Robust Registration and Tracking Using Kernel Density Correlation

Maneesh Singh Himanshu Arora Narendra Ahuja

ECE Department, University of Illinois, Urbana-Champaign, IL 61801
{msingh,harora1,n-ahuja}@uiuc.edu

Abstract

Challenges to accurate registration come from three factors —presence of background clutter, occlusion of the pattern being registered and changes in feature values across images. To address these concerns, we propose a robust probabilistic estimation approach predicated on representations of the object model and the target image using a kernel density estimate. These representations are then matched in the space of density functions using a correlation measure, termed the Kernel Density Correlation (KDC) measure. A popular metric which has been widely used by previous image registration approaches is the Mutual Information (MI) metric. We compare the proposed KDC metric with the MI metric to highlight its better robustness to occlusions and random background clutter—this is a consequence of the fact that the KDC measure forms a re-descending M-estimator. Another advantage of the proposed metric is that the registration problem can be efficiently solved using a variational optimization algorithm. We show that this algorithm is an iteratively reweighted least squares (IRLS) algorithm and prove its convergence properties. The efficacy of the proposed algorithm is demonstrated by its application on standard stereo registration data-sets and real tracking sequences.

1. Introduction

Image registration is the process of establishing correspondence between a given object in one image and its transformed instances in other images. The transformation between images might be local (registration of deformable objects, stereoscopic images etc.) or global (image alignment, pose estimation etc.). This transformation is usually modeled with two components: affine transformation of the shape and location of the object, and transformation of its features (e.g. color, intensity etc). Automatic image registration is important for several vision tasks including tracking [1, 2], structure from motion estimation, optical flow and stereopsis [3, 4], pose estimation and gesture recognition [5, 6] and searching image databases [7, 8] among others.

Difficulties in accurate registration arise due to the presence of occlusions, background clutter and imaging noise. There are two main classes of approaches that have been proposed to address these concerns. The first uses statistical models for the background clutter, occlusions and noise [9, 10] and applies an ML/MAP estimation framework for registration. The problem with these approaches is that, in several applications, it is not possible to have good models for these phenomena. For example, in tracking, an object often passes through diverse background phenomena and it is not a realistic expectation to have *a priori* models for them. Similarly, in content-based image retrieval (CBIR), the object of interest might occur in multiple images, with backgrounds which are too diverse to be modeled meaningfully.

To circumvent these problems, a second class of approaches uses robust formulations for registration. These approaches are predicated on robust features which are invariant to certain transformations (for example, [11, 12]), and use cost functions (e.g. [13, 14]) for registration that are robust to occlusions and clutter without statistically modeling these phenomena. The main advantage of these schemes is that they are not predicated on the *a priori* knowledge of complex phenomena like occlusions and background clutter. The proposed formulation belongs to this class of approaches.

The main contributions of the present paper are as follows: Firstly, we propose a novel probabilistic formulation for robust registration, based on the use of the Kernel Density Correlation (KDC) metric. We demonstrate that the proposed formulation yields better results than the current state-of-the-art. We use statistical simulations to show that our registration metric is more robust to occlusions than the widely popular Mutual Information (MI) metric. This is tied to the property that the KDC metric forms a redescending M-estimator. Secondly, we propose a variational approach to solve the resultant optimization problem. This leads to an itera-

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

tively reweighted least squares (IRLS) algorithm. We also prove the convergence of this algorithm. Finally, we use the KDC metric and the variational optimization framework to propose new estimation algorithms for stereo registration and object tracking. We compare our stereo registration results with those achieved by belief-propagation on standard data-sets with known ground truth disparity maps. We also present results for tracking objects in the presence of occlusions and noise.

In Section 2, we propose the KDC metric for image registration which is based on correlating the reference object and the target image in the domain of probability density functions (pdf). We demonstrate that this metric is robust to occlusions and background clutter. In Section 3, a generic registration problem for non-rigid object motion is formulated as an energy minimization framework with an MRF smoothness constraint on the deformation map. Then, in Section 4, we present an optimization algorithm using variational bounds on the energy functional. We also prove convergence of the proposed algorithm under mild conditions that are easily met. In Section 5.1, we present results for stereo registration of standard test data sets and compare our results with the state-of-the-art available in the literature. In Section 5.2, we present results on tracking of objects in the presence of occlusions and noise.

2. Registration using Empirical Density Estimation

In this section, we propose a framework that addresses the requirements of a robust formulation for the registration problem. Let $I = \{\mathcal{I}_l, \mathcal{I}_r\}$ denote a pair of images, where $\mathcal{I}_l \in \mathbb{S}^{n_l}$ and $\mathcal{I}_r \in \mathbb{S}^{n_r}$ are the reference (template) and target images, of sizes n_l and n_r, and taking values from the set $\mathbb{S}$, respectively. For stereo registration, the images are also referred to as left and right images. Note that in general, $n_l \neq n_r$. For example, for object registration, $\mathcal{I}_l$ may just be the object template while $\mathcal{I}_r$ is a larger target image in which the object is sought to be registered. A common approach to registration is to estimate the mapping function, $\mathcal{D} : \mathbb{N}_l \rightarrow \mathbb{N}_r$, which maps the reference template into the target image such that the template and the samples corresponding to the transformed template in the target image are statistically similar. The sets $\mathbb{N}_l$ and $\mathbb{N}_r$ represent the lattice indices of the two images.

A key requirement is that the registration framework be robust to outliers. This requirement arises from two sources—partial occlusion of the reference template, and, the presence of background clutter. The most commonly used registration metric is the mean squared error (MSE) metric, which naturally arises in the maximum likelihood (ML) formulation if the difference in the registered images is modeled using an i.i.d. Gaussian noise process and there are no occlusions and background clutter. However, this difference (due to imaging noise, occlusions and background clutter) is often non-Gaussian and sometimes, even unknown.

In such cases, it may be desirable to empirically estimate the probability distributions of the two images from the given data and match the two density estimates. Thus, for the *correct* map $\mathcal{D}$, the two sets of samples representing the left and right images, respectively, are statistically similar. Now we formalize this notion: We consider the two sample sets, $S_l \doteq \{(I(i), \mathcal{D}(i)) : i \in \mathbb{N}_l\}$ and $S_r \doteq \{(I(j), j) : j \in \mathbb{N}_r \bigcap \{\mathcal{D}(i), i \in \mathbb{N}_l\}\}$, as two independently generated samples from the same random source $\mathbf{Z}$. The physical explanation of this assumption is that the same underlying physical phenomena (conceptualized as a random source)—lighting, 3D structure of the imaged scene and camera configuration (accounted for by the disparity map), surface reflectances, sensor noise etc., produce the two sample sets. A simple way of estimating densities is by using kernel density estimators.

Definition 2.1 (Kernel density estimator). *Let* $\mathbf{Z}$ *be a d-dimensional random variable and let* $S \doteq \{z_i\}_{i=1}^n$ *be a set of independent observations drawn from the distribution of* $\mathbf{Z}$*, denoted by the pdf* $f_{\mathbf{Z}}(z)$*. Then, the kernel estimator of* $f_{\mathbf{Z}}(z)$*, given the set* S *is,*

$$\hat{f}_{\mathbf{Z}}(z|S_z) = \frac{1}{n|H|} \sum_{i=1}^{n} K_0(H^{-1}[z - z_i]) \qquad (1)$$

where H *is a positive definite* $d \times d$ *bandwidth matrix and* $K_0 : \mathcal{R}^d \rightarrow \mathcal{R}$ *is a kernel such that it is non-negative, has a unit area (*$\int_{\mathcal{R}^d} K_0(z)\, dz = 1$*), zero mean (*$\int_{\mathcal{R}^d} z K_0(z)\, dz = 0$*), and unit covariance (*$\int_{\mathcal{R}^d} z z^T K_0(z)\, dz = I_d$*).*

For grayscale images, $\mathbf{Z}$ is a 3D random vector defined over a 1D intensity domain and a 2D spatial domain (i.e., $\mathbb{S} \subset \mathbb{R}^d$, $d = 3$). A consequence of this definition is that images are defined on a continuous domain and not a discrete lattice[1]. There are several measures that can be used for the task of density matching: mutual information (MI) [14], Jensen-Renyi divergence [15], Bhattacharya coefficient [2] etc. However, we propose to use correlation of the density estimates for two reasons: (1) The resultant optimization problem is easy to solve. In [2], Comaniciu and Meer have used the Bhattacharya coefficient for tracking. However, for

[1] Theoretically, this assumption increases the MSE for the density estimates. However, this is not much of a drawback as we see in the experiments section. On the other hand, it allows for a faster and easily implementable continuous-domain optimization framework.

computational tractability, they solve an approximation to their formulation; (2) The resulting estimator is a redescending M-estimator. M-estimators have been studied for a long time and their properties are well understood. In particular, several authors have demonstrated the robustness of M-estimators for a variety of problems (for example, refer to [16]). Consequently, we present a new measure which we call kernel density correlation (KDC).

Definition 2.2 (Kernel density correlation (KDC)). *Let S_l and S_r be two sets of independent realizations of a random variable Z. Let $K = K_0 \star K_0$ where $\star$ denotes convolution. The kernel density correlation metric between the two density estimates $\hat{f}_Z(z|S_l)$ and $\hat{f}_Z(z|S_r)$, computed as in Definition 2.1, is given by,*

$$KDC(S_l, S_r) = \int_z \hat{f}_Z(z|S_l)\hat{f}_Z(z|S_r)dz$$
$$= \frac{1}{|S_l||S_r||H|^2} \sum_{(x_i,y_j)\in S_l\times S_r} K(H^{-1}(y_j - x_i)) \quad (2)$$

The goodness of the match can thus be quantified in terms of $KDC(S_l, S_r)$. We can now redefine[2] $\mathcal{D} : \mathbb{N}_l \to \mathbb{R}^2$ and $S_r \doteq \{(I(j), j) : j \in \mathbb{N}_r\}$.

2.1. Statistical Analysis

In this section, we present the statistical analysis of the proposed KDC measure. For this purpose, we consider the translation of a 1D signal in the presence of occlusion and additive noise. Figure 1(a) shows the given 1D object template. Figure 1(b) shows an instance of the target image generated by altering the middle segment of the template (to simulate occlusion), adding noise (white Gaussian noise with mean 0 and standard deviation 10 in the figure), and then shifting the pattern five pixels to the right. Presently, the task of registration involves estimating the mapping function $\mathcal{D}$ which represents the translation of the given template. The ground-truth mapping function is given by $\mathcal{D}(i) \doteq i + 5$.

We compare the performance of the proposed measure with (1) sample correlation (SC), and, (2) mutual information (MI) measures. It is well known that under additive white Gaussian noise (AWGN), SC is the minimum variance unbiased estimator (MVUE). However, SC is sensitive to outliers, and, hence, several robust estimators have been proposed including joint entropy, MI, and robust correlation. We compare our results with MI as it is a standard measure used for robust estimation and it is similarly defined in the space of probability functions. In Figure 1(c), we show the response of KDC, MI, and SC measures for the given object template and noisy target image. For this particular example, the KDC and SC measures yield the correct translation estimate, while MI does not.

[2]Note that this does not induce any problems as the bandwidth matrix H can be appropriately selected for matching continuous densities or discrete point sets (by letting each element of H go to zero).

Table 1: *Registration estimates (x-shift) using KDC, MI, and sample correlation measure. Ground-truth value for x-shift is 5. At each σ_n, 200 experiments are carried out to estimate the mean μ and the variance σ of the estimator.*

	KDC		MI		Correlation	
σ_n	μ	σ	μ	σ	μ	σ
5	5.11	0.16	6.99	1.16	5.00	0.00
10	5.05	0.34	6.42	2.44	5.01	0.17
15	5.06	0.50	6.32	2.62	4.98	0.25

(a) AWGN noise (mean 0, standard deviation σ_n)

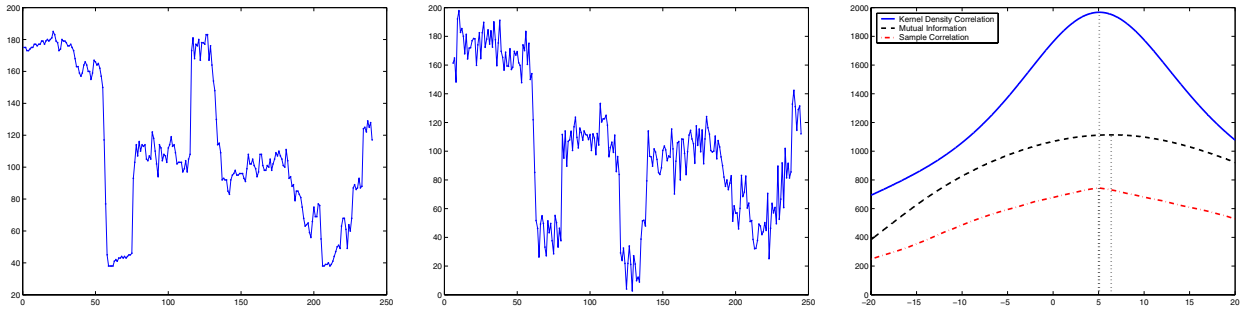
	KDC		MI		Correlation	
σ_n	μ	σ	μ	σ	μ	σ
1	5.09	0.17	6.82	1.65	5.00	0.00
2	5.06	0.24	5.72	2.38	7.02	23.96
3	5.10	0.30	5.53	2.43	8.67	57.05

(b) Additive, i.i.d. standard log-Normal noise (shape parameter σ_n)

Next, we present the response of these measures to i.i.d. Gaussian noise and to i.i.d. standard log-normal noise in Table 1. For the purpose of estimating densities, we use a spatial bandwidth of 5 and an intensity bandwidth of 10 for all experiments. At each variance level, we conduct 200 experiments to calculate the mean and standard deviations of the estimated shifts given by each measure. Table 1(a) shows that sample correlation performs the best for the Gaussian noise case— it has the least bias and variance among the three estimators. This is not surprising as it is the MVUE for such a case. The point to note is that KDC gives better results than MI and thus, is more robust to occlusions (even in the presence of noise). In Table 1(b) we show the response of these measures to heavy tailed noise (used to simulate background clutter). Note that the proposed KDC measure yields the best performance, both in terms of bias and variance. Even in the case of very large noise contamination, the bias and the variance are comparatively small for the KDC measure. This experiment validates the relative robustness of the KDC measure to noise, occlusions as well as random background clutter.

3. Non-Rigid Registration as Energy Minimization

In Section 2.1, we considered a linear map $\mathcal{D}$ for statistical analysis. This corresponded to a rigid-body



(a) (b) (c)

Figure 1: *Comparison of registration performances of KDC, MI and SC measures. (a) 1D object template. (b) Target data realization with additive gaussian noise. Ground-truth registration is a shift of five pixels in the x-direction. Note the difference in the middle segment of the two waveforms, used to simulate occlusion. (c) Responses of KDC (solid curve), MI (dashed curve), and SC (dash-dot curve). The estimated shift is given by the response maxima. Note that KDC and SC yield correct estimates, while MI does not.*

transformation— a translation, for the example considered. In general, the mapping function $\mathcal{D}$ can define a nonlinear warping function on the reference template and thus represent a non-rigid transformation of the given template. Given the reference and the target images, the non-rigid matching problem is to seek the mapping $\mathcal{D}$ which minimizes an appropriate cost function typically defined in terms of the difference between the target image and the warped reference image. The matching problem defined only in terms of this difference is known to be ill-posed. The reason for this is that the intensity for most pixels in the reference image would match the intensity of more than one pixel in the target image. Thus, a regularization criterion is required which is usually expressed in terms of a smoothness prior on the mapping function $\mathcal{D}$ [17, 18].

One common approach is to pose the registration problem in terms of an energy functional [17, 18] where the solution is conceived as a minimizer of the energy functional. The functional has two additive terms: the data term, E_{data}, representing the intensity mismatch between the two images given a map $\mathcal{D}$, and the smoothness term, E_{smooth}, representing the roughness of the map $\mathcal{D}$. Thus, the energy functional may be defined as,

$$E(\mathcal{D}|S_l, S_r) = E_{\text{data}}(S_l, S_r|\mathcal{D}) + E_{\text{smooth}}(\mathcal{D}) \qquad (3)$$

It is important to note the relation of the above formulation to the Bayesian estimation framework. In Bayesian estimation, the energy functional represents the negative logarithm of the *a posteriori* probability $P(\mathcal{D}|S_l, S_r)$ while the data and smoothness terms represent the negative logarithm of the probabilities $P(S_l, S_r|\mathcal{D})$ and $P(\mathcal{D})$ respectively. Thus, the energy formulation subsumes the Bayesian framework.

We propose to use the KDC metric as the data term. Hence, we define

$$E_{\text{data}}(S_l, S_r|\mathcal{D}) \doteq \kappa - KDC(S_l, S_r) \qquad (4)$$

for some constant κ. This term measures the discrepancy associated with the observations, given the model (disparity estimate). A popular smoothness measure used to regularize the matching problem is obtained from an MRF prior for the map $\mathcal{D}$. If the Gibbs energy formulation for the MRF prior is used and E_{smooth} is defined via the negative logarithm of the MRF prior, we get,

$$E_{\text{smooth}} \doteq \alpha_0 \sum_{i,j \in \mathcal{N}} \rho(\mathcal{D}(i) - \mathcal{D}(j)) \qquad (5)$$

where $\mathcal{N}$ is an appropriately defined neighborhood on the lattice, $\mathcal{N}_l$, corresponding to the reference image. We define $\mathcal{N}$ so that horizontally or vertically adjacent pixels in the source image are considered neighbors and use a redescending M-estimator to define $\rho(t) = 1 - \exp(-t^2/\sigma_d^2)$. This choice for the smoothness constraint implies that small discontinuities are smoothed out while large ones are tolerated, and hence the prior knowledge that the objects have smooth surface but large discontinuities can occur between objects. The parameter α_0 controls the relative weighting of the smoothness term and the goodness of the statistical match (the KDC term). The bandwidth parameter σ_d controls the saturation behavior of the M-estimator.

4. Variational Optimization Algorithm

The solution to the non-rigid registration problem is obtained by minimizing the energy functional in (3). The resultant nonlinear optimization problem is solved using an efficient, novel minimization formulation based

on variational upper bounds. We present this solution strategy in the present section.

The approach is based on the inequality $g(x) \geq g(x_0) + g'(x_0)(x - x_0)$ for any convex function $g(x)$. The computational efficiency is based on the following two facts: (1) We always go in the direction of the gradient; and, (2) The step-size in the gradient direction is automatically computed and is guaranteed to decrease the energy functional. At each gradient descent step, the above inequality yields a quadratic upper bound which needs to be minimized. Thus, the variational approach yields an iteratively re-weighted least squares (IRLS) solution to the minimization problem.

We choose the bandwidth matrix in Definition (2.2) to be diagonal, i.e., $H = \mathrm{diag}(\sigma_I^2, \sigma_s^2, \sigma_s^2)$ and the functions $K(\cdot)$ and $\rho(\cdot)$ to be proportional to $\exp(-\|\cdot\|^2)$. Then, using the convexity inequality, it can be shown that for any given estimate $\mathcal{D}_k$, the energy functional is bounded above as follows:

$$E(\mathcal{D}|S_l, S_r) \leq E(\mathcal{D}_k|S_l, S_r) + Q(\mathcal{D}|\mathcal{D}_k) \tag{6}$$

such that,

$$Q(\mathcal{D}|\mathcal{D}_k) \doteq c_1 + c_2 \times \left(\sum_{i \in \mathbb{N}_l, j \in \mathbb{N}_r} w_{i,j}^d(\mathcal{D}_k) \times \right. \tag{7}$$

$$\left. \|\mathcal{D}(i) - j\|^2 + \alpha \sum_{i,j \in \mathbb{N}_l} w_{i,j}^s(\mathcal{D}_k) \|\mathcal{D}(i) - \mathcal{D}(j)\|^2 \right)$$

where data weights, $w_{i,j}^d(\mathcal{D}_k) = \exp(-\|I(i) - I(j)\|^2/\sigma_I^2 - \|\mathcal{D}(i) - j\|^2/\sigma_s^2)$, and smoothness weights, $w_{i,j}^s(\mathcal{D}_k) = \exp(-\|\mathcal{D}(i) - \mathcal{D}(j)\|^2/\sigma_d^2)$, can be easily derived. The constants c_1 and c_2 are inconsequential to the optimization process. The constant α can be computed in terms of α_0, σ_s and σ_d. The next estimate of the mapping function is defined to be,

$$\mathcal{D}_{k+1} \doteq \underset{\mathcal{D}}{\mathrm{argmin}}\, Q(\mathcal{D}|\mathcal{D}_k) \tag{8}$$

Equation (8) requires minimizing a quadratic function by solving a set of linear equations. It is easy to check that such a system of equations is sparse: The smoothness term in the energy functional is a sum of functions involving differences of neighboring (horizontal and vertical) disparities. Each linear equation is derived by differentiating $Q(\mathcal{D}|\mathcal{D}_k)$ with respect to $D(i), i \in \mathbb{N}_l$. Thus, each equation would involve at most five unknown terms: the mappings of the pixel i, its two horizontal and two vertical neighbors. This sparse system of equations at each iteration is solved for the mapping $\mathcal{D}$ using a stabilized, bi-conjugate gradient method [19].

The theorem below shows that the iterative algorithm given by (7) and (8) is convergent.

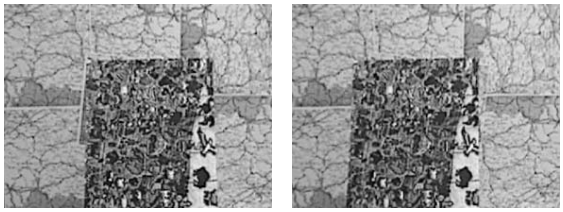


Figure 2: *The rectified, gray-scale map stereo pair.*

Figure 3: *The rectified, gray-scale venus stereo pair.*

Theorem 4.1. *Let $K(\cdot)$ be separable in the intensity and spatial subspaces, i.e., $K(H^{-1}z) \doteq K_I(-\frac{z_I}{\sigma_I})K_s(-\frac{\|z_s\|^2}{\sigma_s^2})$. If K_s is convex and non-decreasing, then the sequences $S_E \doteq \{E(\mathcal{D}_k|S_l, S_r)\}_{k=1,2,\ldots}$ and $S_D \doteq \{\mathcal{D}_k\}_{k=1,2,\ldots}$ defined according to (8), are convergent. The sequences converge to a local minimum of $E(\mathcal{D}|S_l, S_r)$ as defined in equations (3), (4) and (5).*

Proof. See Appendix for the proof. □

5. Applications

In Section 2, we demonstrated, using empirical statistical analysis, the robustness of the KDC measure to occlusions and outliers. In Section 3 and Section 4, we proposed a new algorithm to match two instances of the same object, when these instances are related by a non-rigid transformation $\mathcal{D}$. This algorithm uses the KDC metric in a variational energy minimization framework. In this section, we take two computer vision applications, stereo registration and object tracking, to test the efficacy of the proposed algorithm.

5.1. Stereo Registration

In the stereo registration problem, images $\mathcal{I}_l$ and $\mathcal{I}_r$ respectively represent the left and right images of a stereoscopic pair. The mapping function $\mathcal{D}$ represents the disparity function defined on the left image. The objective of the stereo registration problem is to estimate the disparity map $\mathcal{D}$. Note that in our formulation, $\mathcal{D}$ can take

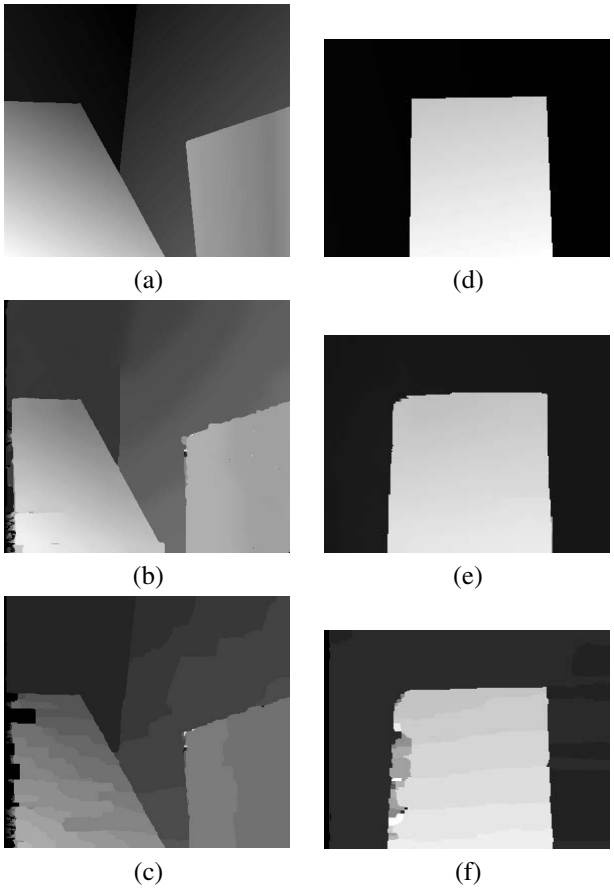


Figure 4: *Disparity maps for venus ((a)-(c)) and map ((d)-(f)) stereo pair. (a),(d): Ground-truth. (b),(e): Proposed algorithm. (c),(f): Belief propagation.*

continuous values and hence, we do not require interpolation for sub-pixel accuracy.

We use the rectified gray-scale *venus* and *map* stereo pairs to evaluate the performance of the proposed formulation. The stereo pairs have known sub-pixel disparity ground truth data, and known occlusion regions [20]. Fig. 2 shows the *map* stereo pair, and Fig. 3 shows the *venus* stereo pair. In each case the disparity map is estimated using the left image as the reference image.

We use the following metrics to quantitatively evaluate algorithm performance, where $D^{KML} = \{d_i^{KML}\}$ denotes the estimated disparity map, and $D^{GT} = \{d_i^{GT}\}$ denotes the ground-truth: (1) B, the fraction of pixels for which $|d_i^{GT} - d_i^{KML}| > 1$, (2) $B_{\overline{O}}$, the fraction of pixels in non-occluded regions, for which $|d_i^{GT} - d_i^{KML}| > 1$, and (3) $M_{\overline{O}C}$, the mean-squared disparity error for non-occluded pixels which additionally satisfy $|d_i^{GT} - d_i^{KML}| \leq 1$. The first two metrics quantify the fraction of pixels with significant disparity error, and do not penalize disparity errors which are solely due to discretization. The third metric incorporates the effect of discretization by quantifying the mean-squared disparity error. The performance of the proposed algorithm is compared to that of a state-of-the-art Potts model based MRF formulation [21], which uses belief propagation for inference. The first two error metrics are used for comparison since they do not penalize discrete disparity values which the algorithm in [21] outputs.

Table 2: *Stereo: Quantitative comparison.*

	$B(\%)$	$B_{\overline{O}}(\%)$	$M_{\overline{O}C}$
Venus Proposed	5.55	2.24	0.056
Venus Belief Prop.	6.19	2.95	0.128
Map Proposed	5.38	0.40	0.026
Map Belief Prop.	6.63	0.20	0.155

The results obtained for the *venus* stereo pair are shown in Figures 4(a)-(c). Figure 4(a) shows the true disparity map, Figure 4(b) the disparity map produced by the proposed algorithm, and 4(c) the disparity map produced by the belief propagation algorithm. As can be seen, unlike the belief propagation algorithm, the proposed algorithm produces a smooth, non-discretized disparity map. 4(d)-(f) shows similar results for the *map* stereo pair.

Table 5.1 quantitatively compares the performance of the two algorithms, using the metrics described above. For the *venus* stereo pair, the metrics B and $B_{\overline{O}}$, which do not penalize discrete-valued estimates, are comparable for the two algorithms. The mean-squared disparity error $M_{\overline{O}C}$ is significantly lower for the proposed algorithm. Similar results are obtained for the *map* pair—$M_{\overline{O}C}$ is again significantly lower for the proposed algorithm, as compared to the belief propagation algorithm.

5.2. Object Tracking

The second application we consider is that of object tracking in image sequences. In this case, an object in the current frame needs to be registered in the next frame. This problem is simpler than stereo registration due to small transformations between consecutive frames. However, the challenge is to accurately segment the object in each frame to iteratively update the object mask. This is imperative since, otherwise, appearance changes accrue over time and the tracker is eventually thrown off.

For the problem of object tracking, let the object transformation from the current frame to the next be modeled by T_Θ that defines an affine transformation in the joint intensity and spatial domains. The following are the steps of our algorithm (given the object samples

in the current frame, denoted by S_X), (1) Estimate T_Θ, the transformation parameters, using the algorithm in Section 4. This gives an estimate of the set of transformed object samples from the current frame, denoted by $S_{T_{\hat{\Theta}}X}$. (2) Using a Parzen Window estimate, $\hat{f}_X(\cdot)$, segment pixels in the current frame in a window around the transformed object. Update S_X using the obtained segment.

The first step of our algorithm gives us the position of the object in the next frame and an estimate of the set of transformed object samples $S_{T_{\hat{\Theta}}X}$. However, due to out of plane rotation and occlusions and the fact that the photometric properties of the object changes over time, the new object template (the set of object pixels in the current frame) is not adequately represented by $S_{T_{\hat{\Theta}}X}$. To further refine the object template in step (2), we extract the subset of samples from the next frame based on the set $S_{T_{\hat{\Theta}}X}$.

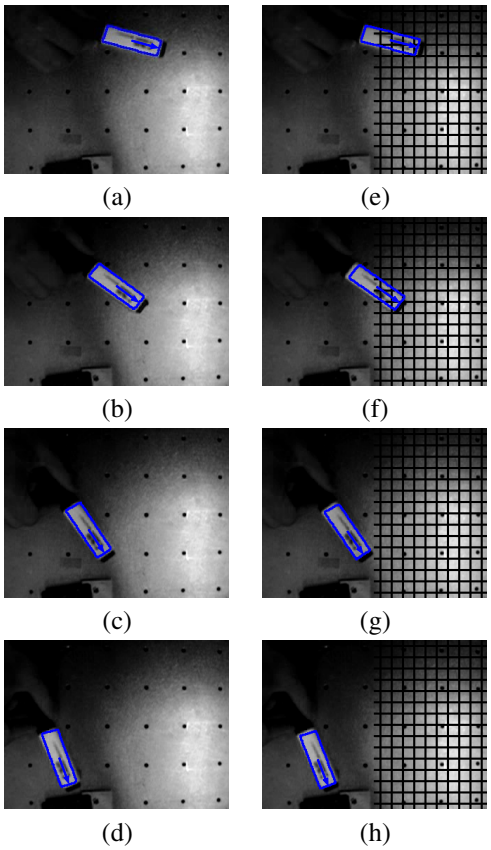


Figure 5: *Results for the marker sequence. Frames of tracked marker with and without an occluding grill are shown respectively: Frame* 146 *in (a) and (e); Frame* 160 *in (b) and (f); Frame* 176 *in (c) and (g); and, Frame* 190 *in (d) and (h). We show overlayed blue boundaries denoting the tracked object and estimated orientation with a blue arrow.*

Given the object size, shape, its distance from the camera and its motion, the object template changes its shape and size in the current frame in a bounded fashion (say, a maximum of Δ_s pixels around the current boundaries)[3]. Based on Δ_s and $T_{\hat{\Theta}}$, we estimate the set of pixels, S', in the current frame that the object is constrained to lie in. Using the Parzen Window estimate of the pdf corresponding to *all image pixels* in the current frame, we can assign pixels belonging to S' to either object or background modes[4] using the Mean Shift algorithm [22] on back-projected pixels from the set S' via the inverse transform T_Θ^{-1}. Thus, for each pixel, y' in S' in the next frame, the starting point for Mean-Shift is $T_\Theta^{-1}y'$ in the current frame. The pixels assigned to the foreground, thus, comprise the new object template.

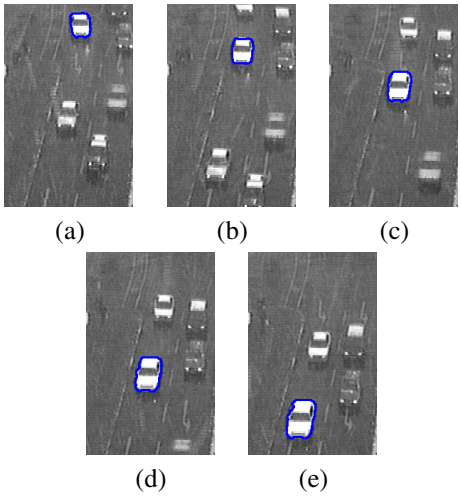


Figure 6: *Results for the car sequence. Tracked car in (a) Frame* 160*, (b) Frame* 180*, (c) Frame* 200*, (d) Frame* 220 *and (e) Frame* 240*. We show overlayed boundaries denoting the tracked object. Pixels within the blue boundary denote the set of updated object samples.*

We now present results on three real sequences. The first two sequences, called Marker sequences, are indoor sequences that we generated in our lab. We use these sequences to validate the first step of our algorithm. Since there are no size changes (an almost-flat object with motion approximately perpendicular to the optical axis), we do not need to update the mask. We present the results in Figure 5. Figures 5 (e)-(h) shows the same frames as in Figures 5 (a)-(d) of a second sequence when a grill occludes part of the object in the initial frames. Thus the set S_X in the first frame, in Figure 5(e), contains pixels belonging to the grill. Results in the two rows are identical—exhibiting robustness to the phenomenon of partial occlusion of the given ob-

[3] We do not address the automatic calculation of Δ_s and take it as a user-input parameter to our algorithm.

[4] These are already available since the object is already segmented in the current frame

ject template and the object in the target frames. Another challenge in this sequence was spatially varying illumination which resulted in the intensities of pixels belonging to the object changed by as much as 25 gray levels. However, our framework was able to track these changes.

In Figure 6, we show results on another sequence, called the Traffic sequence. This result depicts all the vagaries of tracking in uncontrolled, outdoor environments. It shows a traffic sequence taken from a still camera under moderate precipitation. Thus, there is appreciable noise in the sequence. We track a white car (marked out in Figure 6(a)) coming towards the camera through 80 frames. Comparing frames 160 and 240 as shown in Figures 6(a) and (e) respectively, one can see that there is an appreciable change in the size (almost double in area) of the car. It is also clear that there is an appreciable change in the appearance of the tracked car. This requires us to update the set of object pixels (object template) through the sequence. The results validate that the algorithm successfully tracks the car in presence of noise and appearance changes (including change in size).

6. Conclusions

In this paper, we presented a novel approach to object registration. The first key idea developed was to construct an estimate of the pdf for the object and the target image. We then proposed correlation of the density estimates as a registration measure and illustrated its robustness to occlusions and background clutter. We also proposed an energy cost functional, using the above measure, for non-rigid registration. We developed a variational optimization framework for this cost function and proved its convergence. This cost function was then used for registration of stereoscopic images. We implemented the algorithm on standard test images and compared them with the belief propagation algorithm. Both quantitative comparison and visual inspection show the efficacy of the proposed method. We also proposed an algorithm for tracking and showed its performance on real sequences with occlusions and noise. Note that the tracking algorithm is able to estimate rotations as well as handle occlusions, noise and changes in the object size. It further extracts the segmented object from the video stream.

Acknowledgements

We thank Ashish Jagmohan for his help with implementation of the stereo registration algorithm. We also gratefully acknowledge the support of the Office of Naval Research under grant N00014-03-1-0107.

Appendix

Proof of Theorem 4.1. We rewrite (7) as,

$$Q(\mathcal{D}|\mathcal{D}_k)/c_2 \doteq \sum_{i\in\mathbb{N}_l, j\in\mathbb{N}_r} w^d_{i,j}(\mathcal{D}_k)(\|\mathcal{D}(i)-j\|^2 - \|\mathcal{D}_k(i)-j\|^2) + \alpha \times \sum_{i,j\in\mathbb{N}_l} w^s_{i,j}(\mathcal{D}_k)(\|\mathcal{D}(i)-\mathcal{D}(j)\|^2 - \|\mathcal{D}_k(i)-\mathcal{D}_k(j)\|^2) \quad (9)$$

where the weights $w^d_{i,j}(\mathcal{D}_k)$ for an exponential kernel are given in Section 4 after Equation (7). For a general kernel, meeting the requirements of Theorem 4.1, the weights are given by $w^d_{i,j}(\mathcal{D}_k) = K_I(-\frac{z_I}{\sigma_I})K'_s(-\frac{\|z_s\|^2}{\sigma_s^2})$.

For any $l \in \mathbb{N}_l$, let l^N, l^S, l^E, l^W be the north, south, east and west neighbors of l respectively. Then, we construct an $n_l \times n_l$ matrix $\mathcal{A}$ and a vector b with the following entries,

$$\begin{aligned} \mathcal{A}_k(l,l) &= (\sum_{j\in\mathbb{N}_r} w^d_{l,j}(\mathcal{D}_k)) + \alpha\Big(w^s_{l,l^N}(\mathcal{D}_k) + \\ &\quad w^s_{l,l^S}(\mathcal{D}_k) + w^s_{l,l^W}(\mathcal{D}_k) + w^s_{l,l^E}(\mathcal{D}_k)\Big) \\ \mathcal{A}_k(l,m) &= -\alpha w^s_{l,m}(\mathcal{D}_k) \\ &\quad \text{if } m \in \text{ the neighborhood of } l \text{ else } 0 \\ b_k(l) &= \sum_{j\in\mathbb{N}_r} j w^d_{l,j}(\mathcal{D}_k) \end{aligned} \quad (10)$$

Then, we can rewrite (9) as,

$$\begin{aligned} Q(\mathcal{D}|\mathcal{D}_k)/c_2 &= (\mathcal{D}-\mathcal{D}_k)^T\mathcal{A}_k(\mathcal{D}-\mathcal{D}_k) + \\ &\quad 2(\mathcal{D}-\mathcal{D}_k)^T(\mathcal{A}_k\mathcal{D}_k - b_k) \end{aligned} \quad (11)$$

Thus, from (8) and (11), we get,

$$\mathcal{A}_k\mathcal{D}_{k+1} = b_k \quad (12)$$

For this choice of $\mathcal{D}_{k+1}$,

$$\begin{aligned} Q(\mathcal{D}_{k+1}|\mathcal{D}_k) &= -c_2(\mathcal{D}_{k+1}-\mathcal{D}_k)^T\mathcal{A}_k(\mathcal{D}_{k+1}-\mathcal{D}_k) \\ &\le 0 \end{aligned} \quad (13)$$

Note that A_k's are positive-definite matrices since they are strictly diagonally-dominant with positive diagonal entries. Also, the constant c_2 is positive. This implies that the sequence S_E is non-increasing. Since S_E is also bounded from below, it is convergent. Therefore, $Q(\mathcal{D}_{k+1}|\mathcal{D}_k) \to 0$. Since $Q(\mathcal{D}|\mathcal{D}_k)$ is a variational upper-bound, it is also easy to see that $\nabla_{\mathcal{D}}E_k = 2(\mathcal{A}_k\mathcal{D}_k - b_k)$. Thus, we can rewrite (13) as,

$$\begin{aligned} Q(\mathcal{D}_{k+1}|\mathcal{D}_k) &= \tfrac{c_2}{2}(\mathcal{D}_{k+1}-\mathcal{D}_k)^T\nabla_{\mathcal{D}}E_k \\ &= \tfrac{c_2}{2}\mathcal{A}_k^{-1}\|\nabla_{\mathcal{D}}E_k\|^2 \le 0 \end{aligned} \quad (14)$$

Thus, the step taken in each iteration has a negative projection to the gradient and the gradient goes to 0. Hence, S_E goes to a local minimum of $E(\mathcal{D}|S_l, S_r)$.

Convergence of S_D: Positive definiteness of $\mathcal{A}_k$ implies $\delta\mathcal{D}_k \doteq \mathcal{D}_{k+1} - \mathcal{D}_k \rightarrow 0$. Let the sequence S_D have an isolated limit point $\mathcal{D}^*$. Since $\delta\mathcal{D}_k \rightarrow 0$, given small enough r, ϵ, $0 < r < r + \epsilon$ and open balls $B(\mathcal{D}^*, r)$, $B(\mathcal{D}^*, r + \epsilon)$, there exists an index K_1 such that for all $k > K_1$, $\|\delta\mathcal{D}_k\| < \epsilon$, $\mathcal{D}^*$ is the only limit point in $B(\mathcal{D}^*, r + \epsilon)$ and the set $U \doteq \{\mathcal{D}|\mathcal{D} \in B(\mathcal{D}^*, r + \epsilon) \cap B^c(\mathcal{D}^*, r), \mathcal{D} \in S\}$ is non-empty. U has finite items, let their minimum value be m_U. Further, as S_E is non-increasing, there exists an index K_2 such that for all $k > K_2$, $E(\mathcal{D}_k|S_l, S_r) < m_U$. We define $K = \max(K_1, K_2)$. Then, for any $k > K$, $\mathcal{D}_k \in B(\mathcal{D}^*, r) \Rightarrow \mathcal{D}_{k+1} \in B(\mathcal{D}^*, r)$ since $B(\mathcal{D}^*, r) \cup B(\mathcal{D}_k, \delta\mathcal{D}_k) \subset B(\mathcal{D}^*, r + \epsilon)$ and $\mathcal{D}_{k+1} \notin U$. Hence, $\mathcal{D}^*$ is a point of local minimum for $E(\mathcal{D}|S_l, S_r)$. □

References

[1] H. Tao, H. S. Sawhney, and R. Kumar, "Object tracking with Bayesian estimation of dynamic layer representation," *PAMI, IEEE Trans.*, vol. 24, no. 1, 2002.

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *PAMI, IEEE Trans.*, vol. 25, no. 5, pp. 564–575, 2003.

[3] Y. Altunbasak, R. M. Mersereau, and A. J. Patti, "A fast parametric motion estimation algorithm with illumination and lens distortion correction," *IEEE Trans. on Image Proc.*, vol. 12, no. 4, pp. 395–408, 2003.

[4] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," *International Conf. on Computer Vision*, 2001.

[5] M. A. Sipe and D. Casasent, "Feature space trajectory methods for active computer vision," *PAMI, IEEE Trans.*, vol. 24, no. 12, pp. 1634–1643, 2002.

[6] M. J. Black and A. D. Jepson, "A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions," *European Conf. on Computer Vision, ECCV-98*, pp. 909–924, 1998.

[7] A. C. Kak and C. Pavlopoulou, "Computer vision techniques for content-based image retrieval from large medical databases," *7th Workshop on Machine Vision Applications*, 2000.

[8] B. Bhanu, J. Peng, and S. Qing, "Learning feature relevance and similarity metrics in image databases," *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 14–18, 1998.

[9] D. Hasler, L. Sviaz, S. Susstrunk, and M. Vetterli, "Outlier modeling in image matching," *PAMI, IEEE Trans.*, vol. 25, no. 3, pp. 301–315, 2003.

[10] Z. Ying and D. Castanon, "Partially occluded object recognition using statistical models," *IJCV*, vol. 49, no. 1, pp. 57–78, 2002.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *preprint, IJCV*, July 2003.

[12] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," *Proc. ICCV*, pp. 26–33, 2001.

[13] Y. He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *Signal Proc., IEEE Trans.*, vol. 51, no. 5, pp. 1211–1220, 2003.

[14] P. Viola and W. M. Wells III, "Alignement by maximization of mututal information," *IJCV*, vol. 24, no. 2, pp. 137–154, 1997.

[15] A. Hamza, Y. He, and H. Krim, "An information divergence measure for ISAR image registration," *IEEE Workshop on Statistical Image Proc.*, 2001.

[16] C. K. Chu, I. K. Glad, F. Godtliebsen, and J. S. Marron, "Edge-preserving smoothers for image processing," *Journal of the American Statistical Association*, vol. 93, pp. 526–541, 1998.

[17] J. Sun, N. Zheng, and H. Shum, "Stereo matching using belief propagation," *PAMI, IEEE Trans.*, vol. 25, no. 7, pp. 787–800, 2003.

[18] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *PAMI, IEEE Trans.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[19] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*, SIAM, Philadelphia, PA, 1994.

[20] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Comp. Vision*, vol. 47, pp. 7–42, 2002.

[21] O. Veksler, "Efficient graph-based energy minimization methods in computer vision," 1999, PhD Thesis, Cornell Univ.

[22] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *PAMI, IEEE Trans*, vol. 24, no. 5, pp. 603–619, 2002.