VIEW SYNTHESIS FROM A SINGLE UNCALIBRATED IMAGE

U. Castellani, A. Fusiello, N. Mattern
Dipartimento di Informatica - University of Verona, Strada Le Grazie 15, 37134 Verona, Italy
e-mail: {castellani,fusiello}@sci.univr.it

Abstract:

This paper presents a method for generating synthetic views of a soccer ground starting from a single uncalibrated image. The relative affine structure of the players is computed by exploiting the knowledge of the soccer ground geometry and the fact that the players are in vertical positions. Then, novel views are generated using the "plane+parallax" representation to reproject points.

1 Introduction

Nowadays we see an increasing interest in the convergence of Computer Vision and Computer Graphics [4]. One of the most promising and fruitful topics is *Image-Based Rendering* (IBR) [5, 3]. While the traditional geometry-based systems use a 3D model, in IBR novel views are generated by re-sampling one or more example images using appropriate warping functions. A single 2-D image includes 3-D information on scene and camera position. The aim of IBR is the use of these information for the generation of new views while avoiding their explicit extraction. The advantage is that photographs of real scenes can be used as a basis to create very realistic images.
In the case of calibrated camera, algorithms based on image interpolations yield satisfactory results [7]. Uncalibrated techniques utilize image to image constrains such as the fundamental matrix [6], trilinear tensors [1] or the "plane+parallax" [8], to reproject image pixels from a small number of reference images to a given view.
There are few works on view synthesis from a single image. In this case, additional constraints must be used. For example, the symmetry of human faces is exploited in [7].
In this paper we focus on the extraction of the information for the view synthesis starting from a *single, uncalibrated* image. We use the knowledge on the soccer grounds dimension and the fact that the players are in vertical position.

2 Background

In this section we review some background notions needed to understand the paper. A complete discussion and formulation of "plane+parallax" theory can be found in [8, 9]. A more general reference on the geometry of multiple views is [2].
Two views of a planar set of points are related via a homography, that is a non-singular linear transformation of the projective plane into himself. The most general homography is represented by a non-singular 3×3 matrix H.

If $x_i \in I_1$ and $x_i^{'} \in I_2$ are projection in two different views I_1 and I_2 of the same 3D point X_i belonging to some plane Π, we have

$$x_i^{'} \cong H x_i \tag{1}$$

where $\cong$ means "equals up to a scale factor" and points are expressed in homogeneous coordinates. The matrix H has eight degrees of freedom, being defined up to a scale factor: four corresponding points in the two views define a homography (e.g. points x_1, x_2, x_3, x_4 and x'_1, x'_2, x'_3, x'_4 in figure 1).
For a general 3D point X_i (e.g. point X_5 in figure 1), we have

$$x_i^{'} \cong H x_i + k_i e^{'} \tag{2}$$

where H is the homography induced by plane Π, $e^{'}$ denotes the epipole in the second view, and k_i is the *relative affine structure*, which is proportional to the distance of the point X_i from the plane Π.
This equation tells us that points are first transferred as if they were lying on the reference plane Π (like $\bar{x}'_5$ in figure 1), and then their position gets corrected by a displacement $k_i e^{'}$, called *parallax*, in the direction of the epipole with magnitude proportional to the relative affine structure. If $X_i \in \Pi$ then $k_i = 0$ and eq. (2) reduces to eq. (1).
For eq. (2) to hold, it must be suitably normalized, because both the homography matrix H and the epipole are defined only up to a scale factor. To this end, a point x_0 is chosen and scale factors are fixed so as to satisfy:

$$x_0^{'} \cong H x_0 + e^{'}. \tag{3}$$

A very important property is that the relative affine structure is independent of the choice of the second view. Therefore arbitrary "second views" can be synthesized, by specifying a plane homography and the epipole.
This theory is particularly well suited for our case: the reference plane is the soccer ground, and the off-plane points are the players' heads.

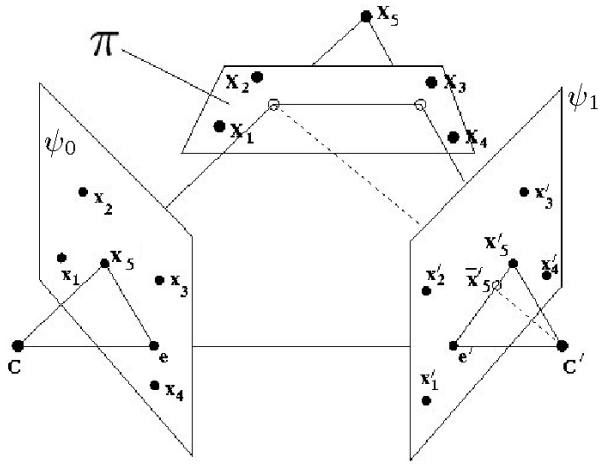

Fig. 1. *Illustration of the relative affine structure.*

3 Players segmentation

Player silhouettes have to be extracted in order to find their position onto the soccer ground. We employ a simple color-based segmentation: non-green regions are labeled as *potential* players, then their size and shape descriptors are used to discard small areas (noise) and terraces in background. We assume that

the players are in a green background (the soccer ground), and there are not intersections between the silhouettes (occlusions).

4 Estimation of the relative affine structure

In order to synthesize geometrically correct arbitrary views we need to first estimate the the relative affine structure of the players with respect to the ground plane.
This could be easily done if one had two or more images. In our case, the key idea to is to synthesize a *zenithal* view[1] of the soccer ground by exploiting the knowledge of its geometry.
The homography matrix H_z between the observed image and the zenithal view is estimated given four point and/or line correspondences. The user interface shows a simplified schematic representation of the soccer ground as seen from the zenith. The user specifies correspondences between the observed image and the schematic one. By applying H_z on the observed image, a synthetic zenithal view is obtained. Supposing that the players are oriented in vertical position, in a *correct* zenithal view the position of the head should corresponds to the feet (if projection is approximately orthographic). However, as we disregarded the the players' 3D structure, they are flattened onto the ground plane, and the segment joining the head and the feet is exactly the parallax.
It is easy to see that the parallax field is radial, and the center is the epipole (see figure 1). Therefore, given the homography between two views and (at least) two off-plane corresponding points one can locate the epipole [2]. Let be $x_5 \leftrightarrow x'_5$ and $x_6 \leftrightarrow x'_6$ two conjugate pairs corresponding to off-plane 3D points (two players' heads, in our case). Let $\bar{x}'_5 = H_z x_5$ and $\bar{x}'_6 = H_z x_6$. Following simple geometric consideration the epipole e' is computed as the intersection between the lines represented (in homogeneous coordinates) by $\bar{x}'_5 \wedge x'_5$ and and $\bar{x}'_6 \wedge x'_6$.
Given homography H_z and the epipole e', we are able to estimate the relative affine structure, but first we need to scale properly the homography. Given a model-point $X_0 \notin \pi$ (the head of one of the players), we set $k_0 = 1$ and then scale H_z by the factor λ_z:

$$\lambda_z = \frac{(x'_0 \wedge H_z x_0)^T (e' \wedge x'_0)}{||x'_0 \wedge H_z x_0||^2} \tag{4}$$

Finally, the relative affine structure k_i for the players is computed. This operation requires the user to specify correspondences between the heads of the players in the two views (observed and zenithal).
From equation (2) we obtain the k_i values:

$$k_i = \frac{(x'_i \wedge e')^T (\lambda_z H_z x_i \wedge x'_i)}{||x'_i \wedge \lambda e'||^2} \tag{5}$$

5 Synthesis of novel views

The relative affine structure k_i are used to synthesize novel views: given a new epipole e'' and a new homography H_n, points are transferred in the third view using an instance of eq. (2):

$$x''_i \cong H_n x_i + e'' k_i. \tag{6}$$

The user interface shows a schematic representation of the soccer ground. The user can rotate and translate a virtual camera in order to decide the new point of view . Then he/she specify the correspondences between the observed image and the schematic view needed to compute the homography H_n.
The interface shows the new view by evidencing the feet positions. The user introduces the positions of the corresponding heads and the new epipole is estimated as described in section 4. The scale factor λ_n is also estimated as before, setting x_0 to be one of the heads of the players used for the epipole.

[1] A zenithal view is an image taken the with optical axis orthogonal to the soccer ground plane.

The positions of the feet and the heads on the new views are computed, then the other parts of the body are linearly interpolated, as players are considered 2D silhouettes (more precisely they are processed as billboards). Feet positions are recovered from equation (6) with $k_i = 0$, whereas the head positions are obtained from the k_i values estimated as described in section 4. Finally the players are overlapped to the new view with the correct dimensions.

6 Examples

Figure 2 shows an example of the segmentation.

Fig. 2. *Actual image (left) and segmented image (right). Circled regions have been classified as players.*

In this section some experiments are presented. Figure 4.b is a synthetic view recovered from image 4.a in which the virtual camera aims toward the goal.

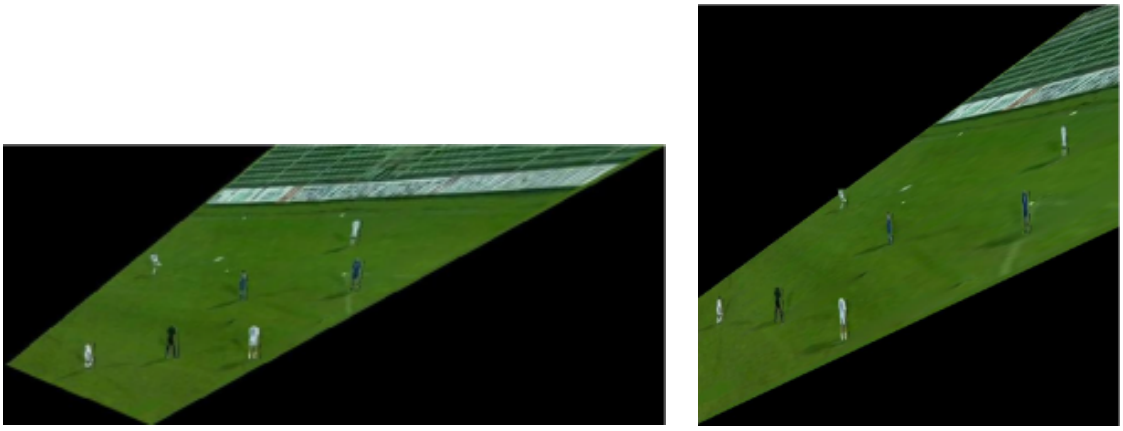

Fig. 3. *Actual image (left) and synthetic view (right)*

Figure 3 shows an example of synthetic views recovered from the image shown in figure 2, in which the new point of view allows us to find the players in offside. Please note as the players appears correctly foreshortened in the synthetic views.

7 Conclusions

In this paper we introduced a method for synthesizing novel views of a soccer ground, starting from a single uncalibrated image. We described the geometric relationships that enables us to first to estimate some projective invariants and then transfer points from the reference view to the synthetic one. Further

Fig. 4. *Actual image (left) and synthetic view (right)*

works could address the improvement of the segmentation using sophisticated classification techniques and the reduction of user intervention. At the moment the manual introduction of the correspondences is a crucial phase.

References

[1] AVIDAN, S., AND SHASHUA, A. Novel view synthesis in tensor space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (1997), pp. 1034–1040.

[2] HARTLEY, R., AND ZISSERMAN, A. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[3] KANG, S. B. A survey of image-based rendering techniques. Tech. Rep. CRL 97/4, Digital Cambridge Research Laboratories, August 1997.

[4] LENGYEL, J. The convergence of graphics and vision. *IEEE Computer 31*, 7 (July 1998), 46–53.

[5] MCMILLAN, L., AND BISHOP, G. Plenoptic modeling: An image-based rendering system. In *SIGGRAPH 95 Conference Proceedings* (August 1995), pp. 39–46.

[6] S. LAVEAU, O. F. 3-d scene representation as a collection of images and foundamental matrices. Technical Report 2205, INRIA, Institut National de Recherche en Informatique et an Automatique, February 1994.

[7] SEITZ, S. M., AND DYER, C. R. View morphing: Synthesizing 3D metamorphoses using image transforms. In *SIGGRAPH 96 Conference Proceedings* (August 1996), pp. 21–30.

[8] SHASHUA, A., AND NAVAB, N. Relative affine structure: Theory and application to 3d reconstruction from perspective views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (1994), pp. 483–489.

[9] SHASHUA, A., AND NAVAB, N. Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*, 9 (September 1996), 873–883.