Signal Processing: *Image Communication* 27 (2012) 849–855

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

3D wide baseline correspondences using depth-maps

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

Image and Sound Processing Group, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy

ARTICLE INFO

Available online 9 February 2012

Keywords:
Machine vision
Feature extraction
3D descriptors

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. In all the other conditions *ad hoc* assumptions on the 3D scene or just weak correspondences through statistical approaches 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. We show how depth information can highly improve matching in wide-baseline contexts with respect to state-of-the-art descriptors for simple images.

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 [1,2]. 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 [3] many other interesting approaches followed: Matas et al. [4] introduced the maximally stable extremal regions (MSER) where stable subset of extremal regions invariant to affine transformations are used to find corresponding *Distinguished Regions* between images, or moment descriptors for uniform regions [5] 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 [6] thanks to its outperforming capabilities, as shown by Mikolajczyk and Schmid [7]. The SIFT algorithm is 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 [8]). 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) [9] 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

* Corresponding author.
E-mail addresses: marcon@elet.polimi.it (M. Marcon), efrigerio@elet.polimi.it (E. Frigerio), sarti@elet.polimi.it (A. Sarti), stefano.tubaro@polimi.it (S. Tubaro).

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2012.01.009

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 presents abrupt geometrical discontinuities (e.g. geometrical edges or corners) the affine approximation is not valid anymore.

A possible solution to these problems could be a rough description of the underlying 3D geometry. In particular, we investigated the opportunity to use scene depth-maps to have a rough estimation of 3D underlying geometry: we use depth-maps to estimate, with respect to the observing camera, the orientation and distance of the plane where the interest point is laying. We then apply a homography in order to transform this plane parallel to the camera's image plane. If depth-map is also metric (i.e. the metric distance of pixels from the image plane is also known) we can also move the feature plane at a specific distance from the image plane. In the first case our descriptors can be just *similarity invariant* with 2 degrees of freedom, scale and rotation while, if the depth-map is also metric, rotation becomes the only variable among different views.

The local descriptors can then be less generic since they have to take into account just rotations and, eventually, scaling, becoming more robust and discriminative with respect to those thought for affine transformations traditionally used in computer vision. Another important aspect where depth-map can be really useful are the geometric discontinuities in objects surface: when the surface presents corners or edges the depth-map presents abrupt changes and local texture will not undergo a planar transformation in different views. The depth-map can then be fruitfully adopted to discard those points in matching search in different views using only points on almost planar surfaces.

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 improving matching capabilities between points even in wide-baseline comparisons.

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 [10] is still the basic element for the localization of feature descriptors: [11]. Applying this algorithm to depth-maps provides us with surface discontinuities like geometrical corners or edges. In particular, accordingly to [10], the analysis of the "Corner Response" applied to depth-maps allows us to automatically find abrupt jumps, edges or corners in the geometrical surface and the texture around those points can then be skipped in the point match research. Once we have the depth-map registered with its corresponding image and we perform the Harris detector both on the depth-map and on the relative image 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. The opportunity to recover univocally the plane where the neighborhood of the significant point lays, 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 we are comparing neighborhood of corners that provide an univocal spatial localization. Furthermore if the same device is used to acquired the analyzed depth-maps or if the acquisition device is also metrically calibrated (it provides us with the metric distance of each point from the camera), we are able to compare all the features as if they are placed at the same distance from the camera and the only remaining degree of freedom is rotation.

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 [12], and also surface curvature from cloud of points has been deeply investigated [13].

In our case we followed a simplified approach to define tangent plane to the surface around the interest point: 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. The first step of the proposed algorithm is then based on the corner localization in the color image, once that possible interest points are located we recover the local tangent plane by the Principal Component Analysis on the depth-map points surrounding the interest point, in particular, accordingly to [14], 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.

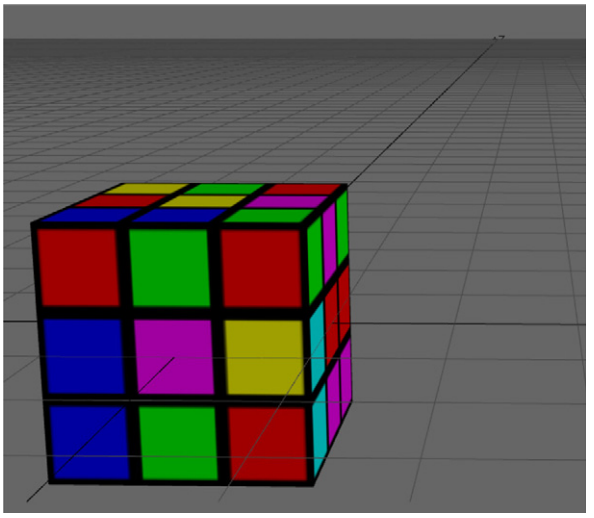

Fig. 1. A synthetic representation of a Rubik cube.

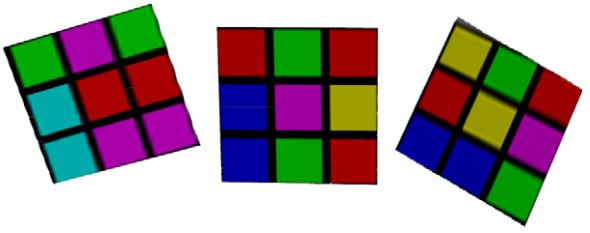

Fig. 2. Texture reprojection from Fig. 1 into planes parallel to the image plane. Planes information was recovered from the depth-map.

In particular, since the lower eigenvalue indicates the dispersion (variance) around the plane normal direction, we account that value as the local "flatness" index of the tangent plane and we accept only values lower than a threshold. In other words we are imposing to consider only texture corners, which neighborhood belongs to a flat surface. Once the laying plane is defined, the homography to recover a frontal and centered image can be easily obtained [15]. Then, through the homography, we can recover a frontal view which is independent from the viewpoint apart for rotation and scaling (Fig. 2).

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 [16], 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 $f_{LP}(\mu,\xi)$

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

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 section. 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 a linear phase contribution while the amplitude will remain unchanged.

5. Taylor and Hessian Invariant Descriptors

Many recent approaches integrate range data or depth-maps together with geometrical descriptors to improve matching rate in wide baseline or scene changing contexts. In particular we point out two quite similar and significant approaches: [17,18], that tackle 3D distortion geometrical corrections, anyway both of them use as final descriptor, after the geometrical corrections, the SIFT descriptors. On the contrary, our aim is to explore, after the rectification of the planar region around the interest point, some less flexible and generic but more robust 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 [19] 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:

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

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.

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

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 [19] is to differentiate the phase spectrum twice to eliminate the linear phase terms, 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 section, we evaluated both descriptors obtaining very similar results anyway the Hessian transform, thanks to its high-pass filter, provides better results with intensity/light variations while Taylor descriptor can be considered slightly better when smoothly changing textures are present.

6. Proposed algorithm

In order to evaluate the matching quality with respect to the state of the art descriptors we compare our Taylor/Hessian Invariant Descriptor with the SIFT, that, accordingly to [7] outperforms other multiple views matching methods.

The proposed algorithm can be summarized as follows:

- For each shot of the scene, significant points are extracted using Harris corner detector applied on the picture.
- The PCA is applied on the neighborhood 15×15 of the corresponding point of each detected point on the depth map and the lower eigenvalue and its associated eigenvector are founded:
  - If the lower eigenvalue is higher that a predefined threshold the interest point is not assumed on a planar region and it is skipped. The algorithm then jumps back to the next point.
  - Else the point is assumed on a planar region and the eigenvector associated to the lower eigenvalue is used to determine the homography that transform the tangent plane into a frontal plane respect to the camera.
- The homography is applied to each point around the interest point, followed by a bicubic interpolation in order to avoid artifacts (rectification process).
- The 2D Fourier-Mellin Transform is applied on the reprojected region:
  - If the depth-map is metrically calibrated the previous homography places the region around our interest point at a constant distant from the image plane, in this case the 2D Fourier Transform is applied to the reprojected region expressed in polar coordinates.
  - Else the scale of the final region is not known and the 2D Fourier-Mellin Transform is applied to the reprojected region (2D Fourier Transform applied on the reprojected region expressed in Log-polar coordinates).
- At last the Taylor or the Hessian Invariant is applied to $F_m(w,k)$.
- 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 minimum.

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

- Maximally Stable Extremal Regions (MSER) [4] 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.

7. Results

We checked the discriminative power of the proposed Taylor/Hessian Invariant Descriptors with respect to state of the art SIFT descriptor applying them on rectified version of the original images acquired with different view points. Obviously SIFT was built to find matches in the more general case of affine/perspective transforms but our aim is to demonstrate that, even when depth-maps are available together with their texture images, more detailed descriptors, robust to rotation and eventually scaling, can be adopted providing us with better matching score.

We performed some experiments using snapshots similar to those visible in Fig. 6. We also analyzed results on synthetic images in order to insulate acquisition noise contributions (see Fig. 3).

In Fig. 4 it is possible to see correspondences found by the Taylor Invariant Descriptor (TID) and the SIFT applied on rectified images after the homography, with metric assumption on the depth-map. The results in terms of correct matches vs. all the matches founded are 98% with TID, while 87% with SIFT. The putative matches are founded using a nearest neighbor classifier.

In order to test our results for real images, without taking into account problems coming from wrong

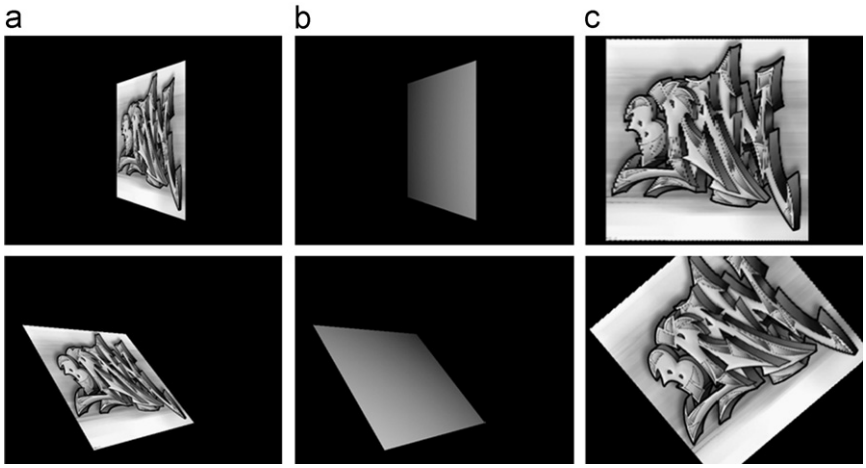

Fig. 3. Examples of (a) synthetic images obtained by a ray-tracing software and (b) the relative depth-maps. (c) Images obtained after the rectification process.

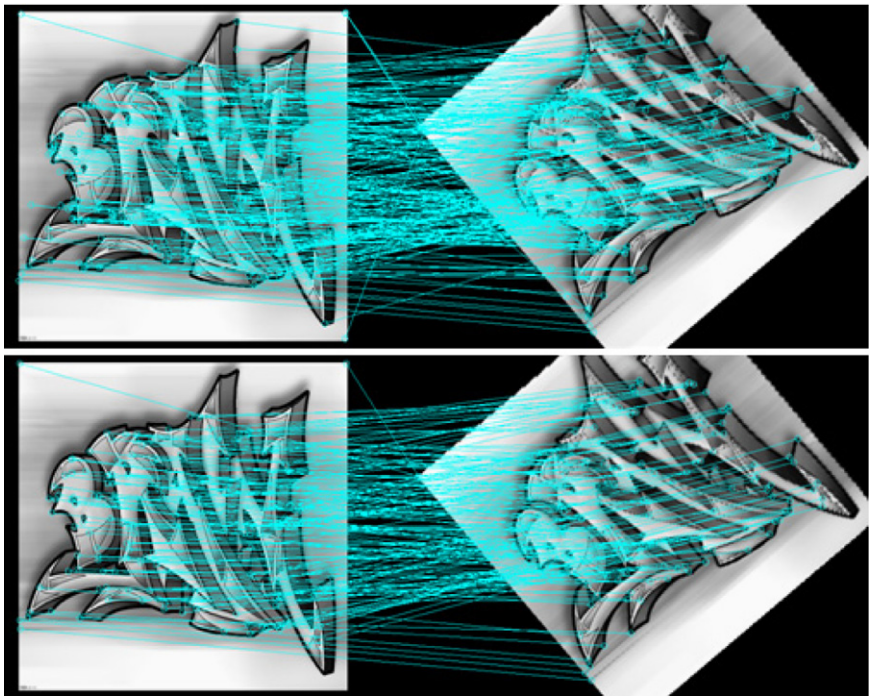

Fig. 4. Point correspondences found by the Taylor Invariant Descriptor and the SIFT applied after rectification process performed on the global image.

estimation of the homography needed to project the acquired scene in frontal view, we analyzed the results on images of the famous wall proposed in the Mikolajczyk database, where also the exact homography between different snapshots is given. With the SIFT descriptor applied to the rectified images of Fig. 5, we obtained a correct match rate of 79%. Even if the nearest neighbor classifier is implemented, we also analyzed the distances ratio between the first nearest neighbor and the second one: the lower is this ratio the higher is the discriminative power of the proposed descriptor. 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 87% with an average ratio of distances for the first match and the second one of 0.65.

No databases of pictures and depth-maps associated with a wide baseline are yet available nowadays, so we decided to test our algorithm taking 10 pictures of the box illustrated in Fig. 6 acquired from different viewpoints. We used a Kinect® device for the acquisitions in an indoor environment and without any restriction except avoid that sun light directly on the IR device's camera. In Fig. 7 we show how the planes, where the interest points lay,

Fig. 5. A famous wall from Mikolajczyk database for feature matching tests acquired under different view-points. Comparisons were made on rectified images.

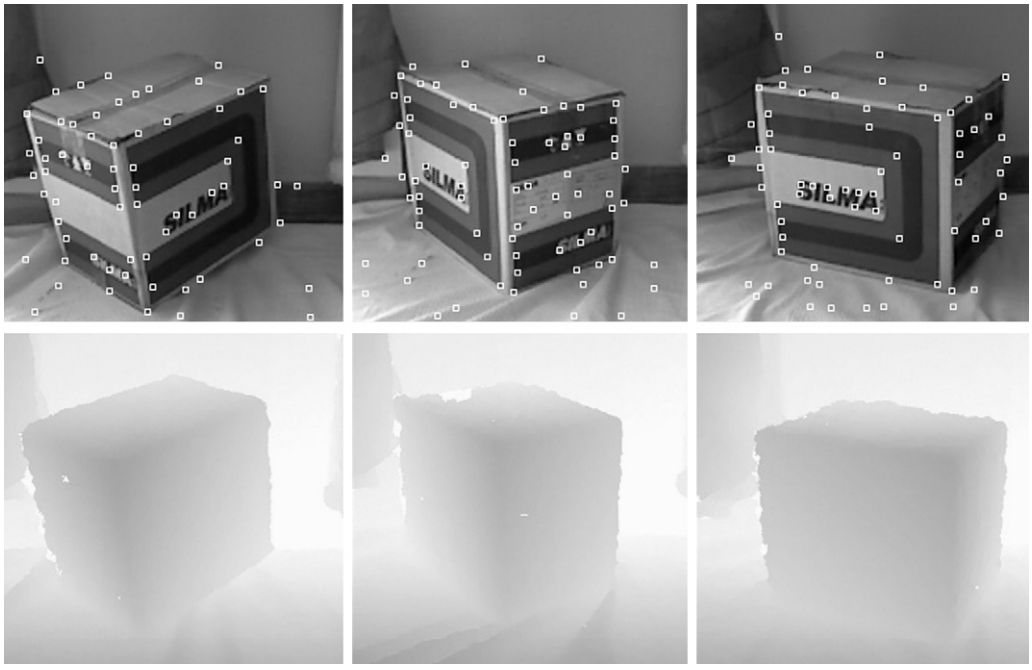

Fig. 6. A box acquired from different viewpoints and its depth-maps.

Fig. 7. Images of interesting points after the homography to obtain a frontal view of framed surface by the depth-map.

are reprojected in frontal views; the homographies have been defined accordingly to the PCA analysis on the underlying depth-map. Matching results are reported in Table 1. The results proposed are obtained using the Taylor Invariant Descriptor, but they are close for both descriptors.

8. 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 (or the Fourier Transform applied on the rectified neighborhood mapped in polar coordinates). We also showed how these descriptors, that are robust only to rotation, or rotation and scaling, provide better results with respect to state of the art descriptors like SIFT

Table 1
Matching results with different descriptors.

Matching results	Synthetic depth-map	Graffiti	Real box
SIFT			
Correct match (%)	81	79	75
1st over 2nd best match	0.79	0.81	0.90
Fourier-polar 2D (calibrated) with Taylor Invariant			
Correct match (%)	88	86	83
1st over 2nd best match	0.61	0.65	0.79
Fourier-Mellin 2D with Taylor Invariant			
Correct match (%)	84	84	80
1st over 2nd best match	0.69	0.71	0.78
Fourier-polar 2D (calibrated) with Hessian Invariant			
Correct match (%)	88	87	85
1st over 2nd best match	0.65	0.65	0.73
Fourier-Mellin 2D with Hessian Invariant			
Correct match (%)	85	84	80
1st over 2nd best match	0.69	0.70	0.79

indicating that the auxiliary information from depth-map improves features matching even for wide baseline views.

The resulting approach demonstrates the profitable integration of depth-maps with acquired images to strengthen matching capabilities and the proposed descriptors hold much more information of the original image patch with respect to more general descriptors. Furthermore the computational complexity of the proposed descriptors is also very close to state of the art feature descriptors. Real examples have been obtained by a low cost Kinect® device. Future work will try to extend actual study to non-planar surfaces and investigate how descriptors robustness can be strengthened to wide illumination changes and to non-Lambertian surfaces.

Acknowledgment

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

References

[1] J. Shi, C. Tomasi, Good features to track, in: Proceedings CVPR'94, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1994, pp. 593–600.
[2] A. Fusiello, E. Trucco, T. Tommasini, V. Roberto, Improving feature tracking with robust statistics, Pattern Analysis & Applications 2 (1999) 312–320.
[3] C. Schmid, R. Mohr, Local grayvalue invariants for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 530–535.
[4] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image and Vision Computing 22 (2004) 761–767.
[5] F. Mindru, T. Tuytelaars, L. Gool, T. Moons, Moment invariants for recognition under changing viewpoint and illumination, Computer Vision and Image Understanding 94 (2004) 3–27.
[6] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.
[7] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence (2005) 1615–1630.
[8] S. Se, H. Ng, P. Jasiobedzki, T. Moyung, Vision based modeling and localization for planetary exploration rovers, in: Proceedings of International Astronautical Congress, Citeseer, 2004, pp. 433–440.
[9] H. Bay, T. Tuytelaars, L. Van Gool, SURF: speeded up robust features, Computer Vision–ECCV, 2006, pp. 404–417.
[10] C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey Vision Conference, vol. 15, Manchester, UK, 1988, pp. 50–60.
[11] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, International Journal of Computer Vision 60 (2004) 63–86.
[12] M. Yang, W. Förstner, 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.
[13] P. Yang, X. Qian, Direct computing of surface curvatures for point-set surfaces, in: Eurographics Symposium on Point-based Graphics, 2007, pp. 29–36.
[14] I.T. Jolliffe, Principal Component Analysis, 2nd ed., Springer-Verlag, New York, 2002, ISBN: 0-387-95442-2.
[15] R.I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed., Cambridge University Press, 2004 ISBN: 0521540518.
[16] R. Mukundan, K. Ramakrishnan, Moment Functions in Image Analysis: Theory and Applications, World Scientific Pub Co Inc, 1998.
[17] K. Koser, R. Koch, Perspectively invariant normal features, in: IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.
[18] C. Wu, B. Clipp, X. Li, J.M. Frahm, M. Pollefeys, 3D model matching with viewpoint-invariant patches (VIP), in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[19] R. Brandt, F. Lin, Representations that uniquely characterize images modulo translation, rotation, and scaling, Pattern Recognition Letters 17 (1996) 1001–1015.