Stochastic Optimization of Multiple Texture Registration Using Mutual Information

Ioan Cleju[1] and Dietmar Saupe[1]

University of Konstanz, Multimedia Signal Processing Group, Fach M697, 78457 Konstanz, Germany

Abstract. We consider the problem of simultaneously registering several images to a 3D model. We propose a global approach based on mutual information that extends previous methods to incorporate the color, and does not require segmentation or feature extraction. We give a stochastic model for joint optimization of multiple image-to-model alignment and we propose a heuristic to solve it. Experiments with synthetic models showed that our algorithm is robust to varying illumination and surface characteristics. Experiments with real data showed that we can achieve very good accuracy even for an object with highly specular surface, in moderate lighting conditions.

1 Introduction

A common framework for creating textured 3D models consists in two steps: firstly the geometry is built, and secondly the texture is mapped from photographs. The texture registration step searches for the projective transformation between the 3D model and the 2D images by solving the camera calibration problem. The parameters that define the projective transformation correspond to the parameters of the camera that acquired the image.

Classical closed-form and iterative numerical solutions for the camera calibration problem use point-feature pair correspondences [4]. Other methods involve more complex features such as silhouettes [6] and lines [2]. If several images are available, it is possible to use 2D-2D pair-features from images to improve the accuracy of the registration [6], [8].

Intensity-based registration techniques rely on global measures such as photo-consistency and mutual information, avoiding feature extraction. Viola and Wells use the mutual information between the normals to the surface and the intensity image to align a 3D model to the image [11]. Several images acquired by a system of cameras with known relative poses were registered to a 3D model using image-model mutual information [7] and photo-consistency [1]. The photo-consistency registration criterion is based on the assumption that any point on a surface with ideal Lambertian reflectance appears with the same color in all images where it is visible. In [5], photo-consistency was used to register two images with unknown relative pose to a 3D model.

We consider the problem of registering several images with unknown relative poses to a 3D model. Our solution extends the intensity-based approaches

from [7], [1], and [5], considering both image-model and image-image mutual information as registration criteria. The contributions of our work are:

- we extend the mutual information registration method to several uncalibrated cameras;
- we propose a stochastic optimization model for the joint registration of several images to a 3D model;
- we show in experiments the advantages of our approach for complex illumination and surface characteristics;
- we experimentally confirm good accuracy of mutual information for texture registration even for a model with a highly specular surface.

In Section 2 we shortly review the texture registration by maximization of mutual information and we introduce the extension to consider color. In Section 3 we present a model for stochastic joint optimization and we give a heuristic solution. In Section 4 we show and discuss the experimental results, and in Section 5 we draw the conclusions and give some outlines for future work.

2 Texture Registration by Maximization of Mutual Information

Let x be a point on the surface of the 3D model that is visible in the texture image and T the 3D-2D projective transformation. Let $u(x)$ be the normal to the surface in x and $v(T(x))$ the intensity value in the image. The value of $v(T(x))$ is given by the rendering equation and depends on the radiance in the scene, the BRDF of the surface in x and the normal to the surface $u(x)$. The goal of the texture registration is to find the transformation T. Since the BRDF and the radiance are not known, Viola and Wells propose to directly exploit the relation between $u(x)$ and $v(T(x))$ by means of mutual information (MI) [11]. A random variable x on the 3D model that is visible in the image allows defining the random variables 'normal' $u(x)$, 'intensity' $v(T(x))$, and 'normal-intensity' $(u(x), v(T(x)))$. From their entropies we can define the MI between the normals to the surface and the intensity image (equations (1)). The MI between $u(x)$ and $v(T(x))$ is maximized when T aligns the model to the image.

Viola and Wells propose a gradient-based search for the optimal transformation T and a fast method to estimate the gradient of the MI with respect to T. If we consider a random variable y, its entropy $h(y)$ can be estimated from two independent samplings of y. One sampling is used to estimate the probability density function with the Parzen window method [3], which is then evaluated on the second sampling. The complexity of the method is quadratic in the size of the samplings. The MI between $u(x)$ and $v(T(x))$ is estimated from small subsamplings of the data (order of tens of points). When defined in this way, the MI can be differentiated with respect to T.

$$\begin{aligned} I(u(x), v(T(x))) &= h(u(x)) + h(v(T(x))) - h(u(x), v(T(x))) \\ \tfrac{d}{dT} I(u(x), v(T(x))) &= \tfrac{d}{dT} h(v(T(x))) - \tfrac{d}{dT} h(u(x), v(T(x))) \end{aligned} \tag{1}$$

Due to the small random subsampling of data, the estimation of the MI gradient is stochastic. Viola and Wells use stochastic gradient descent as the optimization procedure. The subsamplings are changed at each iteration and T is updated in the direction of the gradient. Local maxima of MI can be avoided due to the inherent noise of the gradient.

In [9], a 3D model with reflectance values mapped on its surface was registered to color images using the method of Viola and Wells (the reflectance values were obtained during 3D scanning). The MI has become popular especially in medical image registration.

We extended Viola and Wells' algorithm to register several textures on a 3D model by considering images that contain common patches of the model. If a patch of the surface is visible in two images, we will simply say that the images overlap.

We define image-image MI functions for each overlap. Given two overlapping images i and j with corresponding projective transformations T_i and T_j, let x be a random point on the surface visible in both images. The MI between the colors $v_i(T_i(x))$ and $v_j(T_j(x))$ of the images i and j is then:

$$I(v_i(T_i(x)), v_j(T_j(x))) = h(v_i(T_i(x))) + h(v_j(T_j(x))) - h(v_i(T_i(x)), v_j(T_j(x))) \tag{2}$$

This extension adds the full color information of the images to the registration objective functions. In our implementation we defined the image-image MI from the chrominance components I and Q of the YIQ color space. The image-image MI is parameterized by the projective transformations associated with both images, and it is maximized when both images are aligned to the model. The gradient estimation follows the same procedure as for the image-model MI.

Compared to other registration criteria, the MI does not need the existence of any 3D-2D feature, including visible outlines in the image, and does not make assumptions on the unknown parameters of the rendering function. It is robust to illumination conditions and even to occlusions [11]. One problem when using the MI objective function is that the value of the global maximum cannot be estimated. In contrast, when registration is done with point correspondences, for instance, the global optimum corresponds to 0 projection error.

2.1 Camera Model

The optimization model does not make assumptions on the projective transformation T, and consequently on the camera model. In our implementation we considered the pinhole camera model [4] with four distortion coefficients (two for radial distortion and two for tangential distortion). The intrinsic parameters field-of-view, optical center, and distortions, were calibrated using Zhang's method [12]. Any of the intrinsic parameters can be further optimized using the mutual information objective functions. We considered the intrinsic parameters fixed and we optimized only the extrinsic parameters. The rotation matrix was parameterized by axis-angle form for its advantages over Euler angles in the iterative optimization [10].

3 A Stochastic Optimization Model for Global Texture Registration

For the joint registration of several images to a 3D model we formulated image-model and image-image MI objective functions. When all images are aligned to the model, all objective functions are maximized. If only the image-model MI functions are considered, each set of camera parameters corresponds to one objective function. In this case the iterative gradient-based optimization updates each set of parameters in the direction of the corresponding gradient. When also image-image MI functions are considered, we estimate several gradients for the parameters of each camera, corresponding to the MI with the model and with other overlapping images. In each iteration we must choose the direction for optimization based on the these gradients. In this section we motivate and discuss the fusion of the gradients for the update direction.

Let n be the number of cameras (images), t_i the approximate parameters for camera i, t_i^* the optimal parameters for camera i and δ_i the error.

$$t_i = t_i^* + \delta_i \quad \text{for} \quad i = 1, \ldots, n \tag{3}$$

Let $g_{i,0}$ be the estimated gradient of the MI between the image i and the model, $g_{i,j}$ be the estimated gradient of the MI between the image i and the (overlapping) image j and $g_{i,j}^*$ the true value of the gradient. Let $\epsilon_{i,j}$ be the error introduced in the estimation of the gradient by the data subsampling. Let $\text{overlap}(i, 0)$ state the existence of the overlap between texture i and the model (it is always true), and, for $j \neq 0$, $\text{overlap}(i, j)$ the existence of the overlap between the textures i and j, then:

$$g_{i,j} = g_{i,j}^* + \epsilon_{i,j}, \quad \text{for} \quad i = 1, \ldots, n, \quad j = 0, \ldots, n, \quad \text{if overlap}(i, j) = true \tag{4}$$

Let us consider the objective functions corresponding to the image i. Since all of them are maximized for the correct alignment, any linear combination of these functions with positive weights has the global optimum for the same camera parameters. On the other hand, we expect the other local optima to be less related. For example, the image-model MI objective function relies on normals and intensities in comparison to image-image MI objective functions that are based on colors.

Therefore, we think that, in general, a linear combination of the objective functions has a more emphasized global optimum and faded local optima (that are not global) than any of the individual functions. Since the gradient of the summed objective functions is the sum of the gradients, we look for an update direction as a linear combination of individual gradients.

We observe another positive effect of this formulation. The gradients of the objective functions are perturbed by the estimation errors $\epsilon_{i,j}$ as artifacts of subsampling. Since the subsamplings are independent, the errors are independent, and their effect is not increased after the linear combination.

The update directions have the form:

$$g_i = \sum_{j=0,\text{overlap}(i,j)}^{n} w_{i,j} g_{i,j} \quad \text{for} \quad i = 1, \ldots, n, \quad j = 0, \ldots, n \tag{5}$$

The canonical approach is to assign equal weights for each gradient. Following similar intuitive reasoning as above, an even better update direction may be estimated if the weights of the gradients are correlated with the probability that the current estimation lies in the region of attraction of the global optimum of the corresponding objective function. In the following subsection we give a heuristic for this problem.

3.1 Weighted Gradient Fusion

Before defining the weights from equation (5), we introduce some additional variables. The gradient of MI is 'consistent' if its direction does not change considerably for consecutive iterations. For the gradient of MI between items i and j at iteration k, we define the 'instantaneous consistency' $c_k(i,j)$ and the 'consistency' $C_k(i,j)$:

$$\begin{array}{ll} c_k(i,j) = \frac{1}{2}(\cos(g_{k-1,i,j}, g_{k,i,j}) + 1) & \text{for} \quad k \geq 1 \\ C_k(i,j) = (1-\alpha) C_{k-1}(i,j) + \alpha c_k(i,j) & \text{for} \quad k \geq 1, \quad 0 \leq \alpha \leq 1 \end{array} \tag{6}$$

We start with initial values zero for $c_0(i,j)$ and $C_0(i,j)$. $g_{k,i,j}$ is the estimation $g_{i,j}$ at iteration k. $C_k(i,j)$ measures the consistency over a sequence of iterations, where the most recent instantaneous consistencies have a larger weight. A gradient has low consistency if the MI function has poor convexity, if $\epsilon_{i,j}$ is considerable in (4), or if an optimum is already attained. In the implementation α was set to 0.05.

We introduce the 'alignment' variable $A_k(i,j)$ to measure the alignment between the items i and j. The alignment at iteration k is estimated as the maximum value of consistency for that gradient:

$$A_k(i,j) = \max_{l=1,\ldots,k} C_l(i,j) \tag{7}$$

A large value of $A_k(i,j)$ does not mean that i and j are aligned at iteration k, but rather it indicates that the parameter estimation lies on the region of attraction of a pronounced local optimum. We can estimate that the texture i attained a pronounced local optimum after k iterations if the consistency $C_k(i,0)$ is small but the alignment $A_k(i,0)$ large.

We may now define the weights for equation (5):

$$\begin{array}{lll} w_{i,j} = C(i,j) A(j,0) (1 - \frac{C(j,0)}{A(j,0)}), & \text{for} & i = 1, \ldots, n, \quad j = 1, \ldots, n \\ w_{i,0} = C(i,0), & \text{for} & i = 1, \ldots, n \end{array} \tag{8}$$

For simplicity, we omitted the iteration number k in (8). The weights are recomputed in each iteration. From equations (6) and (7) it follows that $w_{i,j}$ are

between 0 and 1. The weight $w_{i,j}$ $(j > 0)$ is large when the alignment between the texture j and the model is large but the consistency is small (image j and the model are possibly aligned), and when the gradient between i and j is consistent.

4 Implementation Issues and Experimental Validation

We implemented our texture registration method on a point-based framework [13]. We present results for optimization of the extrinsic parameters of the cameras. The probabilities used in the stochastic framework were estimated using 6-dimensional gradients. We used different updating step sizes for rotation and translation. In each iteration, the updating direction was normalized separately for rotation/translation, and we made steps in the updating direction. We are currently working on implementing optimization with adaptive step sizes. For each experiment presented in Fig. 2, the optimization consisted in 2000 iterations.

We compared three registration algorithms. The first algorithm uses only image-model MI (Viola_TR_MI), the second uses equal weights of the gradients in equation (5) (Canonical_TR_MI), and the third algorithm uses adaptive weights computed using equation (8) (Weighted_TR_MI). For a fair comparison, we restricted the computation for the estimation of the gradients (the number of kernel estimations in the Parzen window method) to obtain the same computation effort for all methods. In all experiments we used three texture images, all overlapping. Correspondingly, there are 3 times more MI objective functions for Canonical_TR_MI and Weighted_TR_MI than for Viola_TR_MI. We used subsampling sizes of 100 points for Viola_TR_MI and of 58 for the other two, thus having roughly the same number of kernel estimations in the Parzen window method (the complexity is quadratic in subsampling sizes). Almost all the runtime was spent for gradient estimation. The corresponding speed was roughly 100 iterations per second for each texture, on a Pentium 4 at 3 GHz. From time to time full z-buffer projections have to be done (once for hundreds of iterations). Disregarding cache-memory issues and full z-buffer projections, the running time is independent on the size of the model.

We evaluated the accuracy of the registration using ground truth. For the real case, we estimated the extrinsic camera parameters 5 times, each time with 20-25 point correspondences chosen interactively, and we averaged the parameters. Two types of errors can be defined: parameter error (matrix distance between optimized parameters and ground truth) and projection error. The projection error is the root mean squared distance from the projections of all points of the 3D model on the image plane using optimized parameters to the projections obtained by true parameters [5]. We averaged the errors for the images.

The characteristics of the experiments are summarized in the table. Images of the models are shown in Fig. 1. We randomly perturbed the ground truth (rotation and translation) to simulate initial inexact parameters, and ran the optimization algorithms.

Details	**Experiment Index**					
	1	2	3	4	5	6
Model Name	Square1	Square2	Square3	Square4	Trilobite	Shakyamuni
Model Type	Synthetic				Real	
No. Points	40 000				726 027	1 693 444
Reflectance	Isotropic		Anisotropic			-
Surface Type	Diffuse and Specular				Diffuse	-
Light	Directional	Directional and Ambient			Directional	Light Tube
Photographs	3 Rendered					3 Real
Resolution	800×600				1062×864	2048×1536
Overlap	Total			Partial		
Ground Truth	Known					Estimated

For experiments 1-4 we used planar square models with normals, colors, diffuse and specular surface reflectance coefficients introduced as a mixture of Gaussians (e.g., the normals were bump mapped) with the same distribution of the variances. The first three models have the surface properties distributed all over the surface. One half of the fourth model has perturbed normals and white color; the other half constant normal and varying color (Fig. 1). For the first three experiments we used images with the whole models, and for the fourth experiment each image covers about 60% of the model, containing the two different halves in different ratios.

In experiment 5 we used the Trilobite model from Arius3D (www.arius3d.com). The model was acquired with a special high resolution scanner capable of sampling the color, and we simulated the photographs by rendering.

The last experiment was performed with the Shakyamuni model from the University of Konstanz (www.inf.uni-konstanz.de/cgip/projects/surfac). The photographs were taken in a room illuminated by one light tube (about 2 meters in length), placed at the ceiling at about 5 meters from the model. We show results for registration of three images taken from front, front-left and front-right. The images were used at full resolution, without any preprocessing. For visual assessment of the registration, we drew two vertical and two horizontal lines on the model (see Fig. 1, bottom left) and we show close views with the model textured with initial and optimized parameters.

We show the results in Fig. 2 plotting the average projection error in pixels over all texture images versus the initial projection error before optimization. We observed a decrease of accuracy when ambient light (experiment 2) and varying reflectance coefficients (experiment 3) were introduced. The accuracy of Canonical_TR_MI and Weighted_TR_MI was similar, considerably higher than Viola_TR_MI.

In the experiment 4, Weighted_TR_MI gave the best accuracy for all runs, performing considerably better than Canonical_TR_MI. For the settings of this experiment, the global optima of some objective functions were pronounced, while other objective functions had many local optima. In particular, the image-

Fig. 1. Models used for evaluation. Top row: Square4 (left), Trilobite (middle), and Shakyamuni (right). Bottom row: part of Shakyamuni photo (left), remark the hand drawn lines; renderings with initial registration (middle) and optimized registration (right) of three photos.

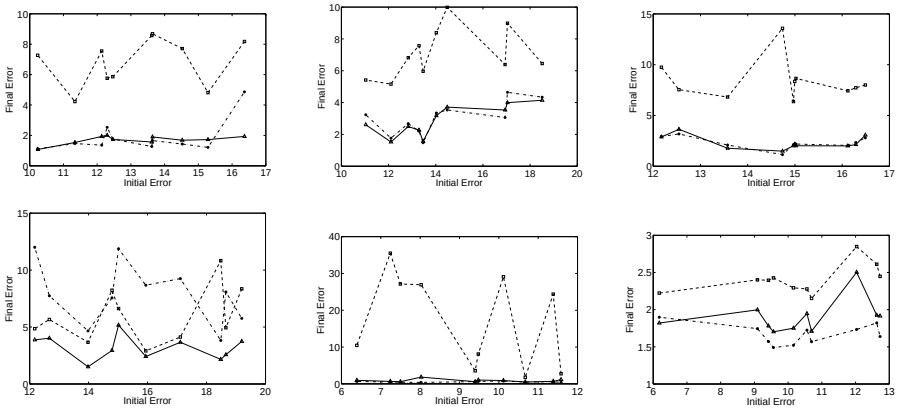

Fig. 2. Comparison between Viola_TR_MI (dashed), Canonical_TR_MI (dashdot), and Weighted_TR_MI (solid) for experiments 1 (upper left) to 6 (bottom right), for several initial parameters. The plots show initial projection error (horizontal axis) versus final projection error (vertical axis) in pixel units. The errors are averaged for all images.

model MI was a good objective function for only one of the images. Our heuristic identified and assigned higher weights to the gradients of the functions with pronounced global optima.

For the Trilobite model, in 6 out of 10 runs of Viola_TR_MI the error increased. The other two algorithms achieved the best results out of our experiments: the projection errors of Canonical_TR_MI were under 0.8 pixels, and the largest error of Weighted_TR_MI was 1.8 pixels, while the other 9 errors were under 1.2 pixels. We see two causes for the better accuracy of Canonical_TR_MI and Weighted_TR_MI for the Trilobite model as compared to the synthetic models. Firstly, specular coefficients associated to the surface of the synthetic models allowed modelling of more complex BRDF. Secondly, the distribution of the surface properties on the synthetic models determined MI objective functions with many local maxima.

The most visible features in the images with the Shakyamuni model are the specular highlights, and this may pose serious problems for other registration algorithms. The compared methods gave very good results, with slightly worse performance of Viola_TR_MI. We were even able to register images from an initial average error of about 60 pixels. The visual difference between renderings with estimated ground truth parameters and optimized parameters cannot be perceived.

It is difficult to present direct comparisons to other registration algorithms. Our accuracy is significantly better than the one reported in [5], mean projection error of 5-6 pixels for 512×512 resolution images, and also much faster. Involving only mutual information, the algorithm is conceptually simpler than [8], and does not require the texture image to contain the entire object for the purpose of silhouette extraction as in [8] and [6].

5 Conclusions and Future Work

We motivated and proved experimentally the advantages of joint registration of several images to a 3D model using the mutual information. By considering the image-image mutual information, introduced in this paper, we improved significantly the registration accuracy in all experiments. Our heuristic for the weighted gradient fusion clearly outperformed the canonical approach in only one experiment. We are looking for other heuristics, for instance choosing the update direction using a voting scheme among gradients. Similar methods may be used to adjust the computational effort, e.g., by allocating more run-time to estimate the gradients of the relevant objective functions.

We will study the performance of our algorithm for optimization of the intrinsic camera parameters. After obtaining a good estimation for extrinsic camera parameters with the method described above, a similar refining optimization step should consider all camera parameters. We want to improve the optimization by adaptive step sizes in the gradient descent, using for instance the consistency of the gradient (defined in this paper). This can be combined with a multi-resolution approach proposed in other registration methods. Finally, we will

complete the texture mapping framework by implementing the texture fusion for 3D point based models.

6 Acknowledgments

The work was supported by the DFG GK 1042 'Explorative Analysis and Visualization of Large Information Spaces' at the University of Konstanz, Germany. The Trilobite model was provided by Arius3D www.arius3d.com.

References

1. M. J. Clarkson, D. Rueckert, D. L. G. Hill, and D. J. Hawkes. Using photo-consistency to register 2D optical images of the human face to a 3D surface model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(11):1266–1280, 2001.
2. P. David, D. DeMenthon, R. Duraiswami, and H. Samet. Simultaneous pose and correspondence determination using line features. In *Proc. Computer Vision and Pattern Recognition*, pages 424–431. IEEE Computer Society, 2003.
3. R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
4. O. Faugeras. *Three-Dimensional Computer Vision A Geometric Viewpoint*. The MIT Press, 1993.
5. Z. Jank and D. Chetverikov. Photo-consistency based registration of an uncalibrated image pair to a 3D surface model using genetic algorithm. In *Second Int. Symposium on 3D Data Processing, Visualization and Transmission (3DPVT'04)*, pages 616–622, 2004.
6. H. P. A. Lensch, W. Heidrich, and H.-P. Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63(4):245–262, 2001.
7. M. Leventon, W. M. Wells III, and W. E. L. Grimson. Multiple view 2D-3D mutual information registration. In *Image Understanding Workshop*, 1997.
8. P. J. Neugebauer and K. Klein. Texturing 3D models of real world objects from multiple unregistered photographic views. *Computer Graphics Forum*, 3(18):245–256, 1999.
9. K. Nishino, Y. Sato, and K. Ikeuchi. Appearance compression and synthesis based on 3D model for mixed reality. In *Proc. IEEE Int. Conf. Computer Vision*, pages 38–45, 1999.
10. C. J. Taylor and D. J. Kriegman. Minimization on the Lie group SO(3) and related manifolds. Technical report, Center for Systems Science, Dept. of Electrical Engineering Yale University, 1994.
11. P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *Int. J. Computer Vision*, 24(2):137–154, 1997.
12. Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proc. Int. Conf. Computer Vision*, volume 1, pages 666–673, 1999.
13. M. Zwicker, H. Pfister, J. van Baar, and M. H. Gross. Surface splatting. In *Proc. SIGGRAPH*, pages 371–378, 2001.