A Scale Independent Selection Process for 3D Object Recognition in Cluttered Scenes

**Emanuele Rodolà · Andrea Albarelli ·
Filippo Bergamasco · Andrea Torsello**

Abstract During the last years a wide range of algorithms and devices have been made available to easily acquire range images. The increasing abundance of depth data boosts the need for reliable and unsupervised analysis techniques, spanning from part registration to automated segmentation. In this context, we focus on the recognition of known objects in cluttered and incomplete 3D scans. Locating and fitting a model to a scene are very important tasks in many scenarios such as industrial inspection, scene understanding, medical imaging and even gaming. For this reason, these problems have been addressed extensively in the literature. Several of the proposed methods adopt local descriptor-based approaches, while a number of hurdles still hinder the use of global techniques. In this paper we offer a different perspective on the topic: We adopt an evolutionary selection algorithm that seeks global agreement among surface points, while operating at a local level. The approach effectively extends the scope of local descriptors by actively selecting correspondences that satisfy global consistency constraints, allowing us to attack a more challenging scenario where model and scene have different, unknown scales. This leads to a novel and very effective pipeline for 3D object recognition, which is validated with an extensive set of experiments and comparisons with recent techniques at the state of the art.

1 Introduction

In the recent past, the acquisition of 3D data was only viable for research labs or professionals that could afford to invest in expensive and difficult to handle high-end hardware. However, due to both technological advances and increased market demand, this scenario has been altered significantly: Semi-professional range scanners can be found at the same price level of a standard workstation, widely available software stacks can be used to obtain reasonable results even with cheap webcams, and, finally, range imaging capabilities have been introduced even in very low-end devices such as game controllers. Given this trend, it is safe to forecast that range scans will be so easy to acquire that they will complement or even replace traditional intensity-based imaging in many computer vision applications. The added benefit of depth information can indeed enhance the reliability of most inspection and recognition tasks, as well as providing robust cues for scene understanding or pose estimation. Many of these activities include fitting a known model to a scene as a fundamental step. For instance, a setup for in-line quality control within a production line could raise the need to locate the manufactured objects that are meant to be measured (Newman and Jain 1995). Range-based SLAM systems can exploit the position of known 3D reference objects to achieve a more precise and robust robot localization (Borrmann et al. 2008). Finally, non-rigid fitting may be used to recognize hand or whole-body gestures in next

E. Rodolà (✉) · A. Albarelli · F. Bergamasco · A. Torsello
Dipartimento di Scienze Ambientali, Informatica e Statistica,
Università Ca' Foscari Venezia, Venice, Italy
e-mail: rodola@dsi.unive.it

A. Albarelli
e-mail: albarell@unive.it

F. Bergamasco
e-mail: bergamasco@dsi.unive.it

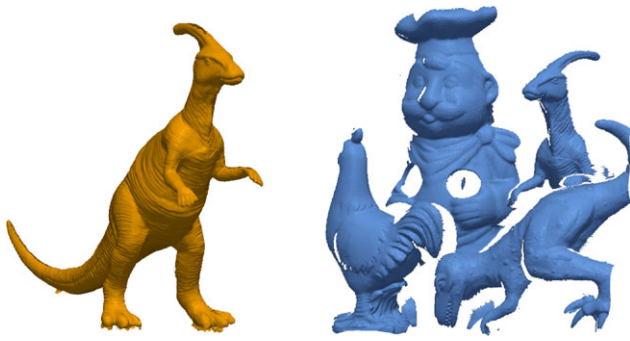A. Torsello
e-mail: torsello@dais.unive.it

Fig. 1 A typical 3D object recognition scenario. Clutter of the scene and occlusion due to geometry of the ranging sensor seriously hinder the ability of both global and feature-based techniques to spot the model

generation interactive games or novel man-machine interfaces (Ahn et al. 2009).

The matching problem in 3D scenes shares many aspects with object recognition and location in 2D images: The common goal is to find the relation between a model and its transformed instance (if any) in the scene. In the 3D case, however, scenes can undergo a variety of non-rigid deformations such as variations in local scale, variation in the topology of the observed mesh, and even global affine deformations or warping effects due, for instance, to miscalibration of the scanning device or to the action of natural forces on the objects in rather specific scenarios (Ghosh et al. 2010). While in general severe deformations of the scene are unlikely to occur, they are commonly present in a measure and should be accounted for in matching applications.

Among the basic approaches to object recognition are feature-based techniques, which adopt descriptors that are associated to single points respectively on the image (in the 2D case) or on the object surface. In principle, each feature can be matched individually by comparing the descriptors, which of course decouples the effect of partial occlusion. In the 2D domain, intensity based descriptors such as SIFT (Lowe 1999) have proven to be very distinctive and capable to perform very well even with naive matching methods. However, the problem of balancing local and global robustness is more binding with texture-less 3D scenes than with images, as no natural scalar field is available on surfaces and thus feature descriptors tend to be less distinctive. In practice, global or semi-global inlier selection techniques are often used to avoid wrong correspondences. This, while making the whole process more robust to a moderate number of outliers, can introduce additional weaknesses. For instance, if a RANSAC-like inlier selection is applied, occlusion coupled with the presence of clutter (i.e., unrelated objects in the scene) can easily lower the probability for the process to find the correct match (see Fig. 1). The limited distinctiveness of surface features can be tackled by introducing scalar quantities computed over the local surface area. This

is the case, for instance, with values such as mean curvature, Gaussian curvature or shape index and curvedness (Akagündüz et al. 2009). Unfortunately, this kind of characterization has proven to be not very selective for matching purposes, since it is frequent to obtain similar values in many different objects and locations.

To overcome these limitations, several approaches try and characterize the whole neighborhood of each point. Such methods can be roughly classified in approaches that define a full reference frame for each point, and techniques that only need a reference axis (usually the normal direction). When a full reference frame is available it is possible to build very discriminative descriptors (Chua and Jarvis 1997; Sun et al. 2003), which, however, are usually not robust with respect to noise and small differences in the meshes. By converse, methods that just require a reference axis trade some descriptiveness to gain greater robustness. These latter techniques almost invariably build histograms based on some properties of points falling in a cylindrical volume centered and aligned to the reference axis. The most popular histogram-based approach is certainly Spin Images (Johnson and Hebert 1999), but many others have been proposed in literature (Chen and Bhanu 2007; Frome et al. 2004; Pottmann et al. 2009).

Lately, an approach that aims to retain the advantages of both full reference frames and histograms has been introduced in Tombari et al. (2010). The authors take the hint from SIFT descriptors in the 2D domain, to build a 3D descriptor that encodes histograms of basic differential entities, which are further enhanced by introducing geometric information of the points within the given support and resulting in a descriptor that is both very robust and descriptive.

Any of the interest point descriptors above can be used to find correspondences between a model and a 3D scene that possibly contains it. Most of the cited papers, in addition to introducing the descriptor itself, propose some matching technique. These span from very direct approaches, such as associating each point in the model with the point in the scene having the most similar descriptor, to more advanced techniques such as customized flavors of PROSAC (Chum and Matas 2005) and specialized keypoint matchers that exploit locally fitted surfaces for computing depth values to use as feature components (Mian et al. 2010).

Recent contributions providing both a descriptor and a matching technique for the specific problem of object recognition in clutter include the works by Bariya and Nishino (2010) and Novatnack and Nishino (2008), where recognition is performed via an interpretation tree whose nodes represent correspondences between a model feature and scene feature. In the matching step, hypotheses are effectively culled based on the scale of the corresponding features, resulting in an effective approach providing good recognition

under moderate clutter and occlusion. In Mian et al. (2006), the authors presented an object recognition and segmentation algorithm based on a tensorial representation for descriptors, where pairs of vertices are randomly selected from the model to construct a tensor, then matched with the tensors of the scene by casting votes using a 4D hash table.

Object recognition in the 3D domain has several similarities with surface registration. In this case, two surfaces representing different points of view (with unknown positions) of the same object are to be rigidly aligned one to another. While there may be apparently many aspects in common with this class of problems, adopting the same techniques to solve both can be far from beneficial. Most surface alignment methods like RANSAC-based DARCES (Chen et al. 1999) or 4-points Congruent Sets (Aiger et al. 2008) (currently at the state-of-the-art for surface registration) first generate pseudo-random, not necessarily point-based matching hypotheses and then validate the match in an attempt to maximize the overall surface overlap. It is clear that in an object recognition scenario, where occlusion and clutter are present and where the object itself cannot even be assumed to be in the scene, such approaches can give completely wrong results, fooled by the structured noise offered by the clutter. Thus, even though in a technical sense similar methods can be adopted for both surface registration and object recognition, the assumptions underlying the two problems, as well as the expected results, are very different. Recently, a novel surface alignment approach was presented in Albarelli et al. (2010a). The approach adopted a natural selection process derived from game-theoretic considerations to drive the selection of sets of corresponding points satisfying a global rigidity constraint. The method selects correspondences from initial guesses based on point descriptors, but was shown to work well even with very loose descriptors that provide a lot of false positives (Albarelli et al. 2010b). While the approach copes well with large false positives in the initial correspondences caused by bad descriptors, as well as large false negatives due to occlusion, it cannot be adapted as is to the recognition setup due to the presence of structured outliers caused by the clutter, the change of scale breaking the isometry assumptions at the basis of the approach, and the inability to deal with the absence of the target. The latter problem was addressed in Albarelli et al. (2011) by adopting a directional mapping, resulting in a robust recognition approach. However, the recognition process was still based on the enforcement of a global isometric transformation, rendering the recognition sensible to changes in scale or small mesh deformations. One of the main characteristics of the game-theoretic framework is its adaptability to different contexts. In fact, it is a general selection approach that confines the domain knowledge to the definition of a *payoff* function that, intuitively, measures how well one hypothesis is supported by

another with respect to the final goal. The competition, then, induces a selection process in which incompatible matches are driven to extinction whereas a set of sparse, yet very reliable correspondences survive. For instance, in Albarelli et al. (2012) a similar approach was used in the context of multiview stereo, where the dynamical process drove the selection towards sets of image points that collectively satisfied the epipolar constraint.

In this paper we introduce a novel feature-based 3D object recognition pipeline crafted to deal in a robust manner with strong occlusion and clutter using the game-theoretic framework to drive selection. While the use of game theory in matching problems has already been explored in previous work, this paper presents at least two novel contributions. First, it introduces a novel pipeline tuned to the specifics of object recognition and which outperforms the state-of-the-art for 3D object recognition in clutter. The approach is based on a recent local surface descriptor to find a set of matching candidates and on the adaptation of the described game-theoretic framework to drive the selection of corresponding features on model and scene. Acting as distinctive priors, the introduced descriptors allow to reduce the problem size and to gain in robustness, while the matching candidates are then let to compete in a non-cooperative game. The resulting pipeline allows the game-theoretic selection process to cope with the large number of distractors due to clutter and with the possible lack of the object from the scene. Secondly, it introduces scale invariance by the adoption of a multi-scale approach in the creation of matching hypotheses and the use of a novel descriptor to drive the selection. With this approach, the change in scale is accounted for by considering geometric information along the paths connecting pairs of points. Specific scale-invariant descriptors are not needed. Rather, we compute scale-dependent descriptors at different scales and then let the selection process extract the correct matches from the generated pool of multi-scale hypotheses. With this approach the payoff function is not enforcing global geometric constraints as in Albarelli et al. (2010a, 2011), but rather only on characteristics of the descriptor, resulting in a selection process that is scale-invariant and generally more robust to changes in the meshes. This new pipeline is tested in a wide range of experiments and is shown to outperform the state-of-the-art for 3D object recognition in clutter.

2 A Game-Theoretic Pipeline for Recognition

Originated in the early 40's from the seminal work of von Neumann and Morgenstern (1953), Game Theory was an attempt to formalize a system characterized by the actions of entities with competing objectives. In this setting, multiple players have at their disposal a set of strategies and their goal
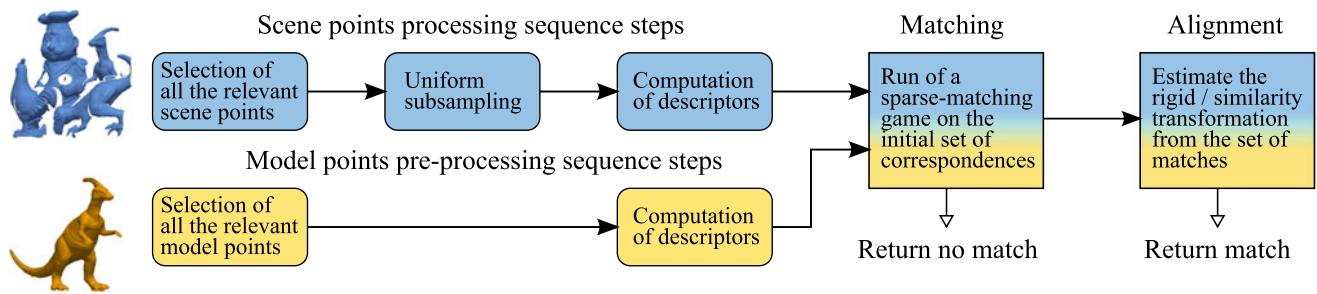
Fig. 2 An overview of the object recognition pipeline presented in this paper

is to maximize a payoff that depends also on the strategies adopted by other players.

Let $O = \{1, \ldots, n\}$ be the set of available strategies (*pure strategies* in the language of game theory), and $\Pi = (\pi_{ij})$ be a matrix specifying the payoff that an individual playing strategy i receives against someone playing strategy j. A *mixed strategy* is a probability distribution $\mathbf{x} = (x_1, \ldots, x_n)^T$ over the available strategies O, and is constrained to lie in the n-dimensional standard simplex

$$\Delta = \left\{ \mathbf{x} \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i \in 1, \ldots, n, \sum_{i=1}^{n} x_i = 1 \right\}.$$

The expected payoff received by a player choosing element i when playing against a player adopting a mixed strategy $\mathbf{x}$ is $(\Pi\mathbf{x})_i = \sum_j \pi_{ij} x_j$, hence the expected payoff received by adopting the mixed strategy $\mathbf{y}$ against $\mathbf{x}$ is $\mathbf{y}^T \Pi \mathbf{x}$. The *best replies* against mixed strategy $\mathbf{x}$ is the set of mixed strategies

$$\beta(\mathbf{x}) = \left\{ \mathbf{y} \mid \mathbf{y}^T \Pi \mathbf{x} = \max_{\mathbf{z}} (\mathbf{z}^T \Pi \mathbf{x}) \right\}.$$

The *support* of a mixed strategy $\mathbf{x} \in \Delta$, denoted by $\sigma(\mathbf{x})$, is defined as the set of elements chosen with non-zero probability: $\sigma(\mathbf{x}) = \{i \in O \mid x_i > 0\}$. A strategy $\mathbf{x}$ is said to be a *Nash equilibrium* if it is the best reply to itself, i.e., $\forall \mathbf{y} \in \Delta, \mathbf{x}^T \Pi \mathbf{x} \geq \mathbf{y}^T \Pi \mathbf{x}$. This implies that $\forall i \in \sigma(\mathbf{x})$ we have $(\Pi \mathbf{x})_i = \mathbf{x}^T \Pi \mathbf{x}$; that is, the payoff of every strategy in the support of $\mathbf{x}$ is constant.

Evolutionary game theory (Weibull 1995) originated in the early 70's as an attempt to apply the principles and tools of game theory to biological contexts. In contrast to traditional game-theoretic models, players are not supposed to behave rationally, but rather they act according to a preprogrammed behavior, or mixed strategy, and are subject to some selection process that favors players that receive high payoffs, thus acting over time on the distribution of behaviors. In this context, the competition is formalized by assuming that individuals are repeatedly drawn at random from a large population to play a game, and the distribution of the population is altered through a dynamic process that depends on the payoffs received by the players.

A strategy $\mathbf{x}$ is said to be an *evolutionary stable strategy* (ESS) if it is a Nash equilibrium and

$$\forall \mathbf{y} \in \Delta \mathbf{x}^T \Pi \mathbf{x} = \mathbf{y}^T \Pi \mathbf{x} \Rightarrow \mathbf{x}^T \Pi \mathbf{y} > \mathbf{y}^T \Pi \mathbf{y}. \tag{1}$$

This condition guarantees that any deviation from the stable strategies does not pay.

Under very general conditions, any dynamics formalizing the selection process in a way that respects the payoffs is guaranteed to converge to Nash equilibria (Weibull 1995) and (hopefully) to ESS's. More precisely, if the process allows the population of players adopting strategies with higher than average payoff to grow, while reduces the share of population adopting worse than average strategy, then the process starting from any state in the interior of the simplex converges to a Nash equilibrium. Moreover, under these conditions, evolutionary stable strategies are hyperbolic attractors for the dynamics.

With this setup, the search for a stable state is performed by simulating the evolution of a natural selection process; further, the actual choice of the process is not crucial and can be driven mostly by considerations of efficiency and simplicity.

Interestingly, in the special case in which the payoff matrix Π is symmetric, there is a relationship with optimization theory (Weibull 1995): Stable states correspond to the strict local maximizers of the average payoff $\mathbf{x}^T \Pi \mathbf{x}$ over Δ, whereas all critical points are related to Nash equilibria. In this context, the simplex constraint is related to the well known L_1 regularizer, imposing similar sparsity conditions, thus motivating its use in an inlier selection process where we do not need to find all compatible candidates, but only a few very good candidates are sufficient.

Following Albarelli et al. (2011), we base our recognition framework on a game-theoretic inlier selection approach. In this framework, the intrinsic matching problem is better interpreted as a problem of selecting a small group of correspondences that are highly coherent according to some notion of compatibility. In this scenario, pairs of players are repeatedly extracted from a (ideally infinite) population to play a symmetric game in which the strategies correspond to the available correspondences. As the game is repeated,

players will adapt their behavior pattern to prefer strategies that yield larger payoffs, thus driving inconsistent hypotheses to extinction.

The complete pipeline proposed in this paper consists in a preprocessing step and two non-cooperative games (see Fig. 2). The preprocessing is performed both on the model and on the scene. This step involves an initial selection of relevant points on the respective surfaces. The relevance criteria are explained in the next section, however, in this context the general meaning of the culling is to avoid surface patches that are not significant for matching, such as flat areas. All the interest points on the model are kept, while those on the scene are uniformly subsampled. This makes sense for many reasons. In many applications the set of models does not change in time, and thus descriptors must be computed just once. In addition, as explained in the following sections, the direction of the matching is from the scene to the model and having less source than target points allows the game to proceed faster without compromising accuracy. Finally, the model tends to be measured with greater accuracy (either because more time can be spent on it or because it comes from a CAD model). A descriptor is computed for all the retained points, and these are used to build the initial candidates that, in turn, are fed to a *matching game*. In general, a matching game (Albarelli et al. 2010a) can be built by defining just four basic entities: a set of model points M, a set of data points D, a set of candidate correspondences $S \subseteq D \times M$ and a pairwise compatibility function between them $\Pi : S \times S \rightarrow \mathbb{R}_+$. The goal of the gameplay is to operate a (natural) selection among the elements in the initial set S. This happens by setting up a non-cooperative game where the set S represents the available strategies and Π the payoffs between them. In this game, a real-valued vector $\mathbf{x} = (x_1, \ldots, x_{|S|})^T$ that lies in the $|S|$-dimensional standard simplex represents the amount of population that plays each strategy i at a given time. The game starts by setting the initial population around the barycenter (to be fair with respect to each strategy). Then, the population can be evolved at discrete steps by applying the replicator dynamics equation:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) \frac{(\Pi \mathbf{x}(t))_i}{\mathbf{x}(t)^T \Pi \mathbf{x}(t)} \qquad (2)$$

where Π is a matrix that assigns to row i and column j the payoff (compatibility) between strategies (correspondences) i and j. As stated before, such dynamics converge to a Nash equilibrium. In addition, the values of the elements of $\mathbf{x}$ are proportional to the degree of compatibility of each strategy with the equilibrium (Albarelli et al. 2010a). In practice, a much faster convergence to the equilibrium can be obtained by replacing the iteration in Eq. (2) with the adaptive exponential replicator dynamics introduced by Pelillo and Torsello (2006) or by adopting the infection-immunization dynamics recently introduced by Rota Bulò and Bomze (2011). Performance considerations regarding the choice of

the dynamics will be made in Sect. 3.4. Since we defined the payoff as the compatibility between candidates, these are all desirable properties from a selection standpoint. In our context, M and D always correspond to the retained model and scene points, while S and Π will be defined differently for the isometric and scale-invariant matching games. Specifically, for the first game the construction of S will be driven by the similarity of descriptors computed at a fixed scale, whereas in the second case the set of strategies includes match hypotheses at many different scales. Likewise, the payoff Π will reflect the different notions of compatibility.

2.1 Feature Detection and Description

For both efficiency and robustness reasons, the proposed matching technique works on a subset of model and scene vertices. Interest point selection is performed by computing for each point a single-component *Integral Hash* (Albarelli et al. 2010b) at a given support scale σ, and retaining only those samples that obtain a value less than a fixed threshold. Being designed as a simple approximation to the integral invariant (Pottmann et al. 2009), this step is very fast and is roughly equivalent to extracting points that belong to concave surface areas, where the measure of concavity is proportional to the absolute value of the Integral Hash at that point. Keeping only small values means, in practice, that we are avoiding flat and convex areas which, empirically, we have seen to be less distinctive in a large variety of cases. By modulating the value of σ, a more or less inclusive sample selection can be carried out (see Fig. 3). All the relevant points extracted from the model surface are kept. By contrast, uniform subsampling is optionally performed on the set of relevant points in the scene. Although more sophisticated detection algorithms could be used for this step (see Mian et al. 2010, or Salti et al. 2011 for a recent survey), we favored efficiency over repeatability since the game-theoretic selection mechanism is very effective at eliminating wrong guesses. Finally, a descriptor vector is computed for each vertex. To this extent, any of the descriptors discussed in the introduction may be used; however, after an initial round of experiments, the SHOT descriptor (Tombari et al. 2010) was chosen as it obtains the best performance overall. Again, these steps are not strictly necessary, but introducing such priors proves to be beneficial both for reducing the problem size (which is proportional to the cardinality of the set of matching strategies) and in terms of inlier ratio, which increases with rejection of unlikely hypotheses. In this regard we emphasize that our method acts as an inlier selector whereas no ex-post verification is performed to validate the matches (Albarelli et al. 2012), and that this inlier selection behavior is put under considerable strain in the specific case of object-in-clutter scenarios, where strong groups of structured outliers can divert the selection process
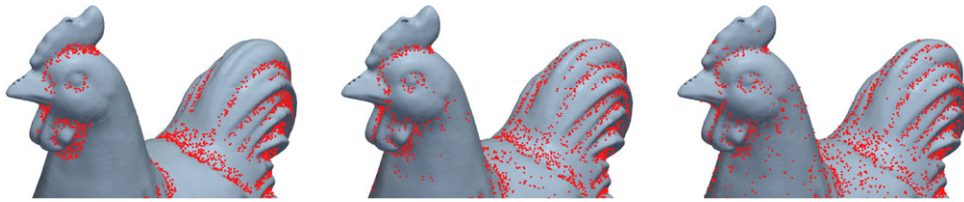
Fig. 3 In order to avoid mismatches and reduce the convergence time it is important to use only relevant points. Model vertices selected with a σ respectively equal to 8, 5 and 2 times the median model edge are shown from left to right

towards the wrong solution. We also note that existing techniques usually tend to forge ad-hoc matching methods for the specific descriptors they propose (Johnson and Hebert 1999; Mian et al. 2006, 2010; Novatnack and Nishino 2008; Bariya and Nishino 2010), while our method is general in this respect. In the experimental section we investigate both the influence of the relevant point selection and of the descriptor adopted.

2.2 Sparse Matching Game

In this section we present a matching game for a 3D object recognition scenario. We assume that relevant points were previously extracted from model and scene, and that every point of interest has a descriptor vector associated to it. We take a correspondence-based approach in that a match, if present, is established by means of point-wise correspondences between the two surfaces. This matching process is similar to the surface registration technique presented in Albarelli et al. (2010a). However, both the scope of the methods and their underlying assumptions are quite different; in fact, preliminary experiments demonstrated the inability of the "pure" surface registration algorithm to deal with the strong structured outliers due to clutter, strong occlusions and possible absence of the object from the scene, which are characteristic of the object recognition scenario.

We start by defining the initial set of strategies S, where each reference point in the scene is associated with the k-nearest model points in the descriptor space:

$$S = \big\{ (a, b) \in D \times M \mid b \in dn_k(a) \big\}, \tag{3}$$

where $dn_k(a)$ is the set of model vertices associated to the k-neighbors of the descriptor at a. This means that each (relevant) sample in the scene is considered to be a possible match with samples in the model that exhibit similar surface characteristics, and we limit the number of "attempts" to k. If the closest model descriptor is deemed too far apart from the data query, the corresponding scene point can be excluded altogether from the matching, so as to operate a form of clutter pre-filtering (although in our experiments we did not perform any filtering of this kind). If the chosen descriptor allows it, using fast search structures such as kd-tree can be beneficial for this step. Note that the direction of the

matching is from scene to model; this is motivated by the fact that the scene likely contains only a partial view of the model object, and that originating candidate matches from the scene helps to reduce the false positive rate for equal number of strategies.

Next we define a pairwise compatibility function among the strategies (in the following we refer to this function as *payoff function*). Since we are interested in finding a correspondence between the model and part of the surface in the scene, we are looking for a subset of candidates that enforce an isometric transformation among the two sets of vertices. Even though we discard connectivity information at this point, we argue that strategies enforcing this isometry constraint are likely to lay on the same surface both in the scene and in the model, and thus to be a viable solution. We define the payoff function $\delta : S \times S \to [0, 1]$ as

$$\delta\big((a_1, b_1), (a_2, b_2)\big) = \frac{\min(\|a_1 - a_2\|, \|b_1 - b_2\|)}{\max(\|a_1 - a_2\|, \|b_1 - b_2\|)}. \tag{4}$$

This function takes pairs of strategies $(a_1, b_1), (a_2, b_2) \in D \times M$ and gives a reward (a value close to 1) if the corresponding source and destination points are separated by the same Euclidean distance up to positional noise. By contrast, the value of δ will be small when the two strategies exhibit very different distances. This kind of check will succeed with correct pairs and will give false positives only for a small amount of cases, those preserving the isometry constraint by chance. However, since our game is seeking a large group of candidates with large mutual payoff, such outliers will be filtered out with high probability by the other strategies that participate to the Nash equilibrium. This makes for a semi-local approach that guarantees a robust global agreement among mating strategies, while operating at a local level.

We note, nevertheless, that Eq. (4) does not guarantee injectivity of the solution, while in this scenario we do not expect any point in the scene to correspond to more than one point in the model. To avoid possible many-to-many matches, we impose a hard constraint by setting to 0 the compatibility between candidates that share the same source or destination vertex (Albarelli et al. 2009). Additionally, we require that the variation in orientation between each pair of data points be maintained on the model. In order to obtain
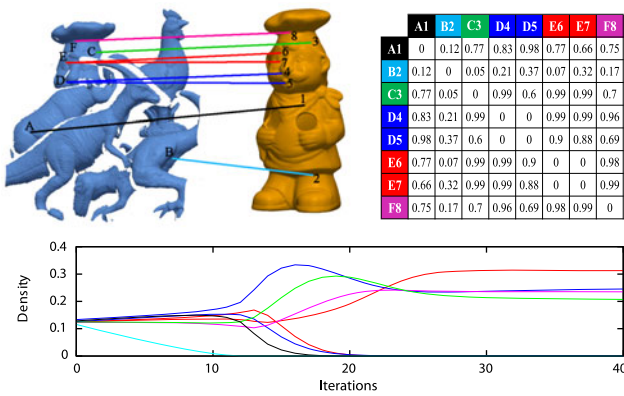
	A1	B2	C3	D4	D5	E6	E7	F8
A1	0	0.12	0.77	0.83	0.98	0.77	0.66	0.75
B2	0.12	0	0.05	0.21	0.37	0.07	0.32	0.17
C3	0.77	0.05	0	0.99	0.6	0.99	0.99	0.7
D4	0.83	0.21	0.99	0	0	0.99	0.99	0.96
D5	0.98	0.37	0.6	0	0	0.9	0.88	0.69
E6	0.77	0.07	0.99	0.99	0.9	0	0	0.98
E7	0.66	0.32	0.99	0.99	0.88	0	0	0.99
F8	0.75	0.17	0.7	0.96	0.69	0.98	0.99	0

Fig. 4 An example of the evolutionary process (with real data). Here we use exponential replicator dynamics for faster convergence (Pelillo and Torsello 2006). A set of 8 matching candidates is chosen (*upper left*), a payoff matrix is built to enforce their respective Euclidean constraints (*upper right*, note that cells associated to many-to-many matches are set to 0) and the replicator dynamics are executed (*bottom graph*). At the start of the process the population is set around the barycenter (at 0 iterations). This means that initially the vector **x** represents a quasi-uniform probability distribution. After a few evolutionary iterations the matching candidate B2 (*cyan*) is extinct. This is to be expected since it is a clearly wrong correspondence and its payoff with respect to the other strategies is very low (see the *payoff matrix*). After a few more iterations, strategy A1 vanishes as well. It should be noted that strategies D4/D5 and E6/E7 are mutually exclusive, since they share the same scene vertex. In fact, after an initial plateau, the demise of A1 breaks the tie and finally E6 prevails over E7 and D4 over D5. After just 30 iterations the process stabilizes and only 4 strategies (corresponding to the correct matches) survive (Color figure online)

a higher stability in the measurement, we characterize the variation in orientation as the angle between the principal axes of the descriptor frames rather than between the normal vectors computed from the mesh. Thus, the final payoff for the sparse matching game that we are defining is

$$\Pi = \begin{cases} \delta((a_1, b_1), (a_2, b_2)) & \text{if } a_1 \neq a_2 \text{ and } b_1 \neq b_2 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Once the candidate set and the payoff matrix are built, the game is started from the barycenter of the simplex (Eq. (2)). When a stable state is reached, all the strategies supported by a large percentage of the population are considered non-extinct and retained as correct matches (see Figs. 4 and 5). Since convergence is only reached in infinite time, we cannot expect the weakest strategies to be completely extinct at the equilibrium. We address the resulting thresholding problem by selecting only strategies whose population is within a fixed proportion of that of the best strategy. Then, if the total number of surviving matches is more than a fixed minimum (set to 8 in our experiments), the object is recognized and, optionally, its pose computed. Note that, unlike other approaches, we do not run any costly hypothesis verification step by making considerations on the resulting surface overlap. Finally, we note that Eq. (4) admits symmetric groups of matches and that reflections are not accounted for a posteriori in the pipeline. Nevertheless, probably due to the strong

inlier selection nature of the method and to the lack of perfectly symmetric shapes in the dataset, we never observed mismatches of this sort in all of our experiments.

2.3 Scale-Invariant Matching

The matching scheme presented in Sect. 2.2 assumes a scenario where model and scene have the same scale (although sampling may be different). This allowed us to devise a game that explicitly enforces pairwise isometries between the two surfaces. In this section we tackle a more general setting by allowing the model object and the scene to have different scales.

It is clear that in an object-in-clutter setting there is no simple way to give a model-data relative scale estimate. For example, considerations on the bounding boxes of the two surfaces have little significance, as basic assumptions on the location and possible presence of the object in the scene involve solving the recognition problem itself. In principle, the game-theoretic framework can be adapted so as to consider triplets rather than pairs of points from model and scene. The change in scale can then be accounted for by introducing triangular distance ratios in the formulation of the payoff function. Using ratios in place of plain distances would eliminate the effect of scale and then allow to extract isometric-compatible groups in a similar way to the previous game. The resulting 3-way payoff tensor can in fact be used in a higher-order selection process via generalization of the replicator dynamics (Rota Bulò and Pelillo 2009). Similar techniques have been applied in hypergraph and probabilistic clustering (Shashua et al. 2006). A major problem with this approach is in its computational complexity, which grows with the third power of the total number of strategies, rendering the game infeasible for medium to large-scale problems; looking at memory usage, an unrealistically simple example with 50 points on both model and scene would produce a 58.2 GB single-precision payoff tensor.

Instead, we wish to fit the scale-invariance property within the current scheme. Our insight is to consider the straight line connecting points in each pair (from the same mesh) and then segment this line into a predetermined number of parts. Being in a rigid setting (up to scale), taking a Euclidean path is appropriate. Then, we construct a pairwise descriptor by enriching each pair of points with geometric information at fixed, equally-spaced steps along the line; since the number of segments into which lines are divided is the same for both model and scene, this effectively removes dependence from scale. There are many ways to characterize the surface along the edge, but thanks to the strongly selective behavior of our framework, we can restrict ourselves to simple measures, without taking into account any additional information that may increase distinctiveness at a computational cost. In Fig. 6 we illustrate three possible choices.
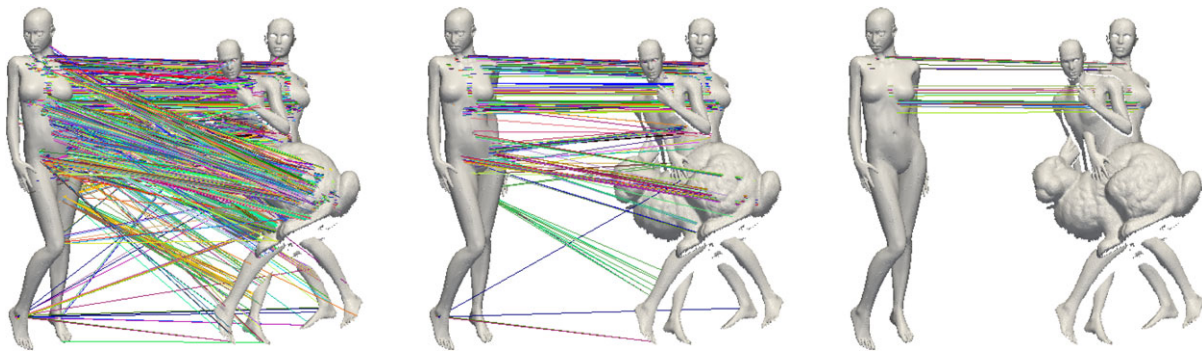
Fig. 5 A real-data example of the selection process. The *first column* shows the set of non-extinct strategies after 10,000 iterations of the evolutionary dynamics. In the *middle column* (12,000 iterations), a set of competing strategies with a common model point (the *green* matches reaching the right thigh on the model) can be clearly seen. Finally, at convergence the evolutionary process selects 30 out of the initial 15,000 candidate matches, effectively resulting in an inlier ratio of 0.2 % over the set of strategies. In this scene, the model is 82.8 % occluded and there is 73.4 % clutter; the scene is especially challenging due to the presence of a very similar object (*david2* from TOSCA dataset). The selection process reached convergence in about 4 seconds (Color figure online)
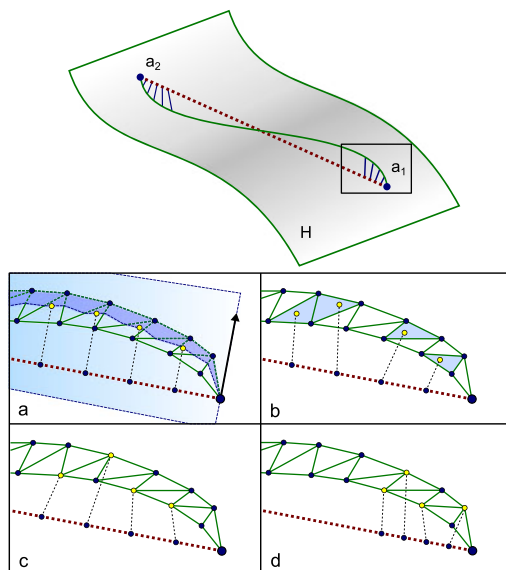

Fig. 6 Definition of a binary descriptor between mesh points a_1 and a_2 for scale-invariant recognition. The n equally-spaced samples along the (*dotted*) segment separating the two points can be projected onto the mesh along a specific direction, such as the normal vector at a_1 (**a**); a minimum-distance projection can instead be computed so as to avoid the choice of a possibly unstable direction and for increased accuracy (**b**); efficiency can be attained by approximating minimum projections with closest points, which is appropriate in the majority of real cases where sampling density is consistent between model and scene (**c**); in order to be robust to occlusions, only the first $m \leq n/2$ samples are considered (**d**)

Given n ordered line samples $s_{i=1,\dots,n}$, we take their projection over the mesh H (Fig. 6b) and build the sequence of minimum (normalized) distances

$$D_n = (d_1, d_2, \dots, d_n),$$
$$d_i = \|s_{i\perp}^H - s_i\| / \|s_1 - s_n\|, \quad i = 1, \dots, n, \quad (6)$$

where $x_\perp^H$ is the minimum-distance projection of point x onto surface H and the denominator acts as a scale normal-ization term. For efficiency reasons we avoid computing the actual projection and approximate it by taking the nearest mesh point to the given line sample (Fig. 6c). We can then associate a descriptor vector $P_{ab} \in \mathbb{R}_+^n$ to each pair of points $a, b \in H$, representing the corresponding distance sequence of length n between them. To be robust against clutter and deal with boundary conditions, we only consider the first $m \leq n/2$ samples from each endpoint (Fig. 6d). The set of strategies S' can be built in a similar way to the isometry-enforcing game, although in this case candidate matches cannot be directly constructed as per Eq. (3), since the descriptors we use are not invariant to scale. To this end, instead of introducing new descriptors, we prefer to rely on gameplay and compute, for each (relevant) point, fixed-scale descriptors at multiple scales; when the game is run, the selection process will operate on the pool of multiple scales and hopefully extract the most (scale-)compatible pairs of strategies. Thus, similarly to Eq. (3), we define the set of strategies as

$$S' = \{(a, b) \in D \times M \mid a \in dn_k(b)\}. \quad (7)$$

Here, descriptors at many different scales are associated with each $a \in D$ and the set of matching strategies will thus include mixed-scale associations between model and scene. We remind that the size of S' depends on parameter k and thus the problem does not necessarily grow in size with respect to the first game; in fact, in Sect. 3 we will use the same k in both games, making the matching step equally efficient in both cases.

Next, we define the new payoff function to be

$$\rho\big((a_1, b_1), (a_2, b_2)\big) = \frac{1}{2} + \frac{A^T B}{2\|A\|\|B\|}, \quad (8)$$

where $A = P_{a_1 a_2} - \bar{P}_{a_1 a_2}$, $B = P_{b_1 b_2} - \bar{P}_{b_1 b_2}$, and $\bar{X}$ denotes the sample mean of X. Although we don't repeat them here,
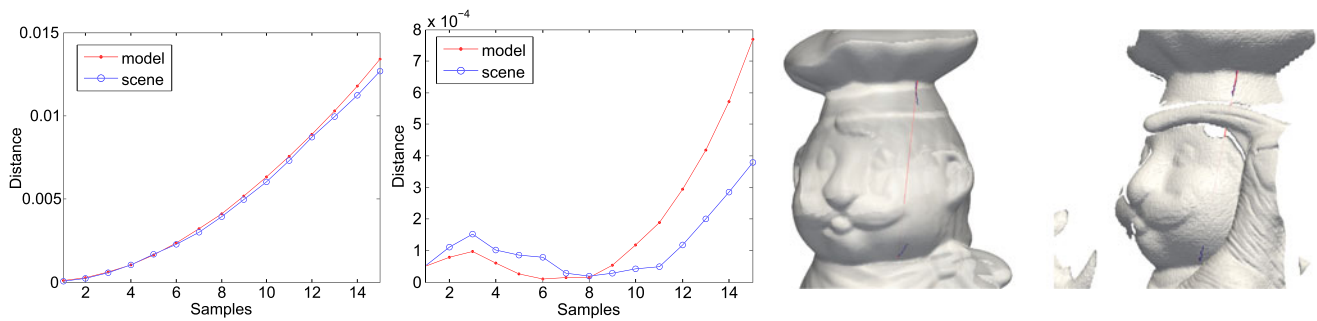
Fig. 7 Example of two (correct) matching distance sequences as extracted by the scale-independent selection process. The path (*red segment* in the images on the *right*) joins the neck of the chef to the hat and has been sampled at a resolution of $n = 100$ and $m = 15$ samples are considered from each endpoint, allowing to mitigate the influence of

clutter on the match. The two graphs plot model and scene descriptors originating at the neck and hat endpoints respectively (corresponding to the *blue paths* in the images). The scene (*last image*) is rescaled to the same scale of the model for visualization purposes (Color figure online)

the same hard constraints from the previous game are also applied in this case. Again, the payoff function takes pairs in $S' \times S'$ and gives values in $[0, 1]$; the payoff in this case is a normalized inner product reflecting the degree of similarity of the distance sequence for a pair in the model with one in the scene. This new formulation is quite different from Eq. (4) in that we choose to avoid enforcing isometries explicitly, and use instead information coming from the pairwise descriptors alone. Of course, there are other ways in which this information can be used in the definition of a payoff function. In the experimental section we present three alternatives and compare the results obtained with each of them. Figure 7 shows an example of two correctly matched strategies and the corresponding distance descriptors.

The scale-invariant game we have just defined favors pairs of matches having compatible distance sequences on each surface, and similar descriptors between the two. While this works well in practice, it can be easily improved by imposing the additional requirement that consistent matches should also give similar estimates of the relative scale between model and scene. We do this by multiplying the payoff function by an additional term favoring pairs with similar scale ratio, expressed as the ratio of the descriptors support radii:

$$\rho'\big((a_1, b_1), (a_2, b_2)\big)$$
$$= \rho\big((a_1, b_1), (a_2, b_2)\big)e^{-\lambda\left|\frac{\sigma(a_1)}{\sigma(b_1)} - \frac{\sigma(a_2)}{\sigma(b_2)}\right|}, \qquad (9)$$

with $\sigma(x)$ indicating the support radius of the descriptor at point x, and λ is a parameter regulating the tolerance level for different scale estimates (a small value for λ indicates high tolerance to different scales). This definition enforces the final group of matches to map the model to the scene at consistent scale. In the experimental section we give a quantitative evaluation of what can be obtained with and without the newly introduced unary term.

Finally, while in our experiments we found that the resolution of the pairwise descriptors (proportional to the number of line samples n) has no significant influence on the matching results, the actual number of samples (from each endpoint) m that take part to the game has a more direct impact. This value can be set to a fixed percentage of n, but this would bring to an imbalance between strategies where spatially close groups of matches become favored.

While this could be desirable in certain applications, we aim at a sparse match covering the target object as much as possible. To do this, we first note that, in general, it is not required that P_{xy} has the same number of components for every pair $x, y \in H$. The number of samples m may be different among pairs of points as long as it is the same on model and scene for each pair of strategies. That is, when calculating the payoff between two pure strategies $(a_1, b_1), (a_2, b_2) \in S'$, it must be $P_{a_1 a_2}, P_{b_1 b_2} \in \mathbb{R}^{2m}$, but any such pair may have a different value for m. We determine this value dynamically for each pair of strategies as the number of steps required to reach a fixed distance d (equal for all pairs) on the model mesh, that is, $m = \lceil d/(\|b_1 - b_2\|/n) \rceil$ with $b_1, b_2 \in M$. This allows to obtain spatially sparse correspondences more easily, and thus increase robustness in presence of occlusions and a more stable pose estimate after a solution is found. Quantitative results comparing the adoption of fixed versus adaptive sampling are presented in the experimental section.

After the payoff matrix is constructed, the game is started from the barycenter as in Sect. 2.2 and the final group of matches, if any, is extracted. Each of these correspondences has associated a value representing its relative degree of participation to the final equilibrium, and can be used to compute the similarity transformation linking model and scene in a weighted fashion (Horn 1987).
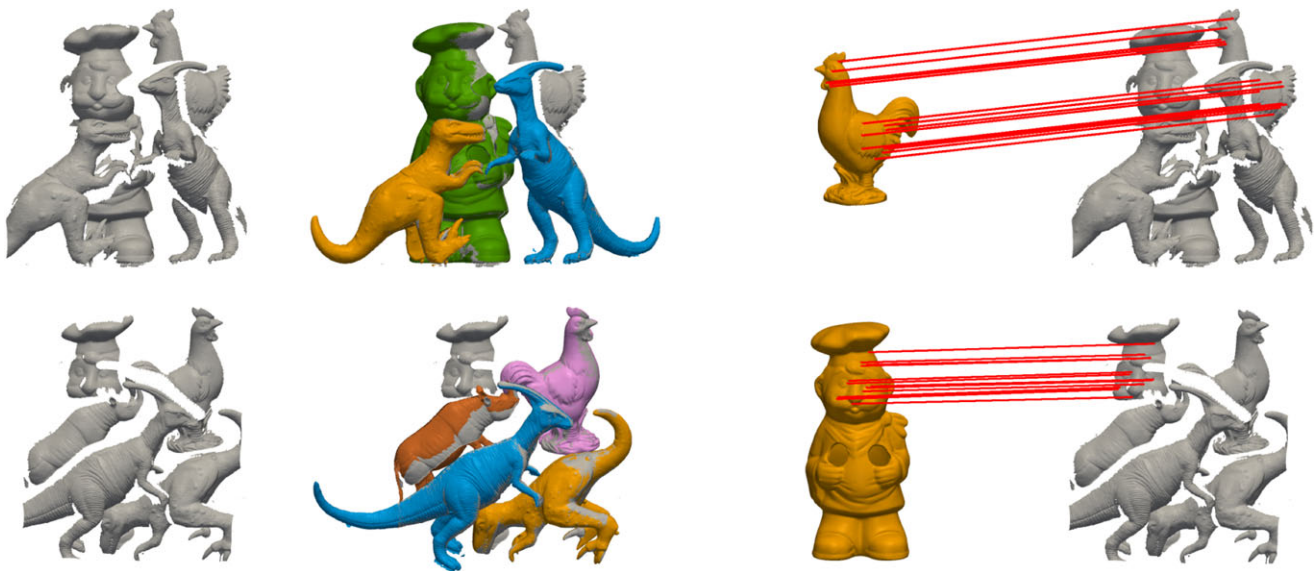
Fig. 8 Two mismatches generated with the method by Mian et al. (*first row*) and the method by Bariya and Nishino (*second row*), which are corrected by our technique. The *first column* shows the range image of the scene, onto which the matched models are successively registered (*second column*). The *chicken* and *chef models* have been missed respectively in the first and second scene, while our method is able to extract correct matches in both cases (*third column*). The figure is best viewed in color (Color figure online)

3 Experimental Results

In order to evaluate the performance of the proposed pipeline we performed a wide range of tests comparing the recognition performance of the proposed approach with recent techniques, as well as characterizing the sensitivity of the approach to the variation of its parameters and its robustness with respect to measurement noise.

3.1 Comparison with the State-of-the-Art

In the first set of experiments we assessed the performance of our approach in the dataset adopted in Bariya and Nishino (2010) and Mian et al. (2006, 2010), which thus acts as a benchmark. This allows us to compare directly our performance against the performances reported in literature, as well as allowing a practitioner to compare our results with those of any other work using the same dataset. This dataset is composed of five high resolution models scanned from real objects (*chef*, *t-rex*, *parasaurolophus*, *chicken* and *rhino*), plus 50 range scans of these objects under various conditions of occlusion (due to the overlap of objects and limits on the field of view of the sensor) and clutter (due to the presence of many objects in the scene). The minimum number of matches to assume the model as recognized in the scene was set to 10 for both fixed scale and scale-invariant matching games. This value is rather conservative as in general it is very unlikely that outliers form consistent groups of more than a few elements; this might be the case, for instance, in situations where strong repeated structure resembling in appearance the object sought for is

present in the scene. In both games, a value of descriptor neighbors of $k = 5$ was used to build the strategy set; relevant points were detected via Integral Hashes with a scale of $\sigma = 8$ edges and then uniformly sub-sampled to 3000 points in the data surfaces, while retaining all relevant points in the model surfaces; 10-bins SHOT descriptors were computed at each relevant point ($\sigma = 8$ edges); the angle separating reference axes at scene points was considered maintained in the model with up to 15 degrees of difference; the final solution at the equilibrium was obtained by thresholding the population vector at 50 % with respect to the most played strategy. Again, this is very conservative as in theory all the matches having a non-zero population share can contribute to the final solution.

Some examples of critical scenes where the proposed technique fixes matches missed by the other methods in the comparison are shown in Fig. 8. The behavior with respect to false positives has not been plotted since the proposed pipeline does not get any in the whole dataset, however Fig. 9 shows an example where our method fails to detect a shape in this set, as well as a false positive match on a much more challenging database we created.

Figure 10 compares our results with recent state-of-the-art algorithms (respectively Bariya and Nishino 2010 and Mian et al. 2006, 2010) and with the well-known 3D Spin Image matching technique (Johnson and Hebert 1999), which is often used as a baseline in literature. The performance of Bariya-Nishino algorithm is that reported in Bariya and Nishino (2010), while the performances of the Tensor, Keypoint and Spin Image approaches are taken from

Table 1 Precision and recall values for the expanded dataset. Out of 20 models, two were left out from detection tests and employed to induce additional clutter

	armadillo	bunny	cat1	centaur1	chef	chicken	dog7	dragon	face
Precision	1.00	1.00	0.78	0.96	0.93	0.93	0.95	1.00	0.91
Recall	0.97	0.97	0.82	1.00	1.00	1.00	0.86	0.89	0.95

	ganesha	gorilla0	horse7	lioness13	para	rhino	t-rex	victoria3	wolf2
Precision	0.89	0.95	0.97	0.88	0.97	0.91	0.97	0.83	0.82
Recall	1.00	0.91	1.00	1.00	0.94	0.91	0.97	0.83	0.95

Fig. 9 Two examples of mismatches generated by our method; in the first case the *dinosaur in the back* was not detected and matches were not generated at all, the occlusion and clutter being respectively 91.4 % and 91.9 %. In the *final example* the *cat model* has been erroneously matched (false positive), due to very similar descriptors being present in the scene (Color figure online)

Mian et al. (2006, 2010). For this reason, the percentages of occlusion and clutter for which the recognition rate is computed are not aligned over all the approaches. Looking at the recognition rate (defined as the true positive rate as in Mian et al. 2006) with respect to model occlusion, the proposed pipeline outperforms even the most recent techniques. Regarding the evaluation of the effects of clutter we were only able to compare our algorithm with Bariya and Nishino (2010), since an implementation for the other approaches and the data they used were not available. Still, it is apparent that the game-theoretic approach obtains good recognition with uniform performance.

While the evaluation on the dataset introduced by Mian et al. (2006) allows us to provide results on a "standard" benchmark which allows direct comparison with the state of the art, this dataset is a bit limited to guarantee statistical significance of the results. For this reason we created a

new dataset which consists of more scenes and models.[1] The dataset is composed of 150 synthetic scenes, captured with a (perspective) virtual camera, and each scene contains 3 to 5 objects. The model set is composed of 20 different objects, taken from different sources and then processed in order to obtain comparably smooth surfaces of almost uniform 100–350k triangles. We opted to construct synthetic scenes in order to have ground truth values for object pose and scene segmentation, but the range extraction process mimics several effects common for real range images, such as the elimination of triangles that are too elongated or at a large angle with the point of view. On top of that, various amounts of random noise was added for all experiments.

Table 1 shows precision and recall values for the synthetic dataset, while Fig. 11 plots the recognition rate as a function of occlusion and clutter for noise values equal to 5 %, 20 %, 30 % and 40 % of the average edge length. It is important to note that the dataset is actually much harder than the other dataset, with several shapes having large flat, featureless areas and several models that are very similar. However, the performance of the algorithm is still very good, and substantially equivalent to that obtained on the dataset by Mian et al. for equivalent noise values, with only a moderate reduction of performance on highly occluded scenes for very high noise values. Finally, Fig. 12 reports the False Positive rate computed at the same noise levels adopted in the previous experiment.

3.2 Sensitivity Analysis

In order to assess the contribution of each component of the pipeline, we substituted each one at a time with alternatives present in the literature. Specifically, we used the same descriptor (Tombari et al. 2010) with the classical matcher proposed in Lowe (2003) (Lowe-SHOT), the game-theoretic matcher without operating the initial relevance-based sampling (GT-Uniform), the descriptors and matching proposed in Albarelli et al. (2010b) (Integral Hashes) and finally the

---

[1]The dataset together with ground-truth information can be downloaded at http://www.dsi.unive.it/~rodola/data.html.
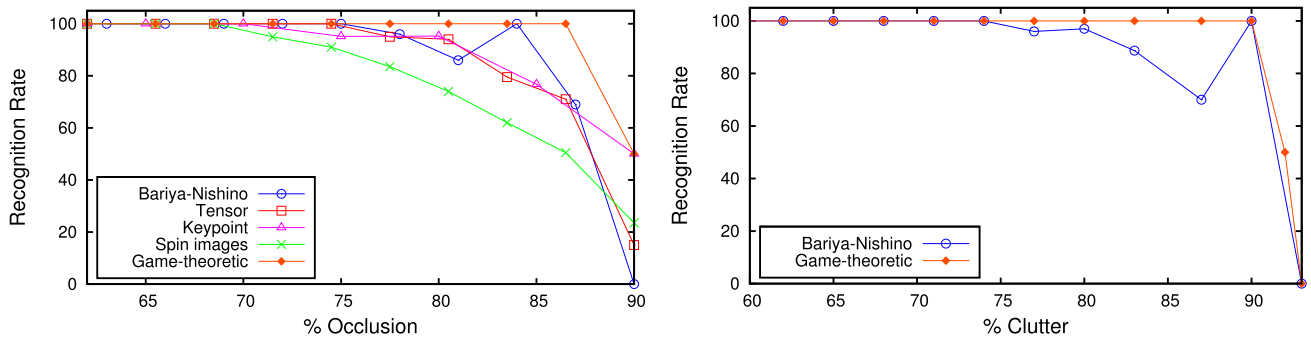
Fig. 10 Comparison with the state of the art on the benchmark dataset introduced by Mian et al. (2006)
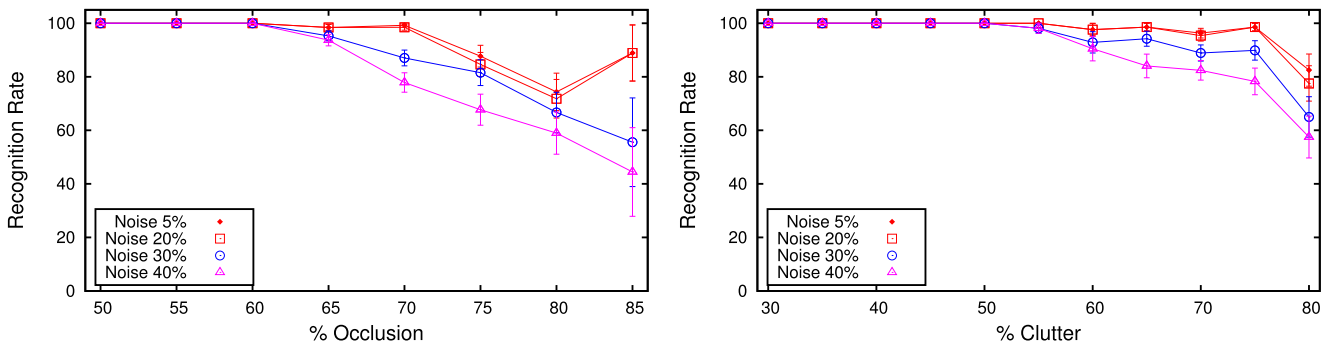

Fig. 11 Recognition rates on the new database at different levels of positional noise. The lowest level represents an attempt to better simulate typical rangemap artifacts on synthetic data
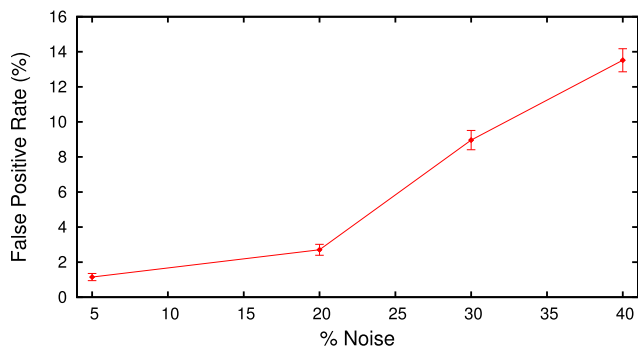

Fig. 12 False Positive rate on the new recognition dataset

full proposed pipeline (GT-Relevant). It is apparent that the proposed pipeline gives its best with all the components in place. The results of this experiment are reported in Fig. 13. Note that the plots in these experiments are more dense than before as we have full control over all the algorithms. This evaluation gives us further insight on the specific setting of object recognition as opposed to other matching scenarios, and confirms some expectations anticipated in the previous sections. First, it is clear that descriptors alone, as robust and descriptive as they may be, are hardly sufficient to guarantee correct matches at moderate levels of occlusion and clutter; they are, in fact, surprisingly good at some challenging scenes while they can fail at apparently simple ones.

This can be caused, for instance, by the presence of repeated structure or featureless objects. The matching process presented in Albarelli et al. (2010a, 2010b) proves to be very effective in a wide range of scenes, but its performance worsens rapidly with increasing levels of clutter. This is symptomatic of the different problem scope of the method, which is tailored to a rigid alignment scenario with symmetric assumptions on the roles of model and data meshes. It is worth noting that two different effects are at the basis of the performance with clutter and noise: Resilience to clutter requires that the approach be unaffected by structured noise. Here it is the selectivity of the game-theoretic inlier selection together with the enforcement of the global constraints that allow the high performance. On the other hand, occlusion is dealt with by the locality of the descriptors. However, it is still the selectivity of the approach that guarantees that even a few good matches obtained in a highly occluded scene provide better payoff than a larger set of less cohesive points.

Skipping only the relevant point selection step yields very high performance, on par with the state of the art (compare with Fig. 10). Nevertheless, at severe levels of occlusion and clutter uniform sampling ceases to be effective as it blindly gives equal importance to all surface regions; this has the effect of drastically reducing the inlier ratio in the construction of hypotheses S, which in turn leads to equilibria where wrong correspondences form larger and stronger
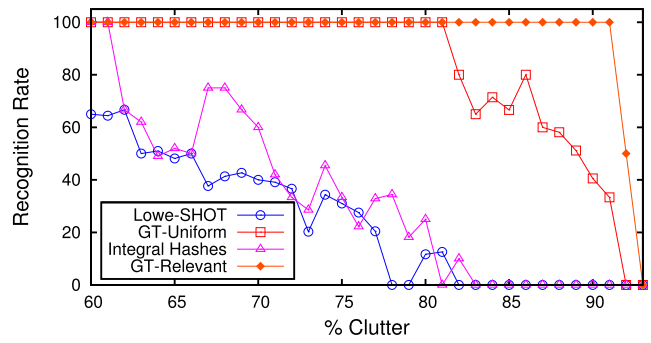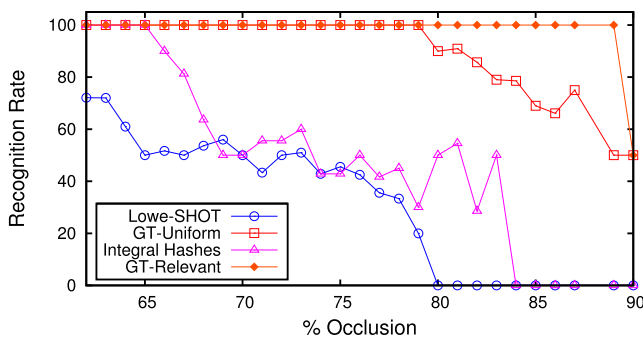
Fig. 13 Contribution of each part of the pipeline tested separately
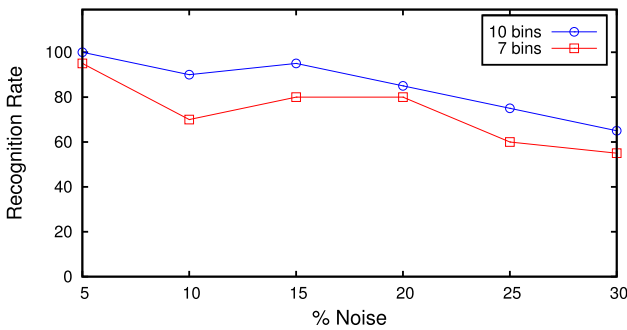

Fig. 14 Evaluation of the robustness of the proposed pipeline with respect to increasing positional noise applied to the scene

groups than the (few) correct ones. Applying relevant sampling to the scene is a simple and fast step, and allows us to obtain excellent results without resorting to more sophisticated interest point detection techniques.

The dataset used in these experiments is made of dense models (300–400k triangles) and slightly less dense scenes produced with a range scanner. Although there is not an exact correspondence between models and scenes, they are rather similar by construction. With the next set of experiments we tried to characterize the performance of the proposed method in presence of positional noise. To do so, we added Gaussian displacement of varying intensity to each vertex in the scenes, and ran the recognition experiments again with the same framework parameters used in the previous evaluations. In order to assess the relative contribution given by descriptors under noisy conditions, we performed this test with two different SHOT parameterizations (the number of bins having a direct effect on resilience to noise, see Tombari et al. 2010 for details). Figure 14 reports the results of this test. As expected, performance gets lower as the noise level increases; still, reasonable recognition rates are maintained also with a moderate amount of noise (with standard deviation equal to 30 % the median edge length). Further, the descriptors do not seem to have a significant impact over the results obtained with additional noise, thus suggesting that robustness to noise is for the ma-

jor part a result of the inlier selection method itself, rather than the specific descriptors used.

3.3 Scale Invariance

In this section we evaluate the effectiveness of the scale-invariant scheme using different definitions for the payoff function and under different parameterizations of the pairwise descriptor. We used the same dataset from the previous experiments, where each scene was randomly scaled from 0.5 to 2.5 times the original scale, and model descriptors spanned over 20 different support radii at each relevant point. All the parameters in common with the isometry-enforcing game are kept at the same values.

The first set of experiments is aimed at determining the best choice for a payoff function. First, we introduce two alternative definitions to ρ (Eq. (8)), giving again a similarity measure based solely on pairwise distance descriptors:

$$\rho_{l2}\big((a_1, b_1), (a_2, b_2)\big) = e^{-\beta \|P_{a_1 a_2} - P_{b_1 b_2}\|_2}; \tag{10}$$

$$\rho_{l1}\big((a_1, b_1), (a_2, b_2)\big) = e^{-\gamma \|P_{a_1 a_2} - P_{b_1 b_2}\|_1}, \tag{11}$$

where $\|\cdot\|_2$ denotes the standard Euclidean norm, $\|\cdot\|_1$ is the L1-norm and parameters β and γ make the functions more or less selective. In our experiments we set $\beta = 1000$ and $\gamma = 1$, values that where empirically seen to yield good results. Note that, after building the strategies set, we do not take into account descriptor information at points a_1, b_1, a_2, b_2 in the definition of the payoff function, although it is certainly possible to introduce another term accounting for their similarity. As in Eq. (9), we wish instead to enforce a common scale mapping by multiplying each payoff function ρ_* by a compatibility term μ based on the local scale of the descriptors:

$$\mu\big((a_1, b_1), (a_2, b_2)\big) = e^{-\lambda \left|\frac{\sigma(a_1)}{\sigma(b_1)} - \frac{\sigma(a_2)}{\sigma(b_2)}\right|}. \tag{12}$$

The introduction of this term helps the selection process by giving small payoff to unlikely hypotheses, thus bringing more stable matches in difficult scenarios, as well as increased efficiency. In the experiments we evaluated all pos-
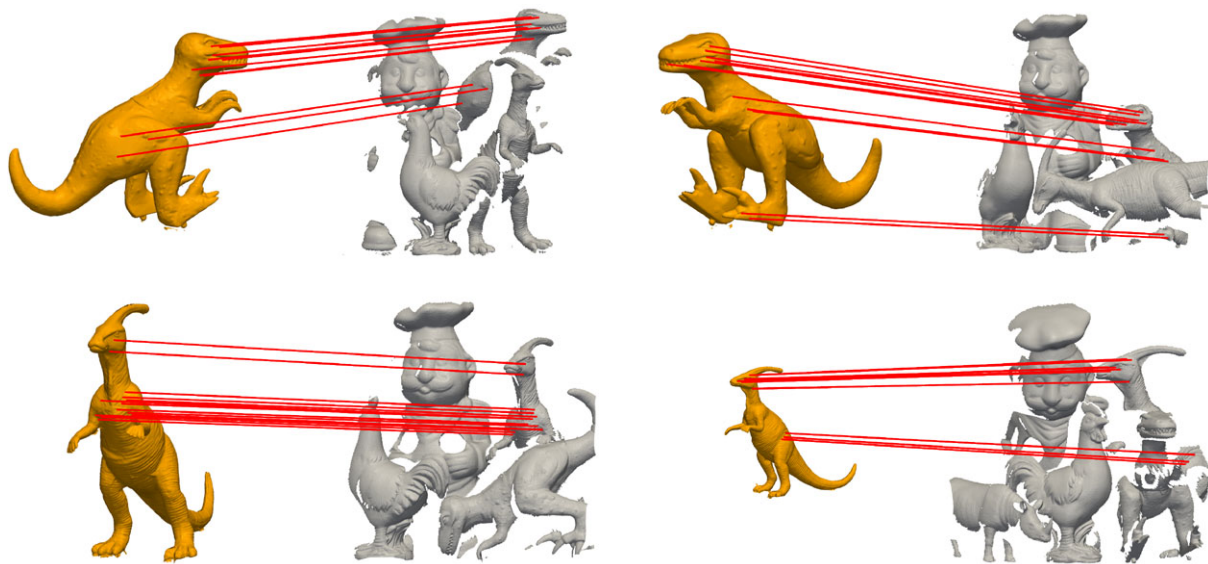
Fig. 15 Examples of scale-invariant object retrieval from cluttered scenes. The foot of *t-rex* in the *top-right image* demonstrates the capability of the method to deal with strong occlusions

Table 2 Scale-invariant recognition rates under different combinations of payoff functions with a scale term

	Dot product	L1-norm	L2-norm
None	88.73 %	83.10 %	84.51 %
Exp	92.96 %	89.75 %	90.14 %
Cut-off	**97.18 %**	87.32 %	91.55 %

Table 3 Average number of matches obtained with different payoff functions, using the same parameters as in Table 2

	Dot product	L1-norm	L2-norm
None	31.70	42.08	43.41
Exp	19.11	12.45	13.54
Cut-off	23.21	12.93	14.10

sible combinations of the payoff functions with three variations on the usage of μ. First, we consider no scale enforcement at all ($\lambda = 0$). Then, we increase its steepness ($\lambda = 30$) so as to make μ very selective and give high values to similar relative scales and very small payoffs to different scales. Finally, we use μ as a cut-off function by putting a hard threshold on the value obtained with $\lambda = 30$; in this case, pairs of strategies receiving $\mu < 0.8$ have the corresponding value of the payoff function set to 0. Table 2 reports the recognition rates obtained with these different combinations on a reduced dataset spanning many levels of occlusion and clutter. We evaluate payoff functions ρ (Dot product), ρ_{l1} (L1-norm) and ρ_{l2} (L2-norm) with no scale consistency (None), large λ (Exp) and hard thresholding (Cut-off). The best results by far are obtained by cutting-off dissimilar relative scales and weighting the result with the inner product of the pairwise descriptors. Scale enforcement is beneficial in all the cases, while L1-norm always gives the worst results. Figure 15 shows some examples of matches obtained with the Dot-Cutoff combination. It should be noted, however, that all the reported recognition rates are rather good considering the scale-invariant setting. In fact, we could gain robustness to mesh sampling and thus achieve better results,

on average, by computing the pairwise descriptors P_{xy} more accurately and not using the closest mesh point in place of projections, as described in Sect. 2.3. The average number of matches for each payoff function on the same dataset are reported in Table 3. Looking at the reported values, it is apparent that the increased selectivity brought by the additional scale constraints (second and third rows) has a direct influence on the size of the solution at the equilibrium.

The second set of experiments analyzes sensitivity to parameters of the pairwise descriptor, namely its resolution (expressed as the total number of line samples n) and the actual number of samples used in the descriptor ($2m$). For these experiments we used the best payoff function as evaluated in Table 2 (ρ' with cut-off). We observed that, while in principle a higher resolution should give better results, in practice the recognition rate is not affected by this parameter: using a small or large value for n (from 10 to 2000 samples) gives the same results on the whole dataset. This is probably due to the fact that, after removing the effect of scale, the game operates a very robust inlier selection in a rigid setting, where a few good hints are sufficient for extracting a consistent group of matches. As a reference, we used $n = 100$ in the following experiments. As described in
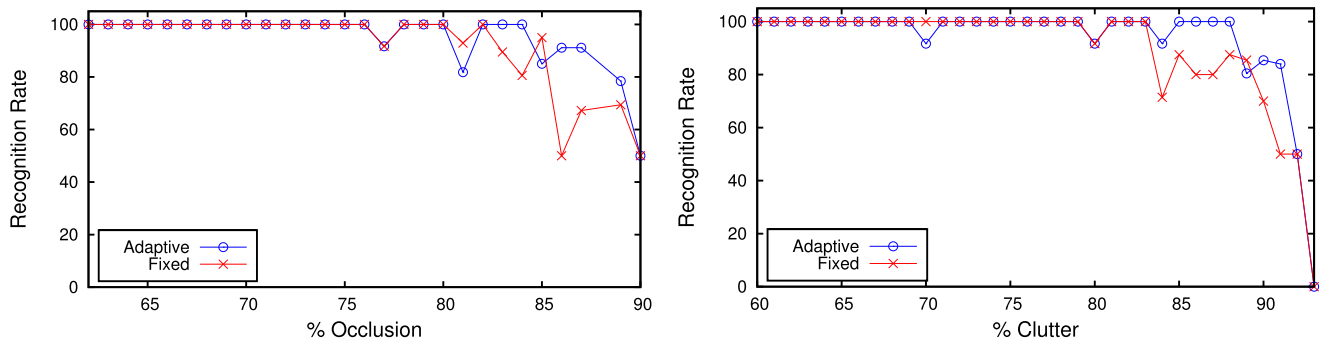
Fig. 16 Recognition rate of the scale-invariant matching game against occlusion and clutter. The *two curves* correspond to different ways of determining the descriptor size
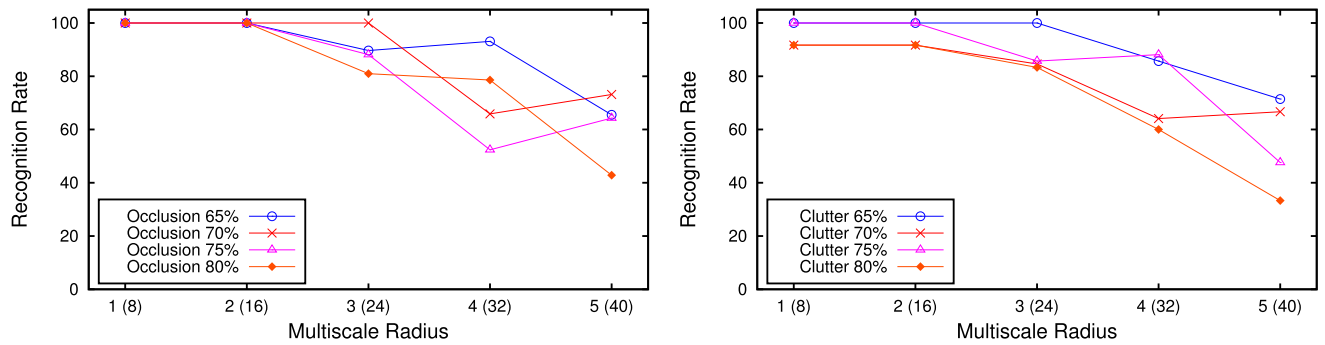

Fig. 17 Recognition rate of the scale-invariant matching game as the number of scales is increased

Sect. 2.3, we analyzed two different approaches to determine a value for m, and thus the size of the descriptor. In Fig. 16 we plot their recognition rate against clutter and occlusion on the full dataset. The first approach (Fixed) takes a fixed number of samples for all the pairs, set in our experiments to the first 7 % samples from each end. As a result, any P_{xy} on model and scene has only 14 components; this value was determined empirically as the smallest number for which performance does not start to decrease. The second approach (Adaptive) is dynamic and each pair of strategies has the value of m set to the number of steps required on the model mesh to reach a (Euclidean) distance of 8 times the model resolution (calculated as its median edge length). Both methods exhibit remarkable performance at high levels of scene noise, with adaptive sampling giving better results on average. Comparing the results with those in Fig. 10, we observe that performance of the scale-invariant pipeline is at least as good as the state of the art for fixed scale recognition on the same dataset.

The final set of experiments is aimed at assessing the robustness of the approach with respect to large changes in scale between model and scene. It is worth noting that, by construction, the multiscale descriptor is invariant to a pure change of scale between model and scene since it is normalized over the mean edge length. It is, however, sensible to variations in mesh resolution, either due to a change

of scale with a fixed resolution scanning system, or due to the use of acquisition processes at different resolutions. In particular, since the approach works under the assumption that scale-compatible assignments are present from the pool of fixed-scaled descriptors computed at multiple scales, the main breaking point is the need to increase the levels of scales considered to account for the expected changes in scales. The effect of this is the reduction of the inlier ratio, i.e., the ratio of correct assignments over all the strategies considered. In particular, since an inlier correspondence can insist only on one level per point, an n-fold increase in scale level induces an n-fold reduction in the inlier ratio, rendering the selection process much harder.

Figure 17 plots the recognition rate of the proposed approach on the benchmark dataset as we increase the scale radius of the descriptor and consequently the number of scales. The first number on the Multiscale Radius axis refers to the radius of the descriptor, while the numbers within parentheses give the actual number of scales induced by the choice of the radius. Hence, we span a five-fold increase in scales from 8 to 40. As it can be seen, the approach exhibits almost optimal performance up to 24 levels, and then has a linear degradation with higher levels, providing performance comparable to the state of the art even with 40 distinct levels, i.e. with a 40-fold reduction in inlier ratio with respect to the fixed scale recognition pipeline.
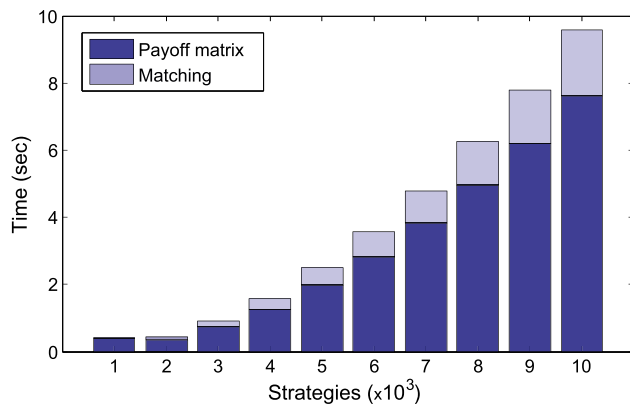
Fig. 18 Time versus the number of strategies in the scale-invariant setting

3.4 Performance Considerations

Following Sect. 2, it is of interest to carry out a performance evaluation of the proposed pipeline. We remind that in a typical matching scenario, only a subset of interesting points from model and scene take part to the matching game (see Sect. 2.2).

Given a payoff matrix and an initial set of candidate correspondences, the selection process is executed by means of evolutionary dynamics, for which it is difficult to give an upper bound for its convergence time. In the case of standard, first-order replicator dynamics (as per Eq. (2)), the computational complexity of each step is $O(N^2)$, with N being the total number of strategies. For this reason replicator dynamics are rarely used in practice, even more so for large-scale problems, where the cardinality of the set of strategies can be in the order of thousands even after strong candidate rejection via descriptor priors. A faster alternative is provided by the infection-immunization dynamics (Rota Bulò and Bomze 2011), which has an $O(N)$ complexity for each step; under this model, the time per iteration is only quadratic with respect to the number of mesh points, allowing to reach convergence in 4–5 seconds (around 15,000 iterations) with tens of thousands of strategies.[2] Figure 18 reports computational times of the pipeline for the scale-invariant game, using infection-immunization to compute the equilibria. The computation is dominated by the construction of the payoff matrix Π, while the matching step takes only a small fraction of time. It can be seen that the selection process attains an equilibrium within seconds even with thousands of strategies. The experiments were written in C++ and run on a Core i7 machine with 12 GB of memory.

4 Conclusions and Future Work

We presented a novel pipeline for model-based 3D object recognition in cluttered scenes obtained with a range scanner. The pipeline starts with the detection of distinctive keypoints in the scene, which in turn is composed of a relevance filter, a subsampling step and the calculation of a descriptor for each sample kept. These relevant points are then matched pairwise with all the model keypoints and a set of candidate pairings is obtained. Finally, a non-cooperative, isometry-enforcing game is played. The gameplay performs the actual recognition step and returns a sparse set of reliable matches. An additional game is then introduced to tackle the more challenging recognition problem where model and scene are allowed to take different scales. To this end, a novel pairwise strategy descriptor utilizing geometric information along the Euclidean path linking surface points is adopted. The scale mapping is further enforced by computing local descriptors at different scales and putting them in the pool of candidate matches, thus letting the selection process extract the most compatible group of correspondences. The two matching approaches fit within the same general framework, and are extensively evaluated through a wide range of experiments under different conditions. The results demonstrate that the proposed pipeline outperforms the most recent state-of-the-art techniques on the same dataset, and are further confirmed on an additional dataset comprising challenging combinations of different objects. Additionally, the method is of easy implementation and can be made very efficient by exploiting recent results in the field of Game Theory.

Interesting directions for future research include the extension of the proposed framework to non-rigid matching and recognition under different classes of deformation. While the problem of non-rigid matching has been extensively tackled in literature, we believe that the presented framework could be easily adapted by defining proper payoff measures taking into account intrinsic (geodesic) surface metrics in place of their Euclidean counterpart. Further, a thorough statistical analysis of the selection process may provide interesting insight on its convergence properties and thus help to define more rigorous ways to characterize matching scenarios.

References

Ahn, Y. K., Park, Y. C., Choi, K. S., Park, W. C., Seo, H. M., & Jung, K. M. (2009). 3d spatial touch system based on time-of-flight

---

[2]Matlab code for infection-immunization dynamics is available at http://www.dsi.unive.it/~rodola/sw.html.

camera. *WSEAS Transactions on Information Science and Applications*, 6, 1433–1442.

Aiger, D., Mitra, N. J., & Cohen-Or, D. (2008). 4-points congruent sets for robust surface registration. *ACM Transactions on Graphics*, 27(3), 85.

Akagündüz, E., Eskizara, O., & Ulusoy, I. (2009). Scale-space approach for the comparison of HK and SC curvature descriptions as applied to object recognition. In *Proc. of the 16th IEEE international conference on image processing, ICIP'09*, Piscataway, NJ, USA (pp. 413–416).

Albarelli, A., Rota Bulò, S., Torsello, A., & Pelillo, M. (2009). Matching as a non-cooperative game. In: *ICCV 2009: Proc. of the 2009 IEEE international conference on computer vision*. Los Alamitos: IEEE Comput. Soc.

Albarelli, A., Rodolà, E., & Torsello, A. (2010a). A game-theoretic approach to fine surface registration without initial motion estimation. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 430–437). doi:10.1109/CVPR.2010.5540183.

Albarelli, A., Rodolà, E., & Torsello, A. (2010b). Loosely distinctive features for robust surface alignment. In *ECCV 2010—11th European conference on computer vision* (pp. 519–532).

Albarelli, A., Rodolà, E., Bergamasco, F., & Torsello, A. (2011). A non-cooperative game for 3d object recognition in cluttered scenes. In *3DIMPVT* (pp. 252–259).

Albarelli, A., Rodolà, E., & Torsello, A. (2012). Imposing Semi-Local geometric constraints for accurate correspondences selection in structure from motion: a Game-Theoretic perspective. *International Journal of Computer Vision*, 97, 36–53. doi:10.1007/s11263-011-0432-4.

Bariya, P., & Nishino, K. (2010). Scale-hierarchical 3d object recognition in cluttered scenes. In *IEEE conference on computer vision and pattern recognition*, CVPR 2010 (pp. 1657–1664).

Borrmann, D., Elseberg, J., Lingemann, K., Nüchter, A., & Hertzberg, J. (2008). Globally consistent 3d mapping with scan matching. *Robotics and Autonomous Systems*, 56, 130–142.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11), 1229–1234.

Chen, H., & Bhanu, B. (2007). 3d free-form object recognition in range images using local surface patches. *Pattern Recognition Letters*, 28, 1252–1262.

Chua, C. S., & Jarvis, R. (1997). Point signatures: a new representation for 3d object recognition. *International Journal of Computer Vision*, 25, 63–85.

Chum, O., & Matas, J. (2005). Matching with PROSAC—progressive sample consensus. In *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 220–226). Washington: IEEE Comput. Soc.

Frome, A., Huber, D., Kolluri, R., Bülow, T., & Malik, J. (2004). Recognizing objects in range data using regional point descriptors. In *ECCV 2004, 8th European conference on computer vision* (pp. 224–237).

Ghosh, D., Amenta, N., & Kazhdan, M. M. (2010). Closed-form blending of local symmetries. *Computer Graphics Forum*, 29(5), 1681–1688.

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

Johnson, A. E., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5), 433–449. http://doi.ieeecomputersociety.org/10.1109/34.765655.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. of the international conference on computer vision* (pp. 1150–1157).

Lowe, D. G. (2003). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 20, 91–110.

Mian, A. S., Bennamoun, M., & Owens, R. (2006). Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 1584–1601.

Mian, A. S., Bennamoun, M., & Owens, R. (2010). On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision*, 89, 348–361.

von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Newman, T. S., & Jain, A. K. (1995). A system for 3d cad-based inspection using range images. *Pattern Recognition*, 28(10), 1555–1574.

Novatnack, J., & Nishino, K. (2008). Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images. In *Proc. of the 10th European conference on computer vision* (pp. 440–453). Berlin: Springer.

Pelillo, M., & Torsello, A. (2006). Payoff-monotonic game dynamics and the maximum clique problem. *Neural Computation*, 18, 1215–1258.

Pottmann, H., Wallner, J., Huang, Q. X., & Yang, Y. L. (2009). Integral invariants for robust geometry processing. *Computer Aided Geometric Design*, 26(1), 37–60. doi:10.1016/j.cagd.2008.01.002.

Rota Bulò, S., & Bomze, I. M. (2011). Infection and immunization: a new class of evolutionary game dynamics. *Games and Economic Behavior*, 71(1), 193–211.

Rota Bulò, S., & Pelillo, M. (2009). A game-theoretic approach to hypergraph clustering. In *Advances in neural information processing conference (NIPS2009)* (Vol. 22, pp. 1571–1579).

Salti, S., Tombari, F., & di Stefano, L. (2011). A performance evaluation of 3d keypoint detectors. In: *International conference on 3D imaging, modeling, processing, visualization and transmission (3DIMPVT)* (pp. 236–243). New York: IEEE Press.

Shashua, A., Zass, R., & Hazan, T. (2006). Multi-way clustering using super-symmetric nonnegative tensor factorization. In *European conference on computer vision* (pp. 595–608).

Sun, Y., Paik, J., Koschan, A., & Abidi, M. A. (2003). Point fingerprint: a new 3-d object representation scheme. *IEEE Transactions on Systems, Man and Cybernetics*, 33, 712–717.

Tombari, F., Salti, S., & di Stefano, L. (2010). Unique signatures of histograms for local surface description. In *ECCV 11th European conference on computer vision* (pp. 356–369).

Weibull, J. (1995). *Evolutionary game theory*. Cambridge: MIT Press.