MODEL-BASED VIDEO CODING WITH GENERIC 2D AND 3D MODELS

Gaël Sourimant, Luce Morin

gael.sourimant,luce.morin@irisa.fr
IRISA / INRIA Rennes, Campus Universitaire de Beaulieu,
Avenue du Général Leclerc, 35042 RENNES Cedex - France

ABSTRACT

3D extraction from video gives a representation adapted to low bitrate coding and provides enhanced functionalities such as 3D cues for augmented reality and interactive navigation in photo-realistic environments. But for degenerated motions of camera, like pure rotation, 3D information can not be retrieved. In this article we propose an original representation based on a hybrid 2D/3D models stream. The idea of this approach is to provide a modelling for all sequences including those with rotations. The sequence is divided into portions and for each one the motion of the camera is identified. Depending on the type of motion a 3D model, a cylindrical mosaic or a spherical mosaic is extracted. They are constructed in order to be suitable for an homogeneous visualization process. Results are shown for synthetic and real video sequences.

1. INTRODUCTION

1.1. Context

3D model based video coding consists of the representation of a video with one or several 3D models of the captured scene. By reprojecting these 3D models one obtains a virtual sequence similar to the original but with enhanced functionalities such as augmented reality, free view point generation or lighting changes. Furthermore, 3D model based representations are more compact than image based-ones.
3D model extraction from images is based on structure-from-motion and thus requires different view points of the same scene. In particular, a camera undergoing a pure rotation does not allow the recovery of 3D information as there is no intersection between different lines of view (see Figure 1). Shape from motion thus requires that camera motion is assumed not to be a pure rotation. On the other hand mosaics are very well suited to represent a video obtained with rotational motion. So here we propose an original 2D/3D hybrid method based on both 3D models and mosaics. The aim is to deal with all types of video representing a fixed scene including those acquired with a camera undergoing pure rotational motion.

1.2. Previous work

1.2.1. 3D Modelling

Retrieving 3D information from videos has long been studied in the field of computer vision [1, 2]. However it usually assumes a video sequence specifically acquired for enabling 3D reconstruction, or it requires a human interaction in the reconstruction or matching steps [3, 4].

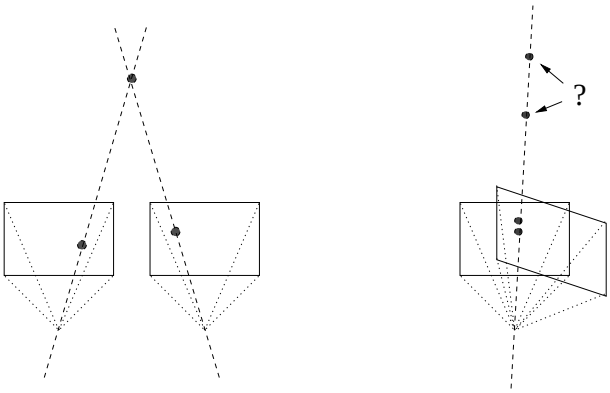


Figure 1: *Camera translation is required for structure-from-motion: in case of pure rotation, viewlines are superimposed*

In 3D model-based coding, acquisition conditions are not constrained but assumptions are made on the scene contents: an a priori known 3D model of the scene contents is available whose pose, texture (and possibly non rigid deformation) are estimated from the video. This approach is very efficient for coding video with specific contents such as visiophony [5].
As our goal is to propose a 3D representation for any video sequence, we thus do not want to make assumptions on camera parameters, scene contents or video length. In this context Galpin proposed a method based on a 3D model stream [6] instead of aiming at a unique realistic model of the scene. Each model is valid for a portion of the original sequence called a GOP (Group of Pictures). These GOPs are delimited with key images that are automatically chosen. For each GOP a 3D model is automatically estimated and inter-GOP coherence is allowed by a sliding adjustment [7].

1.2.2. Mosaics

Mosaics can be obtained by homographic, cylindrical or spherical projection [8, 9, 10].
Homographic mosaics are well adapted to reconstruct planar scenes and can also be used in pure rotational cases. However, cylindrical and spherical mosaics are better adapted to pure rotational cases: they allow large rotations, and avoid the distortion of pictures that are far from the reference image.

1.2.3. Model selection

Model selection in a general case has been studied by Kanatani [11, 12] based on a combined residual/complexity criterion approach. It has been used for motion model selection in video sequences by Berger [13] and Torr [14]. Selecting the motion model based on the model complexity is not suitable for our approach as we want to favor 3D modelling and use 2D modelling only when it fails. We thus use criteria based on residuals only and a selection algorithm favoring 3D modelling.

2. OVERVIEW OF OUR APPROACH

This section presents an overview of our hybrid 2D/3D representation for video, which is based on the scheme proposed by Galpin [7] and Morillon [15].

2.1. General principle

Our approach is based on an analysis-synthesis scheme. In the analysis step, the video is partitioned into *groups of pictures* (GOPs) (figure 2) and for each GOP, a textured 3D model (either general, cylindrical, or spherical) is computed. Each model is associated with a set of camera positions, one for each video frame in the GOP. The stream of 3D models is encoded and transmitted. At the receiver, synthesis is performed on the flow to rebuild the original video sequence (figure 3): each 3D model is textured and rendered using the transmitted camera positions, which provides the reconstructed frames for the current GOP.

2.2. Important notions

- The GOPs are delineated at their extremities by two particular images : the *key-frames*.
- Two consecutive GOPs G_1 and G_2 share respectively their last and first key-frames (see figure 2).
- A 3D mesh is associated to each GOP and its texture is either the first key-frame in the general case, or a mosaic regrouping all the GOP's frames in case of pure rotational motion.
- 2D GOPs refer to GOPs where camera motion is a pure rotation. As in this case 3D information can not be retrieved, these video parts are modelled as 2D mosaics. Mosaics are warped onto a cylindrical or spherical 3D mesh for allowing an homogeneous rendering procedure, but they essentially are 2D models.

2.3. Hypothesis

We remind here the main hypothesis on which is based our method.

- The scene is supposed to be static, or at least motion-segmented.
- It has been shot by a moving monocular camera.
- The camera motion is not constrained.
- Neither the intrinsic parameters nor the extrinsic parameters are known.
- The focal length is fixed (no zoom).
- No hypothesis is made on the scene contents except that there are no, or only few, specular surfaces.

2.4. Algorithm steps

We now describe each step in the proposed algorithm, illustrated in figure 4.

Motion Estimation: The first step is to estimate the motion field throughout the GOP, i.e. the displacement vectors for each pixel between the first and the last frames. This is carried out thanks to a 2D deformable mesh [16, 17, 18]. The motion estimation between two frames I_t and I_{t+1} is performed by netting regularly I_t with a 2D six-valency mesh, and by searching for each node of this mesh the motion that minimizes the quadratic error between a frame and the following, which is motion-compensated. The motion $\vec{u}$ of each pixel inside a triangle is given by the balanced sum of the motion of each vertices for this triangle. Motion throughout a given GOP is finally computed by accumulating the vector fields of the successive images.

Extraction and tracking of interest points: Feature points extraction is performed in the first frame using a Harris corner detector [19] on the set of 2D mesh nodes. Interest points tracking is then deduced from the motion estimation computed earlier.

Camera pose estimation: The camera motion estimation is performed using epipolar geometry. Since we need the intrinsic parameters of the camera to perform such a task, and since self-calibration methods are still unstable and computationally expensive, we prefered to use rough values for the intrinsic parameters.

Parallel mosaic creation: Since we still do not know which type of model will better fit the input images, a mosaic is computed in parallel to motion estimation. Image registration is carried out thanks to the 2D mesh motion estimator. It will be abandoned later if a classical 3D model is better suited.

Depth map estimation: The depth map is deduced from the disparity map (i.e. the motion field) by classical triangulation, using estimated camera parameters.

3D reconstruction: When the next key-frame is selected the corresponding model is built. If a 3D GOP can be constructed, we apply a regular triangular mesh on the depth map and displace the nodes to the corresponding depth. Otherwise a 2D mosaic model is built.

Final reconstructed sequence visualization: The final reconstructed sequence can be viewed by a dedicated software, which takes as input data the different models with their textures and associated camera pose. In addition to viewing some specific post treatments are available, such as global illumination modification, changes in the camera path for virtual navigation or creation of a stereo sequence for immersive visualization.

3. KEY FRAMES AUTOMATIC SELECTION

The size of the GOPs is not fixed. It is determined by the video data. We therefore need to select key-frames in the sequence that will delineate those GOPs. Key-frames choice is very important because it will determine the viability of the final reconstruction. This is done in an fully automatic way, with regard to different

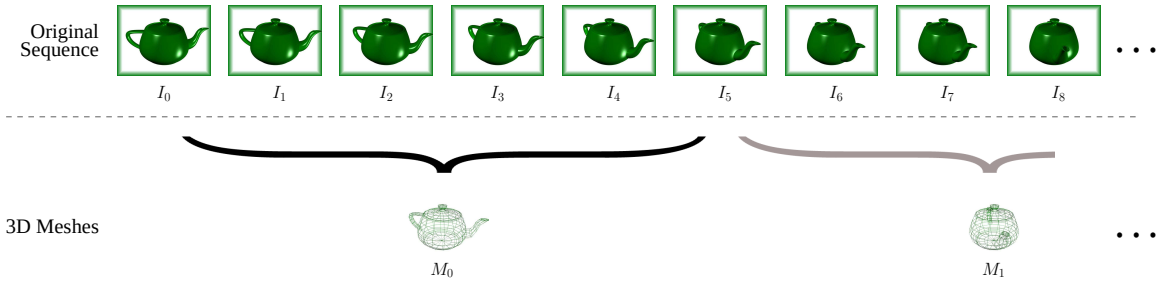


Figure 2: General principle of a 3D models flow representation.

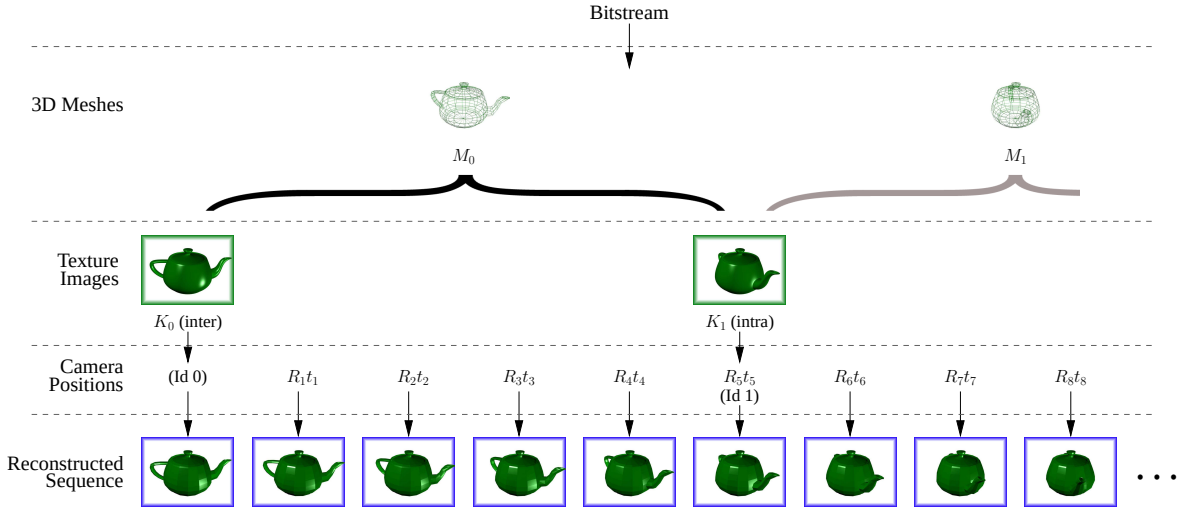


Figure 3: Synthesis step: 3D models, texture images and camera positions are used to reconstruct the original frames.

criteria computed on the basis of the tracked interest points analysis. Moreover we wanted our key-frames selection to fit the following constraints :

- It must provide the GOP type among 3D, panoramic and spherical.
- It has to favor 3D type GOPs and use 2D GOPs only if 3D is not possible.
- In case of a 3D type GOP, a tri-dimensional reconstruction must be possible.
- It also has to maximize the GOPs size in order to avoid redundancy, and ensure that the reconstruction will be made under wide enough baseline conditions.

3.1. Selection criteria

The selection criteria are inspired by the ones defined by Galpin [18] and Morillon [15] in the case of 3D and panoramic GOPs. We have adapted the criteria to the spherical case and completed the algorithm to deal with any sequence of 3D, cylindrical and spherical type of motion. The criteria are estimated for each current frame I_{t+p} and they determine whether or not the current frame is chosen as the next frame K_{t+1} which ends the GOP.

3.1.1. *Apparent displacement* C_d

$D_{t,t+p}$ represents the average motion of the points between the current image I_{t+p} and the key-frame I_t that precedes in the sequence, and the criterion testing $D_{t,t+p}$ is defined as:

$$C_d \Leftrightarrow D_{t,t+p} > S_d$$
$$\text{where} \quad D_{t,t+p} = \frac{1}{N_{t+p}} \sum_{i=1}^{N_{t+p}} \|\vec{u}(m^i_{t,t+p})\| \tag{1}$$

with N_{t+p} the number of tracked points from the last key-frame I_t until current image I_{t+p}, $\vec{u}$ the motion vector of point m^i between I_t and I_{t+p}, and S_d a threshold on the average pixels motion. S_d has been fixed experimentally to 10 pixels. This criterion ensures that the estimation of a depth map will be done under satisfying baseline length.

3.1.2. *Common view field* C_p

C_p supervises the percentage of common points between I_t and I_{t+p}, and is expressed as :

$$C_p \Leftrightarrow \frac{N_{t+p}}{N_t} > S_p \tag{2}$$

with S_p the threshold on the remaining points percentage. This criterion ensures that two key-frames share a sufficient common

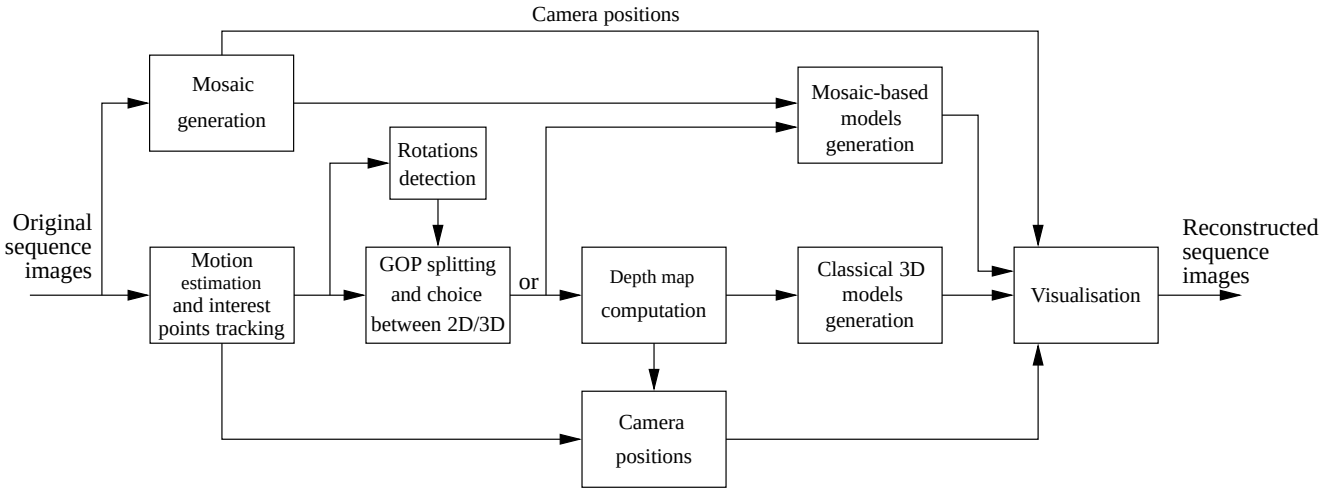


Figure 4: Algorithmic chain

information to build a valid model of the portion of sequence they delineate. It has been fixed to 70%.

3.1.3. Epipolar residual

C_e is defined as:

$$\begin{aligned} C_e \Leftrightarrow \tfrac{1}{N_{t+p}} \sum_{i=1}^{N_{t+p}} (\mathfrak{D}^i_{t,t+p} + \mathfrak{D}^i_{t+p,t}) < S_e \\ \text{where} \quad \mathfrak{D}^i_{t,t+p} = d^2(\tilde{m}^i_t, F_{t,t+p}.\tilde{m}^i_{t+p}) \\ \text{and} \quad F_{t+p,t} = F^{\mathbf{t}}_{t,t+p} \end{aligned} \tag{3}$$

with $\tilde{m}_t$ an interest point in the image I_t expressed in homogeneous coordinates, $\tilde{m}_{t+p}$ its correspondent in I_{t+p}, $F_{t,t+p}$ the estimated fundamental matrix and S_e a threshold on the matching precision, that is fixed to 0.5 pixels. This criterion permits to check the epipolar residual computed from the fundamental matrix and matched points between the last key-frame and the current frame. It therefore ensures that the 3D model will be re-projected on the following key-frame with a sub-pixel error, and that the camera motion as well as the fundamental matrix are consistent with the motion field given by the mesh-based dense estimation.

3.1.4. Rotational motion : C_r

The criterion C_r allows to detect rotational motion. It is defined as:

$$C_r \Leftrightarrow \frac{E_r}{D_{t,t+p}} < S_r$$

The image transform induced by a pure rotation (planar homography) which best fits the interest points motion is estimated. The mean residual is computed as:

$$E_r = \frac{\sum_{i=1}^{N} \sqrt{(x^i_1 - x^i_2)^2 + (y^i_1 - y^i_2)^2}}{N} \tag{4}$$

where (x^i_1, y^i_1) and (x^i_2, y^i_2) are the cartesian coordinates of the interest points m^i_1 and m^i_2 respectively in the first and last image in the current GOP. C_r involves rotational residual with respect to mean image displacement. It thus consider relative contribution of rotational and translational motion in the observed image displacement. Threshold S_r is experimentally fixed to 0.05 pixels. It has to be noticed that when C_r is *true* (i.e. a rotational motion has been detected), then the epipolar criterion C_e is not significant.

3.2. Selection algorithm

We present here the key-frame selection algorithm using the previously defined criteria. It is executed for each current frame, and indicates, if a key-frame is selected, whether it delineates a 2D or a 3D GOP. The current analyzed GOP is by default classified as *unknown*, and if at time t the 3D scene geometry is sufficiently well estimated the GOP is then classified as one of $3D$ type. It can be labelled as $2D$ only if there are not enough tracked points in the sequence and if the epipolar criterion has never been true. The end of a $2D$ GOP occurs as soon as C_r becomes false. We can see in figure 5 the precise scheme of this algorithm.

4. 3D MODEL GENERATION

Once the GOP type is determined the corresponding model is built. For 2D GOPs, 2D planar models would be sufficient. However, in order to get an homogeneous processing at the decoder, we also want to represent 2D GOPs as textured 3D meshes and a set of camera positions, allowing to reconstruct the original frames in the GOP. We build a cylinder in case of vertical oriented rotations, and a geosphere in case of general rotations. Both have a mosaic associated as texture.

4.1. 3D models

In a general motion context, once a key-frame is selected to terminate a 3D GOP, the corresponding model is built. First, the disparity map is used in a retro-projection phase to compute the depth map for each pixel. Note that we want this retro-projection to be perfect on the first image of the GOP. Then a uniform mesh

3Dfeasible = $false$;
GOPtype = $unknown$;

If C_d **Then**
  If C_p **Then**
    If $C_e \wedge \neg C_r$ **Then**
      3Dfeasible = $true$; ContinueGOP;
    Else
      If 3Dfeasible **Then**
        Finalize3D;
      Else ContinueGOP;
  Else
    If $C_e \wedge \neg C_r$ **Then**
      If GOPtype=2D **Then** Finalize2D;
      Else Finalize3D;
    Else
      If 3Dfeasible **Then** Finalize3D;
      If (GOPtype==2D & !C_r **Then** Finalize2D;
      GOPtype=2D; ContinueGOP;
Else $idle$

Figure 5: Key-frames selection algorithm

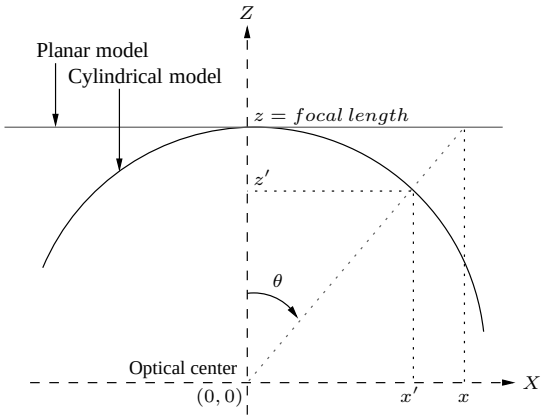


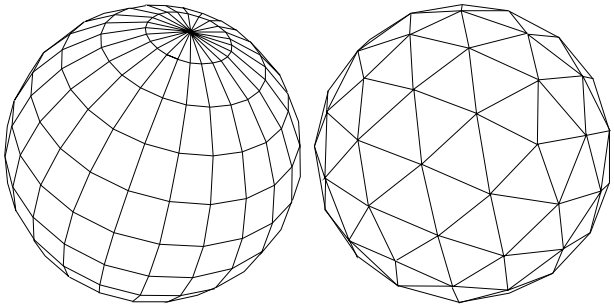
Figure 6: *From planar to cylindrical mesh*

is applied on the depth map, and each of his nodes is moved to the corresponding depth.
The use of a uniform mesh instead of an adaptive one is justified by the fact that in the uniform case we only have to transmit in a predefined way the depths of each node, whereas in the adaptive case we would have to transmit the entire topology of the model.

4.2. Cylindrical model

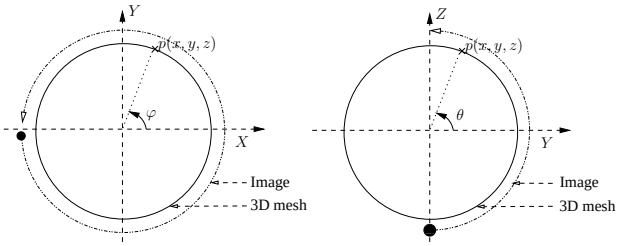
In the case of a panoramic motion, the 3D model generated is a cylinder centered on the camera optical center, and whose radius is equal to the focal length. The computed panoramic mosaic is mapped onto the 3D cylinder in such a way that the reprojections of the model with the virtual camera will generate the original frames. This procedure is very simple and can be seen as a transformation from cartesian to cylindrical coordinates (see figure 6):

$$\begin{cases} x' = fsin\theta \\ z' = \sqrt{f^2 - x'^2} \\ \theta = tan^{-1}(\frac{x}{f}) \end{cases} \quad (5)$$



(a) Sphere (b) Geosphere

Figure 7: *3D spherical model types comparison*



(a) Horizontal mapping (b) Vertical mapping

Figure 8: *Image mapping on a sphere*

4.3. Spherical models

In the case of general rotational motion, the scene is reconstructed as a sphere centered on the camera center. As in the case of panoramic motion, the corresponding texture is a mosaic generated from the different frames of the GOP. In order to provide models which can be efficiently coded, the mesh is of geospherical type, in opposition to classical spherical type. This is justified by the fact that a sphere is latitude and longitude-based, whereas as geosphere is constructed from a refined regular polyhedron (here an icosahedron). As a consequence the vertices of a geosphere are more homogeneously distributed among the mesh surface (especially around the poles on the sphere, see figure 7), and the amount of information is less important for an equivalent viewing quality. As for the cylindrical models, we compute the texture coordinates of the mosaic so as the rendering with the virtual camera is identical to the original frames. A transformation from cartesian to spherical coordinates is then performed for the mosaic texture coordinates. The respective horizontal and vertical transformations are illustrated in figure 8, where

$$\begin{cases} \varphi = tan^{-1}(\frac{y}{x}) \\ \theta = cos^{-1}(\frac{z}{f}) \end{cases} \quad (6)$$

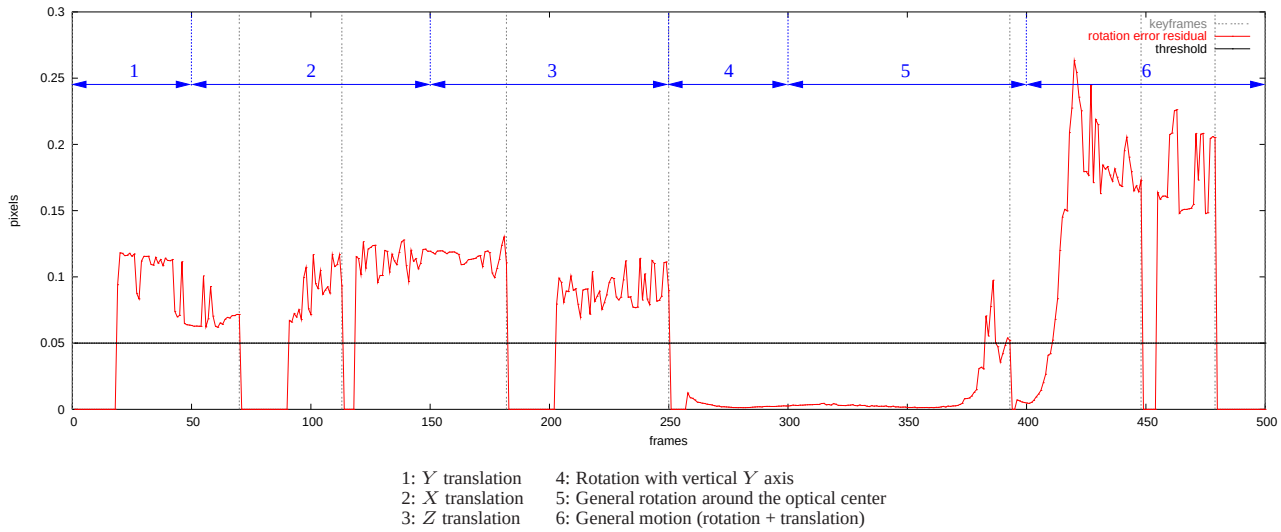


Figure 9: *Evolution of the rotation error residual through the sequence* archi

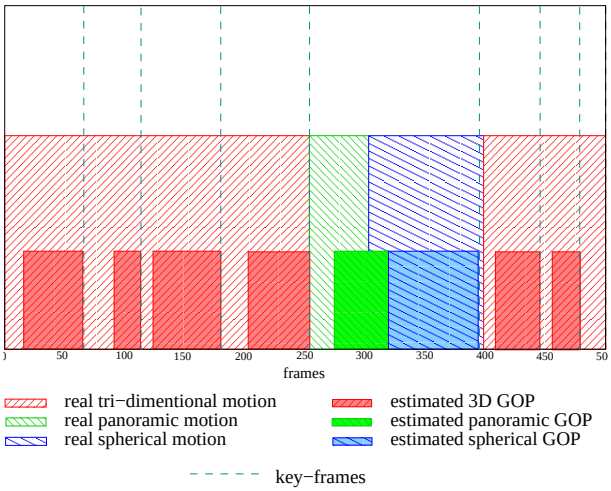


Figure 10: *Sequence* archi *: Comparison between the estimated GOPs and the real motion*

5. RESULTS

The presented approach has been implemented and tested over several real and synthetic videos to be validated. We will show the results obtained over two specific sequences. The first one, *archi*, is a synthetic view of a car parked next to a building, whose main advantage is to describe all the different camera motions to be tested (see figure 11). The second one, *bushes*, is a real exterior panoramic view of a stair within a leafy environment (see figure 13(a)).

The depth map gives a realistic information on the scene geometry. It enables to reconstruct with good quality the images in the GOP. In figure 12, we show the first key-frame of a 3D GOP, generated from the *archi* sequence, and its associated depth map, computed frem the first and last key-frames of the GOP. A rendering of the model from a virtual viewpoint is also presented. It shows globally satisfying, with small artifacts, such as elongated textures in some areas of the model.

In this sequence *archi*, the camera describes a general rotational motion between frames 250 and 400. Our key-frames selection algorithm detected a spherical GOP for frames 250 to 392 (see figure 10). We can also see in figure 9 the amount of the rotation error residual for each frame. Note that it is computed only for frames with sufficient apparent displacement, which explains the $null$ value of the residual at each GOP beginning. In figure 14 we show the corresponding part of the generated geospherical mesh, with its associated mosaic. The difference in figure 14(c) between the original and reconstructed frames (noticeable near the principle edges) is mainly due to an anti-aliasing effect introduced by OpenGL in the reconstructed frame.

In the *bushes* sequence, since the camera is attached to a tripod, it describes an almost pure vertical rotation. As expected, the key-frame selection algorithm detected only one panoramic GOP, illustrated by the corresponding mosaic in figure 13(b) and the generated 3D panoramic mesh (figure 13(c)).

6. CONCLUSION AND FUTURE WORK

We proposed here an algorithmic scheme for representing videos of arbitrary fixed contents by 3D or a 2D model depending on the camera motion. We thus extended an existing scheme to the case of general rotations and a robust key-frame selection algorithm was proposed. Moreover, tests made with real and synthetic videos validated the approach. The method could be validated in several ways. The mesh-based motion estimator should be adapted so that it computes homographic transformations between the images instead of affine ones, so as to produce correct spherical mosaics. This is related to the fact that the motion projected onto a sphere is not globally affine. We also plan to investigate an analysis scheme where there will not be a clear distinction between 3D, panoramic and spherical GOPs, but a smooth degradation of the 3D models into spherical models in

case of rotational motion.

7. REFERENCES

[1] M. Pollefeys, M. Vergauwen, K. Cornelis, J. Tops, F. Verbiest, and L. Van Gool, "Structure and motion from image sequences," in *Proc. Conference on Optical 3-D Measurement Techniques*, Vienna, October 2001, pp. 251–258.

[2] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521623049, 2000.

[3] P.-L. Bazin, J.-M. Vézien, and A. Gagalowicz, "Shape and motion estimation from geometric primitives and parametric modelling," in *Proceedings of the Machine Vision Applications workshop*, Tokyo, 2000.

[4] C. J. Taylor, P. E. Debevec, and J. Malik, "Reconstructing polyhedral models of architectural scenes from photographs," in *Proceedings of the 4th European Conference on Computer Vision-Volume II*. 1996, pp. 659–668, Springer-Verlag.

[5] F. Prêteux and M. Malciu, "Model-based head tracking and 3d pose estimation," in *Visual Conference on Image Processing*, San Jose, California, 1998, pp. 94–110.

[6] F. Galpin, R. Balter, L. Morin, and S. Pateux, "Efficient and scalable video compression by automatic 3d model building using computer vision," in *Picture Coding Symposium, PCS'2004, San Francisco, USA*, 2004.

[7] F. Galpin and L. Morin, "Sliding adjustment for 3d video representation," *EURASIP Journal on applied Signal Processing - Special issue on 3D Imaging and Virtual Reality, volume 2002, No. 10*, pp. 1088–1101, 2002.

[8] R. Szeliski, "Image mosaicing for tele-reality applications," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, 1994, pp. 44–53.

[9] H. Shum and R. Szeliski, "Panoramic image mosaics," Tech. Rep. MSR-TR-97-23, Microsoft Research, 1997.

[10] S. Coorg and S. Teller, "Spherical mosaics with quaternions and dense correlation," *International Journal of Computer Vision*, vol. 37, no. 3, pp. 259–273, 2000.

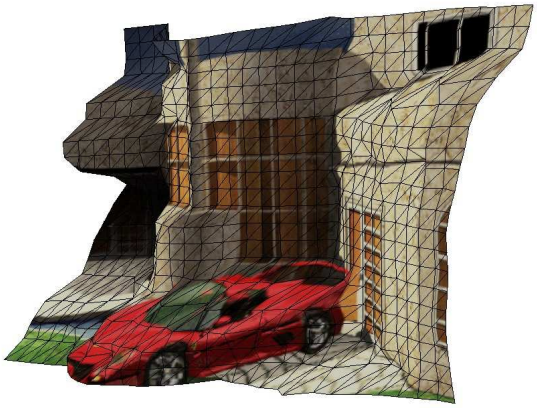
[11] K. Kanatani, "Geometric information criterion for model selection," *Int. J. Comput. Vision*, vol. 26, no. 3, pp. 171–189, 1998.

[12] K. Kanatani, "Model selection for geometric inference," in *Proceedings of the 5th Asian Conference on Computer Vision (ACCV 2002)*, Melbourne, Australia, January 2002, pp. xxi–xxxii.

[13] J.-F. Vigueras-Gomez, M.-O. Berger, and G. Simon, "Calibration multiplanaire d'une caméra : augmenter la stabilité en utilisant la sélection de modèles," in *Journées Francophones des Jeunes Chercheurs en Vision par Ordinateur - ORASIS'2003, Grardmer, France*, May 2003, pp. 147–156.

[14] P.H.S. Torr, "An assessment of information criteria for motion model selection," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*. 1997, p. 47, IEEE Computer Society.

(a) Key-frame (b) Corresponding depth map



(c) Virtual View

Figure 12: *An example of a computed 3D model*

[15] E. Morillon, R. Balter, L. Morin, and S. Pateux, "2d/3d hybrid modeling for video sequence," in *Wiamis 2004,International Workshop on Image Analysis for Multimedia Interavtive Services,april 2004*, 2004.

[16] G. Marquant, S. Pateux, and C. Labit, "Mesh-based scalable image coding with rate-distortion optimization," in *Image and Video Communications and Processing 2000*, San Jose, USA, 2000, vol. 3974, pp. 101–110.

[17] S. Pateux, "Estimation de mouvement par maillages actifs - application au codage vidéo. rapport technique projet cohrainte," Tech. Rep. MSR-TR-97-23, IRISA, 2001.

[18] F. Galpin, *Représentation 3D de séquences vidéo; Schéma d'extraction automatique d'un flux de modèles 3D, applications à la compression et à la réalité virtuelle*, Ph.D. thesis, Thèse de doctorat en Informatique, Université de Rennes 1, France, 2002.

[19] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, 1988, pp. 147–151.

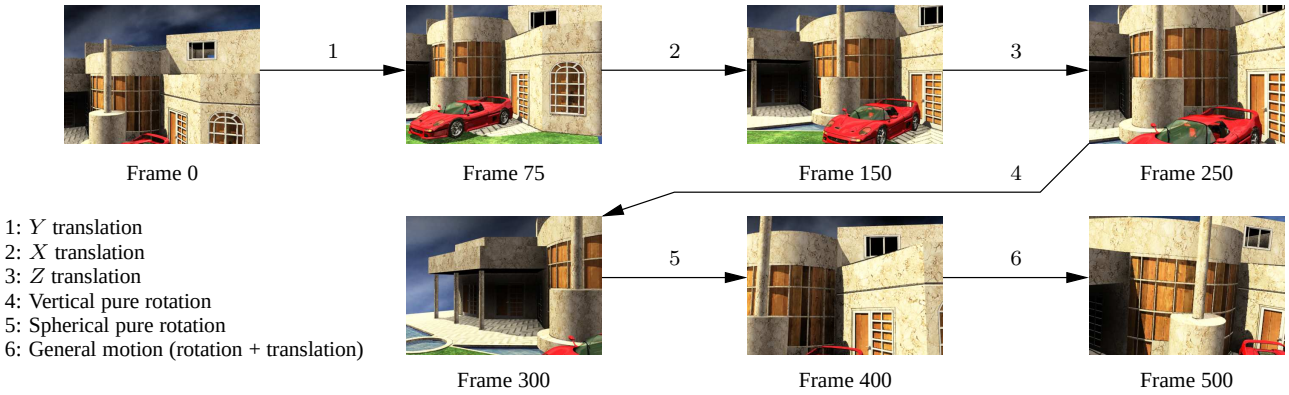
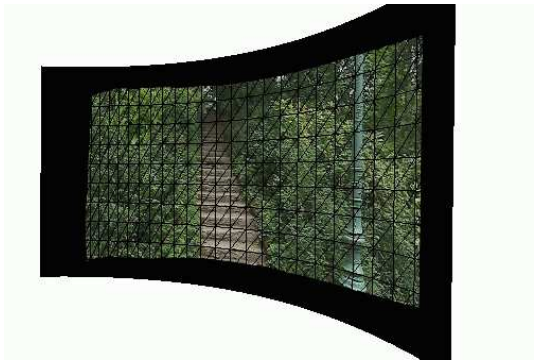


Figure 11: *The* archi *sequence motion decomposition*

(a) Some frames of the sequence

(b) The generated mosaic, covering the entire sequence



(c) Corresponding 3D panoramic mesh

Figure 13: *Sequence* bushes

(a) Mosaic from a spherical GOP

(b) Corresponding 3D geospherical mesh

(c) Comparison between original and reconstructed non key-frame images

Figure 14: *An example of a computed spherical GOP model*