HAL
open science

Motion Panoramas

Adrien Bartoli, Navneet Dalal, Radu Horaud

To cite this version:

Adrien Bartoli, Navneet Dalal, Radu Horaud. Motion Panoramas. [Research Report] RR-4771, INRIA. 2003. inria-00071815

HAL Id: inria-00071815
https://inria.hal.science/inria-00071815v1

Submitted on 23 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Motion Panoramas

Adrien Bartoli, Navneet Dalal and Radu Horaud

First.Last@inria.fr

N° 4771

March 2003

THÈME 3

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4771--FR+ENG

Motion Panoramas

Adrien Bartoli, Navneet Dalal and Radu Horaud
First.Last@inria.fr

Thème 3 — Interaction homme-machine,
images, données, connaissances

Projet MOVI

Rapport de recherche n° 4771 — March 2003 — 23 pages

Abstract: In this paper we describe a method for analysing video sequences and for representing them as mosaics or panoramas. Previous work on video mosaicking essentially concentrated on static scenes. We generalise these approaches to the case of a rotating camera observing both static and moving objects where the static portions of the scene are not necessarily dominant, as it has been often hypothesised in the past. We start by describing a robust technique for accurately aligning a large number of video frames under unknown camera rotations and camera settings. The alignment technique combines a feature-based method (initialisation and refinement) with rough motion segmentation followed by a colour-based direct method (final adjustment). This precise frame-to-frame alignment allows the dynamic building of a background representation as well as an efficient segmentation of each image such that moving regions of arbitrary shape and size are aligned with the static background. Thus a *motion panorama* visualises both dynamic and static scene elements in a geometrically consistent way. Extensive experiments applied to archived videos of track-and-field events validate the approach.

Key-words: video mosaicking, panoramic visualisation, layered representation, motion segmentation, background substraction, texture alignment

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Panoramas de mouvement

Résumé : Nous décrivons une méthode d'analyse de séquences vidéos permettant une repésentation en mosaïques ou panoramas. Les travaux précedents sur les mosaïques d'images abordent essentiellement le cas des scènes statiques. Nous généralisons ces approches au cas d'une caméra subissant un mouvement de rotation pure et observant des objects statiques et dynamiques. Les parties statiques de la scène ne sont pas supposées dominantes. Nous commençons par décrire une technique robuste pour l'alignement d'un grand nombre d'images d'une vidéo, sans connaissance a priori de la rotation et des paramètres internes de la caméra. Cette technique d'alignement combine une approche basée sur des points d'intérêt (initialisation et rafinement) produisant une estimation grossière du mouvement, et une approche directe basée sur les intensités (ajustement final). Cette alignement précis image à image permet une construction dynamique d'une image du fond statique et une segmentation de chaque image permettant de délimiter les objects dynamiques. Un *panorama de mouvement* est constitué des éléments statiques et dynamiques de la scène, arrangés de manière consistante. De nombreuses expérimentations sur des documents vidéos archivés d'événements athlétiques valident notre approche.

Mots-clés : mosaïque d'images, visualisation panoramique, représentation en couches, segmentation de mouvement, soustraction de fond, alignement de texture

1 Introduction

Motion panoramas are visual representations of motion. Traditionally motion is visualised using an image sequence, or a video. Consider, for example the case of a person moving over several tens of meters. A video of such a moving person is gathered, in general, with a rotating and zooming camera such that: (i) the observed person remains in the camera's field of view and (2) preferably at constant image resolution. A compact and convenient representation of such a video is to stitch the individual images into a unique wide-angle panoramic image – a panorama, which is also called a mosaic.

The case of rigid scenes has been thoroughly investigated and a number of methods, algorithms, and software packages are available to produce *static panoramas*. The case that we want to study in this paper is more complex. Indeed, the combination of non-rigid scenes (scenes with multiple and/or articulated moving objects) with camera motion, as well as changes in camera parameters (focus and zoom) raises new difficulties.

The first and main difficulty is to segment each image into regions corresponding to distinct observed motions: multiple object motions and camera motion. The second difficulty is to estimate the camera internal parameters as well as the camera motion parameters such that the mapping from each individual image in the sequence to a single panoramic image can be performed. The third difficulty is to produce a high-quality motion panorama which is basically composed of two layers: a dynamic layer which corresponds to the moving objects and a static layer which corresponds to the static background.

The concept of motion panorama is best illustrated on Figure 1. From an original video (top) two layers are extracted. The static layer, or the background, is used to estimate the camera motion and the camera parameters associated with every image in the sequence. Next a background panorama is built (middle). Finally each individual image is compared to the background panorama in order to extract image regions corresponding to motion – the dynamic layer. Finally, the static and dynamic layers are combined together to form a motion panorama (bottom).

Panoramic photography has received growing interest since a decade [2, 3, 11, 12, 20, 22, 25, 15, 16] resulting in a number of commercial products such as [18]. The idea behind these methods is that there exist a simple invertible transformation between images gathered with a camera rotating around its center of projection [9]. A vast majority of papers (see [20] for a review), concentrates on the static case. While high-quality results are obtained, this assumption prunes many real-life image sequences.

Others have addressed the problem of analysing sequences of one or several moving objects with a static camera [13, 19, 24]. A current approach to detect motion with a static camera is to segment the image into two categories or two layers: a static layer and a dynamic layer, where a layer is a set of pixels. Practical approaches to layered segmentation is background substraction based on pixel-to-pixel comparison between a pre-stored background image and the current image. Of course, these methods work well when background is available.

Methods for analysing videos of moving objects with a moving camera are presented in [14], [12], and [6]. Both are interesting attempts to dynamically build a background image and to find moving object by substracting the background from each individual frame. In [12] the authors propose the use of a direct method to find the camera motion parameters, align frames based on these parameters, and spotting the image regions which do not satisfy these parameters. In [6] block-matching motion detection is first applied to find a motion vector field and this field is clustered to segment dominant motion regions. A direct method (see below) is applied to these regions in order to estimate camera motion. We found that these approaches work well when the moving objects correspond to relatively small image regions. When large portions of the images are occupied by moving objects, direct methods fail to find the camera motion. It is worthwhile to point out that a young company, Dartfish, commercializes software for producing motion panoramas from videos [5]. Their panorama building procedure requires manual intervention both for building the background and for selecting dynamic objects to be eventually overlayed onto the background.

The most crucial characteristics of methods associated with motion panorama construction are (i) the ability to deal with large dynamic image regions and (ii) the accuracy in frame alignment. Generally speaking, two categories of methods are available: Feature-based methods [22] and direct methods [11]. The former consists in extracting image features such as points of interest, matching such features over several images, and estimating the mapping between images based on feature-to-feature correspondences. The latter consists of finding the image-to-image mapping which best aligns the image intensity values (or red, green, and blue values for colour images).

Figure 1: This figure shows 5 images extracted from a 350-frame sequence (top), the static panorama or the background image (middle) showing the static objects used to estimate the time-varying camera parameters (focal length, pan and tilt angles), as well as the motion panorama showing a high-jump athlete in various postures as it would have been filmed with a wide-angle static camera (bottom). Notice the fine image resolution associated with the output images.

Feature-based techniques belong to an interesting class of methods which allow the estimation of an image mapping with as few as two feature-to-feature assignments (see below) and which may be combined with outlier rejection techniques. Therefore these methods can successfully be used to estimate the image mapping corresponding to camera motion while throwing out portions of the image which correspond to a different motion. Nevertheless, the process of extracting features from images introduces artifacts such as offsets in feature localisation. Moreover, features are observed only in the presence of important changes in image intensities. Detected features are not homogeneously distributed across the images which may cause alignment problems.

Direct methods consist of finding the mapping between images by minimising the discrepancy between their pixel values and/or colours, i.e., image correlation techniques. These methods produce the best results in terms of image alignment and hence in terms of the final quality of the mosaic, provided that a good initialisation is available. Robust correlation techniques were suggested in the past, i.e., correlation in the presence of artifacts. However, the idea of combining correlation under such image deformations as plane-projective transformations with robust techniques is not a realistic one.

Both the feature-based and the direct methods outlined above contribute to estimate the image-to-image transformations necessary for aligning the input images onto the panoramic image output. Another important ingredient is the segmentation of each image into two layers, a static one associated with camera motion and a dynamic one associated with moving scene objects. If camera motion has been estimated, one may detect the presence of moving objects by warping two images in the sequence and by detecting intensity or colour discrepancies at each pixel location. Such a method provides a fair initialisation of image regions corresponding to moving objects but fails to provide reliable results for producing motion panoramas. Indeed, consider complex human motion such as running: At any short-time interval some body parts move while some other body parts remain still and homogeneous regions such as skin or cloth are detected only along their contours. For these reasons, a more sophisticated motion detection technique is required.

This paper has the following original contributions:

- First we describe a three step method for building a static panorama in the presence of multiple moving objects. With respect to previous methods, we allow for large objects. (i) We suggest a parameterisation for zooming cameras with two rotational degrees of freedom, pan and tilt. Under the assumption that the focal length smoothly varies over a long image sequence but is almost constant between two consecutive frames we show that the planar homography allowing for frame-to-frame alignment needs only two points to be matched. Therefore the performance of any feature-based method using a robust estimator is improved both in terms of reliability and efficiency. (ii) Based on this initial camera motion estimation we warp the previous and next frames onto the current frame in order to detect moving regions. This three-frame motion detector optimistically detects these regions thus minimising the risk that outliers are included in the static layer. (iii) We describe an efficient direct method which aligns pixels in between two frames based on colour constancy and by minimising over four parameters, three rotational degrees of freedom and focal length. Unlike previous methods, we carefully design the error function such that the most time-consuming processes (such as computing the image gradients) are carried out outside the inner loop of the iterative minimisation procedure;

- Second we describe a background/foreground segmentation method. The static panorama accounts for a background image that is being built dynamically as the video proceeds. Each pixel in this image has statistics associated with it thus allowing simple and reliable comparison with each individual image. Since no assumption is made about the number of objects, the number of motions, etc., highly deformable and/or articulated objects such as humans can be easily detected, and

- Third we describe extensive tests made with 350-frame videos of track-and-field events (high jump and pole vault). We process VHS archived videos for which the camera parameters are unknown. High quality motion panoramas are produced in spite of the poor quality of the input data.

1.1 Paper organisation

The remainder of the paper is organised as follows. Section 2 introduces the un-calibrated camera motion parameterisation. Section 3 formulates the problems of finding image alignments with features and by direct comparison of pixel colours. Section 4 describes the feature-based method and section 5 describes the direct

image alignment method. Section 6 summarises the motion panorama algorithm and describes in detail the dynamic background substraction method allowing the final segmentation of each image into background and foreground. Experiments and their results are shown in section 7 and conclusions and directions for future work are discussed in section 8. Appendix A analyses the sensitivity of the alignment to camera calibration errors and appendix B derives a simple formula useful for the incremental estimation of a homography.

2 Mathematical preliminaries and notations.

We consider the pin-hole camera model which projects the 3-dimensional world onto a 2-dimensional image, represented at each time instant i by a 3×4, rank 3, homogeneous matrix $\mathbf{P}_i$. We express all 3-D entities in a standard camera coordinate frame with its orientation at the first time instant. It is assumed that the camera rotates around its optical center and therefore there is no translation associated with camera motion. The 3×3 matrix $\mathbf{K}_i$ contains the intrinsic parameters of the camera at time i and the 3×3 matrix $\mathbf{R}_i$ defines its orientation. One can write $\mathbf{P}_i \simeq \begin{bmatrix} \mathbf{K}_i\mathbf{R}_i & \mathbf{0} \end{bmatrix}$, where '$\simeq$' denotes equality up to a scale factor.

Let m_{ij} be an image point (the j-th point in the i-th image) and let the 2-vector $\boldsymbol{m}_{ij}$ designate its pixel coordinates while the 3-vector $\boldsymbol{q}_{ij}$ designates its homogeneous coordinates. This is the projection of the 3-D point M_j whose homogeneous coordinates are denoted by the 4-vector $\boldsymbol{Q}_{ij}$. We denote by $\Psi()$ the non-linear function mapping homogeneous image coordinates onto pixel coordinates, $\boldsymbol{m} = \Psi(\boldsymbol{q})$. That is, if the 3-vector $\boldsymbol{q}$ has coordinates q_1, q_2, and q_3, $\Psi(\boldsymbol{q}) = (q_1/q_3,\ q_2/q_3)^\top$. The homogeneous coordinates of the ray passing through this point are given by $\boldsymbol{r}_{ij} \simeq (\mathbf{K}_i\mathbf{R}_i)^{-1}\boldsymbol{q}_{ij}$. This may well be interpreted as a point at infinity R_j with homogeneous coordinates $\boldsymbol{R}_{ij}^\top = (\boldsymbol{r}_{ij}^\top\ 0)$. It is now possible to derive the inter-frame projective model, e.g., between frame i and frame k for a point j, see figure 2:

$$\begin{aligned} \boldsymbol{q}_{kj} &\simeq \mathbf{K}_k\mathbf{R}_k\boldsymbol{r}_{ij} \\ &\simeq \mathbf{K}_k\mathbf{R}_k\mathbf{R}_i^{-1}\mathbf{K}_i^{-1}\boldsymbol{q}_{ij} \\ &\simeq \mathbf{K}_k\mathbf{R}_{ki}\mathbf{K}_i^{-1}\boldsymbol{q}_{ij} \end{aligned}$$

Matrix:

$$\mathbf{H}_{ki} \simeq \mathbf{K}_k\mathbf{R}_{ki}\mathbf{K}_i^{-1} \tag{1}$$

defines a homography and the equation above fixes its parameterisation under the assumption that the camera undergoes a rotational motion around its fixed center of projection. One may use the Rodrigues equation to parameterise the rotation $\mathbf{R}_{ki}$ undergone by the the camera: $\mathbf{R}_{ki} = \mathbf{I} + \sin\phi_{ki}[\boldsymbol{\omega}_{ki}]_\times + (1-\cos\phi_{ki})[\boldsymbol{\omega}_{ki}]_\times^2$, where $[\boldsymbol{\omega}]_\times = (d\mathbf{R}/dt)\,\mathbf{R}^\top$, (the tangent operator) is a skew-symmetric matrix.

We consider the pinhole camera model. Traditionally there are 4 parameters associated with such a model: the focal length, the aspect ratio, and the pixel coordinates of the center of projection. The aspect ratio is fixed by the video standard being used and therefore is known. Throughout the paper we will assume that the center of projection lies at the image center. Appendix A analyses the error associated with this approximation. Only the focal length is unknown and it varies with time: Let f_i be the focal length at time i. Matrix $\mathbf{K}_i$ can now be written as a diagonal matrix:

$$\mathbf{K}_i = \begin{bmatrix} f_i & 0 & 0 \\ 0 & f_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

An useful assumption is that the rotation angle ϕ is small in between two consecutive frames, i and $i+1$. With the approximations $\sin\phi \approx \phi$ and $\cos\phi \approx 1$ we obtain $\mathbf{R} = \mathbf{I} + \phi[\boldsymbol{\omega}]_\times = \mathbf{I} + [\boldsymbol{\Phi}]_\times$ and finally a *small* homography writes, i.e., eq. (1):

$$\mathbf{H}_{i+1,i} = \begin{bmatrix} f_{i+1} & 0 & 0 \\ 0 & f_{i+1} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\phi_{i+1,i}^z & \phi_{i+1,i}^y \\ \phi_{i+1,i}^z & 1 & -\phi_{i+1,i}^x \\ -\phi_{i+1,i}^y & \phi_{i+1,i}^x & 1 \end{bmatrix} \begin{bmatrix} 1/f_i & 0 & 0 \\ 0 & 1/f_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

We consider a sequence of m frames. The parameters associated with each frame are $\mathcal{M}_i \equiv (\mathbf{R}_i, \mathbf{K}_i)$. The layer segmentation consists of a binary classification of pixels lying in the dynamic layer $\mathcal{F}_i$. The parameters of