# A Game-Theoretic Approach to Robust Selection of Multi-View Point Correspondence

Emanuele Rodolà, Andrea Albarelli, and Andrea Torsello
*Dipartimento di Informatica - Università Ca' Foscari*
*Via Torino, 155 - 30172 Venice Italy*

## Abstract

*In this paper we introduce a robust matching technique that allows very accurate selection of corresponding feature points from multiple views. Robustness is achieved by enforcing global geometric consistency at an early stage of the matching process, without the need of subsequent verification through reprojection. The global consistency is reduced to a pairwise compatibility making use of the size and orientation information provided by common feature descriptors, thus projecting what is a high-order compatibility problem into a pairwise setting. Then a game-theoretic approach is used to select a maximally consistent set of candidate matches, where highly compatible matches are enforced while incompatible correspondences are driven to extinction.*

## 1 Introduction

The selection of 3D point correspondences from their 2D projections is arguably one of the most important steps in image based multi-view reconstruction, as errors in the initial correspondences can lead to suboptimal parameter estimation. The selection of corresponding points is usually carried out by means of interest point detectors and feature descriptors. Once salient and well-identifiable points are found on each image, correspondences between the features in the various views must be extracted and fed to the bundle adjustment algorithm. To this end, each point is associated a descriptor vector with tens to hundreds of dimensions, which usually include a scale and a rotation value. Arguably the most famous of such descriptors is the Scale-invariant feature transform (SIFT) [3].Features are designed so that similar image regions subject to similarity transformation exhibit descriptor vectors with small Euclidean distance. This property is used to match each point with a candidate with similar descriptor. However, if the descriptor is not distinctive enough this approach is prone to select many outliers since the approach only exploits 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, features that are close in one view are expected to be close in the other one as well. In addition, if a pair of feature exhibit a certain difference of angles or ratio of scales, this relation should be maintained among their respective matches. This prior information about scene structure can be accounted for by using a feature tracker [4, 6] to extract correspondences, but this requires that the view positions be not far apart. Further, in the presence of strong parallax, a locally uniform 3D motion does not result in a locally uniform 2D motion, and for these reason the geometric constraints can be enforced only locally. A common heuristic for the enforcement of global structure is to eliminate points that exhibit a large reprojection error after a first round of Bundle Adjustment [7]. Unfortunately this post-filtering technique requires good initial estimates to begin with.

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 3D reprojections. The approach selects feasible matches by enforcing global geometric consistency. Specifically, it enforces that all pairs of correspondences between 2D views are consistent with a common 3D rigid transformation. This constraint is in general underspecified, as a whole manifold of pairs of correspondences are consistent with a rigid 3D transformation, as it is well known that at least seven matching points are needed to solve the epipolar equation [2]. However, by accumulating mutual support through a large set of mutually compatible correspondences, one can expect to reduce the ambiguity to a single 3D rigid transformation. In the proposed approach high order consistency constraint are reduced to a second order compatibility where sets of 2D point correspondences that can be interpreted as projections of rigidly-transformed 3D points all have high mutual support. Then, following [8, 1], a game-theoretic approach is used to select a set of candidate matches, enforcing highly compatible matches while driving to extinction incompatible correspondences.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.23

IEEE computer society

## 2 Pairwise Geometric Consistency

There are two fundamental hypotheses underlying the reduction to second order of the high-order 3D geometric consistency. First, We assume that the views have the same set of camera parameters, second, we assume that the feature descriptor provides scale and orientation information and that this is related to actual local information in the 3D objects present in the scene. The effect of the first assumption is that the geometric consistency is reduced to a rigidity constraint that can be cast as a conservation along views of the distances between the unknown 3D position of the feature points, while the effect of the second assumption is that we can recover the missing depth information as a variation in scale between two views of the same point is inversely proportional to variation in projected size of the local patch around the 3D point and, thus, to the projected size of the feature descriptor.

More formally, assume that we have two points $p_1$ and $p_2$, which in one view have coordinates $(u_1^1, v_1^1)$ and $(u_2^1, v_2^1)$ respectively, while in a second image they have coordinates $(u_1^2, v_1^2)$ and $(u_2^2, v_2^2)$. These points, in the coordinate system of the first camera, have 3D coordinates $z_1^1(u_1^1, v_1^1, f)$ and $z_2^1(u_2^1, v_2^1, f)$ respectively, while in the reference frame of the second camera they have coordinates $z_1^2(u_1^2, v_1^2, f)$ and $z_2^1(u_2^2, v_2^2, f)$. Up to a change in units, these coordinates can be rewritten as $p_1^1 = \frac{1}{s_1^1}\begin{pmatrix} u_1^1 \\ v_1^1 \\ 1 \end{pmatrix}$, $p_2^1 = \frac{a}{s_2^1}\begin{pmatrix} u_2^1 \\ v_2^1 \\ 1 \end{pmatrix}$, $p_1^2 = \frac{1}{s_1^2}\begin{pmatrix} u_1^2 \\ v_1^2 \\ 1 \end{pmatrix}$, and $p_2^2 = \frac{a}{s_2^2}\begin{pmatrix} u_2^2 \\ v_2^2 \\ 1 \end{pmatrix}$, where $a$ is the ratio between the actual scales of the local 3D patches around points $p_1$ and $p_2$, whose projections on the two views give the perceived scales $s_1^1$ and $s_1^2$ for point $p_1$ and $s_2^1$ and $s_2^2$ for point $p_2$.

The assumption that both scale and orientation are linked with actual properties of the local patch around each 3D point is equivalent to having 2 points for each feature correspondence: the actual location of the feature, plus a virtual point located along the axis of orientation of the feature at a distance proportional to the actual scale scale of the patch. These pair of 3D points must move rigidly going from the coordinate system of one camera to the other, so that given any two sets of correspondences with 3D points $p_1$ and $p_2$ and their corresponding virtual points $q_1$ and $q_2$, the distances between these four points must be preserved in the reference frames of every view (see Fig. 1).

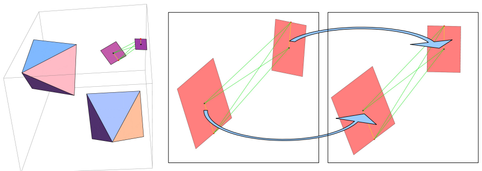


**Figure 1. Scale and orientation offer depth information and a second virtual point. the conservation of the distances in green enforce consistency with a 3D rigid transformation.**

Under a frontal-planar assumption for each local patch, or, less stringently, under small variation in viewpoints, we can assign 3D coordinates to the virtual points in the reference frames of the two images:

$$q_1^1 = p_1^1 + \begin{pmatrix} \cos\theta_1^1 \\ \sin\theta_1^1 \\ 0 \end{pmatrix} \quad q_2^1 = p_2^1 + a\begin{pmatrix} \cos\theta_2^1 \\ \sin\theta_2^1 \\ 0 \end{pmatrix}$$

$$q_1^2 = p_1^2 + \begin{pmatrix} \cos\theta_1^2 \\ \sin\theta_1^2 \\ 0 \end{pmatrix} \quad q_2^2 = p_2^2 + a\begin{pmatrix} \cos\theta_2^2 \\ \sin\theta_2^2 \\ 0 \end{pmatrix},$$

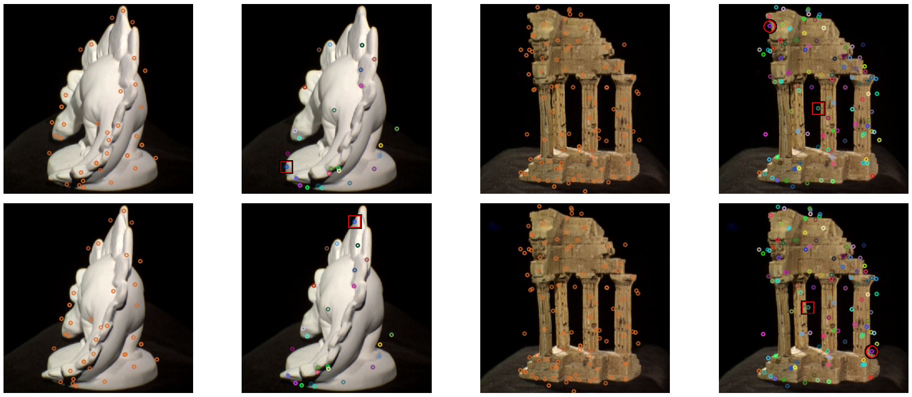
where $\theta_i^j$ is the perceived orientation of feature $i$ in image $j$. At this point, given two sets of correspondences between points in two images, namely the correspondence $m_1$ between a feature point in the first image with coordinates, scale and orientation $(u_1^1, v_1^1, s_1^1, \theta_1^1)$ with the feature point in the second image $(u_1^2, v_1^2, s_1^2, \theta_1^2)$, and the correspondence $m_2$ between the points $(u_2^1, v_2^1, s_2^1, \theta_2^1)$ and $(u_2^2, v_2^2, s_2^2, \theta_2^2)$ in the first and second image respectively, we can compute a distance from the manifold of feature descriptors compatible with a single 3D rigid transformation as

$$d(m_1, m_2, a) = (||p_1^1 - p_2^1||^2 - ||p_1^2 - p_2^2||^2)^2 + (||p_1^1 - q_2^1||^2 - ||p_1^2 - q_2^2||^2)^2 + (||q_1^1 - p_2^1||^2 - ||q_1^2 - p_2^2||^2)^2 + (||q_1^1 - q_2^1||^2 - ||q_1^2 - q_2^2||^2)^2 .$$

From this we define the compatibility between correspondences as $C(m_1, m_2) = \max_a e^{-\gamma d(m_1, m_2, a)}$, where $a$ is maximized over a reasonable range of ratio of scales of local 3D patches. In our experiments $a$ was optimized in the interval $[0.5; 2]$.

## 3 Game-Theoretic Feature Matching

We model the matching process in a game-theoretic framework [1], where two players extracted from a large population select a pair of matching points from two images. The player then receives a payoff from the other players 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. It is supposed that some selection process operates over time on the distribution of behaviors favoring players



Game-Theoretic Bundler Keymatcher Game-Theoretic Bundler Keymatcher

Dino sequence Temple sequence

| | Dino sequence | | Temple sequence | |
|---|---|---|---|---|
| | Game-Theoretic | Bundler Keymatcher | Game-Theoretic | Bundler Keymatcher |
| Matches | $262.5 \pm 61.4$ | $172.4 \pm 79.5$ | $535.7 \pm 38.7$ | $349.3 \pm 36.2$ |
| $\Delta\alpha$ | $0.0668 \pm 0.0777$ | $0.0767 \pm 0.1172$ | $0.1326 \pm 0.0399$ | $0.1414 \pm 0.0215$ |
| $\Delta\gamma$ | $0.4393 \pm 0.4963$ | $0.6912 \pm 0.8793$ | $0.0809 \pm 0.0144$ | $0.0850 \pm 0.0065$ |

**Figure 2. Results obtained with the Dino and Temple data sets (images best viewed in color).**

that receive larger payoffs and driving all inconsistent hypotheses to extinction, finally 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. 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 payoff that an individual playing strategy $i$ receives against someone playing strategy $j$. A *mixed strategy* is a probability distribution $\mathbf{x} = (x_1, \dots, x_n)^T$ over the available strategies $O$, thus lying in the n-dimensional standard simplex $\Delta^n = \{\mathbf{x} \in \mathbb{R}^n \ : \forall i \in 1 \dots n \; x_i \geq 0, \ \sum_{i=1}^{n} x_i = 1\}$. 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 $\mathbf{y}^T C\mathbf{x}$. A strategy $\mathbf{x}$ is said to be a *Nash equilibrium* if it is the best reply to itself, i.e., $\forall \mathbf{y} \in \Delta, \ \mathbf{x}^T C\mathbf{x} \geq \mathbf{y}^T C\mathbf{x}$. A strategy $\mathbf{x}$ is said to be an *evolutionary stable strategy* (ESS) if it is a Nash equilibrium and $\forall \mathbf{y} \in \Delta \quad \mathbf{x}^T C\mathbf{x} = \mathbf{y}^T C\mathbf{x} \Rightarrow \mathbf{x}^T C\mathbf{y} > \mathbf{y}^T C\mathbf{y}$. 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 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, a well-known formalization of the selection process governed by the recurrence $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^t \frac{(C\mathbf{x}^t)_i}{\mathbf{x}^{t^T} C\mathbf{x}^t}$, where $\mathbf{x}_i^t$ is the proportion of the population that plays the $i$-th strategy at time $t$. Once the population has reached a local maximum, all the non-extincted pure strategies can be considered selected by the game.

## 4 Experimental Results

To evaluate the performance of our proposal, we compared the results with those obtained with the keymatcher included in the structure-from-motion suite Bundler [7]. For the first set of experiments we selected pair of adjacent views from the ”DinoRing” and ”TempleRing” sequences from the Middlebury Multi-View Stereo dataset [5]; for these models, camera parameters are provided and used as a ground-truth. For all the sets of experiments we evaluated the differences in radians between the (calibrated) ground-truth and respectively the estimated rotation angle ($\Delta\alpha$) and rotation axis ($\Delta\gamma$). The “Dino” model is a difficult case in general, as it provides very few features; the upper part of Fig. 2 shows the correspondences produced by our method (left column) in comparison with the other matcher (right column). The “Temple” model richer in features and for visualization purposes we only show a subset of the detected matches for both the techniques. The Bundler keymatcher, while still achieving good results, provides some mismatches in both cases. This

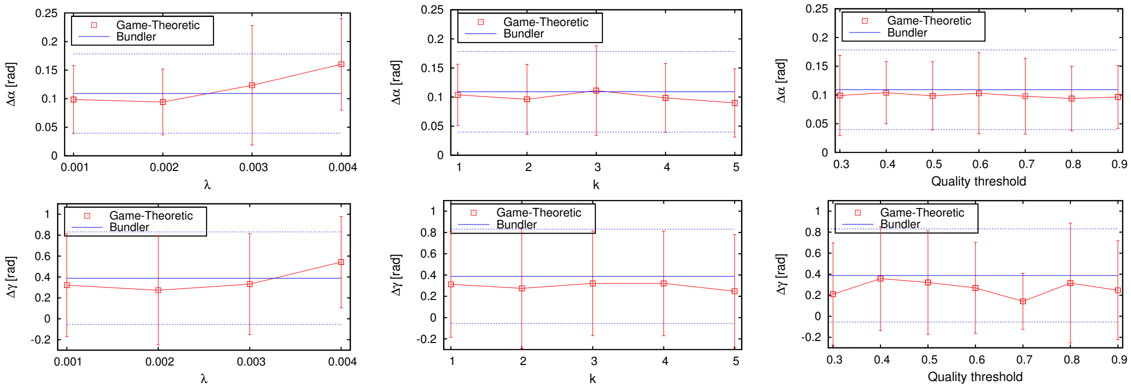


**Figure 3. Analysis of the performance of the approach with respect to variation of the parameters of the algorithm.**

can be explained by the fact that the symmetric parts of the object, e.g. the pillars in the temple model, result in very similar features that are hard to disambiguate by a purely local matcher. Our method, on the other hand, by enforcing global 3D consistency, can effectively disambiguate the matches. Looking at the results we can see that our approach extracts around 50% more correspondence, providing a slight increase in precision and reduction in variance of the estimates. Note that selected measures evaluate the quality of the underlying least square estimates of the motion parameters after a reprojection step, thus small variations are expected.

Next, we analyzed the impact of the algorithm parameters over the quality of the results obtained. To this end we investigated three parameters: the similarity decay $\lambda$, the number $k$ of candidate mates per features, and the *quality threshold*, that is the minimum support for a correspondence to be considered non-extinct, divided by the maximum support in the population. Figure 3 reports the results of these experiments. Overall, these experiments suggest that those parameters have little influence over the quality of the result. However the Game-Theoretic approach achieves better average results and smaller standard deviations for almost all reasonable values of the parameters.

## 5 Conclusions

In this paper we introduced a robust matching technique for feature points from multiple views. Robustness is achieved by enforcing global geometric consistency in a pairwise setting. This is achieved by using the scale and orientation information offered by SIFT features and projecting what is left of a high-order compatibility problem into a pairwise compatibility measure, by enforcing the conservation of distances between the unknown 3D positions of the points. Finally, a game-theoretic approach is used to select a maximally consistent set of candidate matches, where highly compatible matches are enforced while incompatible correspondences are driven to extinction. Experimental comparisons with a widely used technique show the ability of our approach to obtain a more accurate estimation of the scene parameters.

### Acknowledgments

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

[1] A. Albarelli, S. Rota Bulò, A. Torsello, and M. Pelillo. Matching as a non-cooperative game. In *ICCV 2009*, 2009.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[3] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.

[4] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[5] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR '06*, pages 519–528, 2006.

[6] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994.

[7] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2):189–210, 2008.

[8] A. Torsello, S. Rota Bulò, and M. Pelillo. Grouping with asymmetric affinities: A game-theoretic perspective. In *CVPR '06*, pages 292–299, 2006.