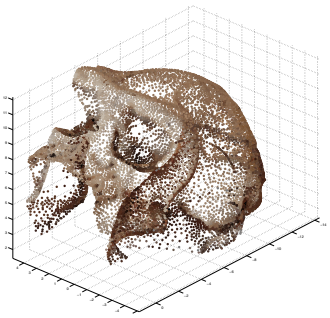
We have used different cluttered scenes and different query objects from the aforementioned dataset, with different shape and texture distributions. Qualitative results are shown in Fig. 18.



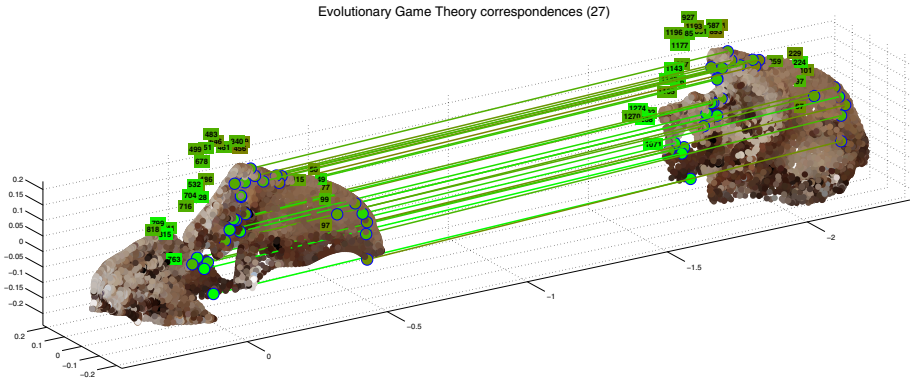
(a) First synthetic view.



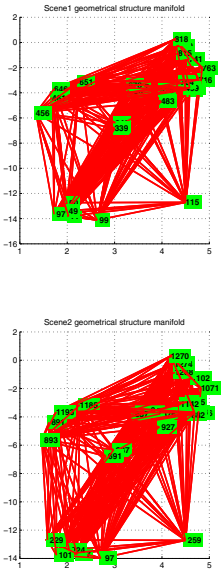
(c) Initial descriptor match candidates via inclusive ratio. The different levels of green on the correspondence lines indicate the likelihood magnitude.



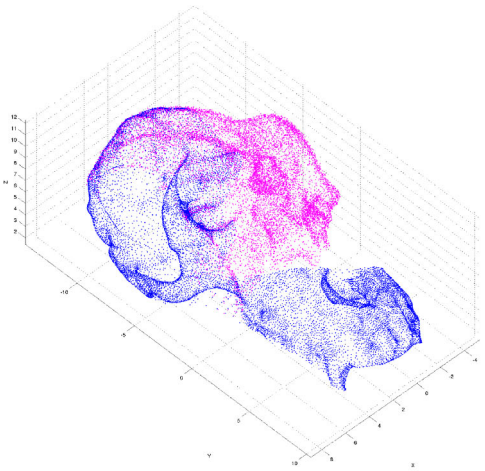
(b) Second synthetic view.



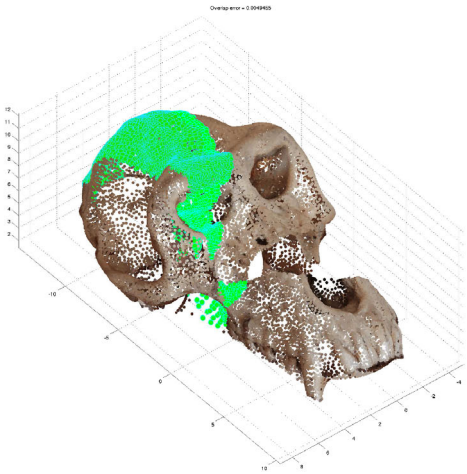
(d) Descriptor match candidates found as the result of the Evolutionary Game Theory solver. Visually, the parallellism level on the candidate lines depicts a good performance on registered point sets.



(e) Geometrically consistent structure.

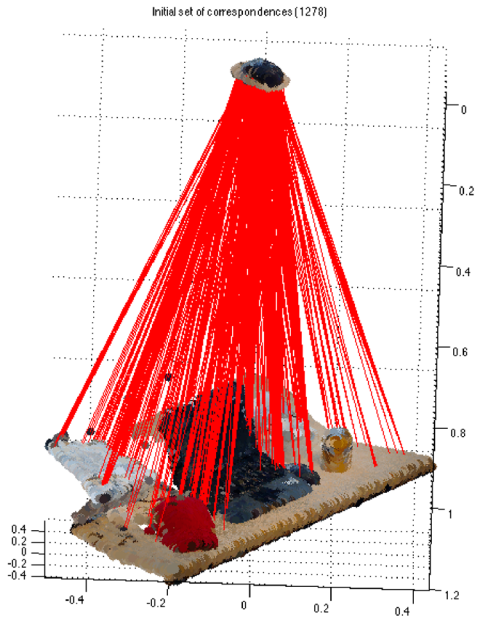


(f) Registered point sets. Each view is plotted on a different color.

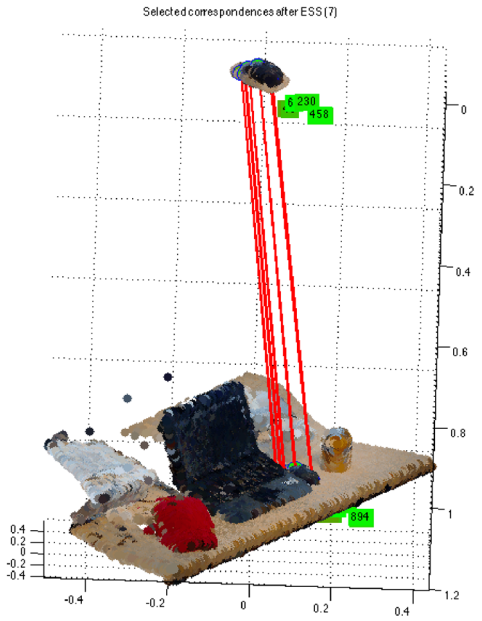


(g) Registered point sets with marked groundtruth points and initial overlap area.

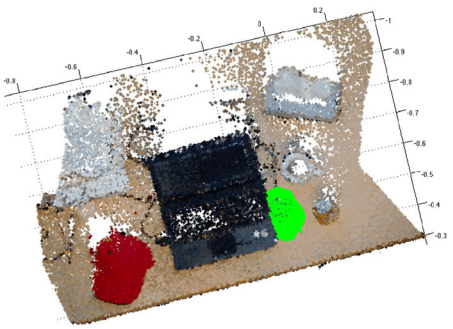
Fig. 17 Example of some steps of the whole presented registration approach on an instance of the *Baboon* model, which poses a challenge due to its homogeneity in color and shape on some areas. **a**, **b** show the synthetic views to be registered. The second view shown in **b** has been altered with an additive noise representing the 3 % of the standard deviation on each of its coordinate and color components; and suffered an arbitrary spatial rigid transformation. The overlapping area between both views represents a 20 % of the scene surface. **c**, **d** represent the amount of match candidates found by the descriptor likelihood and after the Evolutionary Game Theory solution, respectively. **e** shows a graph representation by projecting and connecting the final remaining points at each scene view in order to validate the geometric consistency constraints implicitly encoded in the Game Theory Solution. **f**, **g** show the final registration between scene views. (Color figure online)



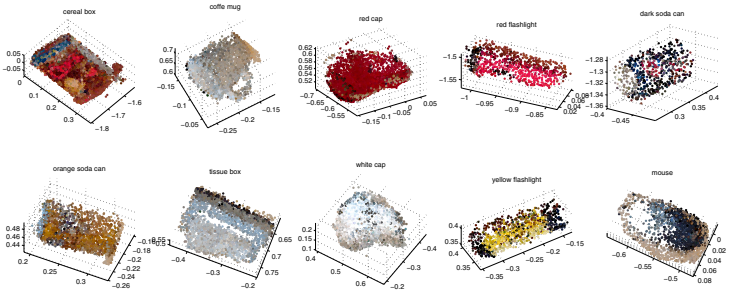
(a) Initial query search matches for instance of "mouse" object in "desk table" scene.



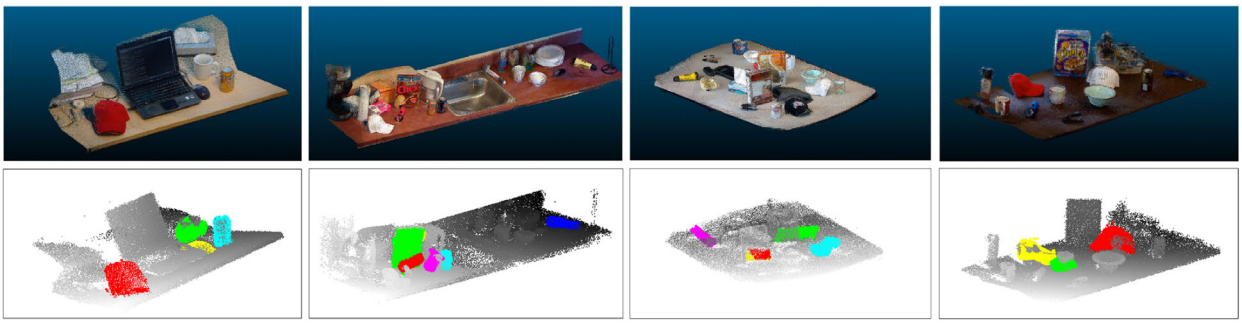
(b) Final query located points. ESS acts as the best subset selector of points of query instance inside the whole scene.



(c) Query instance is transformed according to the rigid transformation found after ESS solution. Inferred query points in the scene are marked in green.



(d) Different object queries in RGB-D dataset. Note the differences on views, resolution and nature of the objects.



(e) Examples of query search results in four cluttered scenes in RGB-D dataset.

Fig. 18 Results from the experimental setup performed on the RGB-D dataset. **a**, **b** show the different stages of our approach, where the set of descriptor candidates (many, due to *low* resolution of the query instance) are selected by the Game-Theoretic approach. Even if the task is challenging due to changes on the quality and different views of the query object, our defined game achieves the goal of selecting that subset of points in the cluttered scene which is considered to belong to the query instance. In **c**, the query instance is projected into the cluttered scene for an inference of its location. **d**, **e** show the set of different query instances available in RGB-D dataset, and some examples of the same search procedure resulting from different scenes (query inferences plotted in *solid* colours on top of grayscale scene point clouds) (Color figure online)

6 Conclusions

We have introduced a novel descriptor for fusion of 3D shape and visual information which is defined to work under spatial rigid transformations and changes in noise and scene resolution. The rather simple formulation of this descriptor has several benefits: it can be extended with additional features in the future besides texture and surface information; it can be used as a salient point selector thanks to its underlying statistical notions; and the computational cost is low as the descriptor calculation does not involve any major operation than vector products and subtractions. Its flexible, compact and statistical-based conception has been analysed as the main reason of its high representation capabilities.

There are also practical advantages in the presented approach. On one hand, MCOV only requires a parameter for radial neighbourhood, which can be set according to each scene nature in a self-contained manner. Other methods will require fine-tuning of parameters for histogram bins, connection neighbourhood, etc. affecting directly to their performance. On the other hand, the Förstner distance defined in Eq. (6) is a geometrically sensitive metric for the inner topology of MCOV, which is coherent with its theoretic geometrical topology. Other methods are based on correlations or histogram distance definitions, which might add some error drift on the likelihood computation.

Our results have been presented in conjunction with a tailored database of twelve scenes which include variant objects in order to represent challenging handicaps of repeated textures, homogeneous regions and symmetric areas. We have demonstrated how the proposed descriptor has a representative and discriminative capability which outperforms other state-of-the-art methods, specially in the case of noise over data, or density variations. The computational and performance benefits of the proposed approach suggest it is a flexible and easy descriptor with many practical applications for representation of scenes with current 3D and color sensors. Its associated keypoint detector feature can also be used on problems which require particular computational efficiency or point reliability.

We want to reflect that the different performance of the descriptor under different matching criteria is not a drawback, but the expected behaviour in this context. The choice between *exclusive* and *inclusive* criteria is a matter of knowing the task where the descriptor will be applied. For object recognition task matches, *exclusive ratio* will be suitable as it reduces the number of possible false positives and is more restrictive. For scene reconstruction problems, for instance, *inclusive ratio* will be more appropriate: in this kind of applications, area repeatabilities are a possible known handicap due to the nature of scenes (as they can contain homogeneous patterns). If a descriptor encodes the nature of an area, and this appears on some places along the scene, the repeatedly found points will be unavoidable (indeed, this asserts that the descriptor is doing what it is supposed to do). So, this reflects the needing of a posterior method which will filter false positive matches regarding a more global set of constraints, i.e. geometric consistencies. This has been hereby one of the motivations for the Evolutionary Game Theory approach introduced in this paper.

Acknowledgments This work has been supported in part by the following research Project Grants: TIN2012-39203 and IPT-2012-0630-020000 awarded by the Spanish Government Ministry of Economy and Competitivity.

References

Albarelli, A., Rodola, E., & Torsello, A. (2010). A game-theoretic approach to fine surface registration without initial motionestimation. In *International conference on computer vision and pattern recognition (CVPR)* (pp. 430–437).

Albarelli, A., Rota Bulò, S., Torsello, A., & Pelillo, M. (2009). Matching as a non-cooperative game. In *International conference on computer vision (ICCV)* (pp. 1319–1326).

Arsigny, V., Fillard, P., Pennec, X., & Ayache, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, *56*(2), 411–421.

Besl, P., & McKay, N. (1992). A method for registration of 3-d shapes. *Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *14*, 239–256.

Brusco, N., Andreetto, M., Giorgi, A., & Cortelazzo, G. M. (2005). 3d registration by textured spin-images. In *International conference on 3D digital imaging and modeling (3DIM)* (pp. 262–269).

Chen, C. S., Hung, Y. P., & Cheng, J. B. (1999). RANSAC-based darces: A new approach to fast automatic registration of partially overlapping range images. *Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *21*, 1229–1234.

Cherian, A., Sra, S., Banerjee, A., & Papanikolopoulos, N. (2011). Efficient similarity search for covariance matrices via the Jensen–Bregman logdet divergence. In *International conference on computer vision (ICCV)* (pp. 2399–2406). IEEE.

Chua, C., & Jarvis, R. (1997). Point signatures: A new representation for 3D object recognition. *IJCV*, *25*, 63–85.

Chum, O. & Matas, J. (2005): Matching with PROSAC-progressive sample consensus. In *International conference on computer vision and pattern recognition (CVPR)* (Vol. 1, pp. 220–226).

Chum, O., & Matas, J. (2008). Optimal randomized RANSAC. *Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *30*, 1472–1482.

Chum, O., Matas, J., & Obdrzalek, S. (2004). Enhancing RANSAC by generalized model optimization. In *Asian conference on computer vision (ACCV)* (Vol. 2, pp. 812–817).

Fehr, D., Cherian, A., Sivalingam, R., Nickolay, S., Morellas, V., & Papanikolopoulos, N. (2012). Compact covariance descriptors in 3d point clouds for object recognition. In *International conference on robotics and automation (ICRA)* (pp. 1793–1798).

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*, 381–395.

Flint, A., Dick, A. R., & Van Den Hengel, A. (2007). Thrift: Local 3d structure recognition. *Digital Image Computing Techniques and Applications*, *7*, 182–188.

Förstner, W. & Moonen, B. (1999). *A metric for covariance matrices. Quo vadis geodesia* (pp. 113–128).

Frome, A., Huber, D., Kolluri, R., Bülow, T., & Malik, J. (2004). Recognizing objects in range data using regional point descriptors. *ECCV*, *3023*, 224–237.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Horn, B. K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. *Optical Society of America*, *4*, 629–642.

Johnson, A. (1997). *Spin-images: A representation for 3-d surface matching.* Ph.D. thesis. Pittsburgh, PA: Robotics Institute, Carnegie Mellon University.

Johnson, A., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *21*, 433–449.

Kovnatsky, A., Bronstein, M. M., Bronstein, A. M., & Kimmel, R. (2012). Photometric heat kernel signatures. In *Scale space and variational methods in computer vision* (pp. 616–627).

Lai, K., Bo, L., Ren, X., & Fox, D. (2011). A large-scale hierarchical multi-view RGB-D object dataset. In *International conference on robotics and automation (ICRA)* (pp. 1817–1824).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, *60*, 91–110.

Luebke, D. P. (2001). A developer's survey of polygonal simplification algorithms. *Computer Graphics and Applications*, *21*, 24–35.

Rodolà, E., Albarelli, A., Bergamasco, F., & Torsello, A. (2013). A scale independent selection process for 3d object recognition in cluttered scenes. *International Journal of Computer Vision, IJCV*, *102*((1–3)), 129–145.

Rusu, R., Blodow, N., & Beetz, M. (2009). Fast point feature histograms (FPFH) for 3d registration. In *International conference on robotics and automation (ICRA)* (pp. 3212–3217).

Tombari, F., Salti, S., & Di Stefano, L. (2011). A combined texture-shape descriptor for enhanced 3d feature matching. In *International conference on image processing (ICIP)* (pp. 809–812).

Tombari, F., Salti, S., & Stefano, L. (2010). Unique signatures of histograms for local surface description. In *European conference on computer vision (ECCV)* (Vol. 6313, pp. 356–369).

Torsello, A., Rodolà, E., & Albarelli, A. (2011). Sampling relevant points for surface registration. In *Conference on 3D imaging, modeling, processing, visualization and transmission (3DIMPVT)* (pp. 290–295).

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *European conference on computer vision (ECCV)* (pp. 589–600).

Tuzel, O., Porikli, F., & Meer, P. (2007). Human detection via classification on Riemannian manifolds. *International conference on computer vision and pattern recognition (CVPR)* (pp. 1–8).

Tuzel, O., Porikli, F., & Meer, P. (2008). Pedestrian detection via classification on Riemannian manifolds. *Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *30*, 1713–1727.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*, 471–494.

Yao, J., Odobez, J., et al. (2008). Fast human detection from videos using covariance features. In *International workshop on visual surveillance, VS2008*.

Zaharescu, A., Boyer, E., & Horaud, R. (2012). Keypoints and local descriptors of scalar functions on 2d manifolds. *International Journal of Computer Vision, IJCV*, *100*, 78–98.

Zaharescu, A., Boyer, E., Varanasi, K., & Horaud, R. (2009). Surface feature detection and description with applications to mesh matching. In *CVPR* (pp. 373–380).