Internation Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Imposing Semi-local Geometric Constraints for Accurate Correspondences Selection in Structure from Motion: a Game-Theoretic Perspective

Andrea Albarelli · Emanuele Rodolà · Andrea Torsello

Received: date / Accepted: date

Abstract Most Structure from Motion pipelines are based on the iterative refinement of an initial batch of feature correspondences. Typically this is performed by selecting a set of match candidates based on their photometric similarity; an initial estimate of camera intrinsic and extrinsic parameters is then computed by minimizing the reprojection error. Finally, outliers in the initial correspondences are filtered by enforcing some global geometric property such as the epipolar constraint. In the literature many different approaches have been proposed to deal with each of these three steps, but almost invariably they separate the first inlier selection step, which is based only on local image properties, from the enforcement of global geometric consistency. Unfortunately, these two steps are not independent since outliers can lead to inaccurate parameter estimation or even prevent convergence, leading to the well known sensitivity of all filtering approaches to the number of outliers, especially in the presence of structured noise, which can arise, for example, when the images present several repeated patterns. In this paper we introduce a novel stereo correspondence selection scheme that casts the problem into a Game-Theoretic framework in order to guide the inlier selection towards a consistent subset of correspondences. This is done by enforcing geometric constraints that do not depend on full knowledge of the motion parameters but rather on some semi-local property that can be estimated from the local appearance of the image features. The practical effectiveness of the proposed approach is confirmed by an extensive set of experiments and comparisons with state-of-the-art techniques.

Keywords Inlier selection · Game-Theory · Structure from Motion

1 Introduction

The common goal of all Structure from Motion (SfM) techniques is to infer as many 3D clues as possible by analyzing a set of 2D images. In general the 3D knowledge that can be obtained by such methods can be classified into two different (but related) classes: *scene* and *camera* information. Scene information is referred to the actual shape of the objects depicted in the images. This often boils down to assigning a plausible location in space to some significant subset of the acquired 2D points. These newly reconstructed 3D points are the "structure" part of SfM. By contrast, camera information includes all the parameters that characterize the abstract model of the image acquisition process. These can in turn be classified into intrinsic and extrinsic parameters. Intrinsic parameters are related to the physical characteristics of the camera itself, such as its focal length and principal point, while the extrinsic parameters define the camera pose, that is its position and rotation with respect to a conventional origin in the 3D space. Unlike the structure part, which is physically bound to a particular 3D configuration, the intrinsic

A. Albarelli
Università Ca' Foscari Venezia, Dipartimento di Scienze Ambientali, Informatica, Statistica, Italy
E-mail: albarelli@unive.it

E. Rodolà
Università Ca' Foscari Venezia, Dipartimento di Scienze Ambientali, Informatica, Statistica, Italy
E-mail: rodola@dsi.unive.it

A. Torsello
Università Ca' Foscari Venezia, Dipartimento di Scienze Ambientali, Informatica, Statistica, Italy
E-mail: torsello@unive.it

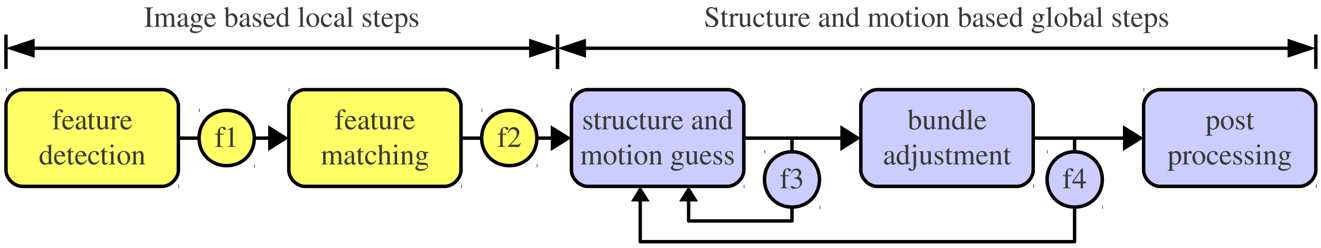


Fig. 1 A simplified schema that captures the general steps found in many SfM approaches. The main loop is usually based on an iterative refinement of the candidate scene points based of their geometric consistency with respect to the estimated motion. Circles between steps represent the applied outlier filtering strategies.

and extrinsic parameters can vary in each shot; for this reason they are usually referred to as "motion".

Given the wide range of practical applications that could take advantage of a 3D reconstruction, it is not surprising that SfM has been a very active research topic during the last decades. In fact, many different approaches have been proposed in literature: some are aimed at solving the most general scenarios, others specialize to sub-domains, both in terms of the number of free parameters allowed and in terms of the assumptions made on some characteristics of the scene to be inferred. While the most relevant SfM approaches will be discussed with more detail in Section 2.3, in this section we will resort to the simplified general workflow presented in Figure 1 in order to introduce the key ideas and contributions of the proposed approach. To this end, the typical pipeline can be roughly split in two subsequent macro steps (respectively dubbed as *Image based* and *Structure and Motion based* in Figure 1). The first step deals with the localization in the source 2D images of salient feature points that are both distinctive and repeatable. Such points are meant to be tracked between different images, thus creating multiple sets of correspondences that will be used in the scene reconstruction step. The use of a reduced set of relevant points is crucial as their repeatable characterization allows us to minimize the chance of including wrong correspondences. Typically, filters are applied to the selection and matching phase in an attempt to make this phase more robust. In Figure 1 the extracted features are further culled by using filter *f1*, which eliminates points that exhibit very common descriptors or that are not distinctive or stable enough. A second refinement can be achieved after the matching: most implementations of filter *f2* remove correspondences that are not reliable enough, that is pairs where the second best match has a very similar score to the first one or that involve too different descriptors. Once a suitable set of point pairs has been found among all the images, the second macro step of the pipeline uses them to perform the actual structure and motion estimation. This happens by building a reasonable guess for both the camera parameters and the spatial locations of the correspondences found, and then, almost invariably, by applying a bundle adjustment optimization to refine them. Also, at this stage, filtering techniques can be adopted in order to remove outliers from the initial set of matches. Specifically, a filter that removes pairs that do not agree with the estimated epipolar constraints can be applied after combining some or all the images into the initial guesses (*f3*), or after bundle adjustment optimized the structure and motion estimates (*f4*). Depending on the result of the filtering a new initial estimation can be triggered, taking advantage of the (hopefully) more accurate selection of corresponding features. This kind of process leads to an iterative refinement that usually stops when the inlier set does not change or becomes stable enough. While this approach works well in many scenarios, it inherently contains a limitation that might drive it to poor results or even prevent it from converging at all: The main criterion for the elimination of erroneous matches is to exclude points that exhibit a large reprojection error or adhere poorly to the epipolar constraint after a first round of scene and pose estimation. Unfortunately this afterthought is based upon an error evaluation that depends on the initial matches; this leads to a quandary that can only be solved by avoiding wrong matches from the start. This is indeed a difficult goal, mainly because the macro step from which the initial matches are generated is only able to exploit strictly local information, such as the appearance of a point or of its immediate surroundings. By contrast the following step would be able to take advantage of global knowledge, but this cannot be trusted enough to perform an extremely selective trimming and thus most methods settle with rather loose thresholds. In order to alleviate this limitation, in this paper we introduce a robust matching technique that allows to operate a very accurate inlier selection at an early stage of the process and without any need to rely on preliminary structure and motion estimations. This is obtained by enforcing properties that are inferable from image regions at a local or semi-local scale and then by extending their validation to a global scale. Similar approaches have already been used to obtain better camera pose estimations when dealing with complex multi-component scenes, where local observations can be handled in a decoupled way, thus leading to a better resilience to outliers (Fermuller et al 1997). In this paper the inlier validation happens by casting the selection process into a Game-Theoretic setting, where feature-correspondences are allowed to compete with one another, receiving support from correspondences that satisfy the same semi-local constraints, and competitive pressure from the rest. The surviving correspondences form a small cohesive set of mutually compatible correspondences, satisfying the semi-local constraint globally. Of course many alternative selection techniques exist and can be adopted to perform the inlier set optimization, nevertheless the proposed Game-Theoretic approach offers the unique advantage of a strong tendency to limit false negatives rather than concentrating on low false positives as most matching techniques in the literature. This propery allows for a strong resilience to the large number of outliers normally encountered in general SfM scenarios. Further, the approach is quit3e general; in fact, in Section 3 we will show how the definition of different payoff functions between strategies leads to optimizers with task-specific goals. Finally, in order to assess the advantage provided by our approach, in the experimental section we compare our technique with a reference implementation of the structure-from-motion system presented in (Snavely et al 2006) and (Snavely et al 2008).

2 Background

Before discussing our robust matching approach we will briefly review the most significant related contributions available in literature and introduce some basic notions about the geometry of the SfM process.

2.1 Features Extraction and Matching

The selection of 2D point correspondences is arguably the most critical step in image based multi-view reconstruction and, differently from techniques augmented by structured light or known markers, there is no guarantee that pixel patches exhibiting strong photo consistency are actually located on the projection of the same physical point. Further, even when correspondences are correctly assigned, the interest point detectors themselves introduce displacement errors that can be as large as several pixels. Such errors can easily lead to sub-optimal parameter estimation or, in the worst cases, to the inability of the optimization algorithm to obtain a feasible solution. For this reasons, reconstruction approaches adopt several specially crafted expedients to avoid the inclusion of outliers as much as possible. In the first place correspondences are not searched throughout the whole image plane, but only points that are both repeatable and well characterized are considered. This selection is carried out by means of interest point detectors and feature descriptors. Salient points are localized with sub-pixel accuracy by general detectors, such as Harris Operator (Harris and Stephens 1988) and Difference of Gaussians (Marr and Hildreth 1980), or by using techniques that are able to locate affine invariant regions, such as Maximally Stable Extremal Regions (MSER) (Matas et al 2004) and Hessian-Affine (Mikolajczyk and Schmid 2002). The affine invariance property is desirable since the change in appearance of a scene region after a small camera motion can be locally approximated with an affine transformation. Once interesting points are found, they must be matched to form the candidate pairs to be fed to the subsequent parameter optimization steps. Most of the currently used techniques for point matching are based on the computation of some affine invariant feature descriptor. Specifically, to each point is assigned a feature vector with tens to hundreds of dimensions, plus a scale and a rotation value. Among the most used feature descriptor algorithms are the Scale-Invariant Feature Transform (SIFT) (Lowe 1999, 2003), Speeded Up Robust Features (SURF) (Herbert Bay and Gool 2006), Gradient Location and Orientation Histogram (GLOH) (Mikolajczyk and Schmid 2005) and more recently the Local Energy based Shape Histogram (LESH) (Sarfraz and Hellwich 2008), the SIFT algorithm being the first of the lot and arguably the most widely adopted. The complete SIFT technique, introduced and patented by Lowe, describes in detail both the detection step and the computation of repeatable descriptors to be associated with the found keypoints. Specifically, the localization of potentially relevant features happens by first applying to the image a Gaussian filter at different scales and then by selecting points that are maxima or minima of the Difference of Gaussians (DoG) that occur at multiple scales. This is done by comparing each pixel in the DoG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales. Subsequently the found candidates are interpolated to nearby data in order to ensure an accurate and repeatable position and thus they are filtered by discarding points that exhibit a low contrast or that are located along an edge

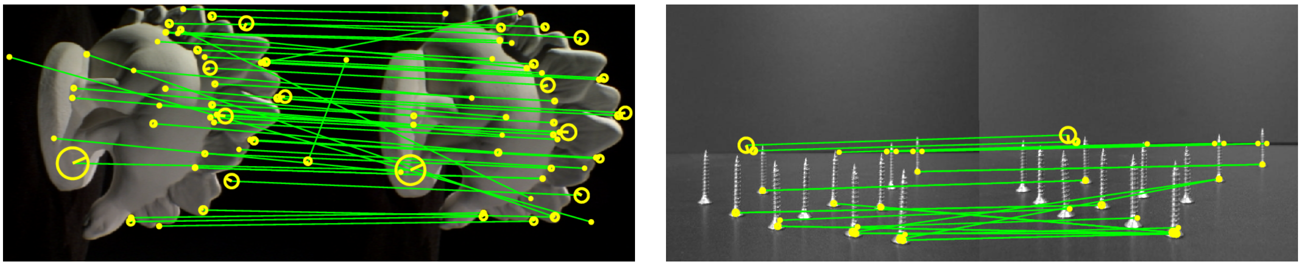


Fig. 2 Example of SIFT features extracted and matched using the VLFeat package. Each feature in the first image has been matched with the feature in the second image that exhibits the most similar descriptor. Note that, while most of the correspondences are correct, many mismatches are still present.

(which could hinder the precision of the localization). Finally, an orientation based on the local image gradient is assigned to each one of the surviving points. The computation of the descriptor vector is then performed on the image closest in scale to the keypoint's scale and rotates accordingly to the keypoint's orientation. To this end, a set of histograms are computed based on the magnitude and orientation values picked from the neighborhood of the feature. The magnitudes are further weighted by a Gaussian function with σ equal to half the width of the descriptor window. The histograms are then packed in a vector which is typically long 128 or 256 elements and that is normalized to unit length in order to enhance invariance to changes in illumination. Given the great success of the SIFT detector/descriptor, several enhancements and specializations were introduced since the original paper by Lowe; for instance, PCA-SIFT (Ke and Sukthankar 2004) applies PCA to the normalized gradient patch to gain more distinctiveness, PHOW (Bosch et al 2007) makes the descriptor denser and allows to use color information, ASIFT (Morel and Yu 2009) extends the method to cover the tilt of the camera in addition to scale, skew and rotation. In all these techniques, the descriptor vector is robust with respect to affine transformations: i.e., similar image regions exhibit descriptor vectors with small mutual Euclidean distance. This property is used to match each point with the candidate with the nearest descriptor vector. However, if the descriptor is not distinctive enough this approach is prone to select many outliers. A common optimization involves the definition of a maximum threshold over the distance ratio between the first and the second nearest neighbors. In addition, points that are matched multiple times are deemed as ambiguous and discarded (i.e., one-to-one matching is enforced). Despite any effort made in this direction, any filter that operates at a local level is fated to fail when the matched regions are very similar or identical, a situation that is not uncommon as it happens every time an object is repeated multiple times in the scene or there is a repeated texture. In Figure 2 we show two examples of SIFT features extracted and matched by using the VLFeat (Vedaldi and Fulkerson 2008) Matlab toolkit. In the first example almost all the correspondences are correct, still some clear mismatches are visible both between the plates of the saurus (which are similar in shape) and on the black background (which indeed contains some amount of noise). In the second example several identical screws are matched and, as expected, features coming from different objects are confused and almost all the correspondences are wrong. It should be noted that such mismatches are not a fault of the descriptor itself as it performs exactly its duty by assigning similar description vectors to features that are almost identical from a photometric standpoint. In fact, this particular example is specially crafted to break traditional matchers that rely on local properties. In the experimental section, we will show how introducing some level of global awareness in the process allows to deal well also with these cases that are indeed very common in the highly repetitive world of human-made objects and urban environments.

2.2 Camera Model and Epipolar Geometry

The pinhole projection (Figure 3) is the most common camera model used in reconstruction frameworks. Its wide adoption is due to its ability to approximate well the behaviour of many real cameras. In practical scenarios radial and tangential lens distortions are the main sources of divergence from the pinole model, however it is easy to fit polynomial models to them and compensate for their effect (Tsai 1987; Weng et al 1992). The most important parameters of this model are the pose of the camera with respect to the world (represented by a rotation matrix R and a translation vector T), the distance of the projection center from the image plane (the focal length f in Figure 3), and the coordinates on the image plane of the intersection between the optical axis and the plane itself (the principal point $c = (c_x, c_y)^T$ in Figure 3). The projection of a world

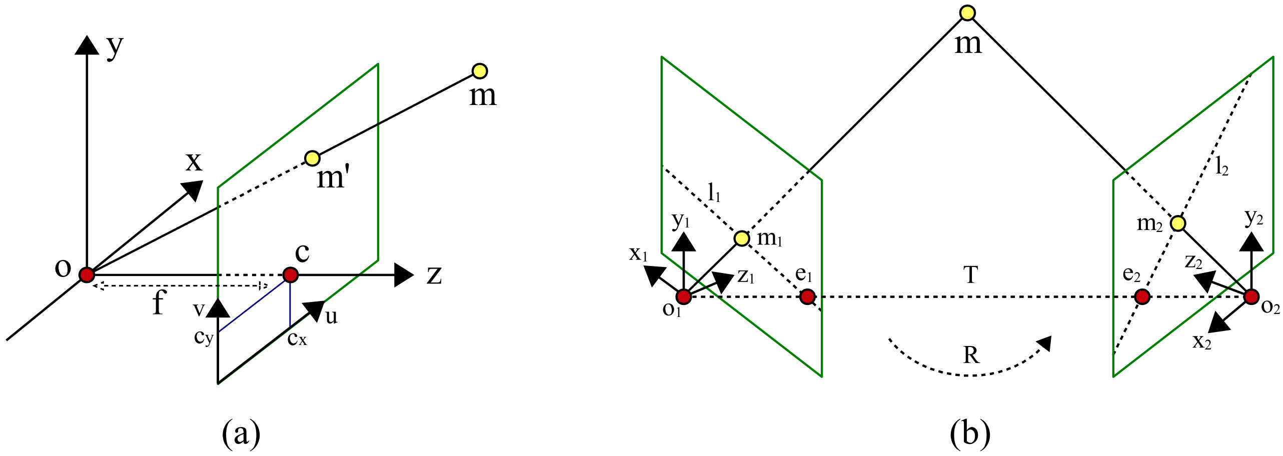


Fig. 3 Illustration scheme of the pinhole camera model (a) and of the epipolar geometry (b). See text for details.

point m on the image plane happens in two steps. The first step is a rigid body transformation from the world coordinate system to the camera system. This can be easily expressed (using homogeneous coordinates) as:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \sim \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

The second step is the projection of the point in camera coordinates on the image planes, which happens by applying a camera calibration matrix $\mathbf{K}$ containing the intrinsic parameters of the model. The most general version of the calibration matrix allows for a different vertical (f_y) and horizontal (f_x) focal length to accommodate for non-square pixels, and for a skewness factor (s) to account for non-rectangular pixels:

$$\mathbf{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

In practice, for most real cameras, pixels can be approximated by perfect squares, thus we can resort to the basic model of Figure 3 and assume $s = 0$ and $f_x = f_y = f$. Usually the camera pose and calibration matrices are combined into a single 3×4 projection matrix $\mathbf{P} = \mathbf{K}[\mathbf{R}\,\mathbf{T}]$. This projection matrix can be directly applied to a point in (homogeneous) world coordinates to obtain its corresponding 2D point on the image plane:

$$\mathbf{m}' = \mathbf{P}\mathbf{m} = \mathbf{K}[\mathbf{R}\,\mathbf{T}]\mathbf{m}\,.$$

When a point is observed by two cameras its projections on the respective image planes are not independent. In fact, given the projection m_1 of point m in the first camera, its projection m_2 on the second image plane must lie on the projection l_2 of the line that connects m_1 to m (see Figure 3). This line is called the *epipolar line* and can be found for each point m_1 in the first image plane by intersecting the plane defined by o_1,o_2 and m_1 (the *epipolar plane*) with the second image plane. The epipolar constraint can be enforced exactly only if the position of m_1 and m_2 and the camera parameters are known without error. In practice, however, there will always be some distance between a projected point and the epipolar line it should belong to. This discrepancy is a useful measure for verification tasks such as the detection of outliers among alleged matching image points, or the evaluation of the quality of estimated camera parameters. The epipolar constraint can be expressed algebraically in a straightforward manner. If we know the rotation matrix and translation vector that move one camera reference system to the other we have that:

$$\mathbf{x_1^T E x_2} = \mathbf{x_1^T} \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \mathbf{R x_2} = 0\,,$$

where the *essential matrix* $\mathbf{E}$ is the product between the cross product matrix of the translation vector $\mathbf{T}$ and the rotation matrix $\mathbf{R}$, and x_1 and x_2 are points expressed in the reference systems of the first and second camera respectively, belonging to the same epipolar plane. If the calibration matrices of both cameras are known, this constraint can also be expressed in terms of image points by applying the inverse of the two calibration matrices to the image points:

$$(\mathbf{K_1}^{-1}\mathbf{m_1})^T\mathbf{E}(\mathbf{K_2}^{-1}\mathbf{m_2}^T) =$$
$$\mathbf{m_1}^T(\mathbf{K_1}^{-1T}\mathbf{E}\mathbf{K_2}^{-1})\mathbf{m_2} = 0\,,$$

Where $\mathbf{F} = \mathbf{K_1}^{-1T}\mathbf{E}\mathbf{K_2}^{-1}$ is called the *fundamental matrix.* It is clear that if intrinsic camera parameters are known the epipolar constraint can be verified on image points by using just the essential matrix, which has only five degrees of freedom; otherwise it is necessary to resort to the use of the fundamental matrix, which has seven degrees of freedom. Many algorithms are known to estimate both **E** or **F** from image point correspondences (Hartley 1995; Zhang et al 1995; Torr and Zisserman 1998).

2.3 Structure from Motion

Structure from Motion (SfM) has been a core Computer Vision topic for a long time and a large number of different problem formulations and algorithms have been introduced over the last few decades (Aggarwal et al 2010; Weng et al 1993; Zhang 1995). The distinctive traits of many SfM techniques recently proposed in literature are usually to be found in the approach used for the initial estimate and in the refinement technique adopted. In general this refinement happens by iteratively applying a bundle adjustment algorithm (Triggs et al 2000) to an initial set of correspondences, 3D points and motion hypotheses. This optimization is almost invariably carried out by means of the Levenberg-Marquardt algorithm (Levenberg 1944), which is very sensitive to the presence of outliers in the input data. For this reason any possible care should be taken in order to supply the optimizer with good hypotheses or at least a minimal number of outliers. When a reasonable subset of all the points is visible in all the images global methods can be used to obtain such initial hypothesis. This approach, commonly called *factorization*, was initially proposed only for simplified camera models that are not able to fully capture the pinhole projection (Tomasi and Kanade 1992; Weinshall and Tomasi 1995). More recently, similar approaches have been presented also for perspective cameras (Sturm and Triggs 1996; Heyden et al 1999), however their need for having each point visible in each camera severely reduces their usability in practical scenarios where occlusion is usually abundant. For this reason incremental methods, which allow to add one or a few images at a time, are by far more popular in SfM applications. Usually such methods start from a reliable image pair (for instance the pair with the higher number of good correspondences), then an initial reconstruction is obtained by triangulation and finally extended sequentially. The extension can happen by virtue of common 2D points between a new camera and one or more images already in the batch. If internal camera parameters are known (at least roughly) rotation and translation direction can be extracted from the essential matrix and translation magnitude can be found using the projection in the new image of an already reconstructed 3D point. In the more general case intrinsic parameters are not known and the new camera can be added by exploiting the correspondences between its 2D features and previously triangulated 3D points to estimate the projection matrix (Beardsley et al 1997; Pollefeys et al 1999). Finally, it is possible to merge partial reconstructions by using corresponding 3D points (Fitzgibbon and Zisserman 1998). Many modern approaches iterate this process by including and excluding point correspondences or entire images by validating them with respect to the currently estimated structure and camera poses (Brown and Lowe 2005; Vergauwen and Van Gool 2006; Snavely et al 2008).

3 Non-Cooperative Games for Inlier Selection

The selection of matching points based on the feature descriptors is only able to exploit local information. This limitation conflicts with the richness of information that is embedded in the scene structure. For instance, under the assumption of rigidity and small camera motion, intuition suggests that features that are close in one view cannot be too far apart in the other one. Further, if a pair of features exhibit a certain difference of angles or ratio of scales, this relation should be maintained among their respective matches. Our basic idea is to formalize this intuitive notion of consistency between pairs of feature matches into a real-valued compatibility function and to find a large set of matches that express a high level of mutual compatibility. Of course, the ability to define a meaningful pairwise compatibility function and a reliable technique for finding a consistent set is at the basis of the effectiveness of the approach. Following (Torsello et al 2006; Albarelli et al 2009), we model the matching process in a Game-Theoretic framework, where two players select a pair of matching points from two images. Each player then receives a payoff proportional to how compatible his match is with respect to the other player's choice. Clearly, it is in each player's interest to pick matches that are compatible with those the other players are likely to choose. In general, as the game is repeated, players will adapt their behavior to prefer matchings that yield larger payoffs, driving all inconsistent hypotheses to extinction, and settling for an equilibrium where the pool of matches from which the players are still actively selecting their associations forms a cohesive set with high mutual support. Within this formulation, the solutions of the matching problem correspond to evolutionary stable states (ESS's), a robust population-based generalization of the notion of a Nash equilibrium. In a sense, this matching process can be seen as a contextual voting system, where each time the game is repeated the previous selections of the other players affect the future vote of each player in an attempt to reach consensus. This way the evolving context brings global information into the selection process. Since the evolutionary process is driven entirely by the payoff between strategies, it is clear that by adopting an appropriate compatibility function it is possible to suit the framework to achieve different goals. In this paper we will introduce two payoff functions to address our multi-view point matching problem. In Section 3.2 we will define a compatibility among pairs of correspondences that is proportional to the similarity of the affine transformation inferred from each match; this is done to exploit the expected local spatial and scale coherence among image patches. In Section 3.3 we will propose a refinement step that filters out groups of matches by letting them play an evolutionary game where the payoff is bound to their mutual ability to comply with the epipolar constraint.

3.1 Game-Theoretic Selection

Originated in the early 40's, Game Theory was an attempt to formalize a system characterized by the actions of entities with competing objectives, which is thus hard to characterize with a single objective function (Weibull 1995). According to this view, the emphasis shifts from the search of a local optimum to the definition of equilibria between opposing forces, providing an abstract theoretically-founded framework to model complex interactions. In this setting multiple players have at their disposal a set of strategies and their goal is to maximize a payoff that depends also on the strategies adopted by other players.

Here we will concentrate on symmetric two player games, i.e., games between two players that have the same set of available strategies and that receive the same payoff when playing against the same strategy. More formally, let $O = \{1, \cdots, n\}$ be the set of available strategies (*pure strategies* in the language of Game-Theory), and $C = (c_{ij})$ be a matrix specifying the payoffs, then an individual playing strategy i against someone playing strategy j will receive payoff c_{ij}. A *mixed strategy* is a randomization of the available strategies, i.e., a probability distribution $\mathbf{x} = (x_1, \ldots, x_n)^T$ over the set O. Clearly, mixed strategies are constrained to lie in the n-dimensional standard simplex

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

The *support* of a mixed strategy $\mathbf{x} \in \Delta$, denoted by $\sigma(\mathbf{x})$, is defined as the set of elements chosen with nonzero probability: $\sigma(\mathbf{x}) = \{i \in O \mid x_i > 0\}$. The expected payoff received by a player choosing element i when playing against a player adopting a mixed strategy $\mathbf{x}$ is $(C\mathbf{x})_i = \sum_j c_{ij} x_j$, hence the expected payoff received by adopting the mixed strategy $\mathbf{y}$ against $\mathbf{x}$ is $\mathbf{y}^T C\mathbf{x}$. The *best replies* against mixed strategy $\mathbf{x}$ is the set of mixed strategies

$$\beta(\mathbf{x}) = \{\mathbf{y} \in \Delta \mid \mathbf{y}^T C\mathbf{x} = \max_{\mathbf{z}}(\mathbf{z}^T C\mathbf{x})\} .$$

The best reply is not necessarily unique. Indeed, except in the extreme case in which there is a unique best reply that is a pure strategy, the number of best replies is always infinite. A central notion of Game-Theory is that of a Nash equilibrium. A strategy $\mathbf{x}$ is said to be a *Nash equilibrium* if it is the best reply to itself, i.e., $\forall \mathbf{y} \in \Delta, \ \mathbf{x}^T C\mathbf{x} \geq \mathbf{y}^T C\mathbf{x}$. This implies that $\forall i \in \sigma(\mathbf{x})$ we have $(C\mathbf{x})_i = \mathbf{x}^T C\mathbf{x}$; that is, the payoff of every strategy in the support of $\mathbf{x}$ is constant. The idea underpinning the concept of Nash equilibrium is that a rational player will consider a strategy viable only if no player has an incentive to deviate from it.

We undertake an evolutionary approach to the computation of Nash equilibria. Evolutionary Game-Theory originated in the early 70's as an attempt to apply the principles and tools of Game-Theory to biological contexts. It considers an idealized scenario where pairs of individuals are repeatedly drawn at random from a large population to perform a two-player game. In contrast to traditional Game-Theoretic models, players are not supposed to behave rationally, but rather act according to a pre-programmed behavior, or mixed strategy. Further, it is supposed that some selection process operates over time on the distribution of behaviors favoring players that receive higher payoffs.

In this dynamic setting, the concept of stability, or resistance to invasion by new strategies, becomes central. A strategy $\mathbf{x}$ is said to be an *evolutionary stable strategy* (ESS) if it is a Nash equilibrium and

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

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

The search for a stable state is performed by simulating the evolution of a natural selection process. Under very loose conditions, any dynamics that respect the payoffs is guaranteed to converge to Nash equilibria (Weibull 1995) and (hopefully) to ESS's; for this reason, the choice of an actual selection process is not crucial and can be driven mostly by considerations of efficiency and simplicity. We chose to use the replicator dynamics (Taylor and Jonker 1978), a well-known formalization of the selection process governed by the following equation

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

where $\mathbf{x}_i$ is the i-th element of the population and C the payoff matrix.

A point x is said to be a *stationary* (or equilibrium) point of our dynamical system, if $\dot{x}_i = 0$, for all $i = 1, \ldots, n$. A stationary point x is said to be *asymptotically stable* if any trajectory starting sufficiently close to x converges to x.

It can be shown (Weibull 1995) that a point $x \in \Delta$ is the limit of a trajectory of the replicator dynamics starting from the interior of Δ if and only if it is a Nash equilibrium. Further, if point $x \in \Delta$ is an ESS, then it is asymptotically stable for the replicator dynamics.

In our approach, we let matches compete with one another, each obtaining support from compatible associations and competitive pressure from all the others. The selection process is simulated by running the recurrence (2) and, at equilibrium, only pairings that are mutually compatible should survive and are then taken to be inliers.

3.2 Affine Preserving Matching Game

Central to this framework is the definition of a *matching game*, or, specifically, the definition of the strategies available to the players and of the payoffs related to these strategies. Given a set M (model) of feature points in a source image and a set D (data) of features in a target image, we call a *matching strategy* any pair (a_1, a_2) with $a_1 \in M$ and $a_2 \in D$. We call the set of all the matching strategies $S \subseteq M \times D$. The total number of matching strategies in S can, in theory, be as large as the Cartesian product of the sets of features detected in the images. Since most interest point detectors extract thousands of features from an image, a suitable selection should be made in order to keep its size limited. To this end we can exploit unary information such as the distance between descriptors or the photo-consinstency of local image patches to select only feasible pairs. Specifically, for each source feature we can generate k matching strategies that connect it to the k nearest destination features in terms of descriptor distance. Since our Game-Theoretic approach operates inlier selection regardless of the descriptor, we do not need to set any threshold with respect to the absolute descriptor distance or the distinctiveness between the first and the second nearest point. In this sense, the only constraint that we need to impose over k is that it should be large enough that we can expect the correct correspondence to be among the candidates for a significant proportion of the source features. In our preliminary work (Albarelli et al 2010) we already analyzed the influence of k over the quality of the matches obtained and we found that a very small amount of candidates (typically 3 or 4) are enough to guarantee a satisfactory performance, however, in the presence of highly repeating patterns, a larger value might be needed. By reducing the number of correspondences per source feature to a constant value, we limit the growth of the number of strategies to be linear with the number of (source) features to be matched.

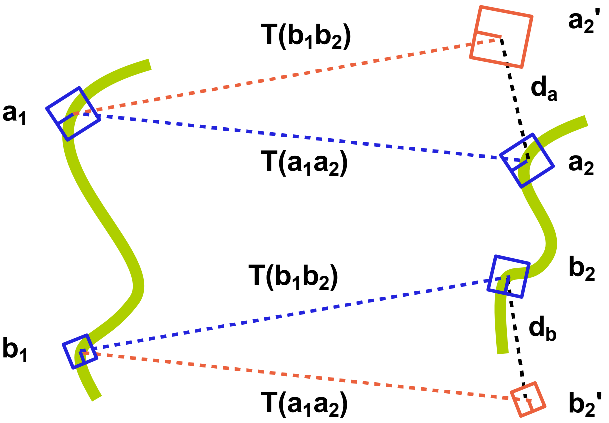
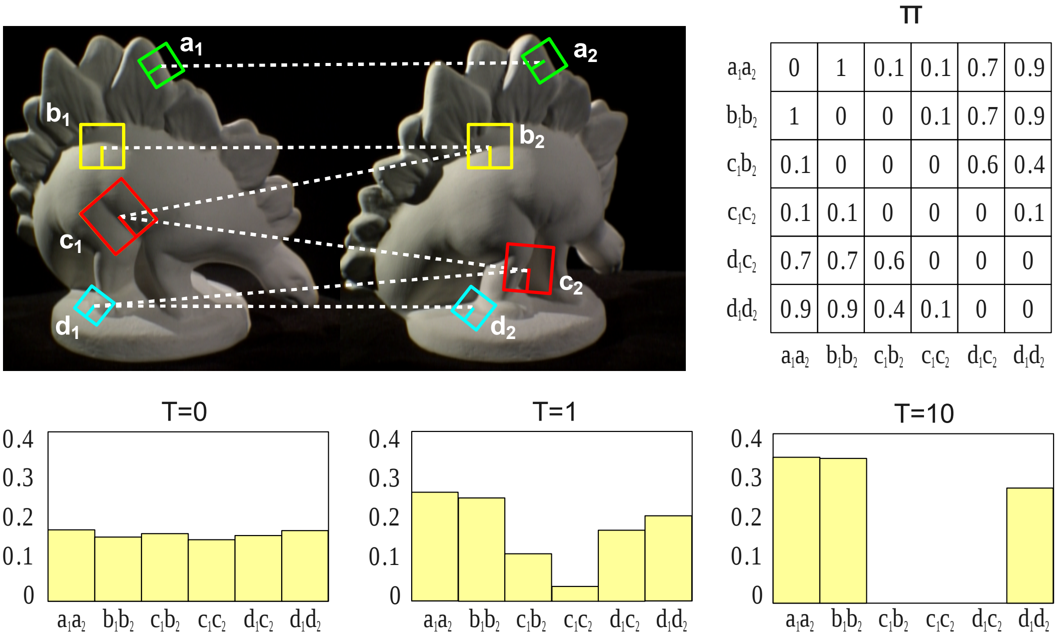


Fig. 4 The payoff between two matching strategies is inversely proportional to the maximum reprojection error obtained by applying the affine transformation estimated by a match to the other.

Once S has been selected, our goal becomes to extract from it a large subset of correspondences that includes only correctly matched features: that is, strategies that associate a physical point in the source image with the same physical point (if visible) in the destination image. To this end, it is necessary to define a payoff function $\Pi : S \times S \rightarrow \mathbb{R}^+$ that exploits some pairwise information available at this early stage (i.e. before estimating camera and scene parameters) and that can be used to impose consistency globally. Since location, scale, and rotation are associated to each feature, we can associate to each correspondence (a, b) between feature a in the source image and feature b in the target image a similarity transform $T(a, b)$ that maps the neighborhood of a into the neighborhood of b, transforming the location, orientation, and scale measured in the source image into the location, orientation, and scale observed in the target image. Under small motion assumptions, we can expect these similarity transformations to be very similar locally. Thus, imposing the conservation of the similarity transform, we aim to extract clusters of feature matches that belong to the same region of the object and that tend to lie at the same level of depth. While this could seem to be an unsound assumption for general camera motion, in the experimental section we will show that it holds well with the typical disparity found in standard multiple view and stereo data sets. Further, it should be noted that with



$$\Pi$$

	a_1a_2	b_1b_2	c_1b_2	c_1c_2	d_1c_2	d_1d_2
a_1a_2	0	1	0.1	0.1	0.7	0.9
b_1b_2	1	0	0	0.1	0.7	0.9
c_1b_2	0.1	0	0	0	0.6	0.4
c_1c_2	0.1	0.1	0	0	0	0.1
d_1c_2	0.7	0.7	0.6	0	0	0
d_1d_2	0.9	0.9	0.4	0.1	0	0

Fig. 5 An example of the affine-based evolutionary process. Four feature points are extracted from two images and a total of six matching strategies are selected as initial hypotheses. The matrix Π shows the compatibilities between pairs of matching strategies according to a one-to-one similarity-enforcing payoff function. Each matching strategy got zero payoff with itself and with strategies that share the same source or destination point (i.e., $\Pi((b_1, b_2), (c_1, b_2)) = 0$). Strategies that are coherent with respect to a similarity transformation exhibit high payoff values (i.e., $\Pi((a_1, a_2), (b_1, b_2)) = 1$ and $\pi((a_1, a_2), (d_1, d_2)) = 0.9$)), while less compatible pairs get lower scores (i.e., $\pi((a_1, a_2), (c_1, c_2)) = 0.1$). Initially (at T=0) the population is set to the barycenter of the simplex and slightly perturbed. After just one iteration, (c_1, b_2) and (c_1, c_2) have lost a significant amount of support, while (d_1, c_2) and (d_1, d_2) are still played by a sizable amount of population. After ten iterations (T=10) (d_1, d_2) has finally prevailed over (d_1, c_2) (note that the two are mutually exclusive). Note that in the final population $((a_1, a_2), (b_1, b_2))$ have a larger support than (d_1, d_2) since they are a little more coherent with respect to similarity.

large camera motion, most, if not all, commonly used feature detectors fail, thus any inlier selection attempt becomes meaningless.

In order to define the payoff function Π we need a way to measure the distance between similarity transforms. In order to avoid the problem of mixing incommensurable quantities, we compute the distance in terms of the reprojection error expressed in pixels. Specifically, given two matching strategies (a_1, a_2) and (b_1, b_2) and their respective associated similarities $T(a_1, a_2)$ and $T(b_1, b_2)$, we calculate virtual points a'_2 and b'_2 by applying the other strategy transformation to the source features a_1 and b_1 (see Figure 4). More formally,

$$a'_2 = T(b_1, b_2)a_1$$
$$b'_2 = T(a_1, a_2)b_1 \,,$$

Given virtual points a'_2 and b'_2, we can measure the similarity between (a_1, a_2) and (b_1, b_2) as:

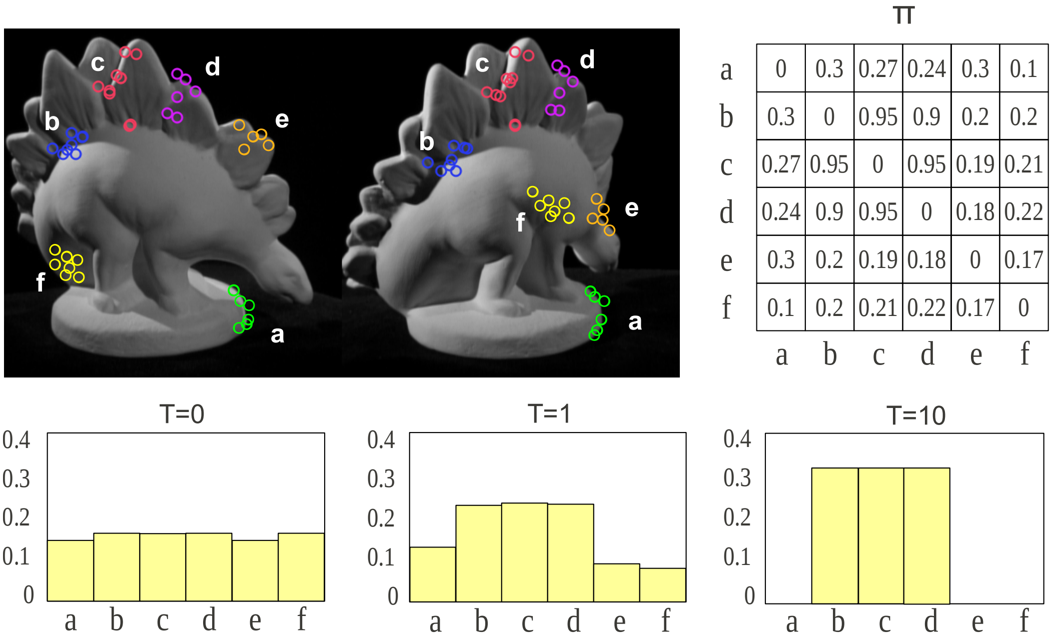
$$\mathrm{sim}((a_1, a_2), (b_1, b_2)) = e^{-\lambda max(|a_2 - a'_2|, |b_2 - b'_2|)} \qquad (3)$$

where λ is a selectivity parameter: If λ is small, then the similarity function (and thus the matching) is more tolerant with respect to deviation in the similarity transformations, becoming more selective as λ grows. Since each source feature can correspond with at most one destination point, it is desirable to avoid any kind of multiple match. It is easy to show that a pair of strategies with zero mutual payoff cannot belong to the support of an ESS (see (Albarelli et al 2009)), thus any payoff function Π can be easily adapted to enforce one-to-one matching by defining:

$$\Pi((a_1, a_2), (b_1, b_2)) = \begin{cases} \mathrm{sim}((a_1, a_2), (b_1, b_2)) & a_1 \neq b_1, \\ & a_2 \neq b_2 \\ 0 & \text{else} \end{cases} \qquad (4)$$

We define payoff (4) a *similarity enforcing payoff function* and we call an *affine matching game* any symmetric two player game that involves a matching strategies set S and a similarity enforcing payoff function Π.

The main idea of the proposed approach is that by playing a matching game driven by a similarity enforcing payoff function such as (4), the strategies (i.e. correspondence candidates) that share a similar locally affine transformation are advantaged from an evolutionary point of view and shall emerge in the surviving population. In Figure 5 we illustrate a simplified ex-



Π

	a	b	c	d	e	f
a	0	0.3	0.27	0.24	0.3	0.1
b	0.3	0	0.95	0.9	0.2	0.2
c	0.27	0.95	0	0.95	0.19	0.21
d	0.24	0.9	0.95	0	0.18	0.22
e	0.3	0.2	0.19	0.18	0	0.17
f	0.1	0.2	0.21	0.22	0.17	0

Fig. 6 An example of the selection of groups of features that agree with respect to a common epipolar geometry. Six matching groups are selected by the affine matching step (labelled from a to f in the figure). Each pair of feature sets is modeled as a matching strategy and the payoff among them is reported in matrix Π. Note that groups b,c and d are correctly matched and thus exhibit a high mutual payoff. By contrast, group a (which is consistent both in terms of photometric and affine properties), e and f are clearly mismatched with respect to the overall scene geometry, which in turn leads to a large error on the epipolar check and thus to a low score in the payoff matrix. At the beginning of the evolutionary process each strategy obtains a fair amount of players (T=0). As expected, after just one iteration of the replicator dynamics the most consistent strategies (b, c and d) obtain a clear advantage. Finally, after ten iterations (T=10) the other groups have no more support in the population and only the correct matches survived.

ample of this process. Once the population has reached a local maximum, all the non-extinct mating strategies can be considered valid, however, technically strategies become truly extinct only after an infinite number of iterations. Since we halt the evolution when the population ceases to change significantly, it is necessary to introduce some criteria to distinguish correct from non-correct matches. To avoid a hard threshold we chose to keep as valid all the played strategies whose population size exceeds a percentage of the most popular strategy. We call this percentage *quality threshold* (q). This criterion further limits the number of selected strategies, but increases their consistency, since the population proportion is linked to the coherence of the strategy with the other surviving strategies. Each evolution process selects only a single group of matching strategies that are mutually coherent with respect to a local similarity transformation. This means that if we want to cover a large portion of the image we need to iterate the process many times, pruning the previously selected matches at each new iteration. Note that by imposing a minimal size for a group to be deemed as valid, the odds of recognizing structured outliers as false positives get lower. In fact, the probability of a large group to be coherent with respect to local affinity by chance is reduced as the minimal group size increases. Of course the usual trade-off between the desired precision and recall parameters must be taken into account when setting this kind of threshold.

3.3 Refinement by Epipolar Constraint Enforcement

The game formulation we just introduced shifts the matching problem to a more global scope by producing a set of correspondences between groups of features. While the affine camera model extracts very coherent groups, making such *macro features* more robust and descriptive than single points, in principle there is nothing that prevents the system to still produce wrong or weak matches. To reduce this chance we propose a different game setup that allows for a further refinement. In this game the strategies set S corresponds to the set of paired feature groups extracted from the affine matching game and the payoff between them is related to the features' agreement to a common epipolar geometry. More specifically, given two pairs of matching groups $a \subseteq M \times D$ and $b \subseteq M \times D$, each one made up of model and data features, we estimate the epipolar geometry from $a \cup b$ and define the payoff among them as:

$$\Pi(a, b) = e^{-\lambda \sum_{(s,t)\in a\cup b} d(t,l(s))} \qquad (5)$$

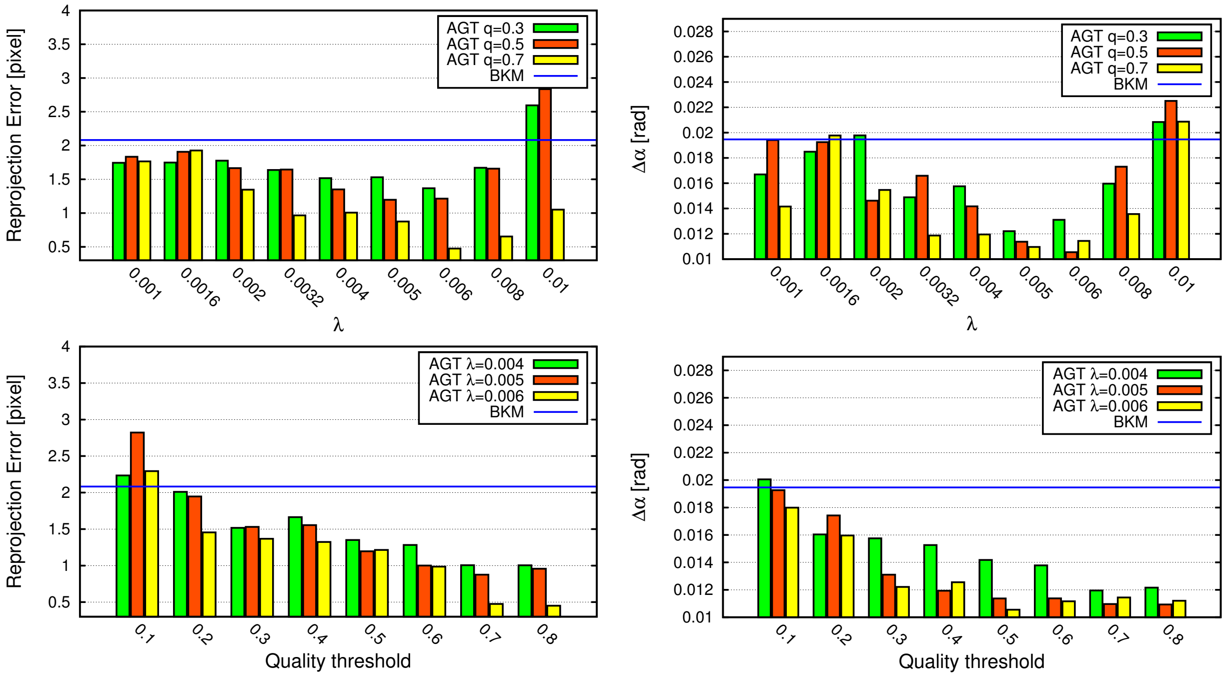


Fig. 7 Analysis of the performance of the Affine Game-Theoretic approach with respect to variation of the parameters of the algorithm.

Where $l(p)$ is a function that gives the epipolar line in the data image from the feature point p in the model image, according to the estimated epipolar geometry, and $d(p, l)$ calculates the distance between point p and the epipolar line l. It is clear that this distance is small (and thus the payoff is big) if the two groups share a common projective interpretation and large otherwise. Of course, different pairs of groups can agree on different epipolar geometry, but the transitive closure induced by the selection process ensures that the strategies in the surviving population will agree on the same (or very similar) projective transformation (see Figure 6 for a complete example of this process). Regarding the estimation of the epipolar geometry, it can be done in two different ways: if we have at least the intrinsic calibration of the camera we can estimate the essential matrix, by contrast, if we do not have any hint about the camera geometry, we must resort to a more relaxed set of constraints and use the fundamental matrix instead. In the experimental section we will test both scenarios.

4 Experimental Results

We performed an extensive set of tests in order to validate the proposed techniques and to explore their limits. Both quantitative and qualitative results are shown and performances are compared with those achieved by a standard baseline method, i.e. the default feature matcher in the Bundler suite (Snavely et al 2008).

4.1 General Setup and Data Sets

All the following experiments have been made by applying a common basic pattern: first a set of features is extracted from the images by using the SIFT keypoint detector made freely available in (Lowe 2003), then these interest points are paired using the matcher we want to test, finally scene and camera parameters are estimated by using the final portion of Bundler pipeline (i.e. the part of the suite that applies Levenberg-Marquardt optimization to a set of proposed matches). We evaluate three different approaches: The first, referred to as Affine Game-Theoretic approach (AGT), uses the affine matching game without the further refinement provided by the enforcement of the epipolar geometry. In this case the iterative extraction and elimination of the groups is image-based, i.e., after a group of matches is selected, all the matches that have sources or targets close to the source and target points of the extracted correspondences are eliminated, and then the evolutionary process is reiterated on the reduced set of strategies. The process is stopped when an extracted group is smaller than a given threshold or has average payoff smaller than a given threshold. This approach is the same described in (Albarelli et al 2010). The second and third approaches, referred to as Calibrated Projective Game-Theoretic approach (CPGT) and Uncalibrated Projective Game-Theoretic approach (UPGT) respectively, make use of the epipolar refinement. CPGT assumes that the camera intrinsic parameters are (ap-

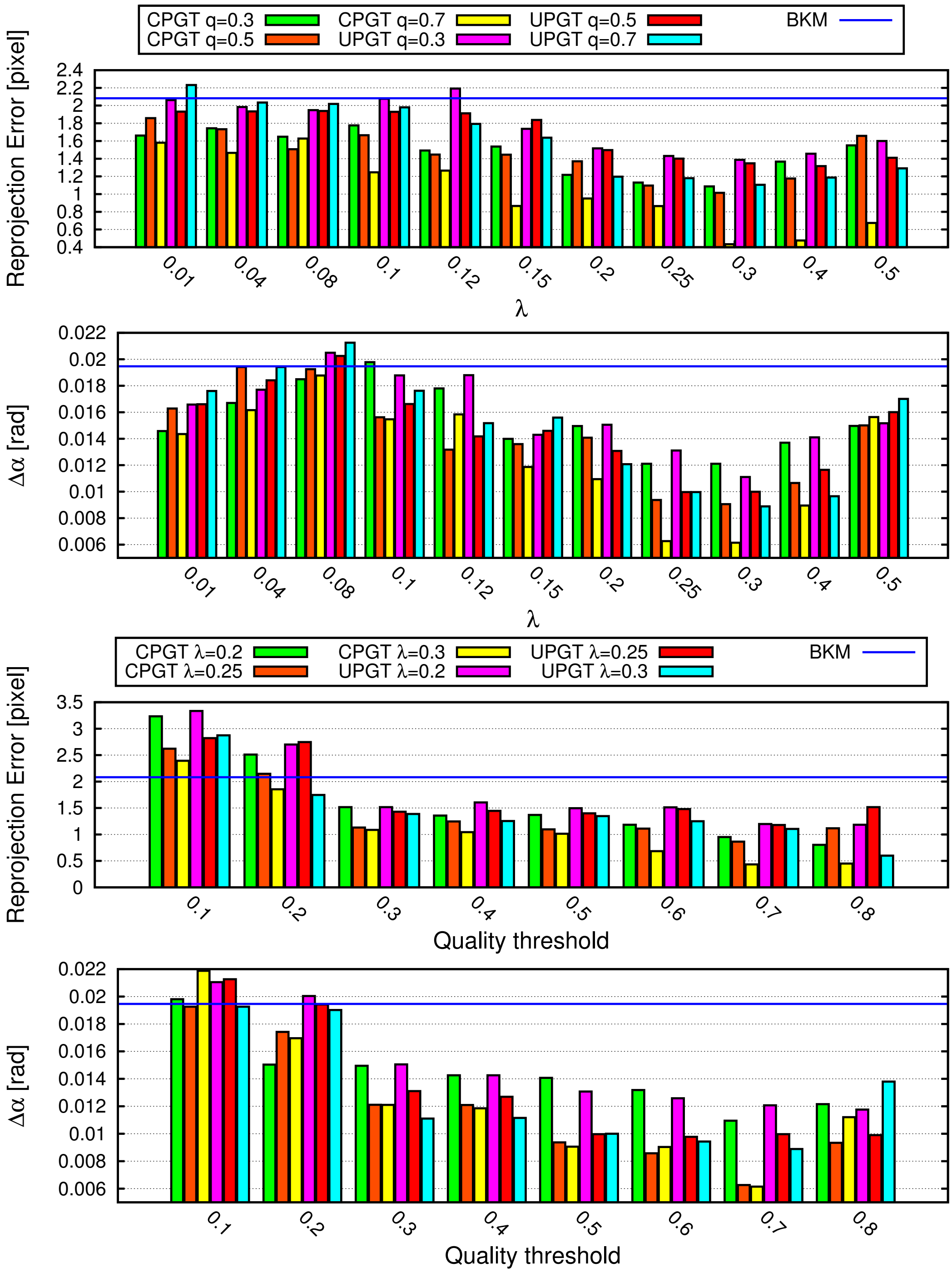


Fig. 8 Analysis of the performance of the Calibrated and Uncalibrated Projective Game-Theoretic approaches with respect to variation of the parameters of the algorithm.

proximately) known and estimate the epipolar geometry through the essential matrix, while UPGT uses the fundamental matrix. In both these approaches the iterative extraction and elimination of the groups is strategy-based, i.e., after a group of matches is selected only those matches are eliminated from the strategy set, thus allowing for the same features to appear in several groups, while the stopping criterion here is the same as that of AGT. In our experiments the intrinsic parameters for CPGT have been estimated from the images EXIF information. The three approaches are compared against the default feature matcher in the Bundler suite (BKM). This is a reasonable choice for several reasons: BKM is optimized to work with SIFT descriptors and, obviously, with the Bundler suite; in addition it is very popular in literature since Bundler itself has been used as the default matcher in many of the recent papers about SfM and dense stereo reconstruction. For each test we evaluated two quality measures: the average reprojection error (expressed in pixels) and the differences in radians between the ground-truth and the estimated rotation angle ($\varDelta\alpha$). The first measure aims to capture the cumulative error made in the reconstruction of the structure and the estimation of the motion, while the second measure aims to decouple the error on the camera orientation from the one related to the scene reconstruction. This is possible since we used images pairs coming from a calibrated camera head or image sets with an available ground-truth. Specifically we used a pair of cameras previously calibrated through a standard procedure and took stereo pictures of 20 different, isolated objects; in addition we also included in the data set the shots coming from the "DinoRing" and "TempleRing" sequences from the Middlebury Multi-View Stereo dataset (Seitz et al 2006). We conducted two main sets of experiments. The goal of the first set is to analyze the impact of the parameters, namely λ and *quality threshold* (q), over the accuracy of the results. Since AGT and CPGT/UPGT have different payoff functions and the selectivity λ is not directly comparable we investigate its influence separately. In addition, all the experiments regarding the refinement methods are made using very relaxed parameters for the AGT step. This is due to the fact that we are willing to accept a slightly higher number of outliers in the first step in exchange for a higher number of candidate groups, in the hope that the refinement process is able to eliminate the spurious groups, but still resulting in a larger number of good correspondences from which to perform parameter estimation. In the second batch of experiments we compare our techniques with the default Bundler matcher. In these experiments the parameters are set to the optimal values estimated previously. We provide both quantitative and qualitative results: the quantitative analysis is based on the errors in reprojection and motion estimation, while the qualitative results are based on a dense reconstruction obtained using the recovered parameters as an input to the PMVS suite (Furukawa and Ponce 2010).

4.2 Influence of Parameters

The AGT method depends on two explicit parameters: the sensitivity parameter λ, which modulates the steepness of the payoff function (4), and q, i.e. the percentage of population density with respect to the most represented strategy that one match must obtain to be considered not-extinct. As stated in Section 3.2, λ controls the selectivity of the selection process, while q allows to further filter the extracted group based on its cohesiveness. Higher values will lead to a more selective culling, while lower values will allow more strategies to pass the screening. Figure 7 reports the results of these experiments averaged over the full set of 20 stereo pairs taken with a previously calibrated camera pair. The first row shows the effect of the selectivity parameter λ. This is evaluated for three different q levels, from 0.3 to 0.7. As expected, both low and high values lead to larger errors, mainly with respect to the estimation of the angle between the two cameras. This is probably due to a too tight and a too relaxed enforcement of local coherence respectively. It could be argued that the estimation of the optimal λ can be tricky in practical situations; however, we must note that, with a reasonable high q, it takes a very large sensitivity parameter to obtain a worse performance than that obtained with the default Bundler matcher. Regarding the quality threshold, we can see in the second row of Figure 7 that the best results are achieved by setting a high level of quality: this is clearly due to the fact that, in practice, the replicator dynamics have converged to a stable ESS and thus most of the non-zero strategies are indeed inliers and are mostly subject only to the (small) feature localization error, thus exhibiting an equally high density. In Figure 8 we show the results obtained by trying different parameters with CPGT and UPGT. As previously stated, these experiments were made by performing an affine matching step with relaxed parameters: namely a λ value of 0.09 and a q of 0.6. The overall behavior with respect to these parameters is similar to what observed with AGT: very low and very high values for λ lead to less satisfactory results (whereas in general better than those obtained with the Bundler key matcher), and high q seems to guarantee good estimates. Overall it seems that CPGT always gives better results than UPGT. We will analyze this behavior with more detail in the next section.

4.3 Comparisons between Approaches

To further explore the differences among the proposed techniques and the Bundler matcher, we executed two sets of experiments. The first set applies the approaches to unordered images coming from the DinoRing and TempleRing sequences from the Middlebury Multi-View Stereo dataset for these models, the camera extrinsic parameters are provided and used as a ground-truth. The rationale for using these sets (in opposite to simple stereo pairs) is to allow Bundler to optimize the

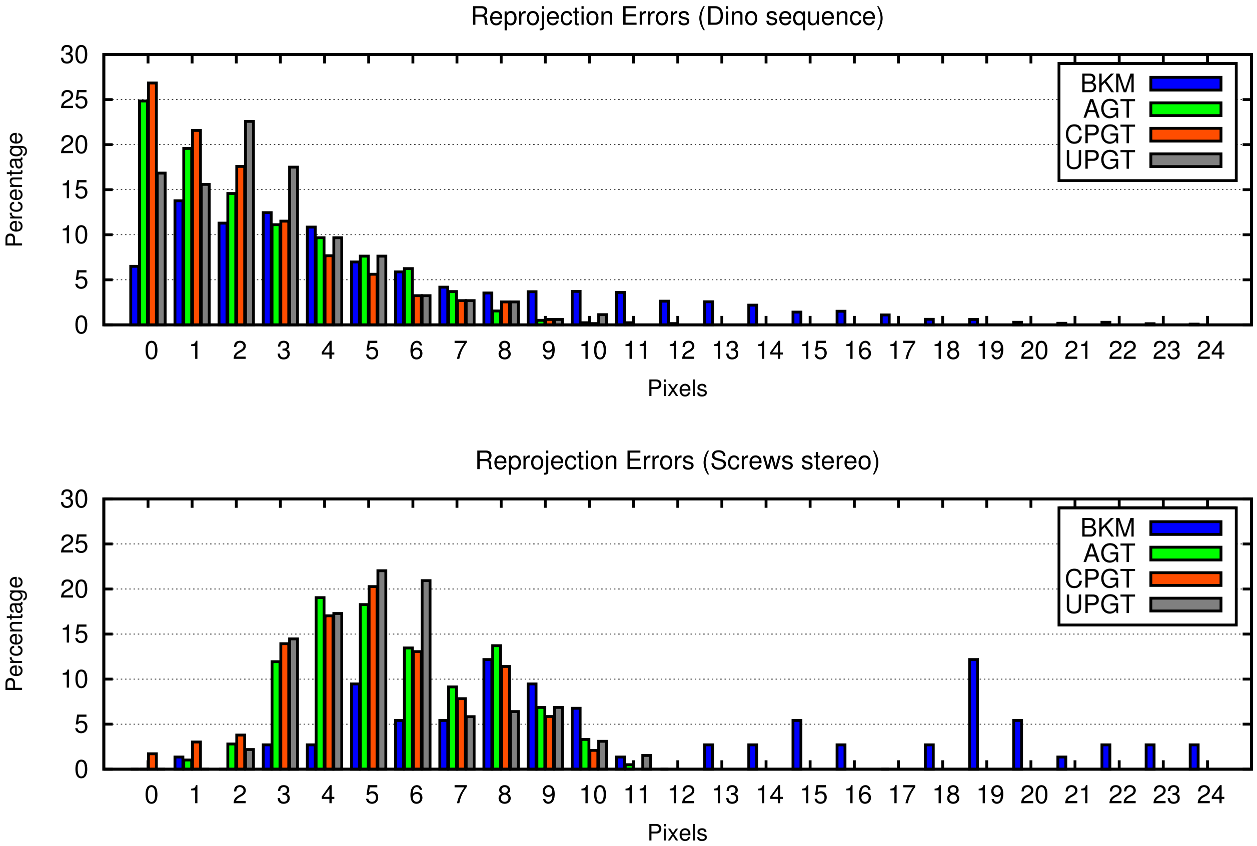
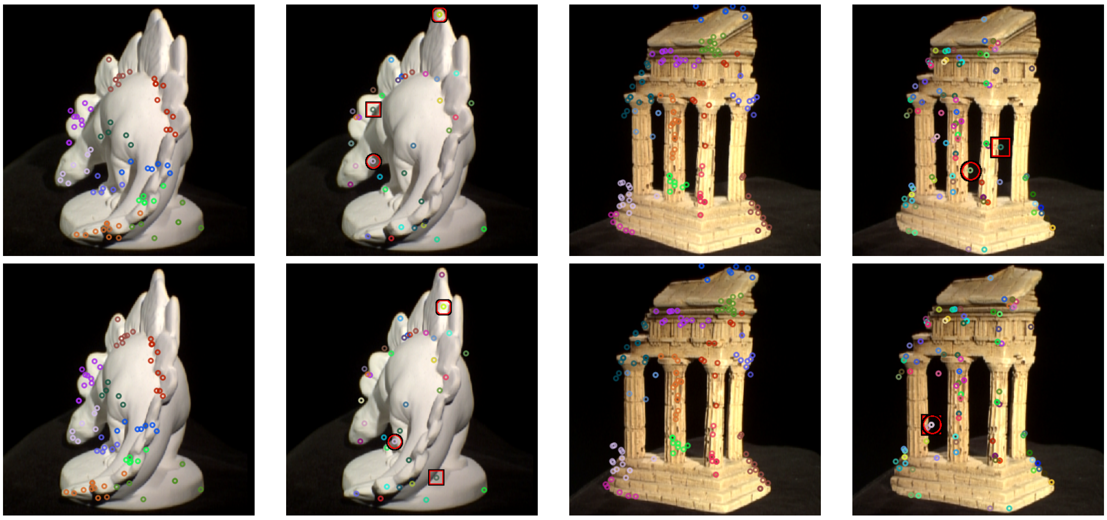


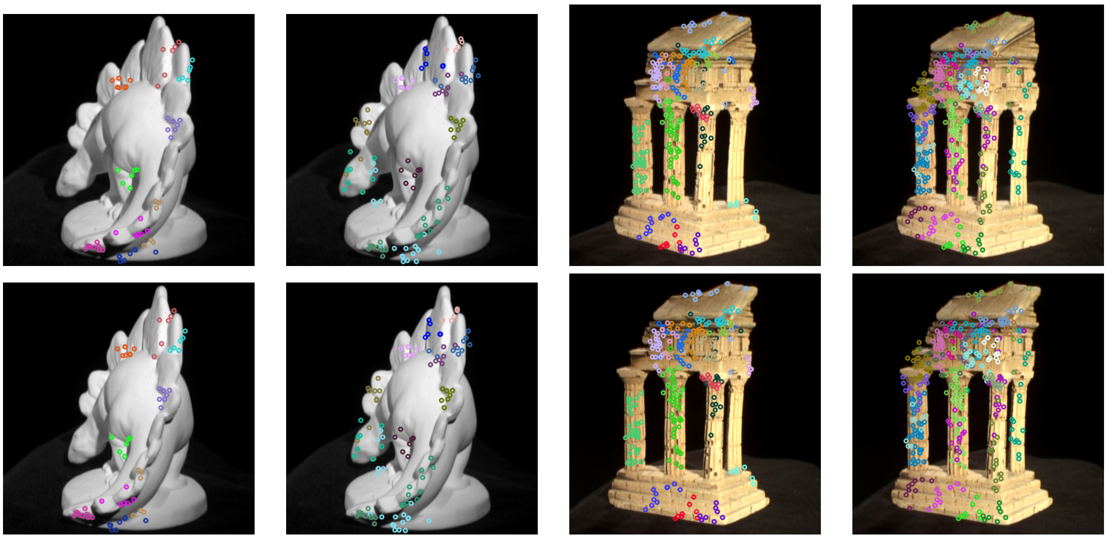
Fig. 9 Distribution of the reprojection error in one multiple view (top) and one stereo pair (bottom) example.

parameters and correspondences over the complete sequence. The second set is composed of two calibrated stereo scenes selected from the previously acquired collection of 20 items, specifically a statue of Ganesha and a handful of screws placed on a table. For all the sets of experiments we evaluated both the rotation error of all the cameras and the reprojection error of the detected feature points. In the Middlebury sets the results are presented as averages. The Dino model is a difficult case in general, as it provides very few distinctive features; the upper part of Figure 10 shows the correspondences produced by AGT (left column) in comparison with BKM (right column). The parameters were set to the optimal values estimated in the previous experiments ($\lambda = 0.06$ and $q = 0.8$). This resulted in the detection of many correct matches organized in groups, each corresponding to a different depth level, and visualized with a unique color in the figure. As can be seen, the different depth levels are properly estimated; this is particularly evident throughout the arched back going from the tail (in foreground) to the head of the model (in background), where clustered sets of feature points follow one after the other. Furthermore, these sets of interest points maintain the right correspondences within the pair of images. The Bundler matcher on the other hand, while still achieving good results in the whole process, also outputs erroneous correspondences (marked in the figure). In the lower part of Figure 10 we can see the results obtained with CPGT and UPGT with $\lambda = 0.3$ and $q = 0.7$ after an affine matching step performed with $\lambda = 0.09$ and $q = 0.9$. We can observe that CPGT gives a significant boost to all the statistics. By contrast UPGT performed worse than AGT (albeit still better than BKM). This is probably due to the higher number of degrees of freedom in the estimation of the fundamental matrix and, thus, to the reduced ability to discriminate incompatible groups. In fact, we can see that the size of the groups obtained with AGT is generally rather small (from 4 to about 10 points), and it is easy to justify such a small number of correspondences under a common fundamental matrix. The quality of reconstruction following the application of all methods can be compared visually by looking at the distribution of the reprojection error in the top row of Figure 9. While most reprojections fall within 1-3 pixels for the Game-Theoretic approaches, the Bundler matcher exhibits a long-tailed trend, with reprojection errors reaching 20 pixels. Unlike the Dino model, the Temple model is quite rich of features: for visualization purposes we only show a subset of the de-



Dino sequence **Temple sequence**

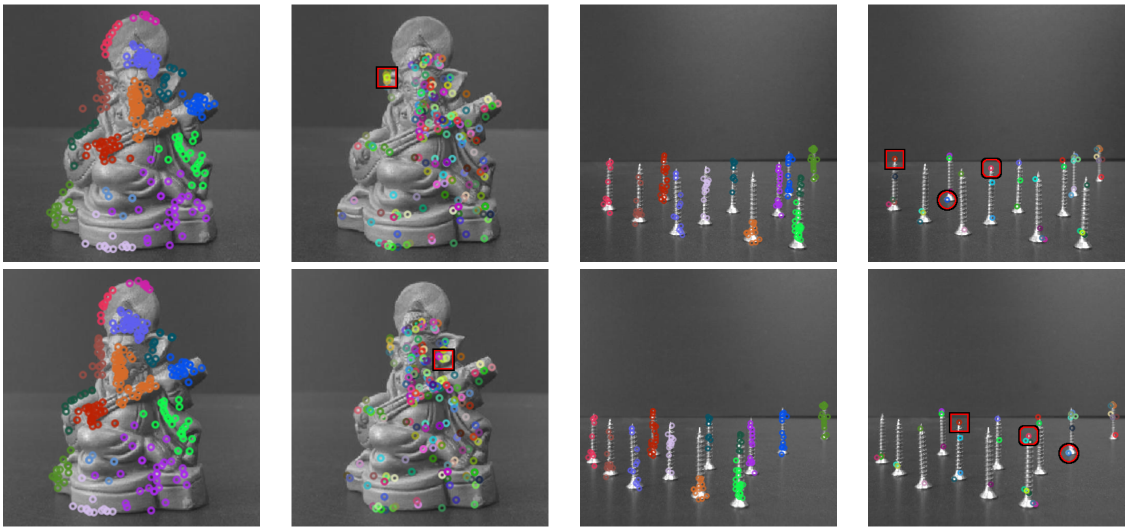
		AGT	BKM	AGT	BKM
Matches		14573	9245	25785	22317
ϵ	$\leq$ 1 pix	24.83	6.49406	22.6049	24.6729
	$\leq$ 5 pix	54.94	48.3659	62.7737	61.8957
	$\geq$ 5 pix	20.21	45.1401	14.6214	13.4314
	Avg.	2.3086	4.5255	2.3577	2.3732
$\Delta\alpha$	Avg.	0.005751	0.005561	0.010514	0.009376
	S. dev.	0.003242	0.003184	0.005282	0.004646
	Max	0.012057	0.011475	0.021527	0.017016
Avg. levels		8.42	-	9.27	-



Dino sequence **Temple sequence**

		CPGT	UPGT	CPGT	UPGT
Matches		15018	15231	28106	28407
ϵ	$\leq$ 1 pix	32.1731	20.0126	25.7232	18.3715
	$\leq$ 5 pix	61.4826	75.4671	64.5294	78.5347
	$\geq$ 5 pix	6.3518	4.5203	9.7474	3.0938
	Avg.	1.7051	2.9841	2.1642	3.6713
$\Delta\alpha$	Avg.	0.004823	0.006437	0.009411	0.01328
	S. dev.	0.003671	0.004514	0.005143	0.006545
	Max	0.013147	0.017421	0.019725	0.027832
Avg. levels		17.21	18.34	20.13	22.05

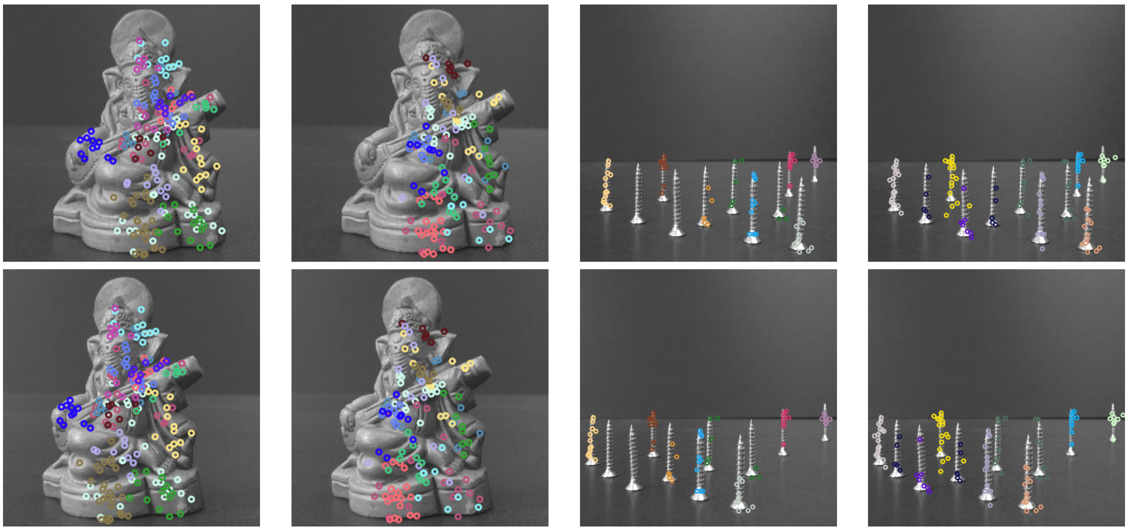
Fig. 10 Results obtained with two multiple view data sets (image best viewed in color).



Ganesha stereo

Screws stereo

		AGT	BKM	AGT	BKM
Matches		280	200	211	46
ϵ	≤ 1 pix	98.2824	20	0	0
	≤ 5 pix	1.7175	80	34.7716	6.75676
	≥ 5 pix	0	0	65.2284	93.2432
	Avg.	0.321248	1.67583	5.86237	10.2208
$\Delta\alpha$		0.001014	0.007424	0.020822	0.030995
Levels		14	-	12	-



Ganesha stereo

Screws stereo

		CPGT	UPGT	CPGT	UPGT
Matches		315	282	72	108
ϵ	≤ 1 pix	99.0017	83.4812	2.1637	0
	≤ 5 pix	0.9983	16.5188	37.5721	26.3417
	≥ 5 pix	0	0	60.2642	73.6583
	Avg.	0.300272	1.2311	3.92133	4.6379
$\Delta\alpha$		0.001623	0.00466	0.025341	0.03945
Levels		15	13	8	9

Fig. 11 Results obtained with two stereo view data sets (image best viewed in color).

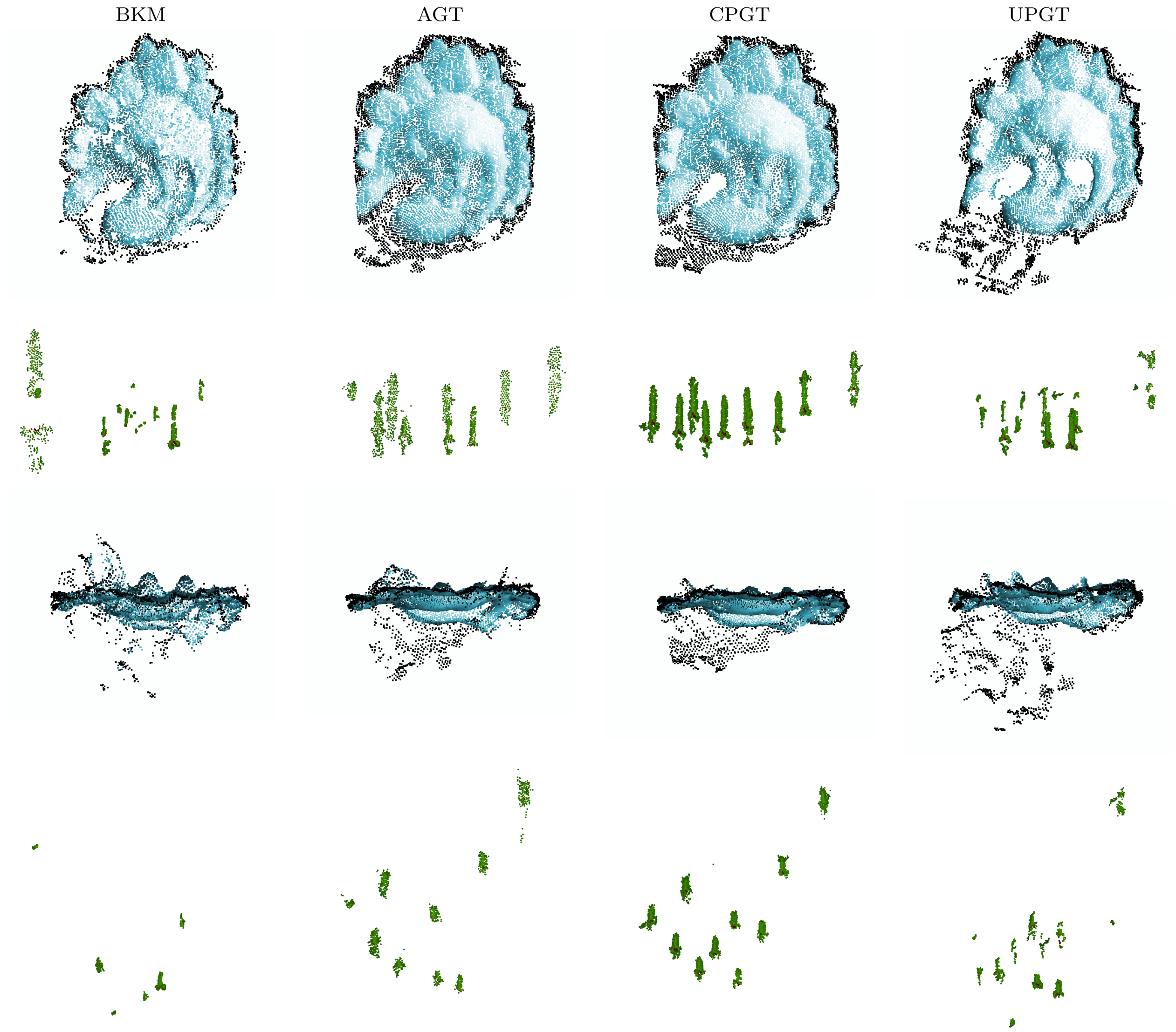


Fig. 12 Comparisons of the point clouds produced by PMVS using the motion estimated with different matching methods. Respectively the Bundler default keymatcher (BKM), the Affine Game-Theoretic technique (AGT) and the calibrated and uncalibrated projective techniques (CPGT and UPGT).

tected matches for all the techniques. While the effectiveness of our approaches is not negatively impacted by the model characteristics, several mismatches are extracted by BKM. In particular, the symmetric parts of the object (mainly the pillars) result in very similar features and this causes the matcher to establish one-to-many correspondences over them. In the calibrated stereo scenario, the Ganesha images are rich of distinctive features and pose no particular difficulty to any of the methods. The Bundler matcher provides very good results, with only one evident false match out of a total of 200 matches (see Figure 11). The resulting bundle adjustment is quite accurate, giving very small rotation errors and reprojection distances. Nevertheless, our methods perform considerably better: reprojection errors dramatically decrease, with around 98 percent of the feature points falling below one pixel of reprojection error for AGT and 99 percent for CPGT. Unfortunately UPGT is unable to refine the results obtained with AGT, but still achieves smaller errors than BKM. The second calibrated stereo scene, “Screws stereo”, is an emblematic case and provides some meaningful insight. The images depict a dozen screws standing on a table, placed by hand at different depth levels. This configuration, together with the abundance of features, should provide enough information for the algorithms to extract significant matches. However, the scene is a difficult one due to the very nature of the objects depicted, which are all identical and highly symmetric, resulting in several features with very similar descriptors and a difficulty in extracting good matches based only on photometric information. Indeed, several false matches are established by the Bundler matcher (see the last column of Figure 11). Still, BKM results in a reasonable estimation of the rigid transformation linking the two cameras, as erroneous pairings are removed *a posteriori* during the subsequent phases of bundle adjustment. By contrast, the AGT approach outputs large and accurate sets of matches, roughly one per object, and even difficult cases, such as the left-right parallactic swaps taking place at the borders are correctly dealt with. It is interesting to note that in this case the boost given by CPGT is even more significant than in the previous experiments, with a lower average reprojection error and an overall better error distribution. Unlike with the previous cases, this happens by reducing the number of total matches rather than increasing it, as the refinement process eliminates correspondences that are not globally consistent. In addition this time even UPGT gives better results than AGT: a histogram of the reprojection errors for this object is shown in Figure 9. Finally, a qualitative analysis of the different approaches is shown in Figure 12, where the estimated parameters and correspondences are fed to the PMVS dense multiview stereo reconstruction tool. The first and the second rows show the Dino and Screws scenes from a frontal view, while the other two show a top view of the same scenes. AGT and CPGT give the best results for Dino with CPGT providing a more correct representation of the hollow area between the neck and the first leg of the figurine and a smaller number of spurious points. With the screws scene CPGT allows by far the more consistent reconstruction, while BKM is substantially unable to offer to PMVS a satisfactory pose estimation.

4.4 Complexity and Running Time

With respect to complexity all the Game-Theoretic approaches are dominated by the steps of the replicator dynamics. Each step is quadratic in the number of strategies, but there is no guarantee about the total number of steps that are needed to reach an ESS. We chose to stop the iterations when the variation of the population was below a minimum threshold. Execution times for the matching steps of our technique are plotted in Figure 13; the scatter plot shows a weak quadratic growth of convergence time as the number of matching strategies increases with a very small constant in the quadratic term, resulting in computation times below half a second even with a large number of strategies.

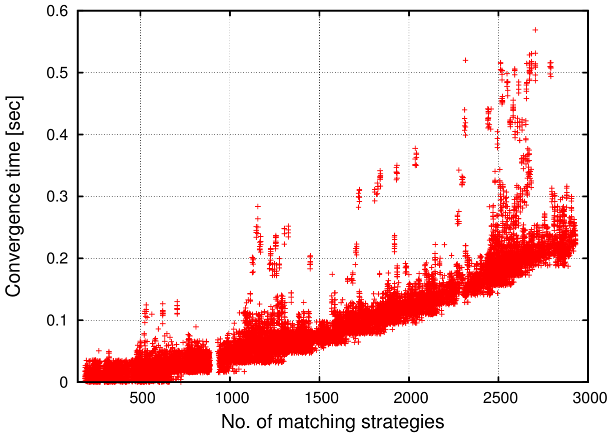


Fig. 13 Plot of the convergence time of the replicator dynamics with respect to the number of matching strategies.

5 Conclusions

In this paper we introduced a novel Game-Theoretic technique that performs an accurate feature matching as a preliminary step for multi-view 3D reconstruction using Structure from Motion techniques. Unlike other approaches, we do not rely on a first estimation of scene and camera parameters in order to obtain a robust inlier selection, but rather, we enforce geometric constraints based only on semi-local properties that can be estimated from the images. In particular, we define two selection games, one that selects local groups of compatible correspondences, enforcing a weak affine camera model, and a second consolidation game that filters out groups of matches by considering their compliance with the epipolar constraint. Experimental comparisons with a widely used technique show the ability of our approach to obtain a tighter inlier selection and thus a more accurate estimation of the scene parameters.

Acknowledgements We acknowledge the financial support of the Future and Emerging Technology (FET) Programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open project SIMBAD grant no. 213250.

References

Albarelli A, Rota Bulò S, Torsello A, Pelillo M (2009) Matching as a non-cooperative game. In: Proc. IEEE International Conference on Computer Vision - ICCV '09

Albarelli A, Rodolà E, Torsello A (2010) Robust game-theoretic inlier selection for bundle adjustment. In: Proc. 3D Data Processing, Visualization and Transmission – 3DPVT '10

Aggarwal JK, Duda RO (1975) Computer Analysis of Moving Polygonal Images IEEE Transactions on Computers, vol 24, pp 966-976

Beardsley PA, Zisserman A, Murray DW (1997) Sequential updating of projective and affine structure from motion. Int J Comput Vision 23(3):235–259

Bosch A, Zisserman A, Munoz X (2007) Image classification using random forests and ferns. In: Proc. 11th IEEE International Conference on Computer Vision – ICCV '07., pp 1–8

Brown M, Lowe DG (2005) Unsupervised 3d object recognition and reconstruction in unordered datasets. In: 3DIM '05: Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling, IEEE Computer Society, Washington, DC, USA, pp 56–63

Brown M, Lowe DG (2005) Unsupervised 3d object recognition and reconstruction in unordered datasets. In: 3DIM '05: Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling, IEEE Computer Society, Washington, DC, USA, pp 56–63

Fermuller C, Brodsky T, Aloimonos Y (1999) Motion segmentation: a synergistic approach Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on (2), pp 637-643

Fitzgibbon AW, Zisserman A (1998) Automatic camera recovery for closed or open image sequences. In: ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I, Springer-Verlag, London, UK, pp 311–326

Furukawa Y, Ponce J (2010) Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 32:1362–1376, DOI http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.161

Harris C, Stephens M (1988) A combined corner and edge detector. In: Proc. Fourth Alvey Vision Conference, pp 147–151

Hartley RI (1995) In defence of the 8-point algorithm. In: Proceedings of IEEE International Conference on Computer Vision, IEEE Comput. Soc. Press, pp 1064–1070

Herbert Bay TT, Gool LV (2006) SURF: Speeded up robust features. In: 9th European Conference on Computer Vision, vol 3951, pp 404–417

Heyden A, Berthilsson R, Sparr G (1999) An iterative factorization method for projective structure and motion from image sequences. Image and Vision Computing 17(13):981–991

Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. In: Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition – CVPR '04, vol 2, pp 506–513

Levenberg K (1944) A method for the solution of certain nonlinear problems in least squares. Quarterly Journal of Applied Mathmatics II(2):164–168

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

Lowe DG (1999) Object recognition from local scale-invariant features. In: Proc. of the International Conference on Computer Vision ICCV, pp 1150–1157

Marr D, Hildreth E (1980) Theory of edge detection. Royal Soc of London Proc Series B 207:187–217

Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22(10):761–767

Mikolajczyk K, Schmid C (2002) An affine invariant interest point detector. In: Proc. 7th European Conference on Computer Vision - ECCV 2002, Springer-Verlag, London, UK, pp 128–142

Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(10):1615–1630

Morel JM, Yu G (2009) ASIFT: A new framework for fully affine invariant image comparison. SIAM J Img Sci 2(2):438–469

Pollefeys M, Koch R, Vergauwen M, Gool LV (1999) Hand-held acquisition of 3d models with a video camera. 3D Digital Imaging and Modeling, International Conference on 0:0014

Sarfraz MS, Hellwich O (2008) Head pose estimation in face recognition across pose scenarios. In: VISAPP (1), pp 235–242

Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp 519–528

Snavely N, Seitz SM, Szeliski R (2006) Photo tourism: exploring photo collections in 3d. In: ACM SIGGRAPH '06, pp 835–846

Snavely N, Seitz SM, Szeliski R (2008) Modeling the world from internet photo collections. Int J Comput Vision 80(2):189–210

Sturm PF, Triggs B (1996) A factorization based algorithm for multi-image projective structure and motion. In: ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume II, Springer-Verlag, London, UK, pp 709–720

Taylor P, Jonker L (1978) Evolutionarily stable strategies and game dynamics. Mathematical Biosciences 40:145–156

Tomasi C, Kanade T (1992) Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision 9:137–154, 10.1007/BF00129684

Torr P, Zisserman A (1998) Robust computation and parametrization of multiple view relations. In: ICCV '98: Proceedings of the Sixth International Conference on Computer Vision, IEEE Computer Society, Washington, USA

Torsello A, Rota Bulò S, Pelillo M (2006) Grouping with asymmetric affinities: A game-theoretic perspective. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition – CVPR '06, pp 292–299

Triggs B, McLauchlan P, Hartley R, Fitzgibbon A (2000) Bundle adjustment – a modern synthesis. In: Triggs B, Zisserman A, Szeliski R (eds) Vision Algorithms: Theory and Practice, Springer-Verlag, Lecture Notes in Computer Science, vol 1883, pp 298–372

Tsai R (1987) A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. Robotics and Automation, IEEE Journal of 3(4):323–344

Vedaldi A, Fulkerson B (2008) VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/

Vergauwen M, Van Gool L (2006) Web-based 3d reconstruction service. Mach Vision Appl 17(6):411–426

Weibull J (1995) Evolutionary Game Theory. MIT Press

Weinshall D, Tomasi C (1995) Linear and incremental acquisition of invariant shape models from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 17:512–517

Weng J, Cohen P, Herniou M (1992) Camera calibration with distortion models and accuracy evaluation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 14(10):965–980

Weng J, Ahuja N, Huang TS (1993) Optimal Motion and Structure Estimation IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(9):864-884

Zhang Z, Deriche R, Faugeras O, Luong QT (1995) A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artif Intell 78(1-2):87–119

Zhang Z (1995) Estimating motion and structure from correspondences of line segments between two perspective images Pattern Analysis and Machine Intelligence, IEEE Transactions on 17(12):1129-1139