# Technical Report 2003-XX-CIRL-CS-JHU

# Real-time Video Mosaicing with Adaptive Parametric Warping

Le Lu

Computational Interaction and Robotics Lab
Computer Science Department
the Johns Hopkins University
Baltimore, MD 21218, USA

## Abstract

*Image registration/video mosaicing can be computed by the intensity difference minimization between two images. The image appearance changes are normally caused by images taken under different viewpoints. Families of spatial transformations are used to compensate for the difference. Different sensors and different illumination conditions can also change image intensities [19, 10]. Mutual information based algorithm is successfully employed to register multi-modal image [47]. In [19, 30], Linear sub-space modelling is used to compute the illumination difference.*

*By stitching a sequence of image frames, we can get a mosaic for the broader field of view. Image mosaic is a very useful visual representation for the 3D world scene [27], medical imaging [8] and so on.*

*There are mainly two approaches to treat image regions, as direct observation measurements in [4, 19, 28, 30, 40], or feature base criterion in [5, 8, 9, 14, 42]. In this paper, we propose a closed form solution based on the linear least-squares method to compute the parametric warping between images. Consequently, the local image intensity differences are approximated as a linear combination of parametric texture warping templates. Further more, we extend the parameterized warp models in [19] from translation, affine transformation to full perspective transformation or quadratic transformation by the requirement of different applications.*

*To accelerate the calculation of Jacobian matrix, we discuss the selectivity of image pixels in template matching. Some real image sequence mosaicing results under different parametric models are shown in this paper and they are processed in real-time. Performance analysis with different parametric models and different scenario are given in detail.*

*We also try to review some new and significant techniques about image registration after Brown's work [6]. Based on these, several potential important research directions are pointed out as future works.*

# 1 Introduction

Image registration/alignment for video mosaicing has many research and real applications. Based on the definition in [6], there are two kinds of image distortions needed to be handled in registration. The first type is the distortion that cause the misalignment which can be minimized by the parametric transformations with fitting coefficients. Non-parametric warping is more suitable, when the scenario is very complex and can not be well presented by the parametric warping. For instance, a patch based optical flow algorithm [40] can be used to find the relative disparities of image features, then the disparities of neighboring pixels can be interpolated. The second type of distortion is not the source of spatial misalignment, but make the registration more difficult. This distortion is the outlier in the registration procedure, according to the first one. For example, occlusions or expressional deformations in template based face tracking [19] belong to the second distortion. The iteratively reweighted least squares (IRLS) technique[1] [19, 49] can be used in conjunction with the image gradient to deal with non-rigid motions and occlusions. This approach decouples the above two types of distortions and solve them in a sequential manner. After the global face motion is tracked or stabilized, eye blinking and mouth motions are then detected [20].

From the viewpoint of image appearance measurement, there are basically two kinds of methods to compute the registrations between successive images or the locations of live frames on the reference imagery. Different images can be stitched together by the certain kind of warping transformations to compensate for the visual motion disparity. Direct approaches [4, 19, 28, 30, 40] use all the image pixels available to compute the image gradient and Jacobian matrices for registration. No corner feature or edge detection is required. The SSD (Sum of Square Differences) error is minimized with a closed form solution from a set of over-determined linear equations. In theory, no iteration is necessary. However, the estimated parameters can be refined locally in an iterative manner, by considering the imperfection of spatial transformation modelling for visual motions and nonlinear image noises. On the contrary, some researchers [5, 8, 9, 14, 42] prefer to first detect certain corner-like or edge-like features[2] from images, then optimize the registration parameters by finding the corresponding spatial relations between image feature pairs in different camera views. By considering the image noises and mismatches, many of robust estimators (for example, Ransac [42, 9], M-estimator [50, 8], MLESAC [43] and IMPSAC [45]) are used in the feature based matching algorithms. As a trade-off, a mixture approach is employed in the work of video alignment between ariel images and the reference imagery [48, 26, 24]. Image edge-like features are firstly aligned, then their neighboring image regions are used as supporting areas to minimize the according SSD errors.

There are some cases where image intensity based registration is unavailable. For example, alignment of two non-overlapping image sequence, images captured by different sensor sources, or under different illumination conditions [10], register the 3D model with the reference image [47] are all need some other coherent information to be discovered from the successive image frames. In [10], two non-overlapped image sequences are aligned together using their camera's correlated temporal motion, assuming that the two cameras are bound together. The coherent appearance requirement among images in the standard registration algorithms is replaced by the constraint of the coherent temporal behavior of the two image sequences which is obtained by Homography or Epipole constraints [23] for the planar or non-planar scenes respectively. This technique can also be applied to significant zooming images or multi-sensor images. From an information-theoretic view, entropy based similarity criterion can be employed to align two object representations from different imaging sources, for example, to register a 3D model with a sequence of 2D images [47]. No direct appearance match-

[1] A well-known optimization of robust estimation.

[2] These kinds of feature can be detected from whether there exists 1 or 2 significant large eigenvalues, after the SVD decomposition [17] from the covariance matrices of the image intensities [38]. The most popular point feature is Harris corner feature [21].

ing exists between 3D and 2D object representations, but mutual information with the model and the image can be formulated as the similarity definition for alignment. Actually, the mutual information is found to be maximized when the two different imaging presentation is correctly registered. This approach is based on stochastic approximation[3], which can be efficiently implemented and provides a robust solution for tracking with clutters and occlusions.

For the mathematical formulation, the image registration methods can be classified as parametric or non-parametric approaches. In projective geometry theory [16, 23], there exist two kinds of Homography matrices for the planar scenes or stationary camera with pure rotations [22]. When the scene is observed by the camera from a certain distance[4], affine transformation model can be near perfectly used to approximate the relative motions between two camera views. In the influential paper [19] for parametric tracking, mostly translation model and affine motion model are explored for the planar surface tracking. The Jacobian matrices are derived from the explicit parametric motion patterns. In [30], full 3D face tracking[5] are considered, thanks to a cylinder head model. The motion vectors are obtained from image intensity changes by local disturbances of warping parameters[6]. On the other side, Joint View Triangulation (JVT) [35] is a typical non-parametric algorithm for image registration and rendering. JVT interpolates the inter-frame visual motions of unmatched image pixels from their semi-dense matched neighbors via the triangle-based texture mapping. All unmatched image portions are automatically segmented as triangles, and each triangle is bounded by three of their matched image feature neighbors via Delaunay triangulation [34].

Visual motions among different image frames in a video sequence are strongly correlated with the camera's motion trajectory (called ego-motion) and the complexity of observed scenes. If the camera motion is geometrically constrained with high precision, the mosaicing task can be performed without explicitly computing the inter-frame visual motions from image observations [39, 33]. A high resolution mosaic (called concentric mosaics [39]) can be generated from the dense sampled plenoptic function [32] captured by a smooth panning video camera. The camera is explicitly calibrated by the well-controlled robotic motions. The final mosaic is composed of hundreds of vertical slots extracted from captured image frames. The index[7] of the slot can be determined by the spatial order between the current rendering viewpoint and camera positions in the image capture circle, which can be read from the image frame number. Furthermore, other non-perspective representations of mosaics, ie. parallel projected camera views can be obtained by rebinning a large sequence of perspective images [11]. A similar For more general types of camera motions, the methodology of mosaicing on adaptive manifolds is proposed in [33]. Mosaicing is performed by projecting thin strips from the images onto manifolds which are adapted to the camera motion. The collected strips should be warped to make the resulting optical flow parallel to the direction in which the mosaic is constructed. The use of more general manifolds overcomes the limitations of camera motions; the use of thin strips is easy to handle the lens distortion, motion parallax and moving objects.

We define our problem as building an image mosaic online from a sequence of video images of planar or quadratic scene with a freely moving camera. The surface can be an approximation of a scene which has insignificant parallax by camera motion. Particularly, one important application of our work is retina image

[3] A superposition of Gaussian densities with Parzen Windows algorithm [15] is used to sample the underling distributions. Besides Gaussian density, any differentiable function could be used to present the distribution. Another good choice is the Cauchy density. On the other hand, non-parametric representation, like histogram, is also in effect here. See the detail analysis in [13].

[4] Empirically, when the distance ratio between the object depth and object width is equal or above 6, affine camera model can be successfully employed.

[5] Totally, LaCascia et al. track the head motions with six degrees of freedom, three rotations and three translations.

[6] Here, the calculation of motion vector is according to the Jacobian matrix computation in [19], but is a little bit of less strict in mathematics. The scale of local disturbances is hard to be determined. An adaptive scale estimation algorithm is analyzed in [7].

[7] The index is composed of the frame number of image and the slot's vertical position in the image.

mosaicing for eye surgery. Age-related macular degeneration is a set of vascular disorders on the human retina. This leads to loss of sight, and in some cases to legal blindness. Laser and radiation treatment is normally not quite effective. The eye surgery plan is insert micro-surgical tools into the lumen of a retinal blood vessel, inject therapeutic agents or remove the excess vascularity, with a part of the retina, preserving the fovea. Our task is to assist the micro-surgical procedure by providing a real-time registration between an intra-operative microscopic image, and a pre-operative angiographic image of the retina. The latter can be used to make a surgical plan. with the registration established, the microscopic images can be overlaid on the plan or other important pre-operative data, enhancing the surgeon's view of the procedure [37].

In this paper, we built a hierarchical and adaptive framework for real-time video registration/mosaicing by extending [19] from 4 parameters without out-of-plane rotation, to 6 or 8 parameter full 3D motion or even 12 parameter quadratic motion models in an incremental order. This adaptive warping model enables us to deal with more complex surfaces and more general camera motions. Another contribution of this paper is that we propose a simple but efficient rule to select part of image pixels that make the most significant contribution of optimization in Jacobian matrix. We describe our paper as follows. In section 2, our video registration framework is presented. Real video sequences are shown in section 3 to demonstrate the validity of our method, and we conclude this paper in section 4.

# 2 Mathematical Formulation of Parametric Warping for Video Registration

In this section, we first describe the general framework of parametric motion estimation [19]. Then, we discuss our adaptive parametric warping implementation with a course-to-fine hierarchical framework.

## 2.1 Parametric Motion Estimation

Let $\mathbf{x} = (x, y)^T$ denote a pixel coordinates in the image plane. Let $I(\mathbf{x}, t)$ be the intensity value for image pixel $\mathbf{x}$ in image $I$ taken at time $t$. The gray-level image gradient at the image point $\mathbf{x}$ is denoted by $\bigtriangledown_{\mathbf{x}} I(\mathbf{x}, t)$. A target region $\mathcal{I}$ with $N$ image points is defined as

$$\mathbf{I} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$$

Let $f$ be a 2D-2D transformation relation: $f : \Re^2 \mapsto \Re^2$, which is parameterized by the coefficient vector $\mu$, therefore, $f(\mathbf{x};\mu(\mathrm{t}))$ denotes the parametric motion of each image pixel $\mathbf{x}$ in terms of $\mu$(t) with m components. Normally, $\mu$(t) is a set of time variant parameters that need to be estimated in real-time in our task. The transformation, sometimes called warping, can be expressed as

$$\mathbf{u} = (u, v)^T = f(\mathbf{x}; \mu)$$

With the *intensity constancy constraint* [25], we have the following equation (1)

$$I_0(\mathbf{x}, t_0) = I(f(\mathbf{x}; \mu(t)), t) \tag{1}$$

Where $I_0$ is the reference frame taken at time $t_0$, $I$ is another image taken at time $t$, and the relative transfor-

mation is $\mu(t)$. For a certain image region $\mathcal{I}$, we denote its pixel intensities using a vector as

$$\mathbf{I}(\mu,t) = \begin{pmatrix} I(f(\mathbf{x}_1;\mu(t)),t) \\ I(f(\mathbf{x}_2;\mu(t)),t) \\ \vdots \\ I(f(\mathbf{x}_N;\mu(t)),t) \end{pmatrix}$$

where N is the pixel size of $\mathcal{I}$. By making the constant intensity assumption, the motion of image pixels can be represented in terms of their spatial and temporal derivatives.

$$\mathbf{I}(\mu+\delta\mu,t+\triangle t) = \mathbf{I}(\mu,t) + \mathbf{M}(\mu,t)\delta\mu + \triangle t\mathbf{I}_t(\mu,t) + h.o.t \tag{2}$$

The above linearization is carried out by expanding $\mathbf{I}(\mu+\delta\mu,t+\triangle t)$ into a Taylor series of $\mu$ and $t$. $\mathbf{I}_t(\mu,t)$ is the partial derivatives of $\mathbf{I}$ with respect to the component of the time parameter $t$, and is written as

$$\mathbf{I}_t(\mu,t) = \begin{pmatrix} I_t(f(\mathbf{x}_1;\mu(t)),t) \\ I_t(f(\mathbf{x}_2;\mu(t)),t) \\ \vdots \\ I_t(f(\mathbf{x}_N;\mu(t)),t) \end{pmatrix}$$

$\mathbf{M}$ is the Jacobian matrix of $\mathbf{I}$ with respect to $\mu$, which is a $N \times m$ matrix of partial derivatives.

$$\mathbf{M}(\mu,t) = \begin{pmatrix} \mathbf{I}_{\mu 1}(\mu(t),t) & \mathbf{I}_{\mu 2}(\mu(t),t) & \dots & \mathbf{I}_{\mu m}(\mu(t),t) \end{pmatrix} \tag{3}$$

$$\mathbf{I}_{\mu i}(\mu,t) = \begin{pmatrix} I_{\mu i}(f(\mathbf{x}_1;\mu(t)),t) \\ I_{\mu i}(f(\mathbf{x}_2;\mu(t)),t) \\ \vdots \\ I_{\mu i}(f(\mathbf{x}_N;\mu(t)),t) \end{pmatrix}$$

From the intensity constant constraint, the motion parameter vector of the target region can be estimated at time $t$ by minimizing the following least squares error function

$$\mathbf{e}(\mu,t) = \sum_{\mathbf{x}\in\Re} (I(f(\mathbf{x};\mu(t)),t) - I_0(\mathbf{x},t_0))^2 \tag{4}$$

By substituting (2) into (4) and ignoring the higher order terms, we obtain

$$\mathbf{e}(\mu,t) \approx \| \mathbf{I}(\mu,t) + \mathbf{M}\delta\mu + \mathbf{I}_t(\mu,t)\triangle t - \mathbf{I}(\mathbf{0},t_0) \|^2 \tag{5}$$

With the additional approximation $\mathbf{I}_t(\mu,t)\triangle t \approx \mathbf{I}(\mu,t+\triangle t) - \mathbf{I}(\mu,t)$, (5) becomes

$$\mathbf{e}(\mu,t) \approx \| \mathbf{M}\delta\mu + \mathbf{I}(\mu,t+\triangle t) - \mathbf{I}(\mathbf{0},t_0) \|^2 \tag{6}$$

In this equation, $\mathbf{I}(\mu,t+\triangle t) - \mathbf{I}(\mathbf{0},t_0)$ can be considered as the distortion errors of warping the image $\mathbf{I}_{t+\triangle t}$ at time $t+\triangle t$ with the former parameters $\mu_t$ to $\mathbf{I}(\mathbf{0},t_0)$. $\delta\mu$ is the vector of offsets of $\mu_t$, to compensate for the above warping errors. Solving equation (6) in a linear least squares manner, we have

$$\delta\mu = -(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T(\mathbf{I}(\mu,t+\triangle t) - \mathbf{I}(\mathbf{0},t_0)) \tag{7}$$

provided the jacobian matrix **M** is full rank. After this, the new set of warping parameters is updated from

$$\mu(t + \triangle t) = \mu(t) + \delta\mu \tag{8}$$

Most importantly, the construction of Jacobian matrix **M** is the critical problem. According to the partial derivative definition, each element of this matrix is given by

$$m_{ij} = \mathbf{I}_{\mu j}(\mathbf{f}(\mathbf{x}_i; \mu), t) = \nabla_f \mathbf{I}(\mathbf{f}(\mathbf{x}_i; \mu), t)^T \mathbf{f}_{\mu j}(\mathbf{x}_i; \mu) \tag{9}$$

where $\nabla_f \mathbf{I}$ is the gradient of $I$ with respect to the components of the motion model **f**. Unfortunately, this gradient can not be computed directly. By differentiating both sides of 1, we get

$$\nabla_x \mathbf{I}(\mathbf{x}, t_0) = \mathbf{f}_x(\mathbf{x}; \mu)^T \nabla_f \mathbf{I}(\mathbf{f}(\mathbf{x}; \mu), t) \tag{10}$$

where $\mathbf{f}_x$ is the $2 \times 2$ Jacobian matrix of **f** treated as a function of $\mathbf{x} = (x, y)^T$,

$$\mathbf{f}_x(\mathbf{x}; \mu) = [\frac{\partial \mathbf{f}(\mathbf{x}; \mu)}{\partial x} | \frac{\partial \mathbf{f}(\mathbf{x}; \mu)}{\partial y}]$$

Combining 9 with 10, we find that **M** can be written as

$$\mathbf{M}(\mu) = \begin{pmatrix} \nabla_x \mathbf{I}(\mathbf{x}_1, t_0)^T \mathbf{f}_x(\mathbf{x}_1; \mu)^{-1} \mathbf{f}_\mu(\mathbf{x}_1; \mu) \\ \nabla_x \mathbf{I}(\mathbf{x}_2, t_0)^T \mathbf{f}_x(\mathbf{x}_2; \mu)^{-1} \mathbf{f}_\mu(\mathbf{x}_2; \mu) \\ \vdots \\ \nabla_x \mathbf{I}(\mathbf{x}_N, t_0)^T \mathbf{f}_x(\mathbf{x}_N; \mu)^{-1} \mathbf{f}_\mu(\mathbf{x}_N; \mu) \end{pmatrix} \tag{11}$$

Further more, suppose that we can choose **f** so that $\mathbf{f}_x^{-1}\mathbf{f}_\mu$ can be factored into the product of a $2 \times k$ matrix $\Gamma$ which depends only on image coordinates, and a $k \times m$ matrix $\Sigma$ which depends only on $\mu$ as

$$\mathbf{f}_x(\mathbf{x}; \mu)^{-1} \mathbf{f}_\mu(\mathbf{x}; \mu) = \Gamma(\mathbf{x}) \Sigma(\mu) \tag{12}$$

As a result, we substitute (12) into (11), and obtain

$$\mathbf{M}(\mu) = \begin{pmatrix} \nabla_x \mathbf{I}(\mathbf{x}_1, t_0)^T \Gamma(\mathbf{x}_1) \\ \nabla_x \mathbf{I}(\mathbf{x}_2, t_0)^T \Gamma(\mathbf{x}_2) \\ \vdots \\ \nabla_x \mathbf{I}(\mathbf{x}_N, t_0)^T \Gamma(\mathbf{x}_N) \end{pmatrix} \Sigma(\mu) = \mathbf{M}_0 \Sigma(\mu) \tag{13}$$

Recently, Baker and Matthews survey the image alignment algorithms and classify them into four categories [2]. The additive approach estimates an additive increment to the parameters [31], while compositional approach compute an incremental warp instead [40]. Hager and Belhumeur's algorithm is considered as the inverse additive method because they invert the role of the image and the template to achieve very efficient performance. Baker and Matthews proposed a image alignment method called the inverse compositional algorithm by considering the computational efficiency and wider class of warps than linear warps in [19]. We refer their paper [2] for more details. More interestingly, our mosaicing task is different with the template image tracking in [19, 2]. In our case, we have local warping between current successive image frames whose initial warping is represented by a identical matrix. It always avoids the various Jacobian matrices in iterations and

save the computing power. For the global warping among live frames respective to the whole mosaic, incremental warping is used. For affine or perspective transformations, the warps form a semi-group where only the multiplication of 2 warping matrices are needed to obtain the final warp. In short, we use inverse additive approach for local warps and compositional approach for global warps[8]. Because we always need to calculate the overlapping region at first for each iteration with updated parameters and then compute the motion vectors based on it, inverse compositional method can not help us to fix the Jacobian matrix during iterations. In our task, compositional or inverse compositional method does not effect differently.

## 2.2 Case Study: Real-time Video Mosaicing with Adaptive Parametric Warping

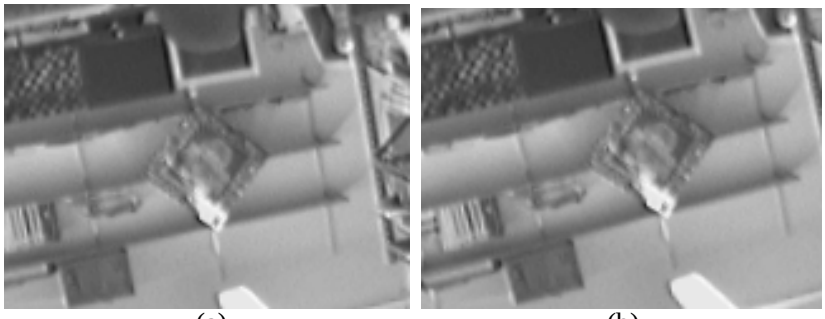
According to our applications, we apply an adaptive framework for the parametric motion estimation. Each building block of the mosaicing process is to register two images. It consists of several steps, and each refines the alignment result of the previous one by adding more parameters into the warping matrix. The number of steps is adaptive to the current warped image difference, or residues. If the residues can not be small enough with the low order warping model, higher order warping is activated for the further compensation.

### 2.2.1 Phase Correlation for Translation Estimation

First of all, a cross correlation based algorithm is used for estimation of 2D translation $d_x$ and $d_y$ in the image coordinates. We use a phase correlation algorithm in [29]. Phase correlation is a frequency domain motion measurement method that makes use of the shift property of the Fourier transform - a shift in the spatial domain is equivalent to a phase shift in the frequency domain. From theoretic analysis, phase correlation is based on the evaluation of the phase of the Cross Power Spectrum (CPS). Therefore the phase correlation is formulated as :

$$\frac{F_1(\zeta,\eta) * F_2^*(\zeta,\eta)}{\mid F_1(\zeta,\eta) * F_2^*(\zeta,\eta) \mid} = e^{j2\pi(\zeta x_0 + \eta y_0)} \tag{14}$$

where $x_0$, $y_0$ are the translational offset of image 1 and image 2, $F_1(\zeta,\eta)$ and $F_2(\zeta,\eta)$ denote the the Fourier transforms of the two images. Normally, there are some high order spatial distortions between these two image, besides translations. Phase correlation is used as the initial alignment for the higher level warping, according to its computational efficiency by avoiding the brute-force search for correlation.



(a) (b)

Figure 1: Two images used for the translational motion estimation by phase correlation : (a) image01 (b) image02

[8]Though we do not call our methods as the names in [2], basically they have the similar functions. Our special 4 parameter affine transformation and 6 parameter perspective transformations in local warps are described in section 2.2.2.

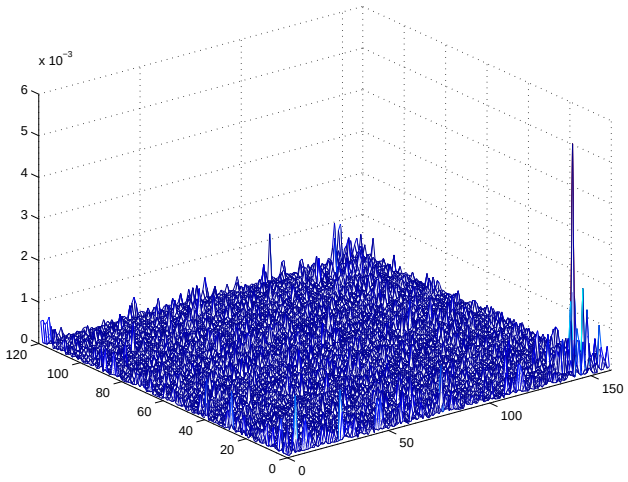



Figure 2: The translational motion estimation with phase correlation.

The two images shown in Figure 1 are used to demonstrate the validity of phase correlation. Clearly, there are not only the translational transformation, but more complex distortions existed between these two images. However, the translation values in X-Y direction are discovered as a distinguished peak from the 2D response function shown in Figure 2.

### 2.2.2 Optimization Formulation and Jacobian Matrix Computation

The basic warping function is the transformation between corresponding pixels of two images. Then $f(\mathbf{x};\mu(t))$ can be expressed by matrix multiplication (15). Assuming the surface is planar or pure rotational camera motion, we have

$$\lambda \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = W \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{15}$$

where $(u, v)$ is the pixel coordinates of a physical point projected in the first image, and $(x, y)$ is the pixel coordinates of its projection in the second image. Here we take the first image as the anchor image for others to be aligned. $\lambda$ is a scaling factor and $W$ is a parameterized $3 \times 3$ matrix presented the parametric warping.

If we only consider scale $s$ and in-plane rotation $\theta$ in addition to $d_x$ and $d_y$, which are corresponding to 3 dimension translations and rotation around Z axis of the camera, the 4 parameter warping matrix is

$$W_a = \begin{pmatrix} s\cos\theta & -s\sin\theta & d_x \\ s\sin\theta & s\cos\theta & d_y \\ 0 & 0 & 1 \end{pmatrix} \tag{16}$$

We denote this as the affine warping.

A more general form of warping transformation $W$ is

$$W_p = \begin{pmatrix} s\cos\theta & -s\sin\theta & d_x \\ s\sin\theta & s\cos\theta & d_y \\ \alpha & \beta & 1 \end{pmatrix} \tag{17}$$

with two extra parameters to represent the image shearing deformation caused by out-of-plane rotations. In other words, more freedom of motion is allowed. We found that $W_p$ is a good approximation of the standard homography transformation (8 parameters with 2 extra constraints) of $W$ and often leads to more stable results.

With two out-of-plane rotations added, $W$ has another formulation as follows.

$$W_p = \begin{pmatrix} s_x \cos\theta & -s_x \sin\theta & d_x \\ s_y \sin\theta & s_y \cos\theta & d_y \\ h & -h & 1 \end{pmatrix} \tag{18}$$

Here we replace $s$ with x-scaling parameter $s_x$ and y-scaling parameter $s_y$, and add a shearing parameter $h$ to meet the two extra degree of freedom. Equations (17) and (18) are considered as alternative formulations of the perspective warping we called. In the following analysis, we take (17) as our instance.

For quadratic surfaces, we approximate the warping as:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} w_{11}w_{12}w_{13}w_{14}w_{15}w_{16} \\ w_{21}w_{22}w_{23}w_{24}w_{25}w_{26} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \\ xy \\ x^2 \\ y^2 \end{pmatrix} \tag{19}$$

Let $p$ be the vector of warping parameters (for example, $(d_x, d_y, s, \theta)^T$), $U = (u, v)^T$, and $X = (x, y)^T$, superscripts indicate the number of image points,

$$e = (I_1(U^{(1)}(p)) - I_2(X^{(1)}), \ldots, I_1(U^{(n)}(p)) - I_2(X^{(n)}))^T$$

The key equation for optimization in our framework is to compensate the misalignment error by the increments of warping parameters.

$$-J(p^* - p) = -J\delta p = e \tag{20}$$

The goal of minimizing $||e||$ can be achieved by $p^* = p - (J^T J)^{-1} J^T e$, given an initial close estimate of $p$, where $J = (J^{(1)}, \ldots, J^{(n)})^T$ and

$$J^{(i)} = \frac{\partial I_1}{\partial p} = \frac{\partial I_1}{\partial U^{(i)}} \frac{\partial U^{(i)}}{\partial p} \tag{21}$$

The $\frac{\partial I_1}{\partial U}$ part of the Jacobian matrix $J$ is the image gradient [9], and the $\frac{\partial U}{\partial p}$ part can be derived from $W$. Compared with Equation (11), we do not need to calculate $\mathbf{f}_x(\mathbf{x}; \mu)^{-1}$, because we pre-warp the image to make this item equal to 1. This process simplifies the computation for parameter updating, but the $\delta p$ is only valid locally. The global parameter updating is achieved by the multiplication of warping matrices, not directly through $\delta p$. In the 4 parameter case, for example,

$$u = s\cos\theta x - s\sin\theta y + d_x$$

$$v = s\sin\theta x + s\cos\theta y + d_y$$

[9] The image gradient can be computed by image gradient operators very easily.

$$\frac{\partial U}{\partial p}|_{d_x=0,d_y=0,s=1,\theta=0} = \begin{pmatrix} 1 & 0 & x & -y \\ 0 & 1 & y & x \end{pmatrix} \tag{22}$$

where $p = (d_x, d_y, s, \theta)^T$ is the parameter vector. For 6 or 12 parameters, we can get Jacobian similarly. According to equation 17, we have

$$u = \frac{s\cos\theta x - s\sin\theta y + d_x}{\alpha x + \beta y + 1}$$

$$v = \frac{s\sin\theta x + s\cos\theta y + d_y}{\alpha x + \beta y + 1}$$

$$\frac{\partial U}{\partial p}|_{d_x=0,d_y=0,s=1,\theta=0,\alpha=0,\beta=0} = \begin{pmatrix} 1 & 0 & x & -y & -x^2 & -xy \\ 0 & 1 & y & x & -xy & -y^2 \end{pmatrix}. \tag{23}$$

where $p = (d_x, d_y, s, \theta, \alpha, \beta)^T$ is the parameter vector. In the same way, we the follows for equation 18.

$$u = \frac{s_x\cos\theta x - s_x\sin\theta y + d_x}{hx - hy + 1}$$

$$v = \frac{s_y\sin\theta x + s_y\cos\theta y + d_y}{hx - hy + 1}$$

$$\frac{\partial U}{\partial p}|_{d_x=0,d_y=0,s_x=1,s_y=1,\theta=0,h=0} = \begin{pmatrix} 1 & 0 & x & 0 & -y & -x^2+xy \\ 0 & 1 & 0 & y & x & -xy-y^2 \end{pmatrix}. \tag{24}$$

where $p = (d_x, d_y, s_x, s_y, \theta, h)^T$ is the corresponding parameter vector.

This parameterization describes the out-of-plane rotation as shear deformations in the image coordinates. As an extension from 4, or 6 parameter case, we can also present the higher order deformations (mostly 2nd order) by introducing the quadratic items in the warping equation.

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \mu_{11}+1 & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{15} & \mu_{16} \\ \mu_{21} & \mu_{22}+1 & \mu_{23} & \mu_{24} & \mu_{25} & \mu_{26} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \\ x^2 \\ xy \\ y^2 \end{pmatrix}, \tag{25}$$

From the above equation, we can see 12 parameters are needed to be estimated in the optimization.

$$\frac{\partial U}{\partial p}|_{p=0} = \begin{pmatrix} x & y & 1 & x^2 & xy & y^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x & y & 1 & x^2 & xy & y^2 \end{pmatrix}. \tag{26}$$

where $p = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \mu_{15}, \mu_{16}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}, \mu_{25}, \mu_{26})^T$ is the parameter vector.

Once we have the alignment $W_{n,n+1}$between current successive image frames $n$ and $n+1$, the next task is to warp image $n+1$ to the global mosaic given the coordinates of image $n$. Denote the global alignment of image $n$ to mosaic is a Warp $W_n^M$, so we get the global transformation for image $n+1$ as follows

$$W_{n+1}^M = W_n^M * W_{n,n+1} \tag{27}$$

Note the updating relation in equation 27 is called compositional approach [2]. Quadratic warps do not form a semi-group, so only additive approach is available. To achieve the more precise mosaic, similar compositional optimization are implemented for global alignment based on the initial warp $W_{n+1}^M = W(\Delta p) * W_{n+1}^M$.

### 2.2.3 Pixel Selection

Image-based direct registration methods are usually more computational costly due to the large number of pixel involved. Therefore, we try to only use a portion of the pixels to achieve very close or same performance, compared with using all image pixels for optimization.

Our pixel selection criterion is very intuitive. We try to minimize the sum of squares of each $J^T J$'s element's change caused by not using a pixel. Observe that

$$J^T J = \sum_{i=1}^{n} {J^{(i)}}^T J^{(i)} \tag{28}$$

and since $J^{(i)}$ is a row vector,

$$\sum_k \sum_l ({J^{(i)}}^T J^{(i)})^2_{k,l} = \sum_k \sum_l J^{(i)}_k J^{(i)}_l = (\sum_k J^{(i)}_k)^2 \tag{29}$$

So those $J^{(i)}$s that have the smallest magnitudes will be discarded. Here we consider both the magnitude of image gradients and the parametric relations encoded in the Jacobian matrix.

For every incoming image frame, it is registered with the last frame, which has known warping parameters. And it is then registered with the current global mosaic image to refine the result. both local and global spatial information are considered. Finally it is warped and added to the mosaic by a proper weight.

# 3 Implementation Details

Unlike the mathematical formulation in section 2, we discuss 2 technical aspects of our work on image acquirement interface and image processing library.

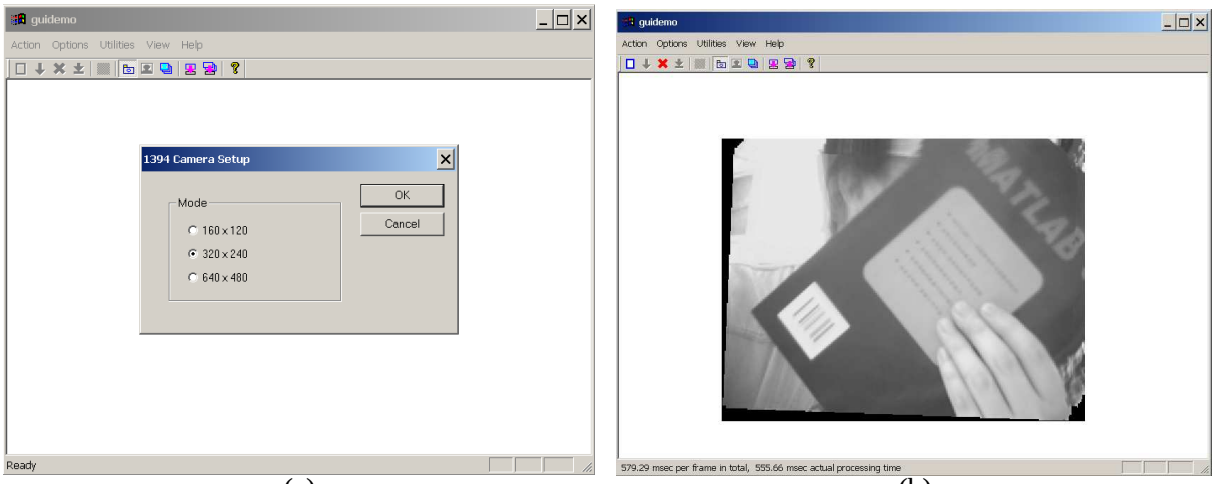
## 3.1 Image Acquirement with IEEE 1394 Fire-wire Interface

In our implementation, the first task is to acquire the live video for mosaicing in real time. Currently, we use IEEE-1394 based digital cameras with the firewire interface. The publicly available IEEE-1394 software driver from Robotics Institute, CMU is employed in our project. In Figure 3, We show an example of image resolution setting and one snapshot of images being processed.

Now we are exploring to control pan-tilt camera to capture video image actively. The pan-tilt head can achieve nearly pure rotation camera motion, which makes our parametric warping assumption more accurate to any environment by rotational homography [22].

## 3.2 Image Processing via Intel IPL Library

With the Image Processing Library (IPL), Intel provides Pentium and MMX-optimized low-level routines for fast image processing. We mainly use three functions for different warping: *iplWarpAffine()* for affine transformation, *iplWarpPerspective()* for perspective transformation (Homography), and *iplRemap()* for arbitrary




(a) (b)

Figure 3: The graphical user interface used for live image acquirement and real time mosaicing: (a) Resolution selection for camera(b) A snapshot of captured image.

geometric transformation, mostly quadratic warping with a pre-computed warping matrix. The image interpolation functions are automatically called from the above geometric transformation functions. We also use ipl library for image filtering and pyramid.

# 4 Experiments and Performance Evaluation

In our mosaicing framework, we introduce different orders of parametric warping model. As shown in Figure 4, Translational transformation between images is taken as the Zero-order warping, and Affine transformation, Quadratic transformation are considered as First-order or Second-order warping model, respectively. Perspective warping is a nonlinear model by reflecting the influence of depth information, which contains different types of distortion with radial Quadratic distortions. For simplicity, the models up to Affine transformation are linear, and below ones are nonlinear.

We first show some snapshots of the image mosaicing process. The image mosaicing process of a projected eye retina image is demonstrated in Figure 5. Quadratic warping is not used because the local retina surface can be well approximated using planar patches. The other experiment shows the image mosaicing result of a complex indoor scene in our laboratory. In theory, the environment is not planar and has multiple layers. We can still get reasonable results with perspective warping model, under moderate motions.

Finally, we analyze the computational time as the measure of performance. In our mathematical formulation for image registration in section 2, we obtain a close-form solution for parameter updates. No iteration is needed. However, due to the high orders of Taylor series in Equation 2, the image noises and limited computing precision, we still employ iterations to search more accurate result. For simplicity, the number of iteration for each frame is fixed as 5 in our experiments. We find that the computing load is a function of time and complexity of the environment.

In our comparison, we perform our algorithm to three different scenarios, under three different resolutions. Besides the retina mosaics shown in Figure 5, we gave some snapshots of mosaics from a 3D indoor environ-

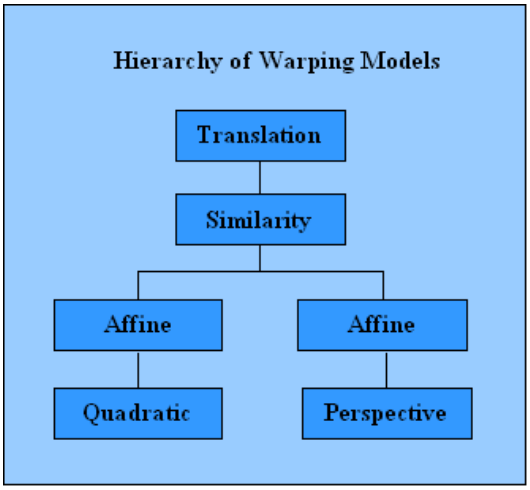



Figure 4: Image mosaicing result for 3D scene image sequence of a laboratory environment.

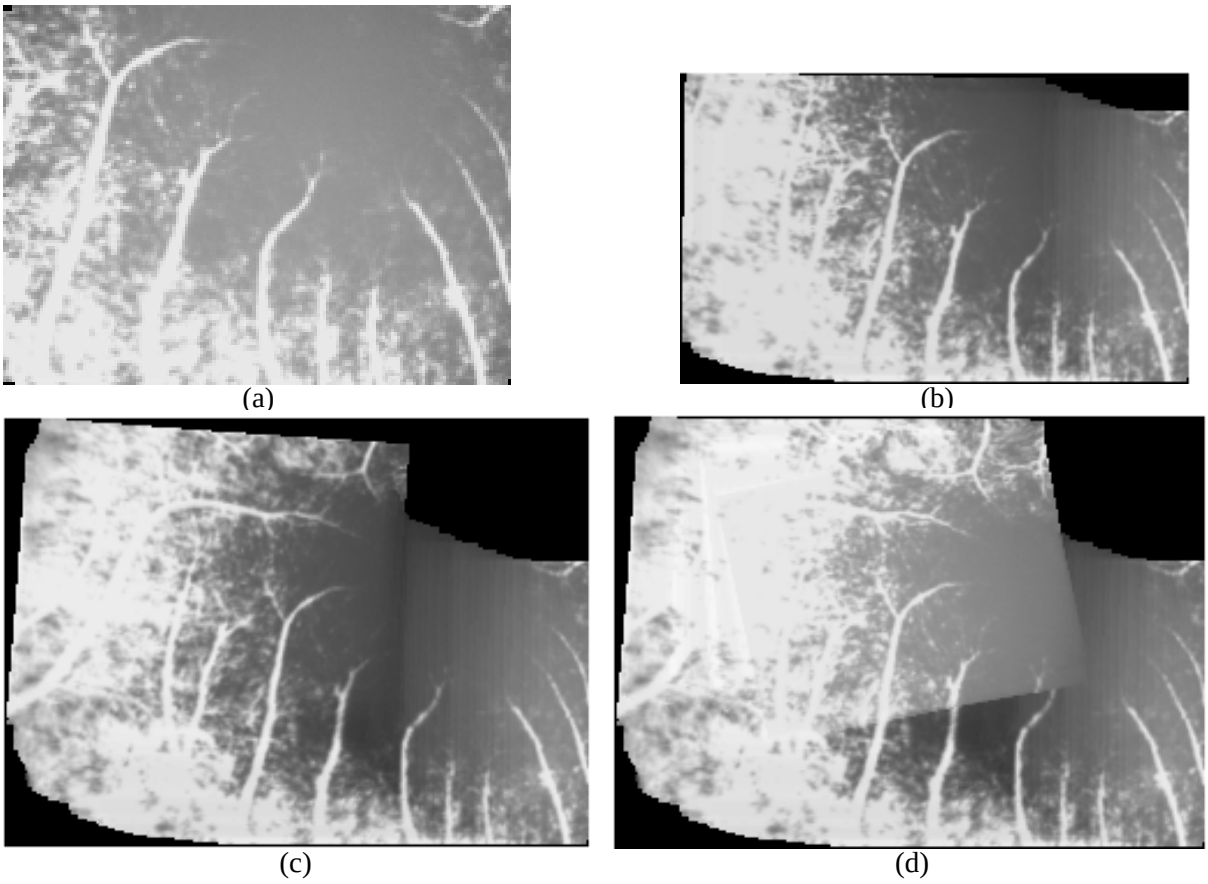
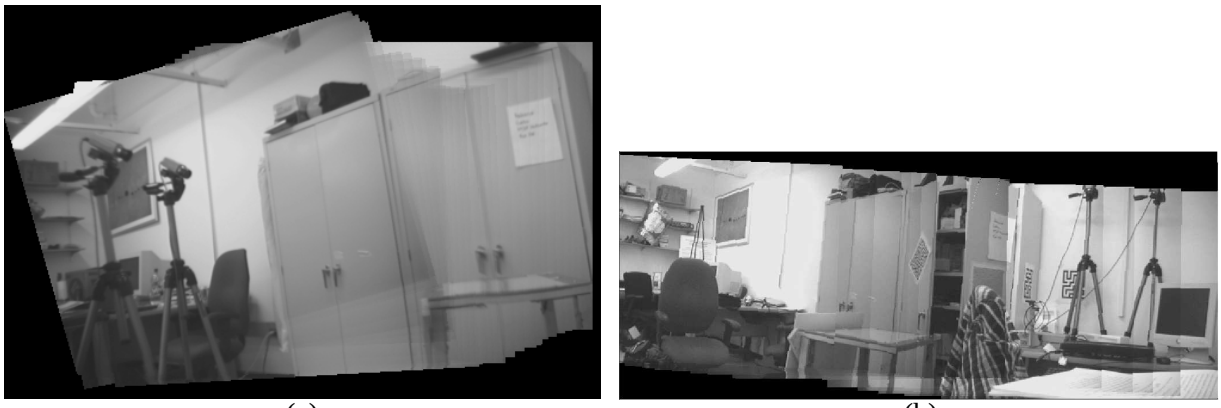



Figure 5: Image mosaicing result for a retina image sequence. (a) Frame 3 (b) Frame 12 (c) Frame 23 (d) Frame 31.



(a) (b)

Figure 6: Image mosaicing result for 3D scene image sequence of a laboratory environment with different warping models. (a) Perspective warping (b) Quadratic warping

(a) (b)

Figure 7: Different scenarios for image registration. (a) 3D indoor environment (b) planar regular pattern.

ment and a planar regular pattern in Figure 7. In theory, the speed of our algorithm depends on the number of pixels filled in the Jacobian matrix which is a function of image resolution, complexity of environment and the camera motion. In Figure 8, we can see the computation time is roughly a linear function with image resolution under three environments. The planar regular pattern normally has more pixels with strong gradient information to be selected, than the indoor scene and retina. The indoor scene roughly has the same order of necessary computations with retina. From Figure 8 (b), the camera motion also has big influence on the calculation load. With more rapid motions, the overlapped portion of successive images are relatively small, so fewer pixels are involved. The oscillated curves in Figure 8 (b) reflect the fast and sudden camera motions. Normally, the computation time does not change much with stable camera motions.

Our algorithm does not work always, so we also investigate when and how it degrades. From our experimental experiences, the failed sequence normally results a wrong and smaller scale factor estimates, so the current images become very small in the global mosaic after warping. Fewer and fewer image pixels are overlapped which make the computation time curve going down in process of time. An example is illustrated in Figure 9.

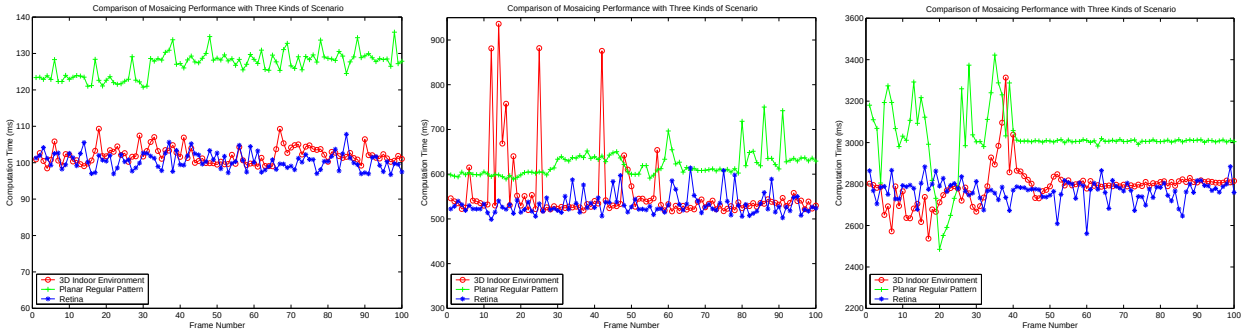



Figure 8: Performance comparison of image registration under different image resolution. (a) $160 \times 120$ (b) $320 \times 240$ (c) $640 \times 480$.

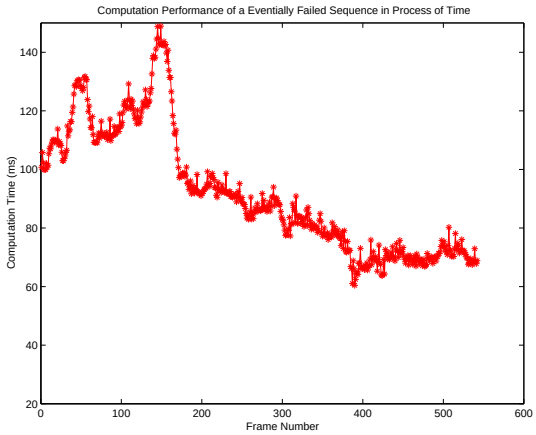



Figure 9: The computation time curve of registration for a gradually degraded video sequence.

# 5 Summary and future work

Our work demonstrated a viable method for real-time image mosaicing by directly minimizing SSD. The method is suitable for a planar or quadratic surface with a free moving camera, or any scene with a rotation-only camera motion. Future work will include a method to detect accumulated errors and dynamically re-distributed them, a better warping model for quadratic surface and other parametric or non-parametric constrained surfaces, and a layered or masked model when motion parallax is significant.

By directly using image intensity patches as alignment criterion, we plan to fuse some stable one-order edge-like feature for registration. Chamfer distance [46] and distance transformation [5] can be applied into our problem solving framework. Combining the SSD information, a pixel weighted or reweighted least squares algorithm can improve our rough pixel selection strategy. On the other hand, we also plan to explore the video frame selection problem. If we take the resulting mosaic image as our final task, not all the frames in the video sequence are necessary, or even have some bad influence on the mosaic registration. We are researching to find a frame selection criterion and algorithm that can be used to eliminate the information redundant frames and difficult frames for registration (We call them ghosting frames) , like frames with suddenly big distortions. We hope both the computational efficiency and image quality can be improved, because we do not try to register

every image frame, especially for ghosting frames. A filtering method on the motion parameter trajectories is under exploring.

In our computer aided surgery application, we use a micro medical robot with camera to get narrow view live video images from the retina surface of the patient. similar with the aero instrument of video camera in the surveillance plane [26], we can read motion parameter values from the medical robot. How to fuse the sensor information with the direct measurement from the image will be our next task. We may have a reference image for a global view of retina, which can also be used to locate the current narrow view live video images with higher resolution. With feature based representation for a long sequence, we can obtain the 3D information via factorization method [41] that will be helpful for 3D surface reconstruction and more complex form image registration. Appearance based layer extraction method [44] are also available for 2.5D vision process, which can decompose the scene by multiple layers for perspective warping. Our parametric warping model assumption will be more practical for multiple layers to approximate the real 3D scene.

Our image registration framework is a direct method to get the close-form solution for a given problem. We plan to analyze the probabilistic aspects about the parameter estimation in this work. The probabilistic modelling can help us the inference the uncertainty of image alignment and find a method to handle the temporal accumulated misalignment error. Our method is efficient for small local distortions. For large frame-in-frame appearance transformation, our method can not guarantee the convergence, even with multiple iterations. Object recognition method [3] can be used for scale, rotation-invariant or affine-invariant automatic initialization or sparse images registration.

# References


[1] P. Anandan. A Computational Framework and An Algorithm for the Measurement of Structure from Motion, *Int. Journal of Computer Vision*, **2:3** pp. 283-310, 1989.

[2] S. Baker and I. Matthews. Equivalence and Efficiency of Image Alignment Algorithms, *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Huwaii, USA, Dec. 2001.

[3] Serge Belongie, Jitendra Malik and Jan Puzicha Shape Matching and Object Recognition Using Shape Contexts *IEEE Trans. Pattern Analysis and Machine Intelligence* **24(4)**:509-522, April 2002.

[4] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation, *Europen Conf. on Computer Vision*, pp. 237-252, 1992.

[5] G. Borgefords. Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **10:6**, Nov. 1988.

[6] L. Brown. A Survey of Image Registration Techniques, *ACM Computing Surveys*, **24:4** pp. 325-376, 1992.

[7] L. Brown. 3D Head Tracking Using Motion Adaptive Texture-Mapping *IEEE Conference in Computer Vision and Pattern Recognition*, Kauai, Hawaii, vol. no.1 pp.998-1005, December 8-14, 2001.

[8] A. Can, C.V. Stewart, B. Roysam, and H.L. Tanenbaum. A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina, *IEEE Trans. on PAMI*, **24:3** pp. 347-364, Mar. 2002.

[9] D.P. Capel. Image Mosaicing and Super-resolution, *PhD thesis*, Oxford University, 2001.

[10] Y. Caspi and M. Irani. Alignment of Non-Overlapping Sequences, *International Journal. of Computer Vision*, **48:1**, pp.39-51, June 2002.

[11] J.-X. Chai and H.-Y. Shum. Parallel Projections for Stereo Reconstruction, *IEEE Conf. Computer Vision Pattern Recognition*, Vol **2**, pp.493-500, Hilton Head Island, SC, USA, Jun. 2000.

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*, John Wiley and Sons, 1991.

[13] X.-T. Dai. Entropy Tracker and Comparison, *Technical Report 2002-XX-CIRL-CS-JHU*, Computer Science Dept. Johns Hopkins Univ. May 2002.

[14] F. Dellaert and R. Collins. Fast Image-Based Tracking by Selective Pixel Integration, *ICCV 99 Workshop on Frame-Rate Vision*, Sep. 1999.

[15] R. Duta and P. Hart. *Pattern Classification and Scene Analysis*, John Wiley and Sons, 2001.

[16] O. Faugeras. *Three-dimensional Computer Vision : a Geometric Viewpoint* MIT Press, Cambridge, Mass. 1993.

[17] G.H. Golub and C.F. Van Loan. *Matrix computations*, Second Edition, Johns Hopkins University Publisher, 1993.

[18] G. Hager and K. Toyama. X Vision: A Portable Substrate for Real-Time Vision Applications, *Computer Vision and Image Understanding*, **69:1**, pp.23-37, Jan. 1998.

[19] G. Hager and P. Belhumeur. Efficient Region Tracking With Parametric Models of Geometry and Illumination, *IEEE Trans. on PAMI*, **20:10**, pp. 1125-1139, 1998.

[20] G. Hager. A Tutorial on Vision-Based Interaction and Control, *Tutorial on AAAI'00*, Austin, Texas, USA, 2000.

[21] C. Harris and M. Stephens. A Combined Corner and Edge Detector, *Fourth Alvey Vision Conference*, pp.147-151, 1988.

[22] R. Hartley. Self-Calibration Of Stationary Cameras, em International Journal of Computer Vision, **22**, pp.5-23, 1997.

[23] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision* Cambridge University Press, June 2000.

[24] D. Hirvonen, B. Matei, R. Wildes and S. Hsu. Video to Reference Image Alignment in the Presence of Sparse Features and Appearance Change, *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Huwaii, USA, Dec. 2001.

[25] B.K.P. Horn and B.G. Schunck. Determining Optical Flow, *Artificial Intelligence Journal*, pp. 185-203, Vol. 16, No. 1-3, Aug. 1981.

[26] S. Hsu. Geocoded Terrestrial Mosaics Using Pose Sensors and Video Registration, *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Huwaii, USA, Dec. 2001.

[27] M. Irani and P. Anandan. Video Indexing Based on Mosaic Representations, *Proceedings of the IEEE*, **86:5**, May 1998.

[28] M. Irani and P. Anandan. About Direct Method, *Vision Algorithms: Theory and Practice*, pp.267-277, Corfu, Greece, 1999.

[29] C. Kuglin and D. Hines. The Phase Correlation Image alignment Method. *IEEE Int. Conf. on Cybernetics and Society*, 1975.

[30] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Robust Registration of Texture-Mapped 3D Models. *IEEE Trans. PAMI*, **22:4**, Apr. 2000.

[31] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. Proc. of IJCAI, pp.674-679, 1981.

[32] L. McMillan and G. Bishop, Plenoptic Modeling: An Image-Based Rendering System, *Proceedings of SIGGRAPH 95*, pp. 39-46, Los Angeles, CA, Aug. 1995.

[33] S. Peleg, B. Rousso, A. Rav-Acha and A. Zomet. Mosaicing on Adaptive Manifolds, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **22:10**, pp.1144-1154, Oct. 2000.

[34] F. Preparata and M.I. Shamos *Computational Geometry, An Introduction*, 1985.

[35] Maxime Lhuillier and Long Quan. Image Interpolation by Joint View Triangulation, *CVPR'99*, pp. 139-145, Fort Collins, Colorado, USA, 1999.

[36] M. Lhuilier, L. Quan, H.-Y. Shum and H.-T. Tsui. Relief Mosaics by Joint View Triangulation, *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Huwaii, USA, Dec. 2001.

[37] O. Sadowsky, Y. Zhan and X. Wu. Tracking Surface Deformation for Image Registration in Retinal Microsurgery, Technical Report, Computer Science Dept. JHU, May 2002.

[38] Jianbo Shi and Carlo Tomasi. Good Features to Track, *IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[39] H.-Y. Shum and L.-W. He. Rendering with Concentric Mosaics, *Computer Graphics Proceedings, Annual Conference Series, Proc. SIGGRAPH'99*, pp.299-306, Los Angeles, CA. USA, August 1999.

[40] H.-Y. Shum and R. Szeliski. Construction of Panoramic Image Mosaics with Global and Local Alignment, *International Journal of Computer Vision* **36(2)**, pp. 101-130, 2000.

[41] C. Tomasi and T. Kanade. Shape And Motion From Image Streams Under Orthography: A Factorization Method, *International Journal of Computer Vision*, **(9)**, pp. 137-154, 1992.

[42] P. Torr and A. Zisserman. Feature Based Methods for Structure and Motion Estimation, *Vision Algorithms: Theory and Practice*, pp. 278-294, Corfu, Greece, 1999.

[43] P. Torr, A. Zisserman. MLESAC: a new robust estimator with application to estimating image geometry, *Computer Vision and Image Understanding*, **78:1**, pp.138-156, Apr. 2000.

[44] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated Bayesian approach to layer extraction from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23(3)**:297-303, March 2001.

[45] P. Torr and C. Davidson. IMPSAC: Synthesis of Importance Sampling and Random Sample Consensus, *IEEE Trans. PAMI*, **25:3**, Mar. 2003.

[46] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, **48(1)**, pp.9-19, 2002.

[47] P. Viola and W. Wells III. Alignment by Maximization of Mutual Information, ICCV 95, pp.16-23, Cambridge, Mass. USA, 1995.

[48] R. Wildes, D. Hirvonen, S. Hsu, R. Kumar, W. Lehman, B. Matei and W.-Y. Zhao. Video Georegistration: Algorithm and Quantitative Evaluation, *Int. Conf. on Computer Vision*, pp.343-350, Vancouver, Canade, 2001.

[49] J. Xiao, T. Kanade, and J. Cohn Robust Full Motion Recovery of Head by Dynamic Templates and Re-registration Techniques, *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FG'02)*, Washington DC, USA, May 2002.

[50] Z. Zhang, R. Deriche, O. Faugeras and Q.-T. Luong. A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry, *Artificial Intelligence Journal*, Vol.78, pp.87-119, Oct. 1995.