UNCALIBRATED VIEW SYNTHESIS FROM RELATIVE AFFINE STRUCTURE BASED ON PLANES PARALLELISM

Stefano Tebaldini, Marco Marcon, Augusto Sarti, Stefano Tubaro

Dipartimento di Elettronica e Informazione - Politecnico di Milano
Piazza Leonardo Da Vinci, 32, 20133 Milano (ITALY)

ABSTRACT

This paper focuses on the generation of physically valid views from two or more uncalibrated images acquired by standard cameras. The problem is faced without trying to yield a three dimensional reconstruction of the imaged scene, which would be unfeasible without the exact knowledge of the positions of the cameras in the Euclidean frame where the scene is to be described. Instead, starting from the previous works of Shashua and Navab on Relative Affine Structure [1] and the article of Fusiello on views synthesis from uncalibrated views [2] we propose a novel approach that does not require the presence of a plane at infinity to define the homography between two views but merely the parallelism between couples of planes. This allows our approach to be applied to numerous scenes where two parallel planes can be defined (indoor scenes, straight streets and avenues). Experiments with synthetic images illustrate the approach.

Index Terms— Calibration, Rendering, Computational Geometry, Geometric modeling

1. INTRODUCTION

In the last years, researches about the synthesis of realistic virtual views have been developed both in the field of Computer Vision and Computer Graphics, mainly due to the increasing attention for virtual reality and its related applications Hence, a wide variety of approaches have been developed, as witnessed by the the huge amount of literature on this topic. Roughly speaking, the techniques for the generation of synthetic views may be classified into three classes, basing on how much the geometric information is exploited Model based rendering techniques rely on an explicit geometrical 3D model of the scene. A first step is required to extract the 3D model from a set of acquisitions, typically obtained by laser scanning or calibrated cameras [3],The novel views may then be rendered from the point of view of the virtual camera. These techniques usually require a well calibrated acquisition system. When no a priori information abut the calibration system is available the generation of novel views is yielded by exploiting the geometrical constraints existing among different images of the same scene [4, 5, 2].In this way, an implicit geometric model of the scene is obtained, and the generation of the novel view may proceed by interpolation or extrapolation from the references images. The appealing aspect of this approach is that no calibration of the acquisition system is required, and only two reference images are needed to generate a novel view. However, a computationally extensive preprocessing step is required to achieve a robust point matching among the reference images, This is the framework in which our technique is embedded. Finally, techniques based on the plenoptic function do not require any geometric information or correspondence. Though, a large set of reference images is needed, in order to achieve a sufficiently dense sampling of the plenoptic function [6].

The proposed process for generating a novel image starting from a set of reference ones may be roughly split into three parts. First a sufficiently dense set of point correspondences among the images must be established.A possible approach is described in [7]. The second step investigates the intrinsic geometry of the reference views and the synthesis of a novel view while the last step accounts for final images rendering mapping textures onto the novel views (we followed the approach proposed in [8]).

2. RELATIVE AFFINE STRUCTURE

A very compact and useful approach to the problem of reconstruction from uncalibrated cameras is the one developed in [1]. The key idea of that paper is to represent the 3D geometry of the scene in a projective frame where an arbitrarily chosen plane is mapped at the infinity. This results in an elegant notation for modeling the point transfer between two reference views, constituted by a 2D projective transform plus an affine term. This term, named *relative affine structure* by the authors, is characterized by the noticeable property of being invariant to the choice of the second view.

Consider the projections in a homogeneous projective space $\mathbb{P}^3$ of a scene onto two images ($\mathbb{P}^3 \rightarrow \mathbb{P}^2$) by the action of two cameras i and j: making explicit the role of the scale factors the 3D points projection can be written as:

$$\begin{aligned} \mathbf{p}_i^T \mathbf{X} \mathbf{x}_i &= \mathbf{P}_i \mathbf{X} \\ \mathbf{p}_j^T \mathbf{X} \mathbf{x}_j &= \mathbf{P}_j \mathbf{X} \end{aligned} \tag{1}$$

where $\mathbf{P}$ represents each camera matrix, $\mathbf{X}$ is a 3D point whose projections on the two cameras are $\mathbf{x}_i$ and $\mathbf{x}_j$ and $\mathbf{p}_i^T$, $\mathbf{p}_j^T$ are the third rows of $\mathbf{P}_i$ and $\mathbf{P}_j$, respectively. Considering the projection of $\mathbf{P}$ on a generic plane π and remembering that the camera centre $\mathbf{C}$ is the right null vector of the projection matrix the previous equation for camera i can be rewritten as:

$$\pi^T\mathbf{X} = \pi^T\mathbf{P}_i^\dagger\mathbf{p}_i^T\mathbf{X}\mathbf{x}_i + \pi^T\mathbf{C}_i\frac{\mathbf{C}_i^T\mathbf{X}}{\mathbf{C}_i^T\mathbf{C}_i} \tag{2}$$

(where $\mathbf{P}_i^\dagger$ represents the pseudo-inverse of the $\mathbf{P}$ matrix).

Combining the equations for the two cameras we can obtain a compact form:

$$\mathbf{x}_j \simeq \mathbf{A}_{ij}(\pi)\,\mathbf{x}_i + \mu\mathbf{e}_{ji} \tag{3}$$

where: $\mathbf{A}_{ij}(\pi) = \mathbf{P}_j\mathbf{P}_i^\dagger - \frac{\pi^T\mathbf{P}_i^\dagger}{\pi^T\mathbf{C}_i}\mathbf{e}_{ji} = \frac{1}{\mathbf{p}_i^T\mathbf{X}}\frac{\pi^T\mathbf{X}}{\pi^T\mathbf{C}_i}$
may be easily interpreted as follows: the point transfer from view i to view j is obtained by a 2D homography, $\mathbf{A}_{ij}(\pi)$, plus an affine term directed as the epipole $\mathbf{e}_{ji}$ and proportional to μ. This term may be given an explicit geometric meaning by recalling that the inner product between the homogeneous representations of a plane and a point is proportional to the normal distance of the point to the plane. Therefore

$$\mu \simeq \frac{d(\pi,\mathbf{X})}{z_i} \tag{4}$$

where z_i is the distance of $\mathbf{X}$ from the principal plane of the $i-th$ camera and $d(\pi,\mathbf{X})$ is the distance of $\mathbf{X}$ from π.

To disambiguate the scaling dependence in (4) Shashua and Navab proposed to normalize both $\mathbf{A}_{ij}(\pi)$ and μ such that $\mu = 1$ for a fixed point $\mathbf{X}^0$. After this normalization, we have:

$$\mu = \frac{\mathbf{p}_i^T\mathbf{X}^0}{\pi^T\mathbf{X}^0}\frac{\pi^T\mathbf{X}}{\mathbf{p}_i^T\mathbf{X}} = \frac{z_i^0}{d(\pi,\mathbf{X}^0)}\frac{d(\pi,\mathbf{X})}{z_i} \tag{5}$$

In this way, the amplitude of the correction given by the affine term in (3) is related only to the first view and the geometry of the scene, without being affected by the choice of the second view nor by scale ambiguities. This property allows to build a 3D representation of the scene geometry as follows:

$$\mathbf{X}^P = \begin{bmatrix} \mathbf{x}_i \\ \mu \end{bmatrix}$$

With this definition, it is immediate to see that:

$$\begin{aligned} \mathbf{x}_i &\simeq [\mathbf{I}|\mathbf{0}]\,\mathbf{X}^P \\ \mathbf{x}_j &\simeq [\mathbf{A}_{ij}(\pi)\,|\mathbf{e}_{ji}]\,\mathbf{X}^P \end{aligned}$$

where $\mathbf{I}$ is the 3×3 identity matrix. Therefore, $\mathbf{X}^P$ is to be regarded as a 3D reconstruction of the scene in the coordinate system where the camera matrices (assuming, for simplicity, the same internal calibration matrix $\mathbf{K}$) are given by:

$$\begin{aligned} \mathbf{P}_i &= \mathbf{K}[\mathbf{R}_i|\mathbf{t}_i] = [\mathbf{I}|\mathbf{0}] \\ \mathbf{P}_j &= \mathbf{K}[\mathbf{R}_j|\mathbf{t}_j] = [\mathbf{A}_{ij}(\pi)\,|\mathbf{e}_{ji}] \end{aligned}$$

For a random choice of the reference plane, the transformation relating the geometry of the scene in the Euclidean frame, $\mathbf{X}$, and its reconstruction, $\mathbf{X}^p$, is projective, as remembered by the superscript ^{P}. A special case is given when the plane at infinity is used as the reference plane. Under this condition, the relative affine structure (5) becomes proportional to the inverse of the normal distance from the principal plane of the $i-th$ camera, and the transformation between $\mathbf{X}$ and $\mathbf{X}^p$ becomes affine.

According to Fusiello [2], the generation of a synthetic view from a smooth motion working in a projective frame, cannot be directly obtained as in an Euclidean frame (keep in mind that there's no reason why a path that is smooth in the projective frame should be smooth as well in the Euclidean one). Following his work he introduced the $\mathbf{G}$ matrix, defined as:

$$\mathbf{G} = \begin{bmatrix} \mathbf{P} \\ [\,\mathbf{0} \quad 1\,] \end{bmatrix} \tag{6}$$

This square matrix belongs to the Special Euclidean Group of $\mathbb{R}^3$) and, being a *Lie Group* , it is equipped with smooth differentiable operators. The infinitesimal variation of $\mathbf{G}$ can then be defined as

$$\mathbf{G}(dt) = \mathbf{I} + dt\frac{d\mathbf{G}}{dt}$$

This displacement K times, yielding:

$$\mathbf{G}(Kdt) = \left(\mathbf{I} + dt\tfrac{d\mathbf{G}}{dt}\right)^K \quad \text{where} \quad \tfrac{d\mathbf{G}}{dt} = logm(\mathbf{G})$$

Therefore, by posing $t = Kdt$ results in:

$$\mathbf{G}(t) = \lim_{K\to\infty}\left(\mathbf{I} + \frac{t}{K}\frac{d\mathbf{G}}{dt}\right)^K = expm(t\cdot logm(\mathbf{G})) = \mathbf{G}^t \tag{7}$$

where $expm$ denotes matrix exponentiation. Further details about the smoothness of interpolated (extrapolated) transformations can be found in [2].

3. VIEW SYNTHESIS

According to the projective reconstruction theorem, if the scene has been reconstructed from the views, it is affected by a projective ambiguity:

$$\mathbf{x}_i \simeq [I|0]\mathbf{X}_i^P \quad , \quad \mathbf{x}_j \simeq [I|0]\mathbf{D}_{ij}\mathbf{X}_i^P$$

where $\mathbf{D}_{ij}$ is the relative transformation between the cameras and $\mathbf{X}_i^P$ is the reconstructed scene for a generic plane π:

$$\mathbf{X}_i^P = \begin{bmatrix} \mathbf{x}_i^T & \mu_i \end{bmatrix}^T = \begin{bmatrix} \mathbf{x}_i^T & \frac{z_i^0}{\pi^T\mathbf{X}^0}\frac{\pi^T\mathbf{X}}{z_i} \end{bmatrix}^T$$

For the invariance of the inner product in homogeneous coordinates, $\pi^T\mathbf{X} = \pi_i^T\mathbf{X}_i$ for any choice of i (where $\pi_i^T = \begin{bmatrix} \mathbf{v}_i^T & c_i \end{bmatrix}$). This results in the following writing:

$$\mathbf{X}_i^P \simeq \mathbf{H}_i\mathbf{X}_i = \begin{bmatrix} [I|0] \\ \frac{z_i^0}{\pi_i^T X_i^0}\pi_i^T \end{bmatrix}\mathbf{X}_i \tag{8}$$

where $\mathbf{H}_i$ is the projectivity between $\mathbf{X}_i$ and $\mathbf{X}_i^P$. By posing the last row of $\mathbf{H}$ is given by:

$$\frac{z_i^0}{\pi_i^T X_i^0}\pi_i^T = \frac{z_i^0}{\pi_i^T X_i^0}\begin{bmatrix} \mathbf{v}_i^T & c_i \end{bmatrix} \tag{9}$$

It is important to note that (9) does not depend on the scale ambiguity related to the homogeneous representation of π_i^T and $\mathbf{X}_i^0$. The transformation between the views in the projective frame, defining $\mathbf{G}_{ij} = \mathbf{G}_j\mathbf{G}_i^{-1}$ can then be written as:

$$\begin{aligned} \mathbf{X}_j^P \simeq \mathbf{H}_j\mathbf{X}_j = \mathbf{H}_j\mathbf{G}_{ij}\mathbf{X}_i \simeq \mathbf{H}_j\mathbf{G}_{ij}\left(\mathbf{H}_i\right)^{-1}\mathbf{X}_i^P \\ \Rightarrow \mathbf{D}_{ij} = \mathbf{H}_j\mathbf{G}_{ij}\left(\mathbf{H}_i\right)^{-1} \end{aligned} \tag{10}$$

Substituting the previous results we obtain:

$$\mathbf{D}_{ij} = \begin{bmatrix} \mathbf{A}_{ij}^\infty - e_{ji}\frac{\mathbf{v}_i^T}{c_i} & \mathbf{e}_{ji}\frac{1}{c_i}\frac{\pi_i^T\mathbf{X}_i^0}{z_i^0} \\ \mathbf{0} & \frac{z_j^0}{z_i^0} \end{bmatrix} \tag{11}$$

where $\mathbf{A}_{ij}^\infty = \mathbf{R}_j\mathbf{R}_j^T$ is the homography of $\mathbb{P}^2$ induced by taking the plane at infinity as the reference plane. Combining eqs. (7) and (3) we obtain:

$$\mathbf{x}_t \simeq [\mathbf{I}|\mathbf{0}]\,\mathbf{D}_{ij} = \mathbf{H}_j\mathbf{G}_{ij}^t\left(\mathbf{H}_i\right)^{-1}\mathbf{X}_i^P \tag{12}$$

Unfortunately we do not know *a priori* the π plane parameters (the $\mathbf{v}_i$ vector). To overcome this limitation, Fusiello proposed in [2] the exploitation of the plane at infinity as the reference plane: $\pi_j = \begin{bmatrix} \mathbf{0}^T & c \end{bmatrix}^T$ resulting in

$$\widehat{\mathbf{D}}_{ij} \simeq \begin{bmatrix} \mathbf{A}_{ij}^\infty & \mathbf{e}_{ji}\frac{1}{c}\frac{\pi_i^T\mathbf{X}_i^0}{z_i^0} \\ \mathbf{0} & \frac{z_j^0}{z_i^0} \end{bmatrix}$$

3.1. View Synthesis through two parallel planes and a calibration point

If the plane at infinity is not available in the reference views, new constraints must be sought in order to calibrate the scene. For example, the relative motion between two planes could be exploited. To this aim, we measure the transformation $\mathbf{D}_{ij}$ induced by two different planes $\pi_i = \begin{bmatrix} \mathbf{v}_i^T & c_i \end{bmatrix}^T$, $\omega_i = \begin{bmatrix} \mathbf{w}_i^T & d_i \end{bmatrix}^T$:

$$\begin{aligned} \widehat{\mathbf{D}}_{ij}(\pi) &= C_\pi \begin{bmatrix} \mathbf{A}_{ij}^\infty - \mathbf{e}_{ji}\frac{\mathbf{v}_i^T}{c_i} & \mathbf{e}_{ji}\frac{1}{c_i}\frac{\pi_i^T\mathbf{X}_i^0}{z_i^0} \\ \mathbf{0} & \frac{z_j^0}{z_i^0} \end{bmatrix} \\ \widehat{\mathbf{D}}_{ij}(\omega) &= C_\omega \begin{bmatrix} \mathbf{A}_{ij}^\infty - \mathbf{e}_{ji}\frac{\mathbf{w}_i^T}{d_i} & \mathbf{e}_{ji}\frac{1}{d_i}\frac{\omega_i^T\mathbf{X}_i^0}{z_i^0} \\ \mathbf{0} & \frac{z_j^0}{z_i^0} \end{bmatrix} \end{aligned} \tag{13}$$

If the reference point is equidistant from the principal planes of the two cameras, then $\frac{z_j^0}{z_i^0} = 1$. This allows to solve for the scale factors C_π, C_ω in (13). Now, let $\mathbf{M}(\pi), \mathbf{M}(\omega)$ denote the upper left blocks of $\widehat{\mathbf{D}}_{ij}(\pi)$ and $\widehat{\mathbf{D}}_{ij}(\omega)$ after solving for the scale factors:

$$\mathbf{M}(\pi) = \mathbf{A}_{ij}^\infty - \mathbf{e}_{ji}\mathbf{v}_i^T \qquad \mathbf{M}(\omega) = \mathbf{A}_{ij}^\infty - \mathbf{e}_{ji}\mathbf{w}_i^T$$

Therefore: $\mathbf{M}(\pi) - \mathbf{M}(\omega) = \mathbf{e}_{ji}\left(\mathbf{w}_i^T - \mathbf{v}_i^T\right)$ from which the difference $\left(\mathbf{w}_i^T - \mathbf{v}_i^T\right)$ may be computed up to a scale factor (the norm of the epipole). If we add the further condition that the planes π and ω are parallel to each other, then $\mathbf{w}_i = \gamma\mathbf{v}_i$ for some $\gamma \neq 1$. Therefore, exploiting two parallel planes and a point equidistant from the cameras, we get to obtain $\mathbf{v}_i$ up to a scale factor. From this knowledge, and under the assumption that the calibration matrices are equal, it is possible to obtain the expression of $\mathbf{A}_{ij}^\infty$ as follows.

Let $\widehat{\mathbf{v}}_i$ and $\widehat{\mathbf{e}}_{ji}$ be two unitary norm vector directed as $\mathbf{v}_i$ and $\mathbf{e}_{ji}$, respectively. It is no hard to see that the one parameter function

$$J(\alpha) = \det\left(\mathbf{M}(\pi) + \alpha\widehat{\mathbf{e}}_{ji}\widehat{\mathbf{v}}_i^T\right) \tag{14}$$

is linear in α. Therefore, it admits a unique $\overline{\alpha}$ such that $J(\overline{\alpha}) = 1$. From the condition on the calibration matrices it follows that $\det\left(\mathbf{A}_{ij}^\infty\right) = 1$, and thus it must be $\mathbf{A}_{ij}^\infty = \mathbf{M}(\pi) + \overline{\alpha}\widehat{\mathbf{e}}_{ji}\widehat{\mathbf{v}}_i^T$ Since $\mathbf{A}_{ij}^\infty$ is the homography induced by the plane at infinity, recovering $\mathbf{A}_{ij}^\infty$ is equivalent to using the plane at infinity as the reference plane. At this point, the algorithm for view synthesis may proceeds following (12). It is worthwhile noticing that the equation $J(\alpha) = 1$ may be solved in a closed form, yielding: $\overline{\alpha} = \left(\frac{1}{|\mathbf{M}(\pi)|} - 1\right)\frac{1}{tr\left(\mathbf{M}(\pi)^{-1}\widehat{\mathbf{e}}_{ji}\widehat{\mathbf{v}}_i^T\right)}$

3.2. View Synthesis through two pairs of parallel planes

If the ratio $\frac{z_j^0}{z_i^0}$ is unknown, then (14) must be rewritten as:

$$\begin{aligned} \mathbf{M}(\pi) &= \frac{z_i^0}{z_j^0}\left(\mathbf{A}_{ij}^\infty - \mathbf{e}_{ji}\mathbf{v}_i^T\right) \\ \mathbf{M}(\omega) &= \frac{z_i^0}{z_j^0}\left(\mathbf{A}_{ij}^\infty - \mathbf{e}_{ji}\mathbf{w}_i^T\right) \end{aligned}$$

Therefore, it is still possible to compute $\mathbf{v}_i^T$ up to a scale factor, but the optimal value for α should be found by imposing:

$$J(\alpha) = \det\left(\mathbf{M}(\pi) + \alpha\widehat{\mathbf{e}}_{ji}\widehat{\mathbf{v}}_i^T\right) = \frac{z_i^0}{z_j^0}$$

where $\frac{z_i^0}{z_j^0}$ is unknown. A solution to this problem is found by considering another pair of parallel plane, π_i', ω_i'. Repeating the same reasoning, we get that the optimal value for α is found when:

$$J'(\alpha) = \det\left(\mathbf{M}(\pi') + \alpha\widehat{\mathbf{e}}_{ji}\widehat{\mathbf{v}}_i'^T\right) = \frac{z_i^0}{z_j^0}$$

Therefore, it is possible to solve for α by posing:
$J(\alpha) = J'(\alpha)$

Fig. 1. The central image is a virtual view obtained from the two side cameras

4. EXAMPLES AND CONCLUSION

In this paper we present a novel approach of view synthesis from uncalibrated cameras based of the Relative Affine Structure proposed by Shasua. We focused on defining a new set of constrains that allows the use of planes parallelism instead of plane at infinity. This allows to use our approach in indoor scenes or in every place where parallel planes can be easily identified (e.g. Fig. 1) without the requirement of a plane at infinity. Lots of tests were performed on synthetic images to evaluate correctness of the algorithm, in Fig.2 we present the virtual camera position following the two proposed approaches (red points for the two parallel planes and black points for the two couples of parallel planes) (Synthetic views are in Fig. 3).

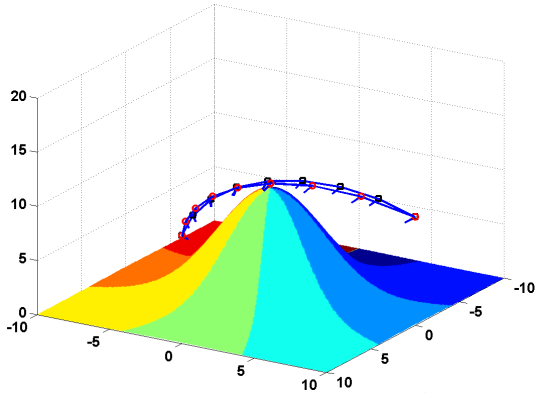

Fig. 2. Virtual Camera Motion estimated from the two different proposed approaches

5. REFERENCES

[1] A. Shashua and N. Navab, "Relative affine structure: Canonical model for 3d from 2d geometry and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 873–883, Sept. 1996.

[2] A. Fusiello, "Specifying virtual cameras in uncalibrated view synthesis," *IEEE Trans. on Circuit and Systems for Video Technology*, vol. 17, pp. 604–611, May 2007.

[3] D. Hogg (coordinator), "The resolv project (reconstruction using scanned laser and video)," *http://www.scs.leeds.ac.uk/resolv*.

[4] S. Seitz and C. Dyer, "View morphing: Synthesizing 3d metamorphoses using image transforms," *SIGGRAPH 96 Conference Proceedings*.

[5] A. Shashua and A. Avidan, "Novel view synthesis in tensor space," *Conference on computer vision and pattern recognition*, pp. 1034–1040, June 1997.

[6] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," *SIGGRAPH 95 Conference Proceedings*.

[7] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. IJCAI, Vancouver, Canada*, pp. 674–679, 1981.

[8] J. Shade, S. Gortler, L. He, and R. Szeliski, "layered depth images," in *SIGGRAPH 98 Conference Proceedings*, 1998.

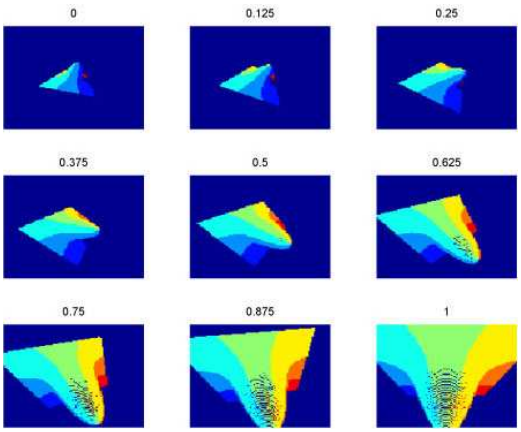

Fig. 3. Virtual views from the virtual cameras in Fig. (2)