3D Wide Baseline Correspondences Using Depth-Maps

Marco Marcon, Eliana Frigerio*, Augusto Sarti, and Stefano Tubaro

Politecnico di Milano - Dipartimento di Elettronica e Informazione,
P.zza Leonardo Da Vinci, 32, 20133 Milano, Italy
marco.marcon@polimi.it, efrigerio@elet.polimi.it
http://home.dei.polimi.it/marcon

Abstract. Points matching between two or more images of a scene shot from different viewpoints is the crucial step to defining epipolar geometry between views, recover the camera's egomotion or build a 3D model of the framed scene. Unfortunately in most of the common cases robust correspondences between points in different images can be defined only when small variations in viewpoint position, focal length or lighting are present between images. While in all the other conditions ad-hoc assumptions on the 3D scene or just weak correspondences can be used. In this paper, we present a novel matching method where depth-maps, nowadays available from cheap and off the shelf devices, are integrated with 2D images to provide robust descriptors even when wide baseline or strong lighting variations are present.

Keywords: Machine vision, feature extraction, 3D descriptors.

1 Introduction

Feature points matching between two shots of a scene from different viewpoints is one of the basic and most tackled computer vision problems. In many common applications, like objects tracking in video sequences, the baseline is relatively small and features matching can be easily obtained using well known feature descriptors [14,4]. However many other applications require feature matching in much more challenging contexts, where wide baselines, lighting variations and non-lambertian surfaces reflectance are considered. Many interesting approaches based on two single images have been proposed in the literature, starting from the pioneering work of Schmid and Mohr [12] many other interesting approaches followed: Matas et al. [8] introduced the maximally stable extremal regions (MSER) where affinely-invariant stable subset of extremal regions are used to find corresponding *Distinguished Regions* between images, or moment descriptors for uniform regions [10] while other approaches are based on clearly distinguishable points (like corners) and affine-invariant descriptors of their neighborhood. One of the most popular approaches in the last few years becomes the Scale Invariant Feature Transform (SIFT) proposed by Lowe [3] thanks to its outperforming capabilities, as shown by Mikolajczyk and Schmid [7]. The SIFT algorithm is

* Corresponding author.

L. Atzori, J. Delgado, and D. Giusto (Eds.): MOBIMEDIA 2011, LNICST 79, pp. 194–203, 2012.

based on a local histogram of oriented gradient around an interest point and its success is mainly due to a good compromise between accuracy and speed (is as also been integrated in a Virtex II Xilinx Field Programmable Gate Array, FPGA [13]). Actually some other approaches, always based on affine invariant descriptors, got growing interest like the Gradient Location and Orientation Histogram (GLOH) [7] which is quite close to the SIFT approach but requires a Principal Component Analysis (PCA) for data compression, or the Speeded-Up Robust Features (SURF) [1] a powerful descriptor derived from an accurate integration and simplification of previous descriptors. All of the aforementioned approaches assume that, even if nothing is known of the underlying geometry of the scene, the defined features, since are describing a very small portion of the object, will undergo a simple planar transformation that can be approximated with an affine homography. This simplification has two main drawbacks, first of all the extracted features are very general and weak since wide affine transformations must provide very similar results, moreover, whenever the framed object present abrupt geometrical discontinuities (e.g. geometrical edges or corners) the affine approximation is not valid anymore. A possible solution to such problems could be a rough description of the underlying 3D geometry. In particular, within the Astute Artemis project, we are investigating the opportunity to use scene depth-maps to have a rough estimation of 3D underlying geometry: We use depth-maps to estimate the orientation of the plane, where the considered feature is laying, with respect to the observing camera and then we apply an homography to make this plane parallel to the camera image plane. Accordingly to this, our descriptors can be just *similarity invariant* with 2 Degrees of Freedom, scale and rotation, with respect to the 4 Degrees of Freedom present in an affine transformation (disregarding in both cases the translation on 2 axes). The proposed descriptors can then be less generic becoming more robust and discriminative. Another important aspect which we have been dealing with is geometric discontinuities in objects surface, in particular, when detected corners or edges are not due to texture of a locally planar surface but to the abrupt folding of the surface itself, affine approximation between two wide baseline views is not valid any more. Projection on the average tangent plane or the unfolding of the discontinuity (edge or corner) can significantly improve matching capabilities. In the following we will show how low-cost depth-map acquisition devices (like Microsoft Kinect) can be fruitfully adopted to prove effectiveness of the aforementioned approach.

2 Surface vs. Texture Relevant Points

Actually the, by far, most used algorithm to define significant points in a picture that can be used to be matched with corresponding points in another image, is the corner Harris detector. This pioneering algorithm from Harris and Stephens [5] is still the basic element for localization of feature descriptors: [9]. Applying this algorithm to depth-maps provides us with surface discontinuities like geometrical corners or edges. In most of cases this features are a sub-set of corner

and edges imputable to texture variations, so, once we have the depth-map registered with its corresponding image and we perform the Harris detector we are able to distinguish between:

- edges and corners due to textural variation but belonging to a flat surface.
- edges in the depth-maps corresponding to a folded or truncated surface.
- corners in the depth-maps (that are usually corners in the image too) corresponding to abrupt variations in the surface: e.g. spikes, corners or holes.

The capability to characterize different Harris features as geometrical or not (i.e. if they are also present or not in the depth-maps) is particularly important for definition of robust invariant descriptors. In particular the knowledge of the underlying geometry allows us to apply geometrical transformations to the textures on each slice in the neighborhood of the identified point in such a way to make their representation invariant from the view point. The opportunity to recover univocally a plane where the features in the neighborhood of the significant point lay is particularly important since it allows us, applying e.g. the proper homography, to obtain a frontal view of the neighborhood of a considered point independently from the viewpoint. The direct effect of this transformation is that the comparison between significant points for images acquired from different viewpoints can be simply performed comparing two frontal views of the regions around the points themselves: these regions can undergo only rotation and scaling: i.e. *similarity transform* where translation is disregarded since comparing neighborhood of two points implies the assumption that we are examining regions spatially already aligned).

3 Fusion of Geometric and Texture Descriptors

Many techniques have been developed to find flat planes in depth-maps, a significant example can be found in [15], and also surface curvature from cloud of points has been deeply investigated [16].

In our case we followed a simplified approach to define tangent plane to the surface around the interest point: it can be adopted even in case of discontinuities like corners, edges or generic surface folds. In fig. 1 there is a sample image where a Rubik's cube presents textural corners and edges on faces and abrupt geometrical corners and edges due to surface folds.

To find the tangent plane we followed a Principal Component Analysis for the spatial dispersion of depth-map points surrounding the interest point, in particular, accordingly to [6], we evaluated the covariance matrix (3×3) of the depth-map around the point (we used a 15×15 neighborhood window centered at the considered point but it can be adapted accordingly to the surface roughness or curvature) and then we performed the eigenvector decomposition. The resulting eigenvector associated to the lower eigenvalue represents the direction cosines for the "tangent" plane where we project the texture from the color image: this plane represents the locus where the points surrounding the interest point are maximally dispersed and it will always be the same independently

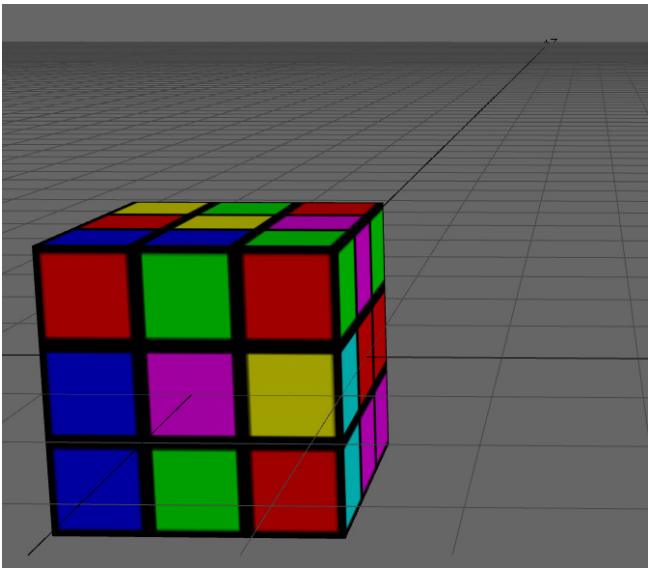

Fig. 1. A synthetic representation of a Rubik cube

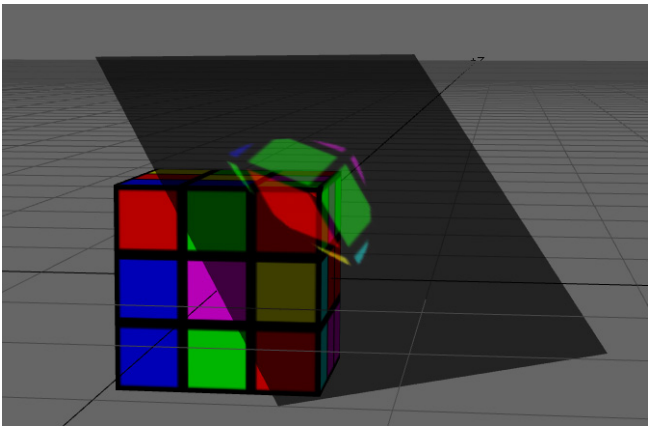

Fig. 2. The definition of the "tangent" plane and the reprojection on it of the neighborhood of the interest point

from the viewpoint (if all the sides are still visible). The image pixels are the projected onto this plane accordingly to their 3D position (recovered from the depth-map). Fig. 2 shows the reprojection of the texture on the "tangent" plane.

Then, through the homography that transforms the tangent plane into a frontal plane (a plane parallel to the image plane of the camera) we can recover a frontal view which is independent from the viewpoint apart for rotation and scaling (in fig. 3).

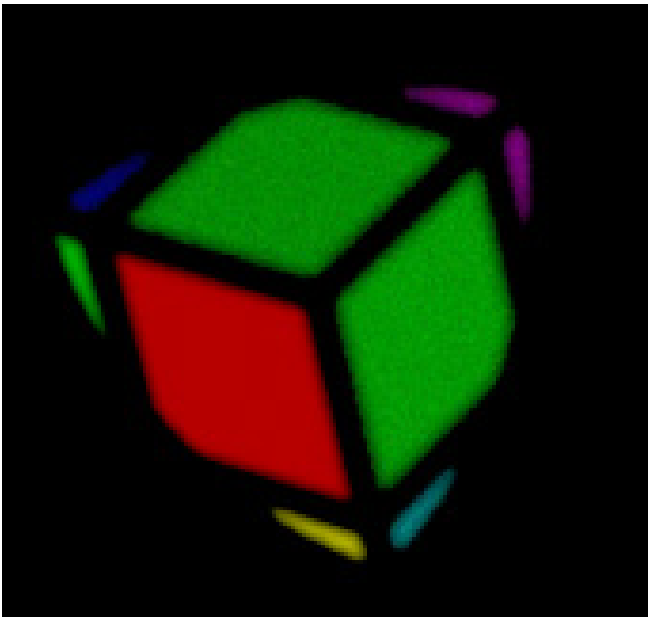

Fig. 3. The neighborhood of the interest point after the homography that provides a frontal view

4 Similarity Invariant Transform

Accordingly to the aforementioned steps we are able to obtain a 2D representation of the same 3D object part whose misalignment can be modeled by a four-parameter geometric transformation that maps each point (x_f, y_f) in F to a corresponding point (x_g, y_g) in G according to the matrix equation (in homogeneous coordinates):

$$\begin{bmatrix} x_g \\ y_g \\ 1 \end{bmatrix} = \begin{bmatrix} \rho\cos\vartheta & \rho\sin\vartheta & -\Delta x \\ \rho\sin\vartheta & \rho\cos\vartheta & -\Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_f \\ y_f \\ 1 \end{bmatrix}$$

Equivalently, defining the two images as two functions denoted by f and g, representing a gray-level image defined over a compact set of R^2, for any pixel (x, y) is true that:

$$f(x, y) = g(\rho(x\cos\vartheta + y\sin\vartheta) - \Delta x, \rho(-x\sin\vartheta + y\cos\vartheta) - \Delta y)$$

where Δx and Δy are translations, ρ is the uniform scale factor, and θ is the rotation angle. In other words, when we speak about similarity transformation we refer to the operations in this order:

$$RST = RS_{\rho,\theta} \cdot T_{\Delta x, \Delta y}$$

Since we are comparing image regions centered around interest points the translation invariance has no relevance in our case and the similarity invariance can be limited to rotation and scaling. Many approaches are present in the literature to tackle this problem [11], anyway most of them are incomplete like geometric moments and complex moments, while we oriented our research toward complete descriptors, that means that only representations retaining all the information of an image, except for orientation and scale, are considered. In particular we used the Fourier-Mellin transform (FMT) that is the Fourier Transform of the image $f(x, y)$ mapped in its corresponding Log-Polar coordinates $f_{LP}(\mu, \xi)$:

$$f_{LP}(\mu, \xi) = \begin{cases} f(e^{\mu} \cos \xi, e^{\mu} \sin \xi) & \xi \in [0, 2\pi) \\ 0 & otherwise \end{cases}$$

The FMT is defined as:

$$F_m(w, k) = \int_0^{\infty} \int_0^{2\pi} f_{LP}(\mu, \xi) e^{-j(w\mu + k\xi)} d\xi d\mu$$

Then we explored two possible invariant for orientation and scale: the Taylor Invariant and the Hessian Invariant, which are described in the following sections. In particular we recall that after a Log-polar transformation a rotation corresponds to a circular shift along the axis representing the angles while a scaling corresponds to a shift along the logarithmic radial axis. Applying the 2D Fourier transform to the Log-polar transform the aforementioned shifts are reflected in phase shifts while the amplitude will remain unchanged.

5 Taylor and Hessian Invariant Descriptors

In this section we depict the two orientation-scale invariant descriptors that we used, both of them are based on the FMT described in the previous section. The Taylor invariant descriptor [2] is focused on eliminating the linear part of the phase spectrum by subtracting the linear phase from the phase spectrum. Let $F(u, v)$ be the Fourier transform of an image $f(x, y)$, and $\phi(u, v)$ be its phase spectrum. The following complex function is called the Taylor invariant:

$$F_{Tl}(u, v) = e^{-j(au+bv)} F(u, v)$$

where a and b are respectively the derivatives with respect to u and v of $\phi(u, v)$ at the origin $(0, 0)$, i.e.:

$$a = \varphi_u(0, 0),$$
$$b = \varphi_v(0, 0)$$

The Taylor invariant is rotationally symmetric, but not reciprocally scaled. It can be modified accordingly to the Laplacian invariant:

$$F_L(u, v) = \left(u^2 + v^2\right) F_{Tl} = \left(u^2 + v^2\right) e^{-j(au+bv)} F(u, v)$$

The effect is then the registration of the input features in such a way that the phase spectrum is flat in the origin, i.e. if we should take the inverse transforms, all of them will be rotated and scaled to accomplish to this constrain.

The idea behind the Hessian Invariant Descriptor [2] is to differentiate the phase spectrum twice to eliminate the linear phase, the invariant parts are then the modulus of the spectrum and the three, second order, partial derivatives of the phase spectrum:

$$F_H(u,v) = [|F(u,v)|, \varphi_{uu}(u,v), \varphi_{uv}(u,v), \varphi_{vv}(u,v)]$$

As described in the following sections, we evaluated both descriptors obtaining very similar results.

6 Results

We applied the previous descriptors to real images together with their depth-maps. The proposed algorithm can be summarized as follow:

- for each shot of the scene, significant points are extracted using Harris corner detector applied on the picture;
- the PCA was applied on the neighborhood 15×15 of the corresponding point of each detected point on the depth map and the eigenvector associated to the lower eigenvalue is used to determine the homography that transform the tangent plane into a frontal plane;
- the Fourie Mellin Transform is applied to the reoriented neighborhood;
- at last the Laplacian invariant is applied to $F_m(w,k)$ (only Laplacian translation invariant is used for these test).
- The resulted vector is used as feature descriptor of the significant point and correct match from different images are selected as those for which the Euclidean distance is minimized.

For completeness we summarize also the main step of the SIFT algorithm implemented for comparing the performances:

- Maximally Stable Extremal Regions (MSER) [8] are found for each shot of the scene;
- all the MSER are approximated as elliptical and oriented so that each major axis is horizontal;
- the ellipsis are deformed in circles and the intensity gradient for each pixel is computed;
- each circular region is divided in rectangular subregions and the histogram of the gradient's direction is computed for each subregion;
- the feature vector is made linking all the histograms computed on the circular neighborhood and, as for the proposed algorithm, correct match from different images are selected as those for which the Euclidean distance is minimized.

We performed some experiments using snapshots similar to those visible in fig. 4. No databases of pictures and depthmaps associated are yet available nowadays, so we decided to test our algorithm taking 20 pictures of the box illustrated in fig. 4 from different viewpoints. We used a Kinect device for the acquisition in an indoor environment and without any restriction except avoid that sun light directly on the IR device's camera. In fig. 5 we show how the planes, where the interest points lay, are reprojected in frontal views; the homographies have been defined accordingly to the PCA analysis of the underlying depth-map.

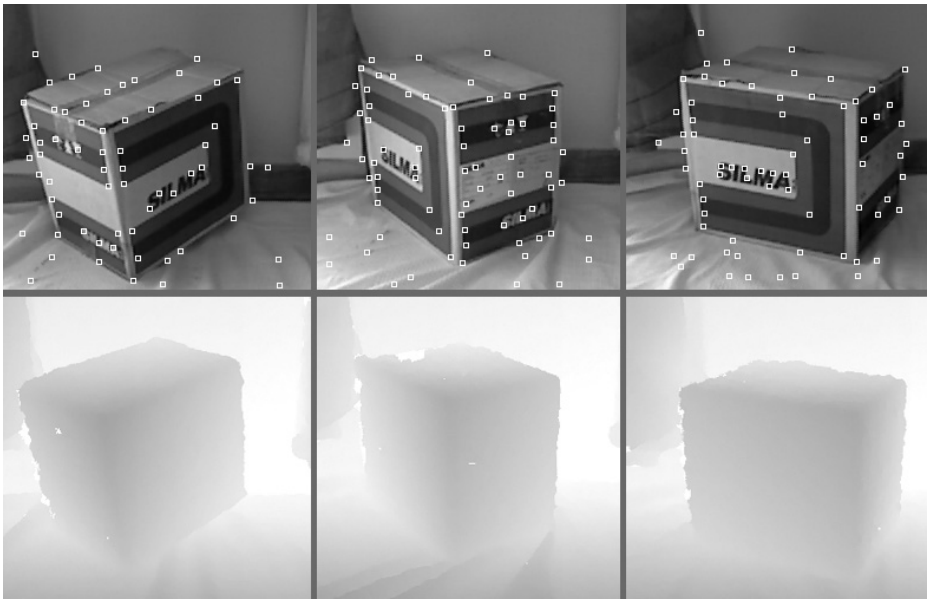

Fig. 4. A box acquired from different viewpoints and its depth-maps

Fig. 5. Images of interesting points after the homography to obtain a frontal view of framed surface by the depth-map

We checked the discriminative power of the proposed descriptors, in particular we compared the correct match rate and the euclidean distance from the closest match and from the second candidate. With the SIFT descriptor applied to the images, we obtained a correct match rate of 73%. For correct matches the mean ratio of the euclidean distances between the correct one and the second one is

around 0.8. Using the proposed approach we obtained a correct match rate of 85% with an average ratio of distances for the first match and the second one of 0.65.

7 Conclusion

In this paper we propose a novel approach to define putative correspondences between images where the information from corresponding depth-maps are fruitfully integrated to reduce variability in the neighborhood around interest points, in particular projective or affine distortions are reduced to similarity transforms making available more robust and complete descriptors like Taylor or Hessian invariants applied to the Fourier-Mellin Transform.

The resulting approach demonstrates the profitable integration of depth-maps with acquired images to strengthen matching capabilities. Examples have been obtained by a low cost Kinect device.

Acknowledgement. This work was supported by the ASTUTE project: a 7 Framework Programme European project within the Joint Technology Initiative ARTEMIS.

References

1. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision and Image Understanding 110(3), 346–359 (2008)
2. Brandt, R., Lin, F.: Representation that uniquely characterize image modulo translation, rotation and scaling. Pattern Recognition Letters 17, 1001–1015 (1996)
3. Lowe, D.: Distinctive image features from scale-infariant keypoints. Int'l J. Computer Vision 2(60), 91–110 (2004)
4. Fusiello, A., Trucco, E., Tommasini, T., Roberto, V.: Improving features tracking with robust statistics. Pattern Analysis and Applications 2, 312–320 (1999)
5. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vision Conf., pp. 147–151 (1988)
6. Jolliffe, I.: Principal Component Analysis, 2nd edn., vol. XXIX - 487. Springer, NY (2002)
7. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(10) (2005)
8. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22(10), 761–767 (2004)
9. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision 60(1), 63–86 (2004)
10. Mindru, F., Tuytelaars, T., Gool, L.V., Moons, T.: Moment invariants for recognition under changing viewpoint and illumination. Computer Vision and Image Understanding 94(1-3), 3–27 (2004)
11. Mukundan, R., Ramakrishnan, K.: Moment Functions in Image Analysis: Theory and Applications. World Scientific Publishing Co. Pte. Ltd., Singapore (1998)

12. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. Pattern Analysis and Machine Intelligence 19(5), 530–535 (1997)
13. Se, S., Ng, H., Jasiobedzki, P., Moyung, T.: Vision based modeling and localization for planetary exploration rovers. In: Proceedings of International Astronautical Congress (2004)
14. Shi, J., Tomasi, C.: Good features to track. In: Computer Vision and Pattern Recognition (1994)
15. Yang, M., Foerstner, W.: Plane detection in point cloud data. Technical Report TR-IGG-P-2010-01, Department of Photogrammetry Institute of Geodesy and Geoinformation University of Bonn (2010)
16. Yang, P., Qian, X.: Direct computing of surface curvatures for point-set surfaces. In: Proceedings of 2007 IEEE/Eurographics Symposium on Point-based Graphics, PBG (2007)