Recovering Camera Motion and Mobile Objects in Video Documents

Damien Paulin, Dinesh Kumar, Raghav Bhaskar and Georges Quénot

CLIPS-IMAG, BP53, 38041 Grenoble Cedex 9, France
Georges.Quenot@imag.fr

Abstract. This paper presents a system that performs the recovery of camera motion parameters and the segmentation of mobile objects in video documents for content indexing. Two different methods are used for the recovery of the camera motion (relatively to the main background), the first for a camera maintained at a fixed location with rotational and zoom degrees of freedom, and the second for a camera of arbitrary motion but assuming a fixed focal length. The first method is based on the search of an optimal projective transform between consecutive images combined with an iterative background / mobile objects segmentation process. The second method is based on a paraperspective factorization method for shape and motion recovery. Both methods rely on the use of a dense and high-quality matching between consecutive images (optical flow). The system also attempts to classify shots or sub-segments of shots into one of the following categories of "no motion", "non mobile camera motion", "mobile camera motion" or "other type of motion". Further subcategorization can be done for each recovered type. Results are presented using sequences extracted from document 8 of the ISIS GDR-PRC GT10/AIM corpus.

1 Introduction

Recovering camera motion parameters and segmenting mobile objects are important tasks for video document content indexing (MPEG-7) [1]. They may also be useful for other applications like video compression (MPEG-4) or robotics. These tasks are generally performed on continuous video shots obtained after temporal video segmentation. They can be used for micro-segmentation (splitting a shot into sub-units with different global motions) prior to object identification. The information can also be used to build synthetic views like reconstructed panoramic views, extracted objects view or three-dimensional views of the background.

Both tasks are related since only objects that do not follow the background global motion can be identified and tracked and similarly the relative motion between the camera and the background can be recovered accurately only if the background and the mobile components have been identified. When the background occupies a significant part of the image and the camera motion is actually of a predicted type, it is often possible to obtain both information simultaneously using an iterative and cooperative background segmentation / background parameters motion estimation process.

In the case of video documents, and unlike in robotics for example, no information is generally available about the intrinsic camera parameters (focal length, optical center) except the pixel aspect ratio which is usually known from the video format. Therefore, systems must be able to recover them or to be robust enough to the fact that they are unknown or approximately known. Also, no clue is available about the possible camera motion and the type of observed scene.

Many methods are available for temporal video segmentation [2]. In this work, we use a system using a combination of several techniques (color histograms, rough contour tracking, motion compensated differences and dissolve detection [3]).

Simple methods have been developed for recovering rough camera motion from medium to low quality vector fields (MPEG prediction vectors for instance) [4]. Such methods may be useful for indexing purposes but the information they provide is poor (for instance it does not allow to distinguish between a camera translation and rotation) and are not very reliable (since MPEG vectors are not computed for this purpose). Intermediate methods, relying on the search for affine transforms between images, are able to provide both a good qualitative camera motion estimation and a panoramic view of the background [5].

More sophisticated camera motion recovery methods have been proposed by making some assumptions about the possible camera motion (limiting the number of degrees of freedom) and about the scene content ("solid" background occupying a large area in the images). For instance, the projective model based approach makes the assumption of a fixed camera location with degrees of freedom only in rotation

and zoom [6] [7] and "motion and structure from motion" approach makes the assumption of a camera with fixed focal length [8].

In this paper, we present a system that integrates specific variants of both type of methods and able to recover both types of camera motion. These methods are presented in the next two sections. They both rely on the use of an optical flow method [9] that provides a dense and high quality matching of images in a continuous sequence.

2 Camera with degrees of freedom in rotation and zoom

The assumption here is that the camera is fixed and has degrees of freedom only in rotation (tilt, pan and roll angles) and zoom. In this case, simple geometry considerations shows that the transformation giving corresponding points between any two images in the sequence has the form of an homography (or projective transform, [10]):

$$\begin{aligned} H_{(a,b,c,d,e,f,g,h)}: \; & R^2 \rightarrow R^2 \\ & (x,y) \mapsto \left(\frac{ax+by+c}{gx+hy+1}, \frac{dx+ey+f}{gx+hy+1}\right) \end{aligned} \tag{1}$$

where the (a,b,c,d,e,f,g,h) parameters depend upon the intrinsic and extrinsic camera parameters. The problem is therefore split into two parts: recovering the optimal homography between consecutive images, and recovering the intrinsic and extrinsic camera parameters from the sequence of found homographies.

2.1 Search for homographies and background regions

Correspondence between four non-aligned points is theoretically enough to recover the eight parameters. However, using as many of them as possible in a statistical combination improves the result. Traditionally, extracted and matched feature points constituting a sparse motion vector set are used for this purpose (like in [7]). However, we have estimated that the dense vector fields obtained with an optical flow technique [9] could further improve the accuracy because there are many more available vectors for the statistic and also because, with the chosen method, the matching is obtained with a sub-pixel accuracy which is expected to be much better than what can be obtained from the matching of extracted feature points.

The homography coefficients are searched for using a least square minimization. The optical flow provides a dense function G:

$$\begin{aligned} G: \; & R^2 \rightarrow R^2 \\ & (x,y) \mapsto (x+\Delta x(x,y), y+\Delta y(x,y)) \end{aligned} \tag{2}$$

and we search the parameter set (a,b,c,d,e,f,g,h) that minimizes the residue function E:

$$E(a,b,c,d,e,f,g,h) = \sum_x \sum_y \alpha(x,y) \left\| G(x,y) - H_{(a,b,c,d,e,f,g,h)}(x,y) \right\|^2 \tag{3}$$

where $\alpha(x,y)$ is a weighting term combining an estimation of the confidence associated to the computed optical flow vectors and an estimation of the probability for the current location to belong to the background. The E function can be rewritten as:

$$E(a,b,c,d,e,f,g,h) = \sum_x \sum_y \alpha(x,y) \left(\left(x' - \frac{ax+by+c}{gx+hy+1} \right)^2 + \left(y' - \frac{dx+ey+f}{gx+hy+1} \right)^2 \right) \tag{4}$$

with $x' = x + \Delta x(x,y)$ and $y' = y + \Delta y(x,y)$. Since within the image area we have: $gx+hy \ll 1$ in most cases, it makes little difference to minimize the following function F instead of E :

$$F(a,\ldots,h) = \sum_x \sum_y \alpha'(x,y) \left(\left((gx+hy+1)x' - (ax+by+c) \right)^2 + \left((gx+hy+1)y' - (dx+ey+f) \right)^2 \right) \tag{5}$$

The advantage of the E to F function substitution is that finding a parameter set (a,b,c,d,e,f,g,h) that minimizes function F is straightforward since that function is quadratic in its variables.

Now that we have an efficient way to obtain an optimal parameter set (a, b, c, d, e, f, g, h) from a vector field G and an confidence estimate α, we can use it in an iterative process to identify the background. In iteration p, we search for the optimal paramter set $(a_p, b_p, c_p, d_p, e_p, f_p, g_p, h_p)$ using a confidence estimate $alpha_p$ defined as follows:

$$p = 0 : \qquad \alpha_0(x, y) = \alpha_c(x, y) \tag{6}$$

$$p > 0 : \qquad \alpha_p(x, y) = \alpha_c(x, y).\alpha_{fp}(x, y)/(g_{p-1}x + h_{p-1}y + 1)^2 \tag{7}$$

where α_c is the confidence estimate associated with the computed vector field G and α_{fp} is a confidence estimate for the current point to belong to the background computed using the previous estimate of the parameter set that is, $(a_{p-1}, b_{p-1}, c_{p-1}, d_{p-1}, e_{p-1}, f_{p-1}, g_{p-1}, h_{p-1})$. The $1/(g_{p-1}x + h_{p-1}y + 1)^2$ factor is the correction for the approximation caused by the E to F function substitution. The α_c function is a combination of three criteria: the first two correspond to the selection of "edge" and "corners" points (identified respectively as having a high modulus of intensity gradient and a high modulus of the gradient of the angle of the intensity gradient), and the second one corresponds to the elimination of points with high velocity gradients. The α_{fp} function is itself the product of two functions, the first, α_{vp}, evaluates how well the predicted motion matches the extractedone:

$$p > 0 : \qquad \alpha_{vp}(x, y) = f_{\theta_v}(\left\| G(x, y) - H_(a_{p-1}, b_{p-1}, c_{p-1}, d_{p-1}, e_{p-1}, f_{p-1}, g_{p-1}, h_{p-1})(x, y) \right\|^2) \tag{8}$$

and the second, α_{ip}, evaluates how well the predicted image intensity matches the extracted one:

$$p > 0 : \qquad \alpha_{ip}(x, y) = f_{\theta_i}(\left\| I_1(x, y) - I_2(H_(a_{p-1}, b_{p-1}, c_{p-1}, d_{p-1}, e_{p-1}, f_{p-1}, g_{p-1}, h_{p-1})(x, y)) \right\|^2) \tag{9}$$

where f_θ is a sigmoid function controlled by a threshold parameter θ. It associates a non-binary membership value with the point to belong to the background and also ensures convergence of the iterative process.

After convergence, we obtain the optimal parameter set (a, b, c, d, e, f, g, h) for the background motion and a function α_f giving an estimation for a point to belong to the background. However, this estimate, though good for the iterative search is not very good for mobile object segmentation. A better way to segment mobile objects from the background is to use the homographic information on the whole sequence to build a panoramic reconstruction. Using combinations of homographies, it is possible to align any view with any other view. All the images of the sequence can be superimposed and a mosaic can be built using the more stable intensity (or color) value. Figure 1 shows an example of such mosaic reconstruction from homography sequence recovery using an image sequence from document 8 of the ISIS GDR-PRC GT10/AIM corpus. Objects can then be segmented (with some filtering) as regions where the intensity (or color) value differs significantly from the background (Figure 2).

2.2 Search for camera parameters

A standard pinhole camera model is used. It comprises four intrinsic parameters: the horizontal and vertical scale factors (a_u, a_v), and the location of the optical axis in the image plane (u_0, v_0), and six extrinsic parameters defining the camera location and orientation. The focal length is the only intrinsic parameter supposed to change during the sequence. u_0 and v_0 are supposed to be fixed (but are unknown). The two scale factors are related by the pixel aspect ratio $(a_v = r.a_u)$, known from the video format, and change with the focal length (zoom factor). The location of the camera is also supposed to remain unchanged and this location will be taken as the world origin. The orientation of the camera is also defined up to an arbitrary rotation and the orientation of the camera in the first image is taken as the reference.

For convenience, for a sequence of N consecutive images (numbered from 0 to $N-1$), the parameters to be recovered are chosen as:

- u_0 and v_0, location of the optical axis in the image plane,
- a_u, horizontal scale factor in the first image,
- z_n, for $1 \leq n < N$, the ratio of the scale factors between images n and $n-1$,
- α_{xn}, α_{yn} and α_{zn}, for $1 \leq n < N$, angles around x, y and z axes between camera orientations for images n and $n-1$.

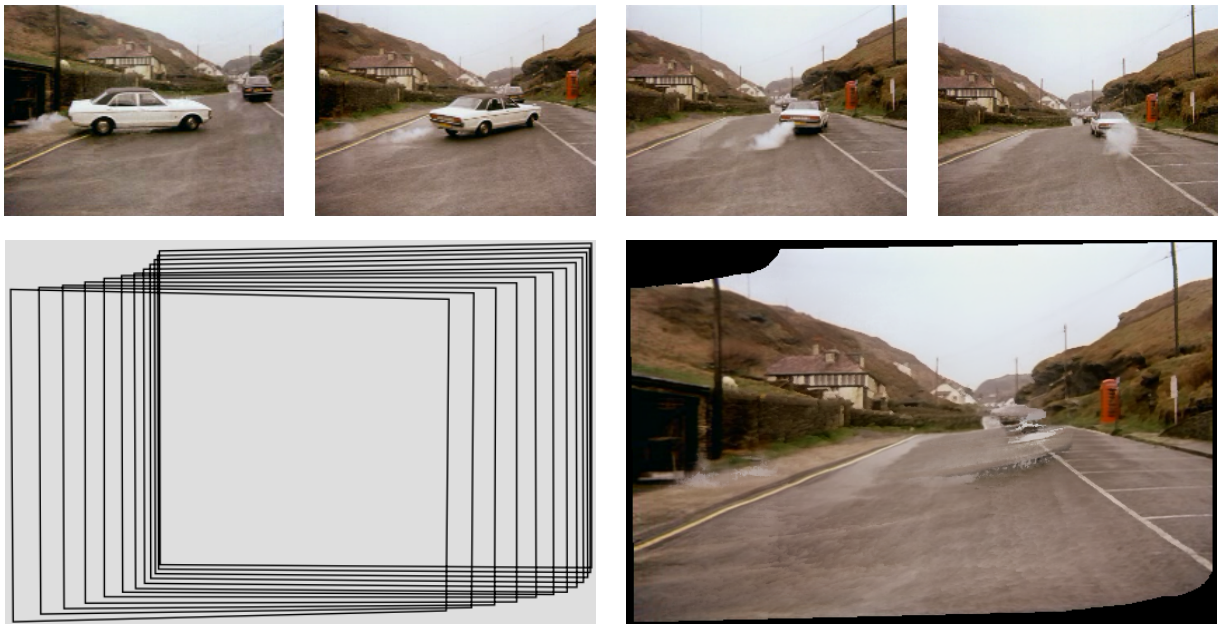

Fig. 1. Car sequence #1 (45-frame, top) Homographic alignment of images (bottom left) and reconstruction of a panoramic view (bottom right).

Fig. 2. Separation of background and mobile objects, car sequence #1. Original image (left), corresponding image from mosaic (middle), extracted mobile objects (right).

This combination has the advantage to split the parameters to be recovered in two categories. "Global" parameters: u_0, v_0 and a_u are related to the whole sequence of recovered homographies and "local" parameters: z_n, α_{xn}, α_{yn} and α_{zn} depend only upon the global parameters and the H_n recovered homography between image $n-1$ and image n. This split allows us to search for the whole set of parameters via an iterative process in which the local and global parameters are alternatively refined.

For any set of values $(z_n, \alpha_{xn}, \alpha_{yn}, \alpha_{zn})$ defining a change in the camera orientation and zoom factor, there is a corresponding homography $h(z_n, \alpha_{xn}, \alpha_{yn}, \alpha_{zn})$ that can be explicitly computed. This function also depends of the global parameters and the sequence of z_p values for $p < n$, obtained from previous iterations. A value for $(z_n, \alpha_{xn}, \alpha_{yn}, \alpha_{zn})$ is obtained as the one minimizing the square difference with the homography H_n recovered between images $n-1$ and n. This is a direct minimization of a four-variable function. It is done independently for each n between 1 and $N-1$. u_0, v_0, a_u and z_p for $p < n$ are considered to be constant during this search.

After the optimal search of the $(z_n, \alpha_{xn}, \alpha_{yn}, \alpha_{zn})$ values, a search is performed for optimal values for u_0, v_0 and a_u. Again, the homographies between consecutive images are built but this time as functions of u_0, v_0 and a_u: $h_n(u_0, v_0, a_u)$. A value for them is obtained as the one minimizing the sum, for the whole sequence, of square differences with the H_n recovered homographies. This is a direct minimization of a three-variable function. $(z_n, \alpha_{xn}, \alpha_{yn}, \alpha_{zn})$ are considered constant during this search except that angles around vertical and horizontal axes are corrected in order to compensate for the change in focal length.

The iterative process is started with initial values set as: center of the image for (u_0, v_0), length of the image diagonal for the a_u scale factor, 1 for all z_n values. Local and global refinements are iterated starting with a local one. After convergence, all the parameters are known. (u_0, v_0) is not a very useful

information but it had to be taken into account for accurate convergence. a_u gives information about the camera angular aperture and is also useful in setting an absolute scale for the recovered angle information.

Figure 3 shows the recovered camera motion parameters for the sequence whose homographic alignment is shown in Figure 1. Recovered motion is very consistent with human estimation. The camera has a main motion toward the right (pan) with a smaller motion upward (tilt). Very little roll motion and zoom change were found where, probably, none were in the original sequence. This gives an indication of the drift which is about 2 % for the whole 45-image sequence. The optical center was found near the image center. Figure 4 shows similar results for a 146-image sequence in which there is a significant motion toward the left (pan), a large zoom out and very little tilt and roll motion. Part of the sequence has no camera motion at all. The scale factor (absolute angle of view) was recovered for both sequences but its accuracy could not be evaluated since the actual value is unknown. However, in other experiments using several photographs taken from the same viewpoint, the known angle of view was found correct with an error less than 5 %.

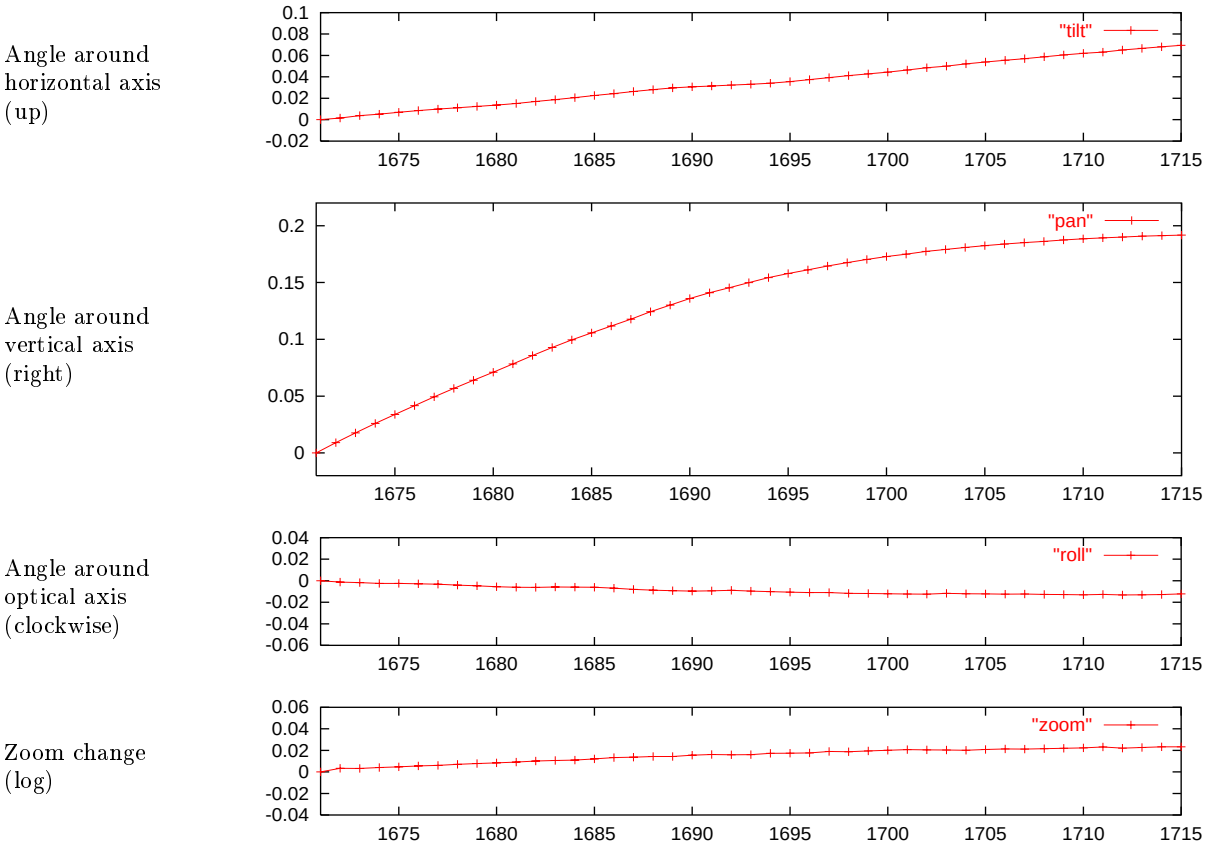

Fig. 3. Recovery of camera motion parameters, rotation and zoom only in car sequence #1

3 Camera with degrees of freedom in rotation and translation

For the recovery of the motion of a mobile camera, the paraperspective factorization method for shape and motion recovery proposed by Poelman and Kanade [8] was used. The paraperspective projection is an approximation to the true perspective projection corresponding to the pinhole camera model. The true perspective projection is a direct point projection of the 3D points of the object on the 2D image plane of the camera. The paraperspective projection is a combination of two projections (Figure 5). The first one is a projection on a plane passing through the center of gravity of the set of considered 3D points and parallel to the image plane (called hypothetical image plane) along a direction defined by the camera optical center and the same center of gravity. The second projection is the same point projection as the one used in the true perspective model.

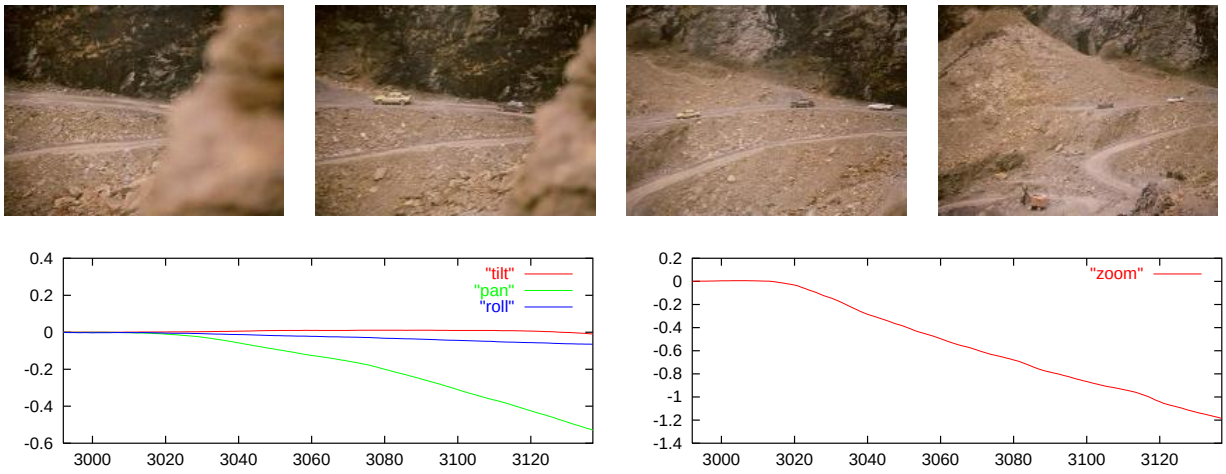

Fig. 4. Car sequence #2 (146-frame, top), recovery of camera motion parameters, non-mobile camera. Pitch, pan and roll angle (in radians, bottom left) logarithm of zoom (bottom right).

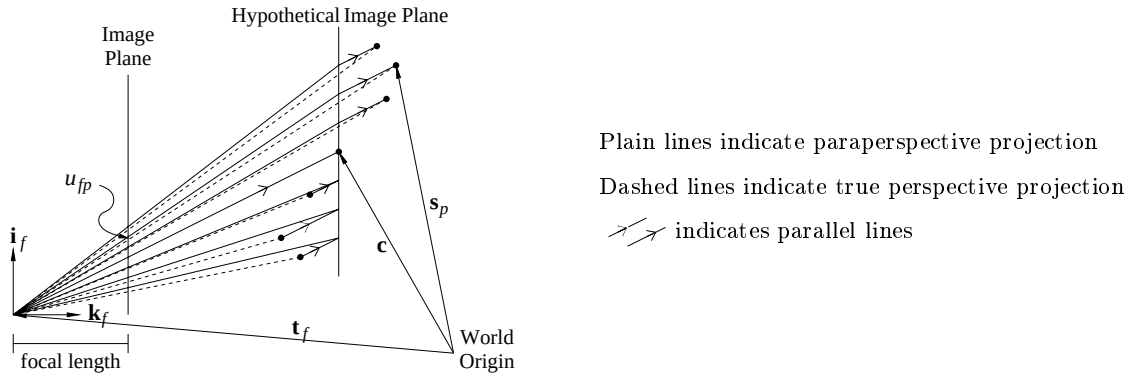

Fig. 5. Paraperspective projection in two dimensions (from [8])

The paraperspective model is somehow equivalent to taking a virtually infinite focal length and distance to the scene but conserving a given ratio between them. There is no difference in this model between zooming and moving towards the scene. Therefore, the recovered camera motion will be up to a scaling factor (linked to the unknown focal length) for the distance to the scene (relative to the scene scale). There is also another unknown scaling factor which is linked to the unknown scene absolute scale.

The approximations of the paraperspective projection are both very good ones (except for wide angle views) and sufficient to keep enough linearities in the system so that the paraperspective factorization method can be used. If we consider P feature points defined by their $\mathbf{s}_p$ 3D coordinates tracked during F frames in which the camera locations are defined by the $\mathbf{t}_f$ 3D coordinates of their optical center and the $\mathbf{i}_f$, $\mathbf{j}_f$ $\mathbf{k}_f$ vectors defining their axes, (u_{fp}, v_{fp}) being the coordinates of the paraperspective projection of point p in frame f (Figure 5), there exists a decomposition of the W (world) $2.F \times P$ (u_{fp}, v_{fp}) matrix:

$$W = M.S + T.[1 \dots 1] \tag{10}$$

in which M is a $2.F \times 3$ matrix, S is a $3 \times P$ matrix, and T is a $2.F \times 1$ vector [8]. The $\mathbf{s}_p$ coordinates can be deduced from the columns of the S (shape) matrix and the $\mathbf{t}_f$, $\mathbf{i}_f$, $\mathbf{j}_f$ $\mathbf{k}_f$ vectors can be deduced from the rows of the M (motion) and T (translation) matrix and vector. All results are defined up to an isometrical transform (world origin and orientation are arbitrary) and a scaling factor (the scene absolute scale is unknown). There is also a scaling factor for the camera distance to the scene due to the unknown focal length.

The originality of this work is again in the tracking and selection of feature points. Again, the tracking is obtained from the computation of dense vector fields using an optical flow technique. All frames in the sequence (or in a sub-segment of the sequence) are aligned with one reference frame for the whole part of the scene visible in all these frames. Feature points are chosen only in the reference frame and tracked

using the optical flow technique. They are directly predicted in other frames and not re-extracted. This has the advantage that, even if a feature point is not exactly located where it should be according to a formal definition, the motion is locally well defined (because of the continuity of the vector field and of a high local curvature in the image intensity) and the matching for the point is obtained with sub-pixel accuracy in the whole sequence.

The use of as many feature points as possible is desirable in order to benefit from the statistical effect on the accuracy of the result. However, it is not possible to use all points visible in all frames with an appropriate weighting like was done in the case of the non-mobile camera because that would amount to too much data for the linear decomposition program and there is no way to weigh the contribution of all the points in that case. Therefore a selection of feature points is necessary.

All points are sorted using a confidence criterion of both a high curvature on the image intensity (high intensity gradient and high gradient for the angle of the intensity gradient) and a low gradient in the recovered vector field. Then, points are selected using a criterion combining both confidence and dispersion. Points are chosen sequentially from those having the best confidence value and are apart by a given minimum distance from the previously selected points. The thresholds on a minimum confidence value and minimum distance between points can be tuned so that the number of selected points remain below a given value (typically a few thousands).

After the selection of the feature points is done, application of the paraperspective factorization is straightforward. It gives the camera motion up to two scaling factors: one for the scene absolute scale and the other for the unknown focal length. Figure 6 shows the result of the camera motion recovery on a sequence with mobile camera from the AIM corpus. The results are presented with reference origin and orientation corresponding to the first camera parameters. Though no ground truth motion is available for evaluation, the result is very consistent with human analysis. The main translation motion is toward the left and the main rotation motion is a rotation to the right (pan) compensating for it so that the scene remains centered. There is also a small translation downwards and toward the scene. A small tilt is also recovered and very little roll is observed.

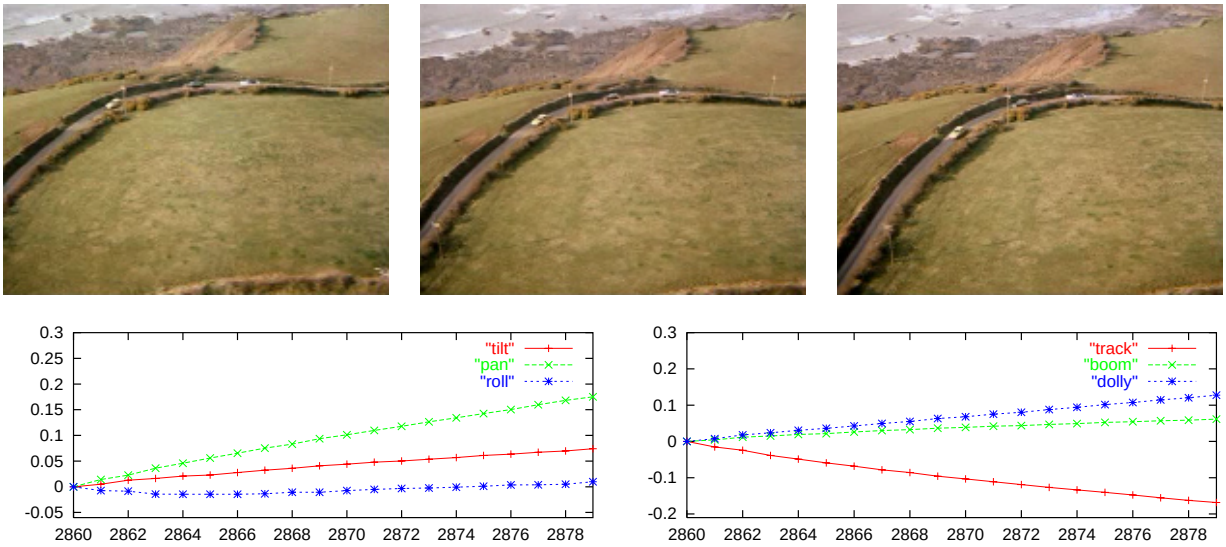

Fig. 6. Car sequence #3 (20-frame subsegment top), recovery of camera motion parameters, mobile camera, fixed focal. Pitch, pan and roll angle (in radians, bottom left), track, boom and dolly motion (in units relative to the mean distance to the scene, bottom right). car sequence #3

The paraperspective factorization also provides the recovered shape of the scene. Feature points locations are readily available and a whole 3D surface can be reconstructed using the recovered camera locations and the matching of all the points visible in all frames (using a least square method). Such reconstruction is not necessarily accurate everywhere (if either the matching is not reliable or there are some moving objects) but it may be a good representation of the three-dimensional aspect of the scene. Figure 7 shows a facet based 3D reconstruction of the scene shown in Figure 6.

Fig. 7. Three-dimensional view of the scene of car sequence #3. Textured (left, middle) and slices (right).

4 Integration

The two methods come with a criterion that assesses the validity of the assumptions they rely upon. This criterion is a combination of the residual reconstruction error with a weight associated with the type of the local texture in the regions identified as background. Using these criteria, the system is able to determine whether any of them is good and, if yes, which one. Classification in one of the "no motion", "non-mobile camera motion", "mobile camera motion" or "other type of motion" is performed. If the camera motion parameters have been successfully recovered, the quantitative information is used to further classify the type of motion (pure zoom, pure rotation or pure translation for example). The criteria can also be obtained on a frame by frame basis and continuous motion type can be assigned to sub-segments defining shots micro-segmentation.

5 Conclusion

Two different methods for the recovery of camera motion (relatively to the main background) and the segmentation of mobile objects have been integrated into a system dedicated to video document content indexing. The first one is dedicated to the case where the camera is non mobile and have degrees of freedom only in rotation and zoom. It is based on the search of an optimal projective transform between consecutive images combined with an iterative background / mobile objects segmentation process. The second method is based on a paraperspective factorization method for shape and motion recovery. Both methods rely on the use of a dense and high-quality matching between consecutive images (optical flow). The system also attempts to classify shots or sub-segments of shots in the appropriate category between "no motion", "non mobile camera motion", "mobile camera motion" or "other type of motion". Subcategories can also be searched for each recovered type using the quantitative data associated to the recovered motion. The first method gives a more complete camera information and can tolerate larger mobile objects. Panoramic reconstructions, three-dimensional representations and segmented mobile objects can be used for content indexing or for generating synthetic views of the document content.

References

1. Cverview of the MPEG-7 standard. In*ISO/IEC JTC1/SC29/WG11 N4031*, Singapore, March 2001.
2. Boreczky, J. S., Rowe, L. A.: Comparison of video shot boundary detection technique. In *IS&T/SPIE Conference on Electronic Imaging Technology and Science*, San Jose, USA, February 1996.
3. Quénot, G.M. and Mulhem P.: Two systems for temporal video segmentation. In *Content Based Multimedia Indexing*, Toulouse, Oct. 1999.
4. Kobla V., Doermann D.S., Lin K-I. and Faloutsos: Compressed domain video indexing techniques using DCT and motion vector information in MPEG Video. In *Proceedings of the SPIE conference on Storage and Retrieval for Image and Video Databases V*, Volume 3022, pp. 200-211, 1996.
5. Nicolas, H.: New Methods for Dynamic Mosaicing. In *IEEE Transactions on Image Processing*, 10(8), August 2001.
6. Irani M., Anandan P. and Hsu S.: Mosaic based representations of video sequences and their applications, In *Proceedings of the International Conference on Computer Vision*, pp 22-30, 1995.
7. Odone F., Andrea Fusiello A. and Trucco E.: Robust motion segmentation for content-based video coding. In *Recherche d'Information Assistée par Ordinateur*, Paris, France, 12-14 Apr. 2000.
8. Poelman C.J. and Kanade T. A Paraperspective Factorization Method for Shape and Motion Recovery, *tech. report CMU-CS-93-219*, Computer Science Department, Carnegie Mellon University, December, 1993.
9. Quénot, G.M.: Computation of Optical Flow Using Dynamic Programming. In *IAPR Workshop on Machine Vision Applications*, pp. 249-252, Tokyo, Japan, 12-14 Nov. 1996.
10. Semple J.G. and Kneebone G.T.: *Algebraic Projective Geometry*, Oxford University Press, 1952.