VISIRE. PHOTOREALISTIC 3D RECONSTRUCTION FROM VIDEO SEQUENCES

Tomas Rodriguez

Eptron SA. Spain
R&D Depart. Madrid
tomasrod@eptron.es

Peter Sturm
Marta Wilczkowiak
Adrien Bartoli
Matthieu Personnaz

INRIA Rhône-Alpes
Montbonnot, France
Peter.Sturm@inrialpes.fr

Nicolas Guilbert
Fredrik Kahl
Martin Johansson
Anders Heyden

Centre for Math. Sc.
Lund Univ. Sweden
heyden@maths.lth.se

Jose M. Menendez
Jose I. Ronda
F. Jaureguizar

E.T.S.I. Telecom.
Univ. Pol. Madrid
jmm@gti.ssr.upm.es

ABSTRACT

Traditionally, building 3D reconstructions of large scenarios such as a museum or historical site has been costly, time consuming and required the contribution of expert personnel. Usually the results showed an artificial look and had little interactivity. However, newly developed technologies in the areas of video analysis, camera calibration and texture fusion allow us to think in a much more satisfying scenario where the user with the only aid of a domestic video camera is able to acquire all the information it is required to construct the 3D model of the desired environment in an easy and comfortable manner.

In this paper, the results obtained in the EC funded project VISIRE are presented. VISIRE attempts to construct photorealistic 3D models of large scenarios using as input multiple freehand video sequences. Once acquired, the computer vision software will process the video information off-line in order to obtain the 3D mesh together with the textures required to obtain a 3D model highly resembling the original.

1. INTRODUCTION

One of the major challenges in the fields of computer vision and computer graphics is the construction and representation of life-like virtual 3D environments within a computer. Many different techniques have been developed up to this moment to achieve this difficult task, like the so-called active methods, which use a combination of a “marking device” (like laser or structured light) and images to produce a 3D model. However, the scientific community immediately saw the convenience of using passive methods like video cameras to infer the 3D structure of the world and build complex models out of image content.

This is in principle possible by just triangulating the corresponding points from two images of the same scene. However, to get exact Euclidean structure, it is required to first off-line calibrate the cameras, i.e. the determination of both internal and external parameters of the camera system. Off-line calibration using a known calibration grid (for instance, Tsai’s method) has been intensively studied in the past. Robust and efficient methods do exist for different applications. However, off-line calibration based on external grids lacks flexibility for real 3D authoring from images, as we just could not put the calibration grid in any scenario.

Autocalibration is the recovery of metric information from only point correspondences of uncalibrated images, using geometric self-consistency constraints. The potential advantages of autocalibration are a reduced need for off-line calibration and greater on-line flexibility.

Several attempts have been made in the past to develop autocalibration based 3D reconstruction systems: Cumuli, Vanguard. However, current systems are mainly based on a small set of images, ranging from 2 to 20 images in today’s very complex systems. VISIRE is a project that aims to break this barrier in several orders of magnitude, mainly because it can use the thousands of images found in video streams to generate the desired results. No system so far has ever tried to accomplish such a complex scenario, in any of the above-mentioned tasks.

Another important goal is to produce photorealistic 3D models, i.e. models that may be realistically rendered from synthetic viewpoints. One way of doing so is to apply IBR (Image-Based Rendering). The other main method is to enhance a geometrical 3D model with texture maps or other information e.g. surface reflectance properties. Up to now, we concentrate on the second solution, which may allow to get a more compact scene description (nevertheless, one of our goals for future research is to combine the 3D and IBR approaches, applying IBR for scene parts whose geometry can not be modelled well). Creation of photorealistic 3D models is done in two major steps here: based on 3D points reconstructed during auto-calibration or subsequently, we first generate a triangular mesh describing the scene’s surfaces. Then, texture maps for the surface patches are extracted using all available images. The challenge for the 3D mesh generation process is that most existing methods are designed for sets of dense and regularly distributed 3D points. This is usually not the case in automatic structure from motion, so we have developed methods that use information provided by the input images, via visibility and photoconsistency constraints.

The computer vision software has come a long way towards process automation. However it is our belief that current state of the art does not allow for fully automatic quality 3D reconstruction. For that reason, it is assumed that user interaction will be necessary at a certain point to compensate for processes that are either too complex or too slow. For that purpose, an adequate authoring tool has been designed to enable effortless user interaction with the system.

The complete chain of the computer vision approach developed in the VISIRE project is based on the achievement of the following steps:

1. Feature Analysis. Described in section 2.
2. On-line calibration. Described in Section 3.
3. 3D reconstruction and mesh generation. Section 4.

Finally, in sections 5 and 6 the VISIRE authoring tool and the conclusions are presented.

2. FEATURE ANALYSIS

Feature extraction and tracking is an important activity in computer vision, as many algorithms rely on the accurate location of a set of points in the image, and on the tracking of those moving points through the image sequence. The selection of the points is a critical task, as they should present any kind of feature that makes them easily detectable, from different points of view. Classically, those points are selected based on some measure of texturedness or cornerness, that is, with high derivative values in more that one spatial direction. Additionally, some a priori hypothesis are applied, such as slow motion of the objects in the scene, high video rate in the acquisition camera (compared to the velocity of the captured objects), and almost neglectable change in the illumination between consecutive frames.

The selected approach for feature extraction in VISIRE relies on the Harris approach, which computes a matrix linked to the image autocorrelation function that takes into account the values of the first spatial derivatives of the image over a limited window. Later, a measure of cornerness is applied that evaluates the determinant and the trace of that matrix. The approach is optimal in the case of pure orthogonal corners, and the implementation of VISIRE follows the enhancement of[1] for the computation of the derivatives.

Feature tracking is applied through a robust method, based on the Kanade-Lucas-Tomasi approach. It relies on the assumption of affine motion field in the projection on the 2D world of the image of the 3D point motion, and the computation of a weighted *dissimilarity* function between consecutive images that is minimized through least-mean squares, over a limited spatial window.

After the trajectories are generated along time, they are validated by checking their compliance to the assumed model of rigid motion of the scene. To this purpose the set of trajectories is processed in two steps corresponding respectively to a local processing and a global processing. For the local processing the sequence is divided into non-overlapping temporal windows and for each of them a RANSAC-based calibration is performed, which is later optimized by bundle adjustment. The temporal windows must be short enough to ensure that enough trajectories remain complete within it but long enough to include enough motion. Local data from different temporal windows are consolidated and reoptimized in the global processing step, which operates iteratively consolidating data from pairs of adjacent windows. After this analysis a trajectory or part of a trajectory results validated when it is successfully approximated by the reprojection of a scene feature.

3. ON-LINE CALIBRATION

Once a sufficient amount of reliable image correspondences have been established the overall structure of the scene may be recovered. This recovery is performed in two major steps, namely the recoveries of projective and Euclidian structure respectively. Projective structure is recovered by extracting camera matrices from the trifocal tensor, see e.g. [2] and followed by a series of the resectionings in order to obtain each of the subsequent cameras. However, the result is only defined up to a projective transformation, and is consequently useless for visualization purposes. However, as shown in [3], very general constraints such as assuming square pixels suffice to establish the in- and extrinsic camera calibration parameters. In VISIRE, the actual implementation makes use of the Cheirality inequalities, see e.g. [2] followed by an identification of the plane at infinity and eventually the recovery of the intrinsic and extrinsic camera parameters and Euclidian structure.

Nevertheless, the estimation of the initial set of cameras depend on the solution of a linearized problem, and are consequently subject to errors. Hence, in order to achieve a maximum likelihood solution, so called bundle adjustment is applied, see eg. [4] for details. This involves minimizing the reprojection error

$$\sum_{i,j} \|(x_{ij} - p(P_i, X_j))\|^2$$

where x_{ij} indicates the j'th image point in the i'th image, and $p : (P, X) \mapsto \mathbb{R}^2$ projects the 3D homogeneous point X using the camera matrix P. In VISIRE, bundle adjustment is implemented using the Levenberg-Marquardt method and a sparse system solver, allowing for significantly more effective processing and for longer sequences than previously, i.e. sequences of up to 300 views and 15000 3D points.

Another unique feature of the VISIRE system is the ability to apply the constraint of *closedness* to a sequence. In a long sequence, the same image feature is likely to appear on several occasions, but will however under normal conditions be reconstructed as a different 3D feature each time. Also, as the error accumulates during the reconstruction process, these multiple instances of the same 3D feature might actually be far from coincident. For the general case, enforcing identity on these features turns out to be indispensable. The basic idea is to distribute the accumulated error equally on all the parameters, although in the norm given by their covariance. Specifically, the reprojection error vector and its associated covariance structure are projected onto their respective lower-dimensional manifolds corresponding to the reduced system, i.e. the system where identity of the parameters has been enforced. Using the resulting values, the optimal reduced parameters are calculated through the equivalent of a Levenberg-Marquardt iteration.

4. 3D REGISTRATION AND MESH GENERATION

After on-line calibration, we are provided with a set of 3D points, projection matrices of a set of images, and 3D-to-2D point correspondences. Usually, only interest points that could be tracked reliably, were used for on-line calibration and metric 3D reconstruction. However, once projection matrices are known, additional point tracks can be checked more easily for outliers and if the correspondingly reconstructed 3D points are reliable. We may thus enrich the set of 3D points before proceeding to mesh generation.

Most existing methods for mesh generation from 3D points rely mainly on 3D geometric reasoning, e.g. proximity constraints between points, see e.g. [5, 6] and references therein. These methods give unusable results for our input data, because they are designed for rather dense and regular point sets. In order to work with more difficult data, other information besides pure 3D geometry should be used: since the 3D points are obtained by reconstruction from images, we have such additional information. First, visibility constraints can be imposed to prune incorrect surface patches, e.g. a hypothetical patch that would lie between a 3D point and the optical center of a view where that point is visible, can be rejected. Other, more complicated, visibility constraints are also used and found to be necessary in practice: for example, a surface patch that partially occludes another one, without occluding an actual 3D point, is rejected. Such visibility constraints were used in [7], were a surface mesh is built incrementally, starting with a mesh obtained by a Delaunay triangulation in one view, and then rejecting and adding triangles based on visibility constraints of one additional view after the other. We proceed differently, by iteratively adding new triangles to manually or automatically selected seed triangles, and thus by letting a mesh grow, directly ensuring all available visibility constraints (and other constraints, see below). This way, we may end up with a better connected surface mesh, which may be easier to edit/complete if necessary.

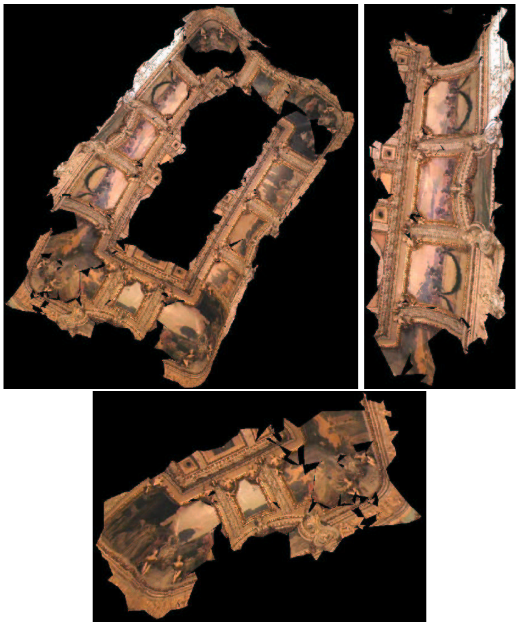

Fig. 1. Ceiling of the Royal Hall. Automatic 3D Model.

Another constraint we use to test hypothetical surface patches, is photoconsistency: a planar patch is acceptable, if its projections into all images where the patch's vertices are visible, correspond to image regions with the same "texture". This is verified by the following process: Image regions corresponding to a planar 3D surface patch, are warped (using homographies associated to the 3D plane) to some common frame (to undo perspective effects). The simplest method is to perform some kind of multi-image cross-correlation using all warped image regions (e.g. compute variance of greylevels) [8]. Optionally, we can apply an extension of this mechanism that takes into account variations in apparent illumination (due to viewpoint and lighting changes) as well as partial occlusions in some input images (occlusions due to unmodelled small objects, but also specularities appearing in a few images). This is done using an M-estimator (IRLS, Iteratively Reweighted Least Squares), whose parameters are a mean texture and coefficients for greylevel correction (different models are implemented, e.g. one affine transformation per color channel and per image). Partial occlusion is dealt with by using a robust influence function to weight residuals (e.g. Huber-function). This process is time-consuming and not always sufficiently discriminative for small patches, thus not routinely used for testing hypotheses of surface patches. However, we use it for the generation of texture maps for the final surface mesh.

5. THE AUTHORING TOOL

Fully automatic 3D reconstruction of complex environments is currently not a practical possibility for a number of reasons. Manual intervention is still required whenever the computer vision software is not able to cope with singular situations, when automatic procedures take too much time to complete or simply when the user desires to add his personal touch.

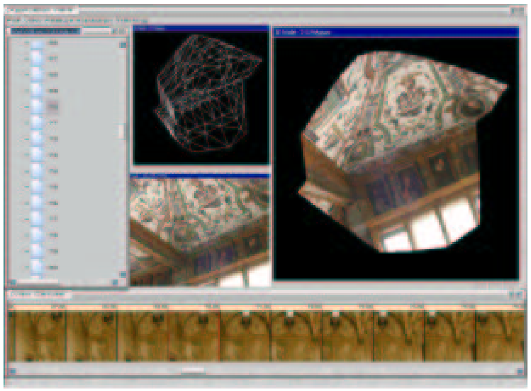

Fig. 2. Screenshot of the VISIRE authoring tool

The VISIRE user interface is a combination of a video editor, a 3D browser and a 3D authoring tool, which has been designed specifically to manage the underlying computer vision software. Optimisation of the production chain has always been the leading objective of VISIRE project. There are many situations when a few keystrokes may save hours of computational time or significantly improve the quality of the results. In that sense VISIRE offers an integrated GUI that allow the user to modify the system parameters at every step of the computer vision chain. When the user modifies a parameter, the system automatically recalculates all related links and updates the resulting 3D mesh. A set of interactive tools have been created to guide the user alter or substitute the automatic mesh generation procedures in view of a best performance. The GUI also includes tools for retouching and correction of possible errors or mismatches automatic procedures may have caused.

The VISIRE Authoring Tool has the following main characteristics: integrated full video editor, support for multiple input video sequences, compatible with most popular video formats, VRML output, integrated 3D browser, direct configuration and control of integrated computer vision software, multiple interactive tools (feature analysis, calibration, 3D rendering, texture processing), hot display capabilities, the result of changed parameters is immediately updated in the 3D model, etc

6. CONCLUSIONS

VISIRE has contributed to advance the state of the art in the computer vision field towards the final objective of fully automatic 3D reconstruction of complex environments. The project succeeded in the use of video information, acquired from domestic camcorders, as input to a near-automatic computer vision based 3D mesh generation system. VISIRE aims at the production of ”complete” photorealistic 3D models rather than the partial reconstructions found today.

Several issues previously unexplored in this type of applications were approached in the project: reliable tracking in sparse environments, introducing constrains such as planarity and closedness, occlusion reasoning, optimised autocalibration, combining and adapting textures from multiple viewpoints, multiresolution, integrating manual and automatic mesh generation methods, etc. The field, however, is still open to further research: better treatment of partial occlussions, handling of reflections, matching in repetitive structures, irregular sampling, considering additional constrains such as pure rotations and planar surfaces, etc.

7. REFERENCES

[1] C. Schmid, *Appariement d’images par invariants locaux de niveaux de gris. Application à l’indexation d’une base d’objects*, Ph.D. thesis, INPG, 1996.

[2] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge U. Press, 2000.

[3] A. Heyden and K. Åström, “Euclidean reconstruction from image sequences with varying and unknown focal length and principal point,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 1997, pp. 438–443.

[4] W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, “Bundle adjustment: A modern synthesis,” in *Vision Algorithms: Theory and Practice*. Springer, 2000.

[5] H. Hoppe, *Surface Reconstruction from Unorganized Points*, Ph.D. thesis, Department of Computer Science and Engineering, University of Washington, 1994.

[6] S. Petitjean and E. Boyer, “Regular and non-regular point sets: Properties and reconstruction,” *Computational Geometry – Theory and App.*, vol. 19, 2001.

[7] A. Manessis, A. Hilton, P. Palmer, P. McLauchlan, and X. Shen, “Reconstruction of scene models from sparse 3d structure,” in *Proc. Conf. Computer Vision and Pattern Recognition, Hilton Head Island, USA*, 2000.

[8] D.D. Morris and T. Kanade, “Image-consistent surface triangulation,” in *Proc. Conf. Computer Vision and Pattern Recognition, Hilton Head Island, USA*, 2000.