Electronic Letters on Computer Vision and Image Analysis 2(1):1-10, 2003

An Appearance-Based Method for Parametric Video Registration

Xavier Orriols, Lluis Barceló and Xavier Binefa

Computer Vision Center, Universitat Autònoma de Barcelona
08193 Bellaterra, Spain

Received 19 December 2001; accepted 29 April 2003

Abstract

In this paper, we address the problem of multi-frame video registration using an appearance-based framework, where linear subspace constraints are applied in terms of the appearance subspace constancy assumption [3]. We frame the multiple-image registration in a two step iterative algorithm. First, a feature space is built through and Singular Value Decomposition (SVD) of a second moment matrix provided by the images in the sequence to be analyzed, where the variabilities of each frame respect to a previously selected frame of reference are encoded. Secondly, a parametric model is introduced in order to estimate the transformation that has been produced across the sequence. This model is described in terms of a polynomial representation of the velocity field evolution, which corresponds to a parametric multi-frame optical flow estimation. The objective function to be minimized considers both issues at the same time, i.e., the appearance representation and the time evolution across the sequence. This function is the connection between the global coordinates in the subspace representation and the parametric optical flow estimates. Both minimization steps are reduced to two linear least squares sub-problems, whose solutions turn out to be in closed form for each iteration. The appearance constraints result to take into account all the images in a sequence in order to estimate the transformation parameters. Finally, results show the extraction of $3D$ affine structure from multiple views depending on the analysis of the surface polynomial's degree.

Key Words: Computer Vision, Image Analysis, Pattern Recognition, 3D Reconstruction, Video and Image Sequence Analysis.

1 Introduction

The addition of temporal information in visual processing is a strong cue for understanding structure and $3D$ motion. Two main sub-problems appear when it comes to deal with motion analysis; *correspondence* and *reconstruction*. First issue (correspondence) concerns the location analysis of which elements of a frame correspond to which elements in the following images of a sequence. From elements correspondence, reconstruction corresponds to $3D$ motion and structure recovery of the observed world. In this paper, we focus on the first issue, and, more specifically, the problem is centered on the observed motion in static scenes onto the image plane which is produced by camera motion: *ego-motion*. In previous work, dense [7, 5] and sparse [9, 6, 4]

Correspondence to: xevi@cvc.uab.es

Recommended for acceptance by Petra Perner.
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

methods to estimate the motion field have been used to this end. Sparse methods strongly rely on the accuracy of the feature detector and not all the information available in the image is employed. Dense methods are based on optical flow estimation which often produces inaccurate estimates of the motion field. Moreover the analysis is instantaneous, which means that is not integrated over many frames. Many authors [15, 1, 12, 2, 8] focus on this registration problem in terms of $2D$ parametric alignment, where the estimation process is still between two frames. Thus, taking into account that the second step, *reconstruction*, requires that all the transformations must be put in correspondence with a certain frame of reference, the accumulation error can be present in these computations.

Authors in [3] introduce the notion of *subspace constancy assumption*, where visual prior information is exploited in order to build a views+affine transformation model for object recognition. Their starting point is that the training set has to be carefully selected with the aim of capturing just appearance variabilities; that is, the training set is assumed to be absent of camera (or motion) transformations. Once the learning step is performed, the test process is based on the computation of the affine parameters and the subspace coefficients that map the region in the focus of attention onto the closest learned image. However, in this paper, the topic that we deal with has as input data the images of a sequence that include a camera (or motion) transformations.

In this paper, we address the problem of multi-frame registration by means of an *eigenfeatures* approach, where linear subspace constraints are based on the assumption of constancy in the appearance subspace. We frame the multiple-image registration in a two-step iterative algorithm. A feature space is built through and SVD decomposition of a second moment matrix provided by the images in the sequence to be analyzed. This technique allows us to codify images as points capturing the *intrinsic degrees of freedom* of the appearance, and at the same time, it yields compact description preserving visual semantics and perceptual similarities [14, 11, 10].

The outline of the paper is as follows: section 2 frames the idea of using the eigenfeatures approach and its relation with the parametric model of transformations. More specifically, we analyze how such an appearance subspace is built according to a previously selected frame of reference. Therefore, a polynomial model is introduced in order to link the appearance constraints to the transformations that occurred across the sequence. In the experimental results, section 3, we show a new manner of encoding temporal information. We point out that when parallax is involved in the problem of video registration, the temporal representation gives a visual notion of the depth in the scene, and therefore it offers the possibility of extracting the affine $3D$ structure from multiple views. The relation between the surface polynomial's degree and $3D$ affine structure is also illustrated. In section 4, the summary and the conclusions of this paper are shown.

2 Appearance Based Framework for Multi-Frame Registration

In this section, we present an objective function which takes into account appearance representation and time evolution between each frame and a frame of reference. In this case, temporal transformations estimation is based on the fact that images belonging to a coherent sequence are also related by means of their appearance representation.

Given a sequence of F images $\{I_1, \ldots, I_F\}$ (of n rows and m columns) and a selected frame of reference I_0, we can write them in terms of column vectors $\{y_1, \ldots, y_F\}$ and y_0 of dimension $d = n \times m$. Both pictures *pixel-based* I_i and *vector-form* y_i of the i-th image in the sequence are relevant in the description of our method. The first representation I_i is useful to describe the transformations that occurred to each pixel. The vector-form picture is utilized for analyzing the underlying appearance in all the sequence.

Under the assumption of brightness constancy, each frame in the sequence I_i can be written as the result of a Taylor's expansion around the frame of reference I_0:

$$I_i(\vec{x}) = I_0(\vec{x}) + \nabla I_0(\vec{x})^T \vec{\omega}_i(\vec{x}) \tag{1}$$

This is equivalent, in a vector-form, to:

$$y_i = y_0 + t_i \tag{2}$$

where t_i is the vector-form of the second summand $\nabla I_0(\vec{x})^T\vec{\omega}_i(\vec{x})$ in eq. (1). First description is exploited in section 2.2, where the parametric polynomial model to describe the velocity field estimates is applied. The vector-form description in eq (2) is employed in the following section 2.1 to develop the appearance analysis respect to a chosen reference frame.

2.1 Appearance Representation Model

First of all, we need to define a space of features where images are represented as points. This problem involves finding a representation as a support for analyzing the temporal evolution. To address the problem of appearance representation, authors in [14, 11, 10] proposed Principal Component Analysis as redundancy reduction technique in order to preserve the semantics, i.e. perceptual similarities, during the codification process of the principal features. The idea is to find a small number of causes that in combination are able to reconstruct the appearance representation.

One of the most common approaches for explaining a data set is to assume that causes act in linear combination:

$$y_i = W\xi_i + y_0 \tag{3}$$

where $\xi_i \in \Re^q$ (our chosen reduced representation, $q < d$) are the causes and y_0 corresponds to the selected frame of reference. The q-vectors that span the basis are the columns of W ($d \times q$ matrix), where the variation between the diferents images y_i and the reference frame is encoded.

With regard to equation (2), and considering the mentioned approximation in (3), we can see that the difference t_i between the frame of reference y_0 and each image y_i in the sequence is described by the linear combination $W\xi_i$ of the vectors that span the basis in W. Notice that in the usual PCA techniques y_0 plays the role of the sample mean. In recognition algorithms this fact is relevant, since there is assumed that each sample is approximated by the mean (ideal pattern) with an added variation which is given by the subspace W. However, in our approach, each image y_i tends to the frame of reference y_0 with a certain degree of variation, which is represented as a linear combination of the basis W.

Furthermore, from eq. (1), the difference t_i, that relies on the linear combination of the appearance basis vectors, can be described in terms of the parametric model which defines the transformation from the reference frame y_0 and each image y_i. This parametric model is developed in the following section 2.2. Besides, from the mentioned description in terms of a subspace of appearance, we can see the form that takes the objective function to be minimized. Indeed, the idea is to find: a basis W, a set of parameters $\{p_1, ..., p_r\}$, (that model the temporal transformations), and a set of registered images where the squared distance between the difference obtained through the taylor's expansion t_i and the projected vector in the appearance subspace $W\xi_i$ is minimum, i.e.:

$$\mathcal{E}(W, \ldots, p_1^i, \ldots, p_r^i, \ldots) = \sum_{i=1}^{F} \mid t_i(p_1^i, \ldots, p_r^i) - W\xi_i \mid^2 \tag{4}$$

The minimization of this objective function requires of a two-step iterative procedure: first it is necessary to build an appearance basis, and therefore, to estimate the parametric transformations that register the images in the sequence. In the following sections introduce closed forms solutions for each step.

2.2 Polynomial Surface Model

In this section we present a polynomial method to estimate the transformation between de reference frame I_0 and each frame I_i in the sequence. To this end we utilize the pixel-based picture. From equation (1) we can see that the difference between a frame I_i and the frame of reference I_0 relies on the velocities field $\vec{\omega}_i(\vec{x})$. A s-degree polynomial model for each velocity component can be written as follows:

$$\vec{w}_i(\vec{x}) = \mathcal{X}(\vec{x})\vec{P}_i \tag{5}$$

where $\mathcal{X}(\vec{x})$ is a matrix that takes the following form:

$$\mathcal{X}(\vec{x}) = \left[\begin{array}{c|c} \Omega(\vec{x}) & 0 \\ \hline 0 & \Omega(\vec{x}) \end{array}\right]$$

with

$$\Omega(\vec{x}) = \left[\begin{array}{ccccccccc} 1 & x & y & xy & x^2 & \dots & (x^l y^k) & \dots & y^s \end{array}\right]$$

where $\Omega(\vec{x})$ is a $d \times 2r$, ($r = (s+1)(s+2)$), matrix that encodes pixel positions, and $\vec{P_i}$ is a column vector of dimension $r = (s+1)(s+2)$, which corresponds to the number of independent unknown parameters of the transformation. In matrix language $\mathcal{X}(\vec{x})$ is a matrix $2d \times r$, $\vec{P}$ has dimensions $r \times 1$, and the velocities corresponding to each pixel can be encoded in a matrix $\vec{w_i}(\vec{x})$ of dimensions $2d \times 1$. The gradient expression in the linear term of the taylor's expansion (1) can be written in a diagonal matrix form as follows:

$$G_x = \left[\begin{array}{cccc} g_x^1 & 0 & \dots & 0 \\ 0 & g_x^2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & g_x^d \end{array}\right] G_y = \left[\begin{array}{cccc} g_y^1 & 0 & \dots & 0 \\ 0 & g_y^2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & g_y^d \end{array}\right]$$

Stacking horizontally both matrices we obtain a matrix G of dimensions $d \times 2d$: $G = [G_x \mid G_y]$. Therefore, according to the vector-form in eq (2), the difference t_i between the i-th frame y_i and the frame of reference y_0, is expressed in terms of the polynomial model through:

$$t_i(\vec{x}, \vec{P_i})_{d\times 1} = G_{d\times 2d}\mathcal{X}(\vec{x})_{2d\times r}\vec{P_i}\mid_{r\times 1} \tag{6}$$

Given that the term $G_{d\times 2d}\mathcal{X}(\vec{x})_{2d\times r}$ is computed once for all the images in iteration, we re-name it as $\Psi_{d\times r} = G_{d\times 2d}\mathcal{X}(\vec{x})_{2d\times r}$. Notice that even when images are highly dimensional, (e.g. $d = 240 \times 320$), the computation of Ψ can be perfomed easily in *Matlab* by means of the operator ".*", without incurring in an out of memory.

2.3 The Algorithm

Given the parametric model for the transformations of the images in a sequence, the objective function (4) can be written explicitly in terms of the parameters to be estimated:

$$\mathcal{E}(W, \vec{P}_1, \dots, \vec{P}_F) = \sum_{i=1}^{F} \mid \Psi\vec{P_i} - W\xi_i \mid^2 \tag{7}$$

In order to minimize this objective function, we need a two step procedure: first given a set of images, the subspace of appearance W is computed, and secondly, once the parameters $\vec{P_i}$ that register each frame y_i to the frame of reference y_0 are obtained, the images are registered in order to build again a new subspace of appearance.

a. Appearance Subspace Estimation. Consider an intermediate iteration in the algorithm, thus, the set of registerd images to be analyzed are: $\{\phi_1(y_1, \vec{P}_1), \dots, \phi_F(y_F, \vec{P}_F)\}$. From this set and the reference frame y_0, the appearance subspace can be performed by means of an Singular Value Decomposition of the second moments matrix*:

$$\Sigma = \sum_{i=1}^{F}(\phi_i(y_i, \vec{P_i}) - y_0)(\phi_i(y_i, \vec{P_i}) - y_0)^T \tag{8}$$

The column vectors of W correspond to the q first eigenvectors of (8), that have been previously ordered from the largest eigenvalues to the smallest one. The projected coordinates onto the appearance subspace are: $\xi_i = W^T(\phi_i(y_i, \vec{P_i}) - y_0)$.

*This can be perfomred following the idea introduced in [10].

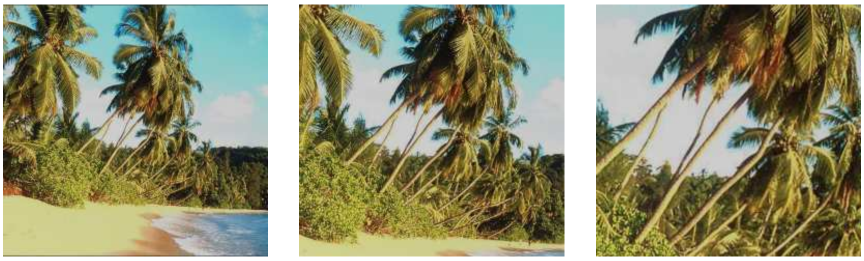


Figure 1: Some selected frames (1st, 3rd, 5th) from a sequence: 1,41,81 form the original one.

b. Transformation Parameters Estimation. Setting derivatives to zero in eq. (7) respect to the transformation parameters, they are computed as follows:

$$\vec{P}_i = \left[\Psi^T \Psi\right]^{-1} \Psi^T W \xi_i \tag{9}$$

Note that the matrix $\left[\Psi^T \Psi\right]^{-1}$ has manageable dimensions $r \times r$, i.e. in the linear polynomial case $r = 3$, in the quadratic case $r = 12$, etc. We can see that while the appearance (global information) is codified in W, the local infomation which is related to the pixels in the images is encoded in Ψ. With this, we can see that their combination in eq. (9) gives a relation between each image's subspace coordinates ξ_i and the parameters that register each frame to the frame of reference. Moreover, this method considers the contribution of all the frames in the sequence to the estimation of each single set of transformation parameters. From these estimates, we compute a new set of registered images $\{\phi_1(y_1, \vec{P}_1), \ldots, \phi_F(y_F, \vec{P}_F)\}$ and repeat step *a*. These two steps are iterated until a certain degree of tolerance in the value obtained through the error function eq. (7).

3 Experimental Results

In order to see the range of applications of this technique, we deal with two sort of problems. First, we study a camera movement, where it is shown the different results that appear when it comes to deal with a specific selected frame of reference. In particular, this camera movement is a zoom that can be interpreted in terms of registration as zoom-in or zoom-out operations depending on the selection of the reference frame. Secondly, the significance of the polynomial's degree is analyzed through a sequence that includes a moving object due to a parallax effect.

3.1 Selecting a Reference Frame. Consequences in the Registration

This topic is about camera operations with a single planar motion. Figure 1 shows three frames from a sequence of 100 frames, where a zoom-in is originally perfomed. In this particular case, we selected 5 frames $(1^{st}, 21^{st}, 41^{st}, 61^{st}, 81^{st})$ from the original sequence to perform this analysis. This was motivated in order to exploit the fact that the images have not to be taken continuously; the key point is that they are related by the same underlying appearance. Here, we analyze three cases depending on the selection of the reference frame: zoom-in registration fig.2 and zoom-out registration fig.3.

Figure 2 shows a zoom-in registration that has been obtained selecting as reference frame the left side image in fig. 1. To this end, we utilized a linear polynomial model (1 degree), and the subspace of appearance has been built using just one eigenvector, given that appearance is mainly conserved in the sequence. The point is that the dimension not only depends on the error reconstruction as in a recognition problem [14, 11, 10], but also relies on the selection of the frame of reference.

Figure 2 (a) shows a time evolution of the registered sequence images, while figure 2(d) the registration picture also explains the module of the velocity field in each pixel. Latter figure gives a notion of the situation

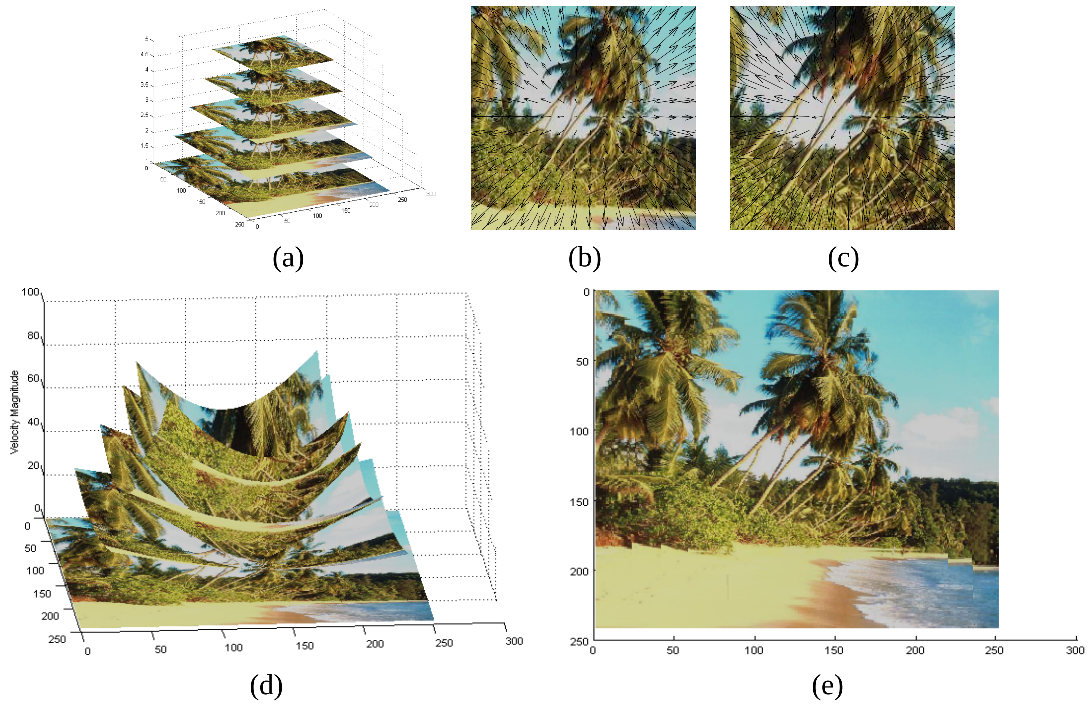


Figure 2: Zoom in: (a) Registered images according to a 1 degree polynomial model, where the first frame has been taken as reference frame. Optical flow field corresponding to the third frame (b), and to the last frame (c). (d) Velocity field module representation of the sequence of images. (e) Top view of (d).

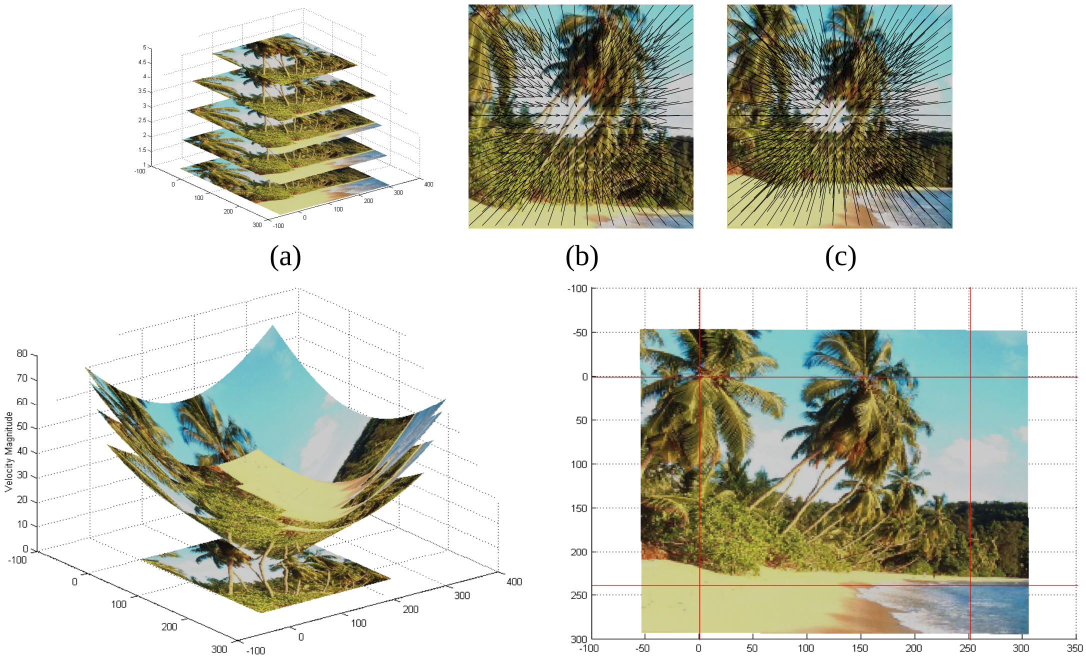
of the camera's center. This is highly useful to perform an analysis of camera operations from this registration technique. Figures 2(b) and (c) show the estimate optical flow field, which is computed respect to the reference frame, in some frames of the sequence. When it comes to register from this vector field, we have to take the inverse direction that is indicated in each arrow.

Besides, even though the sequence evolution showed a zoom-in camera operation, we can register selecting as reference frame the last frame, (see right side image in fig. 1). The main difference between the registrations in figure 2 and figure 3 is the size of the final mosaic (top views of fig. 2(a) and fig. 3(a)). Actually, the size of the final mosaic selecting as reference frame the first frame is equal to the reference frame. However, taking as reference frame the last frame (case fig3) the size of the final mosaic is bigger than the size of the reference frame. This is clearly reflected in the module representations of the sequence registration, figures 2(d) and 3(d).

3.2 Analyzing the Complexity in the Polynomial Model. Towards $3D$ Affine Reconstruction

In order to get an insight into the relation between the complexity of the polynomial estimation of the velocity field and the $3D$ affine structure which is encoded in the image sequence, we deal with three sort of experiments. The idea is to see the variety of possibilities that the polynomial surface model offers in this registration framework. Three cases present different relative motions across the image sequence.

First sequence of images corresponds to a camera panning operation, where the target is an object with different depths respect to the camera position. This fact produces a parallax effect onto the image plane, which means that the affine model (degree 1) to estimate the velocities field is not sufficient. Figure 4 shows three frames of a sequence of ten images, which have been used to perform the first analysis of $3D$ motion. To estimate the introduced parametric optical flow, we used a third degree polynomial model, which according to eq. (5) represents 20 parameters in the estimation process.



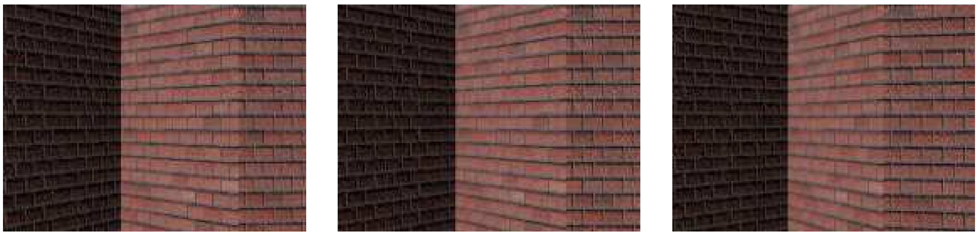
(a) (b) (c)

(d) (e)

Figure 3: Zoom out: (a) Registered images according to a 1 degree polynomial model, where the last frame has been taken as reference frame. Optical flow field corresponding to the third frame (b), and to the first frame (c). (d) Velocity field module representation of the sequence of images.(e) Top view of (d), where the red lines show the original size of the reference frame.

Registration results are shown in figure 5 (a) and (b), where the first frame has been taken as reference frame. First one is a velocity field module representation of the image sequence, where is can be seen that the edge between the dark region and the light one is in the same pixel reference coordinate position in each frame. We use the method described in [13] to estimate the $3D$ affine structure from the registered images. To this end we utilized all the pixels in the images to perform the factorization method. This fact is present in the $3D$ reconstruction results (see figs. 5(c) and (d)) since the union edges between planes are smoothly reproduced. To reproduced properly these mentioned high frequency regions, it is necessary to consider hard constraints in the $3D$ recovery step. This topic remains a task for our future research.

Second experiment deals with a translational camera motion. Two main motion layers are present in this sequence due to a parallax effect. Figure 6 shows three frames of a sequence of five, where the tree belongs to a different motion layer than the background (houses). Apparently, the sequence can be interpreted as a moving object with moving background as well. Nevertheless, the cause is the difference in depth that the tree is situated from the background, and, moreover, the specific movement of the camera. The registration has been performed using 2 eigenvectors of basis appearance and a 3^{rd} degree polynomial model for the motion field. The result of this can be seen in figures 7 (a) and (b). More specifically, figure 7 (a) gives a certain notion of the relative depth among different regions in the images, due to the module representation of the velocity field; regions with higher velocity module are meant to be nearer the camera than regions with a lower module. Figure 7 (b) shows a top view of (a), where the result of registering is regarded in terms of a mosaic image. Finally, figure 7(c) shows the $3D$ affine structure estimation using [13], where all the images pixels in the sequence have been employed. With this, we can see that the final $3D$ smooth surface shows this mentioned



(a) (b) (c)

Figure 4: Three frames of a sequence of ten images. First image (a) corresponds to the first frame, (b) is the fifth and (c) is the tenth.

depth difference due to parallax.

4 Summary and Conclusions

The problem of multi-frame registration has been presented through an *eigenfeatures* approach, where linear subspace constraints are based on the assumption of constancy in the appearance subspace. One of the main contributions of the appearance subspace encoding is that the appropriate scale in each problem is captured from the images themselves, i.e., robust time derivatives of the optical flow are obtained from eigenfeatures. As mentioned in section 2.1, this fact is due to the consideration of both pictures, *pixel-based* and *vector-form*, into the same formulation. First picture exploits local information, while the vector-form is utilized for global information purposes. The aim of this is to point out that image time derivatives are computed coupling the linear combination of the eigenfeature basis and the spatial information which is provided by the polynomial surface model (pixel-based picture). This coupling is performed in a objective function that is minimized in order to obtain the registration of a sequence.

This approach is combined with a polynomial model for estimating the transformation that has been produced across the sequence. Although the objective function, that corresponds to the connection between the global coordinates in the subspace representation and the parametric optical flow estimates, requires a two step procedure, the minimization steps have been reduced to linear least squares subproblems, whose solutions turned out to be in a closed form for each iteration.

We dealt with a variety of experiments in order to analyze the range of applications of this registration technique. One of the purposes is to see that the contribution of a parametric multiframe optical flow estimation provides a smooth reconstruction of the $3D$ affine structure the is imaged in the sequence, where all the pixels information is employed. Besides, from section 3.2, the relation between the polynomial model and the $3D$ reconstruction has been observed qualitatively. It is a task of future work to give a formal description of this relation. Also, the idea of including hard constraints to the reconstruction method in this polynomial framework is encouraging. The purpose is to keep the advantageous motion analysis estimation in terms of a few number of parameters, and, at the same time, the future goal is to introduce prior knowledge in order to indicate where the curvature is locally higher.

Acknowledgments

This work was supported by CICYT grant TEL99-1206-C02-02.

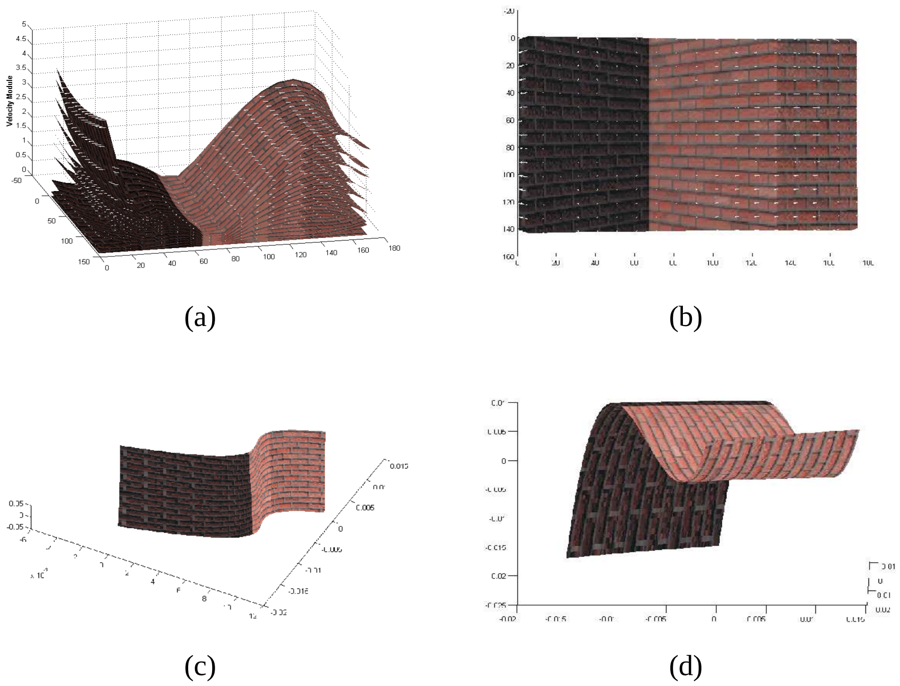


Figure 5: Velocity field module representation (a) of the registered images, where 2 eigenvectors of appearance and a polynomial model of 3^{rd} degree have been used to this estimation. Fig. (b) is the top view of (a). Two views, (c) and (d), of the $3D$ affine structure of the sequence.

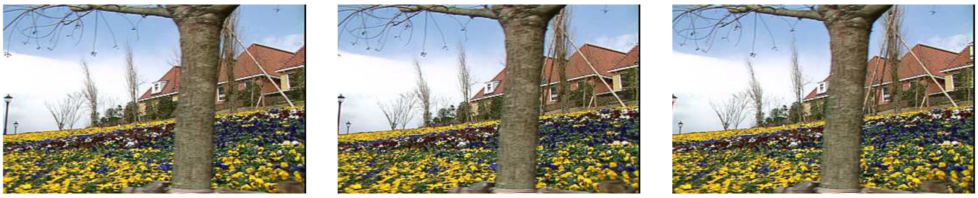


Figure 6: Three frames of a sequence of five images. These images correspond to 1^{st}, 3^{st} and 5^{st} (from right side to left side).

References

[1] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. *ICCV*, pages 777–784, 1995.

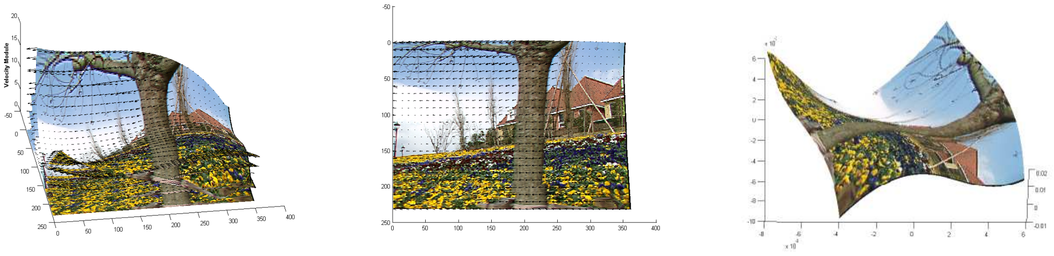
[2] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. *ECCV*, pages 237–252, 1992.

[3] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *ECCV*, pages 329–342, 1996.

[4] R. Chipolla, Y. Okamoto, and Y. Kuno. Robust structure from motion using motion parallax. *ICCV*, pages 374–382, 1993.

[5] H. Heeger and A. Jepson. Simple method for computing 3D motion and depth. *ICCV*, pages 96–100, 1990.

[6] B. Horn. Relative orientation. *IJCV*, pages 58–78, 1988.



(a) (b) (c)

Figure 7: Velocity field module representation (a) of the registered images, where 2 eigenvectors of appearance and a polynomial model of 3^{rd} degree have been used to this estimation. Fig. (b) is the top view of (a). A view (c) of the $3D$ affine structure of the sequence.

[7] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using region alignment. *IEEE trans. on P.A.M.I.*, 19(3), 1997.

[8] S. Ju, M. Black, and A. Jepson. Multilayer, locally affine optical flow, and regularization with transparency. *CVPR*, pages 307–314, 1996.

[9] F. Lustman, O. Faugeras, and G. Toscani. Motion and structure from motion from point and line matching. *ICCV*, pages 25–34, 1987.

[10] H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *IJVC*, 14(5):5–24, 1995.

[11] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[12] H. Sawhney and S. Ayer. Compact representation of videos through dominant and multiple motion estimation. *IEEE Trans. on PAMI*, 18:814–829, 1996.

[13] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9:137–154, 1992.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[15] J. Wang and E. Adelson. Layered representation for motion analysis. *CVPR*, pages 361–366, 1993.

PAGE INTENTIONALLY LEFT BLANK