3D CORRESPONDENCES IN TEXTURED DEPTH-MAPS THROUGH PLANAR SIMILARITY TRANSFORM

Marco Marcon, Eliana Frigerio, Augusto Sarti, Stefano Tubaro

Politecnico di Milano, Dipartimento di Elettronica e Informazione
P.zza Leonardo Da Vinci, 32
20133 Milano (ITALY)

ABSTRACT

Many low cost realtime depth-map reconstruction devices recently appeared on the market opened new opportunities for the Computer Vision community to integrate these information in many research areas. The knowledge of the underlying depth-map together with a visual snapshot of the scene can greatly improve the robustness of points matching between different views even for wide baseline acquisitions. In this paper we are presenting how visual correspondences from different views can be identified by robust Similarity Invariant Descriptors (SID) once their laying plane is known. Furthermore depth-map, providing with a rough geometrical description of the underlying scene, allows to select only feature points belonging to almost planar regions, skipping geometrical corners or edges that undergo non-linear distortion for viewpoint changes. The proposed SIDs keep much more information of the original area with respect to commonly used affine invariant descriptors, like SIFT of GLOH, making the proposed approach much less prone to false matches even for wide viewpoint changes.

Index Terms— Range data, 3D/stereo scene analysis, Pattern Invariants

1. INTRODUCTION

Feature points matching between two shots of a scene acquired 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]. 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 literature. Starting from the pioneering work of Schmid and Mohr [2] many other interesting approaches followed: Matas *et al.* [3] introduced the maximally stable extremal regions (MSER) where affine-invariant stable subset of extremal regions are used to find corresponding *Distinguished Regions* between images, while Van Gool *et al.* [4] introduced moment descriptors for uniform regions. Other approaches are based on clearly distinguishable points (like corners) and affine-invariant descriptors of their neighborhood. Thanks to its outperforming capabilities, as shown by Mikolajczyk and Schmid [5], one of the most popular algorithms in recent years is the Scale Invariant Feature Transform (SIFT) proposed by Lowe [6]. 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 [7]). Actually some other approaches, always based on affine invariant descriptors, got growing interest like the Gradient Location and Orientation Histogram (GLOH) [5], 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) [8] 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/projective 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 the latter problem could be obtained having a rough description of the underlying 3D geometry. In particular, within the Astute Artemis project, we investigated how depth-maps can be used to estimate almost planar regions within the image where the considered interest points are laying. 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, [9] and [10], that tackle 3D distortion

geometrical corrections. Anyway both of them use SIFT as final descriptor, after the geometrical corrections. 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 descriptor. In particular our descriptors are just *similarity invariant* bringing with them more information about the analyzed region.

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 described approach. The rough geometrical description of the scene obtained using the depth-map, together with Similarity Invariant Descriptor applyed on the corresponding visual snapshot, allows to find robust stereo corespondences in images acquired with wide baseline.

2. LOCAL PLANE EXTRACTION

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 [11] is still the basic element for localization of feature descriptors [12]. Applying this algorithm on a region on the depth-map around an interest point provides us with an analysis of the planarity of that region. Once the planarity of a region around an interest point is verified, the parameters of the plane passing through that region could be estimated in order to define the homography transformation that generate its frontal view.

Many techniques have been developed to find flat surfaces in depth-maps, a significant example can be found in [13], and also surface curvature from cloud of points has been deeply investigated [14].

To find the tangent plane we applyed Principal Component Analysis (PCA) on the depth-map regions surrounding each interest point, in particular, accordingly to [15], 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 eigenvector associated to the lower eigenvalue represents the tangent plane direction cosines (n_x, n_y, n_z). Accordingly to the equation for a generic point of plane $\mathbf{p}$:

$$\rho = \mathbf{p} \cdot \mathbf{n} = p_x n_x + p_y n_y + p_z n_z \tag{1}$$

The parameter ρ, can be recovered with metrical exactness only if the depth-map is intrinsically calibrated (pixel size and focal length are known) but this is not crucial if the same acquisition device with same focal length is used for all the acquisition set.

2.1. The Radon Transform (RT) for 3D planes

To improve estimation of plane parameters and to group interest points belonging to the same plane, we adopted a uniform quantization for planes parameters with different orientations and distances from the origin in a spherical coordinate system, in particular, given a point $\mathbf{p}$ on the plane with equation (1) and defining θ as the azimuth angle and φ as the elevation, we can rewrite eq.(1) as:

$$p_x \cdot cos\theta \cdot sin\varphi + p_y \cdot sin\theta \cdot sin\varphi + p_z \cdot cos\varphi = \rho \tag{2}$$

In each cell of the Radon Transform: $C(\rho, \theta, \varphi)$, we store the number of depth-map points belonging to that plane. The regions of cells in the RT with the main number of elements represent the planes present in the acquired scene. For the same quantization of the three parameters ρ, θ, φ, Hough and Radon transform obtain very similar results for the 3D discrete function C (slight differences are just due to different rounding off).

3. SIMILARITY INVARIANT TRANSFORM

Once we found planes in the depth-map we can easily define the homography to rotate each plane in order to obtain a frontal view: the vector normal to the plane has to align to the optical axis and the image projected on the depth-map is then be transformed into a frontal-view image (we suggest bi-cubic interpolation on the rectified images in order to minimize artifacts due to the quantization on the spherical coordinates that could mislead point descriptors). The misalignment between two images of the same plane acquired under different view points after the rectification can be modeled by a 2D similarity transformation, completely defined by translation, rotation and scaling. 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 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 representations retain all information of an image, except for orientation and scale. 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 $f_{LP}(\mu, \xi)$:

$$f_{LP}(\mu, \xi) = \begin{cases} f(e^{\mu}\cos\xi, e^{\mu}\sin\xi) & \xi \in [0, 2\pi) \\ 0 & otherwise. \end{cases} \tag{3}$$

The FMT is defined as:

$$F_m(w, k) = \int_0^{\infty} \int_0^{2\pi} f_{LP}(\mu, \xi)\, e^{-j(w\mu + k\xi)} d\xi d\mu. \tag{4}$$

In particular we recall that after a Log-polar transformation a rotation corresponds to a circular shift along the angle axis 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 linear phase contribution while the amplitude will remain unchanged.

3.1. Taylor and Hessian Invariant Descriptors

We explored two possible invariant for translation: the Taylor Invariant and the Hessian Invariant. Applying one of them to the image mapped in its Log-Polar coordinates, we achieve an orientation-scale invariant descriptor.

The Taylor Invariant Descriptor (TID) [17] is focused on eliminating the linear part of the phase spectrum by subtracting the estimated linear phase contribution from the phase spectrum. The effect is the registration of the input features in such a way that the phase spectrum is flat in the origin, i.e. their inverse transforms will be rotated and scaled to accomplish to this constrain. Let $F_m(u,v)$ be the Fourier Mellin transform of an image $f(x,y)$, and $\varphi(u,v)$ be its phase spectrum. The following complex function is called the Taylor invariant:

$$F_{Tl}(u,v) = e^{-j(au+bv)} F_m(u,v) \tag{5}$$

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

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

The idea behind the Hessian Invariant Descriptor (HID) [17] 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:

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

4. RESULTS

To verify different steps of the proposed approach we used the Kinect device on a 3D calibration pattern: a half-cube (3 contiguous faces) with sides length of $0.8m$. Fig.1 shows three different shoots of the calibration pattern with their respective depth-maps. Using the RT different plane surface from the depth-map could be localized in space (as illustrated in Fig.2). We performed some experiments, an example is visible in Fig.3, where, all the interest points belonging to the same plane (estimated using RT on the depth-map) undergo the same transformation. On each rectified view (shown in Fig.4), we applied the previously described similarity-invariant descriptors: for each interest point, the Taylor Invariant Descriptor is applied on the FMT computed on a region around it. The descriptor is the concatenation of these values obtained on its neighborhood and the distance between compared descriptors is done by a simple cross correlation.

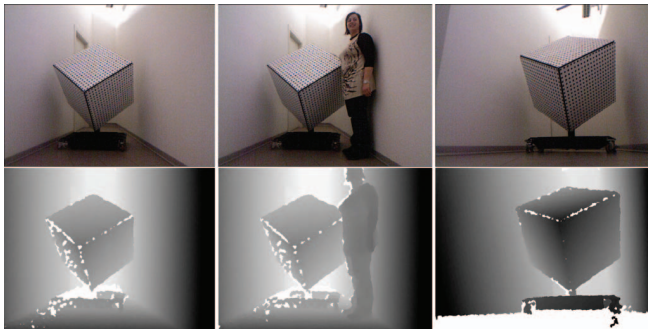

Fig. 1. Three color images acquired with the kinect (top raw) and their respective depth-maps (bottom row)

Fig. 2. Clusterization of different depth-maps planes

We then check the discriminative power of the proposed descriptors, in particular we compare our approach with the SIFT. We evaluate the correct match rate and the ratio between the euclidean distance from the closest match and the second candidate. Using SIFT, we obtained a correct match rate of 73% and, for correct matches, the mean ratio of the euclidean distances is around 0.8. Using the proposed approach we obtained a correct match rate of 85% with an average ratio of distances of 0.65.

5. CONCLUSION

In this paper we proposed 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 recover 3D planes and to strength matching capabilities using complete descriptors. Examples have been obtained by a low cost Kinect device.

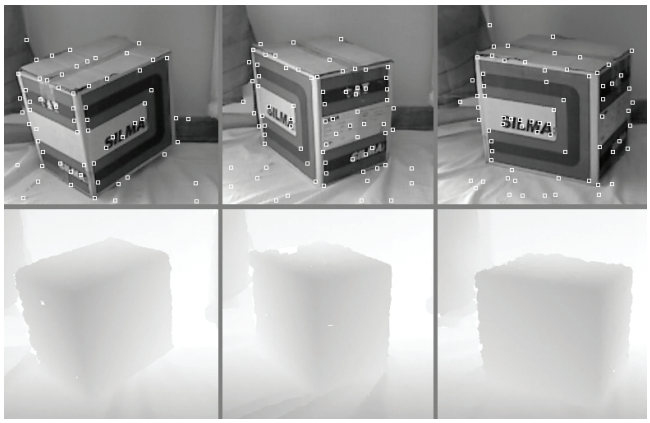

Fig. 3. Gray scale images of a box acquired acquired from different viewpoints and its depth-maps

Fig. 4. Images of the rectified version of the same plane estimated from the depth-map

6. ACKNOWLEDGEMENT

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

7. REFERENCES

[1] A. Fusiello, E. Trucco, T. Tommasini, and V. Roberto, "Improving features tracking with robust statistics," *Pattern Analysis and Applications*, vol. 2, pp. 312–320, 1999.

[2] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.

[3] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004, British Machine Vision Computing 2002.

[4] L. Van Gool F. Mindru, T. Tuytelaars and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Computer Vision and Image Understanding*, vol. 94, no. 1–3, pp. 3–27, 2004.

[5] K.Mikolajczyk and C.Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, 2005.

[6] D.Lowe, "Distinctive image features from scale-infariant keypoints," *Int'l J. Computer Vision*, vol. 2, no. 60, pp. 91–110, 2004.

[7] S. Se, H.K. Ng, P. Jasiobedzki, and T.J. Moyung, "Vision based modeling and localization for planetary exploration rovers," in *Proceedings of International Astronautical Congress*, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[9] K. Koser and R. Koch, "Perspectively invariant normal features," in *IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[10] C. Wu, B. Clipp, X. Li, J. M. Frahm, and M. Pollefeys, "3d model matching with viewpoint-invariant patches (vip)," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vision Conf.*, 1988, pp. 147–151.

[12] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[13] M.Y. Yang and W. Foerstner, "Plane detection in point cloud data," *Technical Report TR-IGG-P-2010-01*, 2010, Department of Photogrammetry Institute of Geodesy and Geoinformation University of Bonn.

[14] P. Yang and X. Qian, "Direct computing of surface curvatures for point-set surfaces," in *Proceedings of 2007 IEEE/Eurographics Symposium on Point-based Graphics(PBG)*, 2007.

[15] I.T. Jolliffe, *Principal Component Analysis*, vol. XXIX, Springer, NY, 2nd ed. edition, 2002.

[16] R. Mukundan and K.R. Ramakrishnan, *Moment Functions in Image Analysis: Theory and Applications*, World Scientific Publishing Co. Pte Ltd, Singapore, 1998.

[17] R.D. Brandt and F. Lin, "Representation that uniquely characterize image modulo translation, rotation and scaling," *Pattern Recognition Letters*, vol. 17, pp. 1001–1015, 1996.