Optimization and Learning for Registration of Moving Dynamic Textures

Junzhou Huang
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854, USA
jzhuang@cs.rutgers.edu

Xiaolei Huang
Lehigh University
19 Memorial Drive West
Bethlehem, PA 18015, USA
xih206@lehigh.edu

Dimitris Metaxas
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854, USA
dnm@cs.rutgers.edu

Abstract

We address the problem of registering a sequence of images in a moving dynamic texture video. This involves optimization with respect to camera motion, the average image, and the dynamic texture model. This problem is highly ill-posed and almost impossible to have good solutions without priors. In this paper, we introduce powerful priors for this problem, based on two simple observations: 1) registration should simplify the dynamic texture model while preserving all useful information. It motivates us to compute a prior for the dynamic texture by marginalizing over specific dynamics in the space of all stable auto-regressive sequences; 2) the statistics of derivative filter responses in the average image can be significantly changed by registration, and better registration should lead to a sharper average image. This offers us the prior of requiring the derivative distribution of the estimated average image to be close to that learned from the input image sequence. With these priors, a new registration approach is proposed by marginalizing over the "nuisance" variables under a Bayesian framework. And superior motion estimation results are obtained by jointly optimizing over the registration parameters, the average image, and the dynamic texture model. Experimental results on real video sequences of moving dynamic textures show convincing performance of the proposed approach.

1. Introduction

Video registration and motion analysis are required by many video analysis applications when a video sequence is captured by a moving camera. In the past decade, significant progress has been made in this area. Traditional methods for image alignment generally assume that the scenes are static, rigid and have brightness constancy [1, 2, 9, 23]. Then the optimal values of motion parameters are estimated by minimizing some error function of sample variance of each pixel over time. However, these assumptions do not hold in dynamic scenes with moving dynamic textures. Examples of moving dynamic textures are shown in Figure 1. Thus, most registration methods are likely to fail to process dynamic scenes with moving dynamic textures.

Figure 1. Examples of moving dynamic textures, see *http://www.robots.ox.ac.uk/ awf/iccv01/.*

Some efforts have been made in the literature to relax these assumptions [5, 21, 15, 11]. The rigidity assumption is relaxed to deal with dynamic scenes with dynamic textures [4, 22, 3]. A nonrigid scene is called a dynamic texture when it is captured by a static camera and its temporal evolution exhibits certain stationarity, such as flowers, water, steam and so on. In these papers, the average image can be directly computed because the dynamic texture is captured by a static camera. After subtracting the average image from each frame in the video, the temporal evolution of the dynamic texture is modeled as the output of a time invariant linear dynamical system (LDS). The key point is to separate image appearance and the underlying dynamics of a scene into two processes. With classical system identification techniques, the joint model for both the appearance and dynamics of a scene can be recovered. However, these approaches can not handle the registration problem in moving dynamic textures, which is a video sequence of dynamic texture, but captured by a moving camera.

Registration of moving dynamic textures involves optimization with respect to both camera motion and the dynamic texture model. It is a typical Chicken-and-Egg problem. If we know the exact camera motion, the dynamic texture model can be easily estimated [4]. If we know the exact model of the captured dynamic texture, the camera motion also can be computed using classic motion estimation approaches [1, 2]. Since we know neither camera motion nor the dynamic texture model, however, we face three key challenges in the problem of registering moving dynamic textures: 1) the mean image over the whole image sequence can not be simply computed as the average image and subtracted from each image for computing the LDS model because the correspondence across image frames for each pixel is not known due to camera motion; 2) LDS can not be used to represent the dynamic texture model because the appearance is no longer invariant due to camera motion; 3) the brightness constancy assumption can not be used to estimate camera motion because the intensity of a pixel changes due to temporal evolution of the dynamic texture. All these challenges lead to a highly ill-posed problem.

To the best of our knowledge, there are only a few attempts for the problem of moving dynamic texture registration until now. Fitzgibbon made a pioneering attempt for this problem [6]. In his work, stochastic rigidity is introduced to search for the optimal global camera motion for registering a moving dynamic texture. The motion parameters are optimized simultaneously by minimizing the entropy function of an auto regressive process. This attempt for optimizing both the motion model and dynamic texture model leads to a difficult non-linear optimization problem. No explicit framework was introduced in [6] to estimate the global motion model or the appearance and dynamic model of moving dynamic textures.

In [17], Dynamic Texture Constancy Constraint (DTCC) is introduced for camera motion estimation instead of brightness constancy. In order to capture both the rigid camera motion and nonrigid motion in dynamic scenes, a time-varying LDS model is proposed to represent the dynamic scenes:

$$\begin{aligned} x(t+1) &= Ax(t) + Bv(t) \\ I(t) &= C(t)x(t) + w(t) \end{aligned} \tag{1}$$

However, time-varying LDS optimization is a difficult problem. In order to make this problem more tractable, it is divided into two sub-problems: 1) optimizing dynamic texture model parameters assuming the camera motion model parameters are known; 2) estimating motion model parameters assuming the dynamic texture model parameters are known according to DTCC. Some assumptions are made: 1) the camera motion is small and thus $C(t)$ is assumed to be constant in each small time window $[t - w + 1, t]$; 2) C does not depend on the view point. With the first assumption, a time invariant LDS $(A(t), B(t), C(t))$ can be learned for each time window using the method in [4]. The time-varying LDS $(A^*(t), B^*(t), C^*(t))$ for the whole sequence can be obtained after normalizing the model parameters $(A(t), B(t), C(t))$ with respect to the same basis. With the computed model parameters of time-varying LDS, the rigid camera motion can be estimated according to DTCC. Compared to Fitzgibbon's method [6], this method results in an optimization formulation that leads to a simpler computation framework to obtain better solutions for the problem of moving dynamic texture registration.

While the DTCC-based camera motion estimation method is simple and general, it greatly depends on the accurate estimation of $C(t)$ and A and it is thus imperative to find a good method to efficiently obtain accurate estimates of $C(t)$ and A. Unfortunately, under the first assumption in Vidal's method [17], the camera is static within each small time window although camera motion does exist in such time windows. Therefore, the estimation of time invariant LDS model $(A(t), B(t), C(t))$ for each time window includes not only the nonrigid motion due to dynamic texture but also part of the rigid camera motion. This causes the nonrigid motion in dynamic textures to be always overestimated and the rigid camera motions to be always underestimated. Their experimental results also showed this limitation of the method [17].

In this paper, we propose a new algorithm for registering a sequence of images containing moving dynamic textures. Our method is inspired by the remarkable capability of human vision to decompose the image motion in a moving dynamic texture into rigid camera motion and nonrigid motion in the dynamic scene. Underlying this ability are two key observations: 1) registration according to the accurate camera motion should simplify the dynamic texture model while preserving all useful information; 2) registration according to the accurate camera motion should lead to a sharp average image whose statistics of derivative filters are similar to those of the original image frames. The first observation motivates us to compute a prior for the dynamic texture by marginalizing over the dynamics in the space of all stable auto-regressive sequences. The second observation offers us a prior that the derivative distribution of the estimated average image should be close to the expected distribution learned from image frames in the video sequence. And the more accurate the estimated camera motion parameters are, the closer the average image distribution should be to the expected distribution. With these two priors, we have both the dynamic texture and the average image of the video imposing some constraints on the camera motion. To the best of our knowledge, no previous method had explored the prior based on the second observation for image registration although the prior based on the first observation has been implicitly used in [6, 17].

In our approach, the prior model for the average image is learned first from the image sequence. Then, the prior model for the dynamic texture is learned by marginalizing over the dynamics in the space of all stable auto-regressive sequences. With these two priors, superior estimates are obtained by optimizing over the camera motion model, the average image, and the dynamic texture model. This results in a general model for a moving dynamic texture, which explicitly models the rigid camera motion and the nonrigid motion in a dynamic scene. Furthermore, the algorithm automatically learns parameters for the average-image prior from the image sequence. Considering that each scene will have different underlying image statistics, the proposed algorithm preserves as much richness and detail as possible from the original scene without encountering problems with conditioning. This will aid our approach to obtain accurate estimates of camera motion and dynamic texture model.

The remainder of the paper is organized as follows. Section 2 introduces the problem formulation. Our powerful prior models are detailed in section 3. Section 4 introduces how to perform the joint optimization for registration. Experimental results and discussions are presented in section 5 and we conclude this paper in section 6.

2. Problem Formulation

The registration of moving dynamic textures is formulated as the problem of finding global motion between image frames in a moving dynamic texture. After each image frame is transformed under desired global motion, the average image of the whole sequence should have a distribution that is similar to those of the input image frames. Moreover, the image sequence after desired transformations should be optimally represented by a dynamic texture model.

We assume that a known dynamic scene ***Y*** can be modeled by a time invariance LDS as a dynamic texture [4]. A video sequence $I(1), I(2), ..., I(n)$ of this known dynamic scene is captured by a moving camera. Given transformation parameters for registration, $\Theta(t)$, the generative image model is introduced as follows:

$$\begin{aligned} I(t) &= W(\Theta(t))(y_0 + y(t)) + w1(t) \\ y(t) &= Cx(t) + w(t) \\ x(t+1) &= Ax(t) + v(t) \end{aligned} \tag{2}$$

Where, $t = 1, 2, ..., n$, $W(\Theta(t))$ is a system transformation matrix for registration, $I(t)$ and $y(t)$ are $l \times 1$ vectors (l is the number of pixels in any image frame), $x(t)$ is a $k \times 1$ vector which is typically the result of applying some filtering or dimension reduction on $y(t)$, $w1(t)$, $w(t)$ and $v(t)$ are Gaussian white noises, y_0 is the mean image for the sequence $y(1), y(2), ..., y(n)$, A and C are the dynamic matrix and the observation matrix respectively, for the dynamic texture model [4].

Figure 2 shows the generative model for a video sequence of moving dynamic texture. In this model, we assume the observed image $I(t)$ is i.i.d and it depends on: the desired average image y_0, the camera motion related $W(\Theta(t))$, and the appearance $y(t)$ of the dynamic texture. In this model, $y(t)$ depends on the dynamics of the dynamic texture, $x(t)$; $x(t)$ is dependent on $x(t-1)$, guided by a first order autoregressive (AR) model with initialization $x_0 \sim \mathcal{N}(0, 1)$.

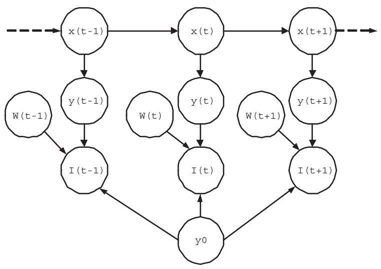

Figure 2. A directed graphical model representing a generative image model for a moving dynamic texture.

For simplicity, the translation motion model is used for camera motion to illustrate our approach [1]. We suppose the exact y_0 and $y(t)$ can be obtained after transforming $I(t)$ with the translation motion parameter $M(t)$:

$$y_0 + y(t) = I(t) + B(t)M(t) \tag{3}$$

where $B(t) = [I_x(t), I_y(t)]$, and $I_x(t)$ and $I_y(t)$ are the first order derivatives in the horizontal direction and vertical direction respectively. In a short video sequence, this approximation is reasonable due to the small amount of motion. In order to well represent the dynamics with a first order AR model [4], $x(t)$ should have zero mean values in equation 2. According to this, the average image y_0 can be derived:

$$y_0 = I_0 + \frac{1}{n}BM + u(t) \tag{4}$$

where $B = [B(1), ..., B(n)]$, $M = [M(1), ..., M(n)]$, I_0 is the mean image of the sequence, $I(1), ..., I(n)$, and $u(t)$ is Gaussian $i.i.d.$ noise with precision θ.

After incorporating noise $w1(t)$ into noise $w(t)$, the proposed generative image model becomes:

$$\begin{aligned} I(t) + B(t)M(t) &= y_0 + y(t), \\ y(t) &= Cx(t) + w(t), w(t) \sim \mathcal{N}(0, \beta^{-1}I) \\ x(t+1) &= Ax(t) + v(t), v(t) \sim \mathcal{N}(0, \alpha^{-1}I) \\ y_0 &= I_0 + \frac{1}{n}BM + u(t), u(t) \sim \mathcal{N}(0, \theta^{-1}I) \end{aligned} \tag{5}$$

where v and w are Gaussian noises with precisions α and β respectively. This is the generative image model for registering images in a moving dynamic texture video. It is a highly ill-posed problem and it is almost impossible to obtain a reasonable solution without strong priors. In the following section, we introduce strong priors into the generative image model based on two key observations.

3. Prior Models

As mentioned in the introduction, strong priors for our generative moving dynamic texture model are derived from

[1] Using a transformation matrix $W(\Theta(t))$ with generic transformation parameters $\Theta(t)$, the framework can be easily extended to other motion models, such as rigid and affine motion models.

two key observations: 1) a good registration according to the accurate camera motion should simplify the dynamic texture model while preserving all useful information; 2) a good registration according to the accurate camera motion should lead to a sharp average image whose statistics of derivative filters are similar to those of the input image frames.

3.1. The Average Image Priors

We consider the average image of a moving dynamic texture video sequence. We observe that, the average image changes significantly given different registration parameters, and the correct registration leads to a sharp average image, whose statistics of derivative filters are close to those in the input image frames. Figure 3 gives an example and illustrates this observation. In the figure, statistics on the average image are shown in red color, and statistics on the input images are shown in blue color. One can see that, before applying registration to correct camera motions, the two distributions are far away as shown in Figure 3 (a). In the middle of the registration process, they become closer as shown in Figure 3 (b). When the correct registration is reached, the statistics of the average image and that of the input images are very similar as shown in Figure 3 (c). From Figure 3, one can observe that the statistics of derivative filters in the images are similar to a Gaussian but have a heavy tail.

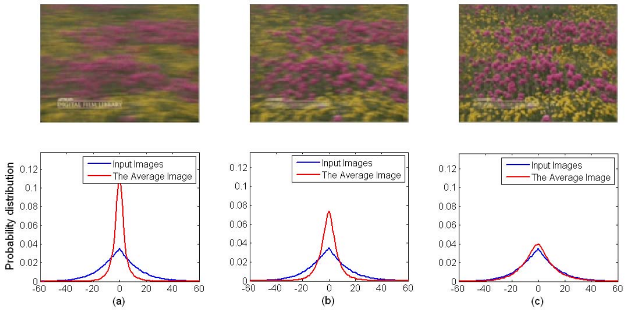

Figure 3. First-order derivative distributions of average image vs. input images. Before registration (a), in the middle of registration (b), and after correct registration (c).

Huang and Mumford have shown that the Student-t distribution can model this heavy tailed image prior very well [8]. This distribution has been successfully used under the Products-of-Experts framework [19, 16, 20]. It has the following form:

$$p(\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \prod_{i=1}^{N} \phi_i(J_i^T \mathbf{x}; \sigma_i)$$
$$\phi_i(J_i^T \mathbf{x}; \sigma_i) = [1 + \frac{1}{2}(J_i^T \mathbf{x})^2]^{-\sigma_i} \quad \sigma_i > 0 \qquad (6)$$

In this paper, we also use this distribution to model the image priors. As the desired average image can be thought as one image of the same scene from an unknown view, it should have similar statistics to the input image frames. Therefore, it is sufficient to learn the distribution of directional image gradients (in the horizontal, vertical, two diagonal directions) of input image frames as the prior model for the desired average image. Hence we only need to learn the parameters $\{\sigma_i\}_{i=1\sim 4}$ in Eq. 6, and we do not need to learn J_i^T as in [19, 16, 20]. This greatly accelerates the learning speed of the proposed algorithm. Experimental results that we will present in Section 5 confirm this point.

The prior model parameters $\{\sigma_i\}_{i=1\sim 4}$ of the average image y_0 are learned from the input frames $I(1), ..., I(n)$ by maximizing the data likelihood. Following [16, 20], the parameter updating is performed using the Contrastive Divergence method [7]:

$$p(y_0) = exp(-E_{y_0})/Z_{y_0}$$
$$E_{y_0} = \sum_{i=1}^{4} \sigma_i \log(1 + \frac{1}{2}(J_i^T y_0)^2)$$
$$\delta\sigma_i = \eta(\langle \frac{\partial E_{y_0}}{\partial \sigma_i} \rangle_{p^k} - \langle \frac{\partial E_{y_0}}{\partial \sigma_i} \rangle_{p^0}) \qquad (7)$$

where η is a user-defined step size, p^0 is the data distribution, p^k is the distribution after k MCMC iterations, and $\langle f(y) \rangle_{\mathbf{P}}$ denotes the average of $f(y)$ over the samples $\mathbf{P}$.

3.2. The Motion Priors

As shown in Equation 4, the registration parameter is M and the desired average image is y_0. Given observed data I_0, we can write the posterior distribution over the unknowns M and y_0 using Bayes' Rule:

$$p(y_0, M|I_0) \propto p(I_0|y_0, M)p(y_0)p(M) \qquad (8)$$

The data likelihood function is as follows:

$$p(I_0|y_0, M) = (\frac{\theta}{2\pi})^l exp\{-\frac{\theta}{2}\|I_0 + \frac{1}{n}BM - y_0\|^2\} \qquad (9)$$

where l is the number of pixels in I_0.

If the camera motion is small in a short sequence, the mean translation estimation M_0 can be thought as 0. Otherwise, pre-registration can be done with existing methods [17, 1] to obtain an initial estimate of M_0. Motivated by the work [14], we assume that the uncertainty in the registration parameters can be modeled by a Gaussian perturbation about the mean estimation M_0 with covariance matrix S, which we restrict to be a diagonal matrix. This assumption is reasonable because we can always find an average image and align each image frame with this average image. During the alignment, if some image frames are translated/rotated to the left, then other images should be translated/rotated to the right. Thus, we can approximate the uncertain motion parameters M with a Gaussian distribution.

$$M = M_0 + m(t), m(t) \sim \mathcal{N}(0, \mathbf{S})$$
$$p(M) = (\frac{|\mathbf{S}^{-1}|}{(2\pi)^n})^{\frac{1}{2}} exp\{-\frac{1}{2}m(t)^T\mathbf{S}^{-1}m(t)\} \qquad (10)$$

3.3. The Dynamic Priors

When we know $y(t)$, the dynamic texture model can be derived from equation 2 as follows:

$$y(t) = Cx(t) + w(t), \quad w(t) \sim \mathcal{N}(0, \beta^{-1}I)$$
$$x(t+1) = Ax(t) + v(t), \quad v(t) \sim \mathcal{N}(0, \alpha^{-1}I) \quad (11)$$

where $y(t)$ is a $l \times 1$ vector and $x(t)$ is a $k \times 1$ vector. Define Y as a $l \times n$ matrix, $Y = [y(1), ..., y(n)]$. Define X as a $k \times n$ matrix, $X = [x(1), ..., x(n)]$. In previous solutions for dynamic texture [4, 6, 17], no dynamic prior has been explicitly proposed. It means that the dynamic matrix A and the mapping matrix C have to be explicitly estimated. However, accurately estimating these matrices is not an easy problem without knowing some priors. Wang [18] et al. developed a dynamic prior model by marginalizing over the uncertain dynamic parameter A and the mapping parameter C. This results in a nonparametric model for dynamical systems that account for uncertainty. With the dynamic priors, Gaussian Process Dynamical Models (GPDM) are developed for tracking people and data-driven animation. Using similar ideas, the Marginal Auto-Regressive (MAR) models are developed as dynamic prior models for tracking in [12]. In the extreme case, the MAR models describe all stable AR models. Thus, it is weakly-parametric and can also be used as a prior for any image sequence of dynamic textures.

From equation 11, we know that X is embedded in Y according to the linear mapping matrix C with dimension $l \times k$. Motivated by the ideas of GPDM and MAR model [12, 18], we assume C is a stochastic matrix whose elements $c_{ij} \sim \mathcal{N}(0, 1)$. To recover the embedded sequence X from the data sequence Y, we consider all possible mappings C instead of a specific mapping C. By marginalizing over all possible mappings C, a marginal Gaussian Process mapping is as follows:

$$p(Y|X, \mu) = \int_C p(Y|X, C)p(C|\beta)dC$$
$$= (2\pi)^{-\frac{ln}{2}}|K_{yx}|^{-\frac{l}{2}} exp\{-\frac{1}{2}Y^T K_{yx}^{-1} Y\}$$
$$K_{yx} = X^T X + \beta^{-1} I \qquad (12)$$

In this formulation, the mapping between X and Y only depends on K_{yx}.

We also can assume A is a stochastic matrix whose elements $a_{ij} \sim \mathcal{N}(0, 1)$ and the initial condition $x_0 \sim \mathcal{N}(0, I)$. suppose $X = [x(1), x(2), ..., x(n)]$ and $X_\triangle = [x_0, x(1), ..., x(n-1)]$. Then, a marginal distribution of the AR model is:

$$p(X|x_0, \omega) = \int_A p(X|A, x_0)p(A|\alpha)dA$$
$$= (2\pi)^{-\frac{kn}{2}}|K_{xx}|^{-\frac{k}{2}} exp\{-\frac{1}{2}X^T K_{xx}^{-1} X\}$$
$$K_{xx} = X_\triangle^T X_\triangle + \alpha^{-1} I \qquad (13)$$

Intuitively, this model favors smooth sampling in the space of X. Hence, the joint distribution of X and Y is:

$$p(Y, X|\alpha, \beta) = P(X|\alpha)P(Y|X, \beta) \qquad (14)$$

where α and β are the hyperparameters of this joint distribution.

4. Joint Optimization for Registration

The registration of moving dynamic textures can be modeled as a joint optimization problem. It involves optimizing the camera motion, the average image, and the dynamics. With the priors introduced in equations 7, 10 and 14, a straightforward approach to this problem is to solve for the maximum a-posteriori (MAP) solution.

Given an input image sequence $I(1), I(2), ..., I(n)$ and the mean image I_0 of this image sequence, we can write the posterior distribution over the unknowns using Bayes' Rule from equations 4 and 5:

$$p(M, y_0, Y, X|I, I_0) \propto p(I_0, I|y_0, M, Y, X)$$
$$p(y_0)p(M)p(Y)p(X) \qquad (15)$$

where I is $[I(1), ..., I(n)]$. If we suppose that the average image y_0 and the dynamic latent X of the dynamic texture independently impose constraints on the possible camera motion for registration, we can obtain the following approximation:

$$p(I_0, I|y_0, M, Y, X) \approx p(I_0|y_0, M)p(I|y_0, M, Y, X) \quad (16)$$

This is equivalent to solving a regularized-least squares problem. It attempts to register images in the sequence to obtain an average image with desired image statistics and a dynamic texture with stable dynamics. The data likelihood given the camera motion, the average image and others is described as follows:

$$p(M, y_0, Y, X|I, I_0) \propto p(I_0|y_0, M)p(I|y_0, M, Y)$$
$$p(y_0)p(M)p(Y|X)p(X) \quad (17)$$

We tried to optimize the above with conjugate gradient searching but found that it failed. One likely reason is that directly optimizing so many uncertain variables makes the MAP objective function very susceptible to local minima.

To solve this problem, we adopt a Bayesian approach to marginalize out the unknown motion parameters. This gives the marginal likelihood as follows:

$$p(y_0, Y, X|I, I_0) \propto q(y_0, Y)p(y_0)p(Y|X)p(X)$$
$$q(y_0, Y) = \int_M p(I_0|y_0, M)p(I|y_0, M, Y)p(M)dM \quad (18)$$

Maximizing the posterior probability of the proposed generative model is not trivial. In order to make the optimization practical, we refrain from using complex inference algorithms. Instead we perform a gradient ascent on the logarithm of the marginal likelihood. This maximization of the marginal likelihood is done using the scaled conjugate gradients algorithm (SCG) [13]. With the finally estimated average image y_0 and the latent appearance $[y(1), ..., y(n)]$ of the dynamic texture, the explicit AR model can be easily reconstructed using ML estimation on the sequence X:

$$A^* = X^T X_\triangle (X_\triangle^T X_\triangle)^{-1} \quad (19)$$

In the following, we summarize the proposed approach for registration of moving dynamic textures:

1. Compute the first-order image derivatives in four directions for all n frames in I(t). As introduced in section 3.1, the Contrastive Divergence algorithm is used to learn our image prior model parameters for the input image sequence.
2. If the input sequence is long, divide it into many short image sequences, each with length τ along the time axis, as done in [17].
3. For each time window $[t - \tau + 1, t]$, we apply an existing texture registration algorithm (such as [17] or [1, 2]) to perform initialization.
4. For each time window $[t - \tau + 1, t]$, we perform model optimization with the priors introduced in sections 3 and 4 until model convergence.
5. With the optimal y_0, Y and X estimated from the previous step, the motion parameters M is then found as the mode of the full posterior, which can be obtained iteratively by Maximum Likelihood estimation using SCG optimization.

5. Experiments

The first set of experiments uses 3 real video sequences in [10]. They are image sequences of Waterfall A, Grass and Pond, respectively. These image sequences are captured by a fixed camera. In order to evaluate the proposed approach, we generate 3 new video sequences by transforming each frame with known motion to simulate the moving dynamic textures. For convenience, each generated video sequence has 100 frames and the size of each image frame in the sequences is 120×160 (shown in Figure 4).

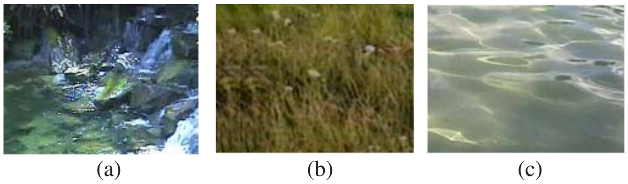
(a) (b) (c)

Figure 4. The synthesized moving dynamic textures from video sequences of (a) Waterfall A, (b) Grass, (c) Pond.

For comparison, we implemented Vidal's motion estimation method ourselves [17] because there is no public code available. We also implemented the robust motion estimation method [2] with a hierarchical motion model [1]. In the following, we call these two previous methods Vidal'method and Hybrid method, respectively. Since Vidal's method does not perform motion estimation in the beginning frames of a video sequence, we only compare the motion estimations from the 21st to the 80th frame in this experiment. Moreover, the ground truth of camera motions between neighboring frames are the same in all three testing video sequences and the ground truth motion profiles are shown in figure 7 (a). One can see that, in the video segments of interest that contain 60 frames each, there is no camera motion from the 1st frame to the 20th frame and from the 41st to 60th frame, while there are constant translation camera motions with speed $[1, 0]$ from the 21st to 40th frames.

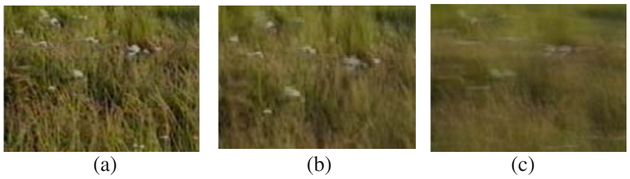
(a) (b) (c)

Figure 5. The average image before and after registration with the proposed approach. (a) one image frame, (b) the average image after registration, (c) the average image before registration.

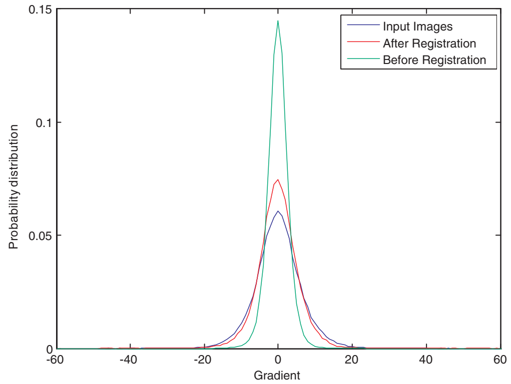

Figure 6. The statistics on input image frames, on the average image after registration, and on the average image before registration.

We first test the proposed approach on the synthetic Grass sequence. The testing results are shown in figure 5. One can see that the average image after registration is much sharper than that before registration. Moreover, the statistics of derivative filter responses in the average image after registration is much closer to that in the input images than the average image before registration, as shown in figure 6. This not only illustrates the efficiency of the proposed method but also shows that our priors regarding registration and the average image statistics are correct.

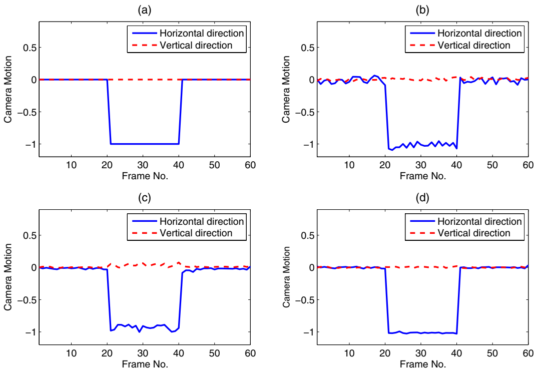

Figure 7. The comparison of estimated motions with the ground truth in the Waterfall A sequence. (a) ground truth, (b) by the hybrid method,(c) by Vidal's, (d) by the proposed method.

Figure 7 shows the motion estimations for the Waterfall A sequence by three different algorithms: hybrid method, Vidal's method, and our method. All three methods generated motion estimations that are close to the ground truth in this sequence. However, the solution given by the Hybrid method has larger variance than others while the recovered motions by Vidal's method are obviously smaller than the ground truth between the 21st and 40th frames. It is easy to interpret these trends in the results by the hybrid and Vidal's methods. The hybrid method is based on the assumption of Brightness Constancy, hence it considers that the local/nonrigid motion in dynamic textures is also caused by camera motion, which makes the estimated camera motion fluctuant. On the contrary, Vidal's method assumes that there is no camera motion in a short sequence hence it unconsciously contributes part of the effect by camera motion to nonrigid motion caused by dynamic textures; this makes Vidal's method tend to under-estimate camera motion. Only the proposed method obtains a solution that is not only the closest to the ground truth but also has small variance, thanking to the proposed generative image model that can model both camera motion and dynamic texture simultaneously.

After recovering the camera motion parameters, we recompute the average image to evaluate the performance of the registration method. As introduced above, the average image should be sharper if the image sequence has been registered with more accurate motion estimation. Figure 8 shows the average images of the video sequence from the 21st frame to the 40th frame after registration by the three methods. Figure 8 also shows the corresponding statistics of their derivative filter responses. It is quite obvious that the average image by our method is sharper than the average images produced by the other two methods. This is due to the explicit prior constraints between registration and statistics of the average image imposed by the proposed model.

Sequence	Waterfall A	Grass	Pond
Hybrid [1, 2]	9.29%	16.86%	13.25%
Vidal's [17]	6.18%	13.56%	10.25%
Proposed	4.63%	9.32%	6.07%

Table 1. False Estimation Fraction (FEF) of motion parameters.

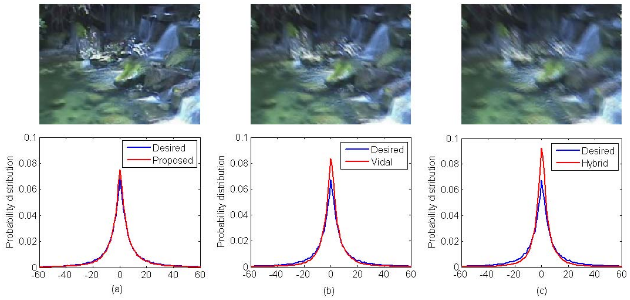

Figure 8. The average images and related distributions (21st frame to 40th frame) after registration by different approaches. (a) by the proposed method, (b) by Vidal's, (c) by the hybrid method.

In order to quantitatively evaluate the performance of camera motion estimation, we define the false estimation fraction (FEF) to indicate the difference between the ground-truth camera motion M_{True} and the estimated camera motion M_{Est}:

$$FEF = \frac{|sum(|M_{Est}|) - sum(|M_{True}|)|}{sum(|M_{True}|)} \qquad (20)$$

Table 1 records the FEF of registration results on the three image sequences (i.e. Water Fall A, Grass, Pond) by different implementations. Some parts of the scene captured in the Waterfall A sequence are static, hence registration is relatively easier for this sequence; it is why all three methods achieved better motion estimation results on this sequence, although the hybrid method's performance is a little worse. The scenes are dynamic everywhere in the Grass and Pond sequences. The Grass sequence is especially challenging to register because the grass waves rapidly in the sequence, which causes large appearance variance and uncertainty. For the Grass sequence, the hybrid method didn't give good results because the brightness constancy assumption is greatly violated. Better results are obtained with Vidal's method and the proposed method because they explicitly model the grass dynamics with a dynamic texture model. In particular, our proposed method achieved the best results because of its utilization of powerful priors and the generative image model.

Figure 9. A sequence of moving flower bed [6, 17].

Besides the three video sequences with synthetic camera motion, we also did experiments on the flower bed sequence with real camera motion shown in Figure 9. This sequence was used as an example for registering moving dynamic textures in [6, 17]. The camera motion in this sequence is a horizontal translation. The ground truth of the translation motion is obtained by manually labeling the locations of one red flower in the first frame and the last frame. The whole sequence includes 554 image frames. While quantitative motion estimation results were not reported in [6], Vidal et al. [17] showed quantitative results on a subsequence with 250 frames out of the whole sequence. It was reported that the cumulative displacement on that subsequence estimated by their approach [17] is 60 pixels along the horizontal direction while the ground truth is 85 pixels; thus the FEF of the cumulative motion there is 29.41%. Since we do not have access to the exact 250-frame subsequence, we ran our algorithm on the entire sequence of 554 frames. The cumulative motion along the horizontal direction is estimated as 104.52 pixels by our approach while the ground truth is 110 pixels based on manual motion labeling of one red flower; thus the FEF of cumulative motion by our approach is 4.98%.

While our method consistently achieves the best accuracy among all three methods, our approach has close ties with the other two. For instance, we depend on either Vidal's or the hybrid method to initialize our model; we take the dynamic texture constancy constraints (DTCC) idea from [17]. Our approach is slightly more computationally expensive than Vidal's because of the initialization and optimization. The main contributions of our method for better accuracy are the novel use of average image, motion and dynamic priors, and the avoidance of assumptions that may lead to over- or under- estimations by marginalizing over registration and latent dynamic texture parameters instead of explicitly estimating them.

6. Conclusions

In this paper we have proposed a new approach to registration of moving dynamic textures, based on two proposed criteria for registration: 1) registration should simplify the dynamic texture model while preserving all useful information; 2) better registration should lead to a sharper average image, whose statistics are closer to those of the input image frames. With the proposed image priors and dynamic priors about moving dynamic textures, we are able to effectively perform joint optimization by marginalizing over the unknown registration parameters as well as the dynamics of latent dynamic textures. Experiments on various real video sequences demonstrate the performance of our method and show marked improvement over previous approaches.

References

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

[2] M. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proceedings of ICCV*, pages 231–236, 1993.

[3] A. Chan and N. Vasconcelos. Layered dynamic textures. In *Proceedings of NIPS*, 2005.

[4] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.

[5] G. Doretto and S. Soatto. Editable dynamic textures. In *Proceedings of CVPR*, pages 137–142, 2003.

[6] A. Fitzgibbon. Stochastic rigidity: Image registration for nowhere-static scenes. In *Proceedings of ICCV*, pages 662–669, 2001.

[7] G. Hinton. Training products of experts by minimizing contrastive divergence. In *Neural Computation*, pages 1771–1800, 2002.

[8] J. Huang and D. Mumford. Statistics of natural images and models. In *Proceedings of CVPR*, pages 1541–1547, 1999.

[9] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *Proceedings of ICCV*, pages 959–966, 1998.

[10] V. Kwatra, A. Schdl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: Image and video synthesis using graph cuts. In *Proceedings of SIGGRAPH*, pages 277–286, 2003.

[11] W. Lin and Y. Liu. A lattice-based mrf model for dynamic near-regular texture tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 29, pages 777–792, 2007.

[12] K. Moon and V. Pavlovic. Impact of dynamics on subspace embedding and tracking of sequences. In *Proceedings of CVPR*, pages 198–205, 2006.

[13] I. Nabney. Netlab algorithms for pattern recognition, 2002.

[14] L. Pickup, D. Capely, S. Roberts, and A. Zisserman. Bayesian image super-resolution, continued. In *Proceedings of NIPS*, 2006.

[15] A. Rav-Acha, Y. Pritch, and S. Peleg. Online registration of dynamic scenes using video extrapolation. In *Workshop on Dynamical Vision at ICCV*, pages 151–164, 2005.

[16] S. Roth and M. Black. Fields of experts: A framework for learning image priors. In *Proceedings of CVPR*, pages 860–867, 2005.

[17] R. Vidal and A. Ravichandran. Optical flow estimation and segmentation of multiple moving dynamic textures. In *Proceedings of CVPR*, pages 516–521, 2005.

[18] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *Proceedings of NIPS*, 2005.

[19] M. Welling, G. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. In *Proceedings of NIPS*, pages 1359–1366, 2003.

[20] O. Woodford, I. Reid, P. Torr, and A. Fitzgibbon. Fields of experts for image-based rendering. In *Proceedings of BMVC*, 2006.

[21] A. Yezzi and S. Soatto. Deformotion: deforming motions and shape averages. In *International Journal of Computer Vision*, volume 53, pages 153–167, 2003.

[22] L. Yuan, F. Wen, C. Liu, and H. Shum. Synthesizing dynamic texture with closed-loop linear dynamic system. In *Proceedings of ECCV*, pages 603–616, 2004.

[23] L. Zelnik-Manor and M. Irani. Multi-frame alignment of planes. In *Proceedings of CVPR*, pages 1151–1156, 1999.