

FEATURELESS GLOBAL ALIGNMENT OF MULTIPLE IMAGES

Bernardo Esteves Pires

The Boston Consulting Group
Lisbon Office, Portugal
E-mail: pires.bernardo@bcg.com

*Pedro M. Q. Aguiar**

Institute for Systems and Robotics / IST
Lisboa, Portugal
E-mail: aguiar@isr.ist.utl.pt

ABSTRACT

The majority of the approaches to the automatic recovery of a panoramic image from a set of partial views are suboptimal in the sense that the input images are aligned, or registered, pair by pair, *e.g.*, consecutive frames of a video clip. These approaches lead to propagation errors that may be very severe, particularly when dealing with videos that show the same region at disjoint time intervals. Although some authors have proposed a post-processing step to reduce the registration errors in these situations, there have not been attempts to compute the optimal solution, *i.e.*, the registrations leading to the *panorama that best matches the entire set of partial views*. This is our goal. In this paper, we use a generative model for the partial views of the panorama and develop an algorithm to compute in an efficient way the Maximum Likelihood estimate of all the unknowns involved: the parameters describing the alignment of all the images and the panorama itself.

1. INTRODUCTION

In this paper, we address the problem of recovering, in an automatic way, a panoramic image, or a mosaic, from a set of uncalibrated partial views, *e.g.*, a set of video frames. Modern digital video systems demand efficient solutions for this problem, *e.g.*, for image stabilization [1, 2] and content-based representations [3]. Other application fields include virtual reality and remote sensing. The key step to the success of the automatic mosaic building is the accurate registration, or alignment, of the input images.

1.1. Related work

Although some authors have approached the registration problem using classical signal processing techniques, such as Fourier transforms [4], or current image analysis tools, such as integral projections [5], the majority of the papers in the literature are mostly distinguished by either requiring a low-level pre-processing step (feature-based methods) or attempting to register the images directly from their intensity levels (featureless methods).

Feature-based methods, *e.g.*, [6], align the images by first detecting and matching a set of pointwise features. Since reliable feature points must correspond to sharp intensity corners [7, 8], this first step is hard to accomplish in a fully automatic way when processing real videos, particularly when the images are noisy, have low texture, or exhibit a small overlap among them.

In opposition, featureless methods are optimal, in the sense that they estimate the registration parameters by minimizing the

difference between the image intensities in a large region, thus leading to more robust solutions to the registration of a pair of views, *e.g.*, [9, 10]. However, when building a panorama from a large set of images, practitioners usually register them sequentially, one at a time. This leads to propagation errors that may become visually noticeable if non-consecutive images cover the same region of the panorama, which is common in applications such as seabed mapping. Although some authors proposed to post-process the registration parameters to deal with this problem [9, 11], there have not been attempts to generalize the highly successful featureless methods to the multi-frame case.

1.2. Proposed approach: featureless global estimation

The robustness of the featureless approaches to the registration of two views motivated us to develop a featureless method to align a larger set of frames. However, it is not obvious how the two-frame cost function, usually the sum of the image square differences [9, 10], should be generalized to the multi-frame case. We were able to derive the appropriate cost function, which is an original contribution of this paper, by including as unknown, jointly with the registration parameters, the panoramic image itself.

Our approach in this paper is then to formulate the automatic recovery of mosaics from a set of partial views, as a classical parameter estimation problem. The input images are modelled as noisy observations of limited regions of the unknown panorama. Naturally, since the images are uncalibrated, the problem includes as unknowns the parameters describing the registration, or alignment, of the entire set of input images. We then use *Maximum Likelihood* (ML) estimation. To minimize the ML cost with respect to the large set of unknowns, we propose an efficient method. First, we derive the closed-form solution for the estimate of the panorama in terms of the other unknowns (the registration parameters). Then, we plug-in the estimate of the panorama into the ML cost, obtaining an error function that depends on the registration parameters alone. This error function is a weighted sum of the square differences between all possible pairs of input images. We derive a gradient-descent algorithm to minimize this cost.

Like in the current featureless approaches to the registration of two images [9, 10], the derivatives involved in the gradient-descent algorithm to minimize our ML cost, are computed in a simple way in terms of the image gradients.

Paper organization In section 2, we formulate the registration of multiple images as a classical estimation problem. Section 3 deals with ML estimation for this problem. In section 4, we derive the gradient-descent algorithm to minimize the ML cost. Section 5 describes experiments and section 6 concludes the paper.

*The work of P. Aguiar was partially supported by the (Portuguese) Foundation for Science and Technology, grant POSI/SRI/41561/2001.

2. PROBLEM FORMULATION

In this section, we develop a generative model for the partial views of an unknown panorama, and use ML to derive the estimation criterion that will allow us to recover the observed panorama, as well as the registration parameters, *i.e.*, the viewing positions.

2.1. Generative model

We model each pixel of each image \mathbf{I}_i , as a noisy sample of the panorama \mathbf{P} . For simplicity, we consider the image domain to be the entire plane \mathbb{R}^2 and, to take care of the limited field of view, we define a window \mathbf{H} as $\mathbf{H}(x, y) = 1$ in the region observed in the images and $\mathbf{H}(x, y) = 0$ in the regions outside the camera field of view. The observation model is then

$$\mathbf{I}_i(\mathbf{x}_i) = [\mathbf{P}(\mathbf{x}_0) + \mathbf{R}(\mathbf{x}_i)]\mathbf{H}(\mathbf{x}_i), \quad (1)$$

where \mathbf{R} denotes the noise, assumed i.i.d. zero-mean Gaussian, \mathbf{x}_i are the image coordinates (x, y) , expressed in the coordinate system of the generic image \mathbf{I}_i , and \mathbf{x}_0 are the corresponding coordinates of the panoramic image \mathbf{P} , expressed in its own coordinate system (which we will refer to as the reference coordinate system). Image models related to (1) have been used in the context of segmenting and tracking moving objects in video sequences [12, 13].

The reference coordinate system and the coordinate system of any of the images are related by a generic parametric mapping

$$\mathbf{x}_i = \mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0). \quad (2)$$

The parameter vector $\boldsymbol{\theta}_i$ in (2) determines thus the mapping between each pixel of the panorama, with coordinates \mathbf{x}_0 , expressed in the reference coordinate system, with the corresponding pixel of image \mathbf{I}_i , with coordinates \mathbf{x}_i . Common parameterizations include translation (2 degrees of freedom (dof)), rotation (1 dof), rigid motion (3 dof), translation+rotation+zoom (4 dof), affine (6 dof), and the projective, or homography (8 dof), see, *e.g.*, [9, 14]. Although our derivations are intentionally left fully generic, in the experiments, we have used the affine mapping.

2.2. Estimation criterion

Given a set of n images, $\{\mathbf{I}_1, \dots, \mathbf{I}_n\}$, our goal is to recover all the unknowns involved: the panorama \mathbf{P} and the set of parameter vectors $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$ that define the viewing positions. We use ML. From the observation model (1), after simple manipulations, we express the symmetric of the log-likelihood function as

$$L(\mathbf{P}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) = \frac{nN}{2} \ln(2\pi\sigma^2) + (2\sigma^2)^{-1} \cdot \sum_{\mathbf{x}_0 \in \mathbb{R}^2} \sum_{i=1}^n [\mathbf{I}_i(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0)) - \mathbf{P}(\mathbf{x}_0)]^2 \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0)), \quad (3)$$

where N is the number of pixels in each image and σ^2 is the variance of the observation noise.

3. MAXIMUM LIKELIHOOD ESTIMATE

To compute the ML estimate of all the unknowns, *i.e.*, to carry out the minimization of the ML cost, given by the symmetric log-likelihood (3), with respect to (wrt) $\{\mathbf{P}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$, we start by noticing that the estimate of the panorama \mathbf{P} can be expressed in closed-form as a function of the remaining unknowns.

3.1. Estimate of the panorama \mathbf{P}

We derive the expression for the ML estimate $\hat{\mathbf{P}}$ of the panorama by minimizing (3) wrt a generic pixel value $\mathbf{P}(\mathbf{x}_0)$. By making zero the derivative of (3) wrt $\mathbf{P}(\mathbf{x}_0)$, the estimate $\hat{\mathbf{P}}$ at pixel \mathbf{x}_0 is easily obtained as a function of the set of unknown registration parameters, which we will compactly denote by $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$:

$$\hat{\mathbf{P}}(\mathbf{x}_0, \boldsymbol{\Theta}) = \frac{\sum_{i=1}^n \mathbf{I}_i(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0)) \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0))}{\sum_{i=1}^n \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0))}. \quad (4)$$

This expression shows that the estimate of the intensity of each pixel \mathbf{x}_0 of $\hat{\mathbf{P}}$ is given by the average of the intensities of the corresponding pixels of all the input images that captured \mathbf{x}_0 , *i.e.*, all the images \mathbf{I}_i for which $\mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0)) = 1$.

3.2. Estimate of the registration parameters $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$

Replacing the ML estimate $\hat{\mathbf{P}}$ of the panorama, given by (4), in the symmetric log-likelihood (3), we express this ML cost L as a function of the unknown registration parameters $\boldsymbol{\Theta}$ alone. After algebraic manipulations, we get:

$$L(\boldsymbol{\Theta}) = \frac{nN}{2} \ln(2\pi\sigma^2) + (4\sigma^2)^{-1} \sum_{\mathbf{x}_0 \in \mathbb{R}^2} \mathbf{W}^{-1}(\mathbf{x}_0, \boldsymbol{\Theta}) \cdot \sum_{i,j=1}^n \mathbf{E}_{ij}^2(\mathbf{x}_0, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0)) \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_j; \mathbf{x}_0)), \quad (5)$$

where \mathbf{E}_{ij} is the error between the co-registered images \mathbf{I}_i and \mathbf{I}_j ,

$$\mathbf{E}_{ij}(\mathbf{x}_0, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \mathbf{I}_i(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x}_0)) - \mathbf{I}_j(\mathbf{m}(\boldsymbol{\theta}_j; \mathbf{x}_0)), \quad (6)$$

and $\mathbf{W}(\mathbf{x}_0, \boldsymbol{\Theta})$ is a weight that counts the number of images that have captured the pixel \mathbf{x}_0 of the panorama, according to the registration parameters in $\boldsymbol{\Theta}$, *i.e.*,

$$\mathbf{W}(\mathbf{x}_0, \boldsymbol{\Theta}) = \sum_{k=1}^n \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_k; \mathbf{x}_0)). \quad (7)$$

By discarding from (5) the constant terms, *i.e.*, the terms that do not depend on the unknown registration parameters $\boldsymbol{\Theta}$, we conclude that the ML estimate for the problem of global multi-frame registration, is equivalent to the following minimization:

$$\hat{\boldsymbol{\Theta}} = \arg \min_{\boldsymbol{\Theta}} \sum_{i,j=1}^n \sum_{\mathbf{x}_0 \in \mathcal{R}_{ij}} \frac{\mathbf{E}_{ij}^2(\mathbf{x}_0, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j)}{\mathbf{W}(\mathbf{x}_0, \boldsymbol{\Theta})}. \quad (8)$$

For simplicity, when deriving (8) from (5), the sums were interchanged and the spatial region of summation was re-defined to take care of the windows $\mathbf{H}(\cdot)$ in (5), *i.e.*, \mathcal{R}_{ij} in (8) is the region where the images \mathbf{I}_i and \mathbf{I}_j overlap,

$$\mathcal{R}_{ij} = \{\mathbf{x} : \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_i; \mathbf{x})) \mathbf{H}(\mathbf{m}(\boldsymbol{\theta}_j; \mathbf{x})) = 1\}. \quad (9)$$

Expressions (7) and (8) condense one the contributions of this paper—they show that the ML estimate $\hat{\boldsymbol{\Theta}}$ of the registration parameters $\boldsymbol{\Theta}$ is given by the minimum of a particular weighted sum of the square differences between all possible pairs of co-registered input images.

4. MINIMIZATION ALGORITHM

Our algorithm to the minimization of the ML cost (8) uses an iterative scheme inspired in the common approaches to the two-frame problem [9, 10]. In each step, the algorithm updates a current estimate that we denote by $\Theta^0 = \{\theta_1^0, \dots, \theta_n^0\}$.

4.1. Iterative minimization of the ML cost

Instead of updating the entire set of parameters Θ in a single step, which would be computationally complex, we propose a coordinatewise minimization: we update each vector θ_q at a time, keeping fixed the remaining registration parameters $\{\theta_i = \theta_i^0, i \neq q\}$. The update is $\theta_q = \theta_q^0 + \hat{\delta}$, where $\hat{\delta}$ is obtained from (8), after discarding the terms that do not depend on θ_q :

$$\hat{\delta} = \arg \min_{\delta} \sum_{i=1}^n \sum_{\mathbf{x}_0 \in \mathcal{R}_{iq}} \frac{\mathbf{E}_{iq}^2(\mathbf{x}_0, \theta_i^0, \theta_q^0 + \delta)}{\mathbf{W}(\mathbf{x}_0, \Theta^0)} \quad (10)$$

To obtain a closed-form solution for the update $\hat{\delta}$, we approximate the error \mathbf{E}_{iq} by its first-order Taylor series expansion,

$$\mathbf{E}_{iq}(\mathbf{x}_0, \theta_i^0, \theta_q^0 + \delta) \approx \mathbf{E}_{iq}(\mathbf{x}_0, \theta_i^0, \theta_q^0) + \delta^T \cdot \nabla_{\theta_q} \mathbf{E}_{iq}(\mathbf{x}_0, \theta_i^0, \theta_q^0).$$

From the definition of \mathbf{E}_{iq} in (6), the gradient in the Taylor series expansion is easily computed in terms of the spatial gradient of image \mathbf{I}_q . Furthermore, that gradient does not depend on θ_i^0 , thus we will denote it more compactly by $\nabla(\mathbf{x}_0, \theta_q^0)$,

$$\nabla(\mathbf{x}_0, \theta_q^0) = \nabla_{\theta_q} \mathbf{E}_{iq}(\mathbf{x}_0, \theta_i^0, \theta_q^0) \quad (12)$$

$$= -\nabla_{\theta_q} \mathbf{m}(\theta_q^0; \mathbf{x}_0) \cdot \nabla_{\mathbf{x}} \mathbf{I}_q(\mathbf{m}(\theta_q^0; \mathbf{x}_0)). \quad (13)$$

By inserting the Taylor series approximation in (10) and making zero the derivative wrt δ , we get the update $\hat{\delta}$ as the solution of a linear system

$$\Gamma(\Theta^0) \cdot \hat{\delta} + \gamma(\Theta^0) = \mathbf{0}. \quad (14)$$

The matrix $\Gamma(\Theta^0)$ and the vector $\gamma(\Theta^0)$ are obtained as

$$\Gamma(\Theta^0) = \sum_{\mathbf{x}_0 \in \mathcal{R}_q} \nabla(\mathbf{x}_0, \theta_q^0) \cdot \nabla^T(\mathbf{x}_0, \theta_q^0), \quad (15)$$

$$\gamma(\Theta^0) = \sum_{\mathbf{x}_0 \in \mathcal{R}_q} \nabla(\mathbf{x}_0, \theta_q^0) \left[\hat{\mathbf{P}}(\mathbf{x}_0, \Theta^0) - \mathbf{I}_q(\mathbf{m}(\theta_q^0; \mathbf{x}_0)) \right], \quad (16)$$

where we used expression (4) for $\hat{\mathbf{P}}$. The sums in (15,16) are over the region observed by image \mathbf{I}_q , $\mathcal{R}_q = \{\mathbf{x} : \mathbf{H}(\mathbf{m}(\theta_q^0; \mathbf{x})) = 1\}$.

4.2. Interpretation in terms of current algorithms

Since the iterations in standard featureless two-frame alignment algorithms [9, 10] also lead to a system like (14), we now interpret our solution (14,15,16) in terms of those approaches. Define \mathbf{E}_{0q} as the difference between image \mathbf{I}_q and the previous estimate of the panorama, obtained with the registration parameters Θ^0 ,

$$\mathbf{E}_{0q}(\mathbf{x}_0, \Theta^0) = \hat{\mathbf{P}}(\mathbf{x}_0, \Theta^0) - \mathbf{I}_q(\mathbf{m}(\theta_q^0; \mathbf{x}_0)). \quad (17)$$

Since the gradient of this error wrt θ_q is equal to the one defined in (12), we can re-write expressions (15,16) in terms of \mathbf{E}_{0q} ,

$$\Gamma(\Theta^0) = \sum_{\mathbf{x}_0 \in \mathcal{R}_q} \nabla_{\theta_q} \mathbf{E}_{0q}(\mathbf{x}_0, \Theta^0) \cdot \nabla_{\theta_q}^T \mathbf{E}_{0q}(\mathbf{x}_0, \Theta^0), \quad (18)$$

$$\gamma(\Theta^0) = \sum_{\mathbf{x}_0 \in \mathcal{R}_q} \nabla_{\theta_q} \mathbf{E}_{0q}(\mathbf{x}_0, \Theta^0) \mathbf{E}_{0q}(\mathbf{x}_0, \Theta^0). \quad (19)$$

Expressions (18,19) are equal to the ones that arise from aligning the previous estimate $\hat{\mathbf{P}}$ of the panorama with image \mathbf{I}_q , by using standard featureless methods, see *e.g.*, [9, 10] or [8]. We thus conclude that our global approach lead to an algorithm that refines the estimate of the registration parameters of each image by using the methodology developed to register a single pair of images.

4.3. Convergence—initialization and multiresolution

Our algorithm starts by aligning the images sequentially, using the standard two-frame approach [9, 10]. Then, we compute an initial estimate of the panorama by using (4). After this, we cyclically refine the registrations parameters of each image. The stopping criterion may either be the error below a small threshold or reaching a maximum number of iterations.

Since the truncated Taylor series is a good approximation only when the vector θ_q is close to its initial value θ_q^0 , estimating the update δ from (14,15,16) leads to the convergence to the globally optimal ML estimate, only when the initial estimate is close enough to it. However, in practice, *e.g.*, in the first experiment described below, it is common that the initial estimate of the panorama is very rough, due to the propagation of (two-frame based) registration errors. To cope with these situations, we use a coarse-to-fine approach similar to the one proposed in [15, 9]: the parameters are first estimated in the coarsest resolution level, then used as an initialization to the next finer level, until the full image resolution is attained. As illustrated in the following section, this multi-resolution approach succeeds in correcting large miss-registrations.

5. EXPERIMENTS

We describe two experiments. The first experiment compares our global approach with the current sequential registration methods. In the second experiment, we illustrate with automatic mosaic building in a seabed mapping context.

5.1. Sequential alignment versus proposed method

To have an exact knowledge of the ground truth, we “synthesized” the input images by cropping a real photo and adding noise. In Fig. 1, we represent the evolution of the standard two-frame featureless sequential alignment (*e.g.*, [9, 10]) of those images. Note that the fourth image is miss-aligned and how that error propagates to the alignment of the remaining images. The (highly incorrect) panorama this way obtained, see the bottom right image of Fig. 1, was then used as the initialization for the global method we propose in this paper. After few iterations, our algorithm converged to the panoramic image shown in Fig. 2, which is visually indistinguishable from the ground truth image.

5.2. Underwater mosaic for seabed mapping

As a final example, we illustrate our method with automatic mosaic construction from video images, captured by an underwater

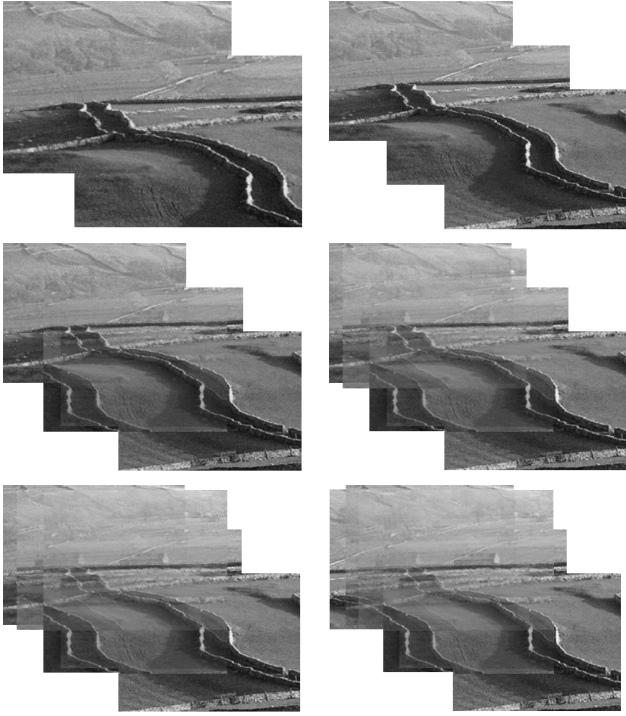


Fig. 1. Sequential registration. Note how the miss-alignment of the fourth image (middle left) propagates to the remaining ones.

camera in the sea. Although underwater images are particularly difficult to align, due to the absence of salient features, the mosaic recovered by our algorithm is visually correct, see Fig. 3.

6. CONCLUSION

We proposed a new method to build a panoramic image from a set of partial views. Rather than composing the input images in an incremental way, our approach seeks the global solution to the estimation problem, *i.e.*, it computes the panorama that best matches all the partial observations. To minimize the global cost, we derived an efficient gradient descent algorithm that generalizes the current most robust two-frame featureless registration approaches.

7. REFERENCES

- [1] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE T-IP*, 2000.
- [2] N. Petrovic, N. Jojic, and T. Huang, "Hierarchical video clustering," in *IEEE MMSP*, Siena, Italy, 2004.
- [3] P. Aguiar, R. Jasinschi, J. Moura, and C. Pluempitiwiriawej, "Content-based image sequence representation," in *Digital Video Processing*, Todd Reed, Ed. CRC Press, 2004.
- [4] B. Reddy and B. Chatterly, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. on Image Processing*, 1996.
- [5] J. Lee and J. Ra, "Block motion estimation based on selective integral projections," in *IEEE ICIP*, Rochester, USA, 2002.



Fig. 2. Proposed approach. Final estimate of the panorama, when our algorithm is initialized with the bottom right image of Fig. 1.

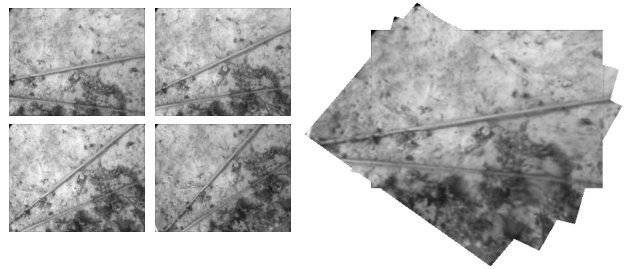


Fig. 3. Seabed mapping: video frames and underwater mosaic.

- [6] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.
- [7] J. Shi and C. Tomasi, "Good features to track," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1994.
- [8] P. Aguiar and J. Moura, "Image motion estimation – convergence and error analysis," in *IEEE ICIP*, Greece, 2001.
- [9] S. Mann and R. Piccard, "Video orbits of the projective group: a simple approach to featureless estimation of parameters," *IEEE Trans. on Image Processing*, 1997.
- [10] B. Pires and P. Aguiar, "Registration of images with small overlap," in *IEEE MMSP*, Siena, Italy, 2004.
- [11] D. Hasler, L. Sbaiz, S. Ayer, and M. Vetterli, "From local to global pparameter estimation in panoramic photographic reconstruction," in *IEEE ICIP*, Kobe, Japan, 1999.
- [12] P. Aguiar and J. Moura, "Detecting and solving template ambiguities in motion segmentation," in *IEEE ICIP*, 1997.
- [13] N. Jojic and B. Frey, "Learning flexible sprites in video layers," in *IEEE Int. Conf. on CVPR*, Hawaii, 2001.
- [14] D. Kim and K. Hong, "Fast global registration for image mosaicing," in *IEEE ICIP*, Barcelona, Spain, 2003.
- [15] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *European Conf. on Computer Vision*, Santa Margherita Ligure, Italy, 1992.