Diversified Dynamical Gaussian Process Latent Variable Model for Video Repair

Hao Xiong, Tongliang Liu and **Dacheng Tao**
Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney
hao.xiong@student.uts.edu.au
tliang.liu@gmail.com
dacheng.tao@uts.edu.au

Abstract

Videos can be conserved on different media. However, storing on media such as films and hard disks can suffer from unexpected data loss, for instance from physical damage. Repair of missing or damaged pixels is essential for video maintenance and preservation. Most methods seek to fill in missing holes by synthesizing similar textures from local or global frames. However, this can introduce incorrect contexts, especially when the missing hole or number of damaged frames is large. Furthermore, simple texture synthesis can introduce artifacts in undamaged and recovered areas. To address aforementioned problems, we propose the diversified dynamical Gaussian process latent variable model (D^2GPLVM) for considering the variety in existing videos and thus introducing a diversity encouraging prior to inducing points. The aim is to ensure that the trained inducing points, which are a smaller set of all observed undamaged frames, are more diverse and resistant for context-aware and artifacts-free based video repair. The defined objective function in our proposed model is initially not analytically tractable and must be solved by variational inference. Finally, experimental testing illustrates the robustness and effectiveness of our method for damaged video repair.

Introduction

Videos, from black and white films to modern blockbusters, are an indispensible part of contemporary life. The celluloid or hard disks used to store videos inevitably suffer from natural deterioration or, sometimes, deliberate damage. Such damage normally appears as scattered missing blobs or large holes in video frames, which has an adverse impact on the appreciation and research value of the videos. It is, therefore, useful to recover damaged parts of videos. However, this is not trivial, especially when the missing areas are large.

One major difficulty in video repair lies in the uncertainty and variety of video scenes and the fact that objects are often highly dynamic in the video sequence. A naive approach to video repair is to regard the damaged frame as a single image and then applying image inpainting, such as in (Criminisi, Perez, and Toyama 2004), to recover the missing part. Although such inpainting techniques retain linear structures like object contours, they tend to fill in the holes by synthesizing neighboring texture without considering temporal information from other video frames. Since neighboring information is limited, these methods are likely to repair the video in the wrong context. Other methods (Wexler, Shechtman, and Irani 2004) have been proposed to fill in missing holes by sampling and synthesizing a number of patches from neighboring frames. However, this is still not always reliable, since the video sequences are likely to be dynamic or complex. In such circumstances, a scene presents in one frame may disappear in other frames and, sometimes, the patches for synthesis are too small to satisfy larger holes. Furthermore, the aforementioned approaches are highly dependent on texture synthesis techniques, which can create saw-shaped artifacts at the boundaries of the original region and the recovered part in a single frame.

Recently, Damianou *et al.* (Damianou, Titsias, and Lawrence 2014) developed a dynamical Gaussian process latent variable model (DGPLVM) to smoothly repair video without saw-shaped artifacts in the presence of missing pixels. To achieve this, a Gaussian process (GP) prior based on auxiliary inducing points was introduced so that the variational Bayes approach was tractable. The latent variables were then variationally integrated out and a closed-form lower bound on a marginal likelihood function computed. The original purpose of inducing points in (Csato 2002; Csato and Opper 2002; Seeger, Williams, and Lawrence 2003; Snelson and Ghahramani 2006; Candela and Rasmussen 2005; Titsias 2009) was to speed up computation by regarding it as a smaller set representing the entire observed frames. In this approach, the inducing point was also critical to obtaining a closed-form lower bound of the defined log likelihood function to render the model robust to overfitting.

However, it transpires that (Damianou, Titsias, and Lawrence 2014) tends to generate relatively similar inducing points that can introduce ghost effects in recovered frames. This is because the inducing points trained in DGPLVM have a tendency to focus on frequent shots or those with salient features within the scenes. We propose D^2GPLVM to address this problem, in which a diversity encouraging prior is applied to the inducing points so that they are more diverse and capture more distinct scenes from the observed complete frames. Since inducing points are pivotal to predicting missing pixels, more resistant inducing points en-

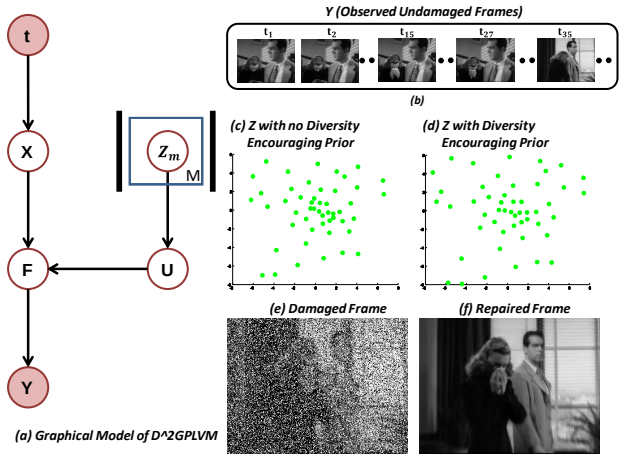

Figure 1: A schematic of the porposed D^{2}GPLVM. (a) A graphical model of the proposed method. (b) The training frames. (c)-(d) A comparison of the generated Z with and without the diversity encouraging prior. (e) A random damaged frame. (d) The image repaired using our method.

hance video repair performance. However, directly integrating out latent variables is infeasible in our model since it is nonlinear. Thus, the variational inference approach is applied to approximate the marginal likelihood by maximizing a Jensen's lower bound and then integrate the latent variables. The parameters of D^{2}GPLVM are optimised by maximizing the Jensen's lower bound using scaled conjugate gradient (SCG). Afterwards, a given damaged video can be repaired by the trained D^{2}GPLVM regardless of its damaged area and form. In the remainder of this paper, we first review the DGPLVM and then introduce our D^{2}GPLVM. To demonstrate the robustness of D^{2}GPLVM, we perform experiments on a **hundred** of various video clips from black and white films to modern movies.

Related Work

Completion approaches are broadly classified into image inpainting and video repairing. In this section, we review works related to both of these approaches.

Image Inpainting

Most inpainting approaches (Drori, Cohen-Or, and Yeshurun 2003; Sun et al. 2005; Bertalmio and Sapiro 2000; Bertalmio et al. 2009; Mo, Lewis, and Neumann 2004; Sobiecki et al. 2013; Min and Dugelay 2013) exploit consistency in neighboring pixels or textures to recover the missing parts. However, these methods are inclined to fail when the missing region is inhomogeneous with its surroundings. Other methods (Hosoi et al. 2011; Amano and Sato 2002; Amano 2004) have attempted to deal with various real-world objects inpainting tasks by defining a similarity term between the input image and the base eigenvector derived by applying PCA to a subspace of the training samples. Based on (Amano and Sato 2002; Amano 2004), (Hosoi et al. 2011) was able to inpaint any object without first specifying the object class beforehand and could inpaint in real time. However, these methods are still imperfect since the damaged objects class for inpainting is much more diverse than expected.

In general, image inpainting exploits neighboring texture information to fill in the missing pixels. These inpainting techniques can work effectively under the circumstances that the missing area in an image is small. This is to say that image inpainting is not applicable to video repair since video context can be complex and the damaged region may be too large. As a result, the recovered videos may have wrong contexts using image inpainting (see Fig.3).

Video Repair

The video repair method proposed in (Jia et al. 2004) inferred occluded background and large motion by sampling and aligning movels (structured moving objects) from the captured video. Missing static background was repaired by constructing a layered mosaic in addition to image repair (Jia and Tang 2004). To repair moving pixels, an optimal alignment based on a homographic transform was computed by assuming that the moving pixels were projections of cyclic motions ((Seitz and Dyer 1997)), where cyclic motions were detected by time-frequency analysis (Cutler and Davis 2009). As an extension of (Jia et al. 2004), (Jia et al. 2006) used tensor voting (Medioni, Lee, and Tang 2000) to address the pertinent spatio-temporal issues in background and motion repair. Variable illumination and moving cameras were also studied. However, although (Jia et al. 2006) utilized texture synthesis to recover the occluded area, it is possible that the occluded region texture may not appear in the neighbouring pixels or frames, resulting in inaccurately repaired pixels with artifacts.

In (Wexler, Shechtman, and Irani 2004), Wexler *et al.* attempted to fill in missing portions by sampling spatio-temporal patches from the input video. Their method defined and used similarity measurements in space and time domains. This method was successful due to the judicious extension of (Efros and Leung 1999) in which non-parametric sampling was used to handle spatial and temporal information simultaneously. They demonstrated that the patches selected for completion may contain errors if the background is complex (e.g., non-textured) and the result will not preserve speed irregularities and may destroy complex structures.

Unlike image inpainting, these method unexceptionally exploited the global temporal information to achieve video repair. However, this is still problematic since a video may include complex objects and scenes. In that case, simply searching and synthesizing similar patches from neighboring frames can introduce severe artifacts in the observed and recovered areas. Hence, for the scenes containing complicated objects, such methods may not generate a smooth repaired video by preserving the structure of objects.

Dynamical GPLVM

In this section, the basic concepts of dynamical Gaussian process latent variable model (DGPLVM) are first introduced for better comprehension of our model.

In DGPLVM, $Y \in R^{n\times p}$ (with columns $\{y_{:,j}\}_{j=1}^{p}$) denotes the observed data where n is the number of data points and p is the dimensionality of each data point in Y. Here, Y is associated with latent variables $F \in R^{n\times p}$ (with columns $\{f_{:,j}\}_{j=1}^{p}$), which is the same size as Y and

$$p(Y|F) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|f_{:,j}, \sigma^{-1} I_N). \quad (1)$$

For the sake of dimensionality reduction, another latent variable $X \in R^{n\times q}$ ($q \leq p$) is introduced, such that

$$p(F|X) = \prod_{j=1}^{p} \mathcal{N}(f_{:,j}|0, K_{ff}). \quad (2)$$

Here, K_{ff} is a $n \times n$ kernel matrix and the kernel function here is an exponentiated quadratic (RBF) as follows:

$$k(x_{i,:}, x_{k,:}) = \sigma_f^2 \exp\Big(-\frac{1}{2}\sum_{j=1}^{q} \alpha_j (x_{i,j} - x_{k,j})^2\Big), \quad (3)$$

where each $x_{i,:}$ is the i^{th} row of X and $K_{ff} = k(x_{i,:}, x_{k,:})$.

In DGPLVM, each datapoint $y_{i,:}$ is observed at corresponding time t_i. Therefore, the prior distribution of X is:

$$p(X) = \prod_{j=1}^{q} \mathcal{N}(x_{:,j}|0, K_x), \quad (4)$$

where $x_{:,j}$ refers to one column of X and $K_x = k(t_i, t'_i)$ is the covariance matrix obtained by evaluating the covariance function k on the observed times t.

Furthermore, (Damianou, Titsias, and Lawrence 2014) introduced the auxiliary inducing variable $U \in R^{m\times p}$, which is a set of m inducing points $u_{i,:} \in R^p$, evaluated at their associated inducing input locations $Z \in R^{m\times q}$(with columns $\{z_{:,j}\}_{j=1}^{q}$). Likewise,

$$p(U) = \prod_{j=1}^{p} N(0, K_{uu}). \quad (5)$$

Here, $K_{uu} = k(z_{i,:}, z_{k,:})$. The introduced inducing points can not only speed up computation but also render the objective function tractable. Interest readers may refer to (Damianou, Titsias, and Lawrence 2014) for more details.

Diversified Dynamical GPLVM

In DGPLVM, Damianou *et al.* proposed a Bayesian approach to train the GPLVM. By taking inducing points into consideration, the input variables of Gaussian process could be robustly integrated out. Then, the trained model was used to recover videos in the presence of missing pixels. However, the inducing points trained from DGPLVM were highly similar and could not fulfill the repair of videos with more complex scenes.

Here, we propose the D^{2}GPLVM with repect to the repulsion property of inducing points due to the fact that real world videos may have much more diverse scenes than expected. To be exact, given a video with an unknown number of damaged frames, let $Y \in R^{n\times p}$ denote all undamaged frames, where n and p are the number of undamaged frames and pixels in the video respectively. F is the noise-free version of Y, whilst X is the reduced dimension version of Y. Since a video is time sequential, the aforementioned time t_i is referred to as the frame serial number.

Remember that the inducing points $U \in R^{m\times p}$ are a small set representing the entire undamaged frames. Meanwhile, the distribution $p(U)$ has a covariance matrix K_{uu} defined in Eq. (5). Here, a diversity prior of covariance matrix K_{uu} would be modeled by:

$$p(U \in Y) = |K_{uu}|, \quad (6)$$

where $|K_{uu}|$ refers to the determinant of matrix K_{uu}. The inducing points selected with respect to such prior can cover multiple distinct scenes of a video instead of focusing on the most salient ones.

Therefore, the new objective function is:

$$F(\theta) = \log P(Y) + \lambda \log |K_{uu}|, \quad (7)$$

where $\theta = \{\sigma_f^2, \sigma, \alpha_1, ..., \alpha_j\}$ are the hyperparameters in our proposed model using the same symbols as in DGPLVM. Furthermore, $\lambda > 0$ is used to balance the weights between measurements of likelihood and the diversity encouraging prior.

Variational Inference

The new objective function $F(\theta)$ can be factorized as:

$$F(\theta) = \log \int |K_{uu}|^{\lambda} \prod_{j=1}^{p} p(y_{:,j}|f_{:,j}) \Big(p(f_{:,j}|u_{:,j}, X)p(X)\mathrm{d}X\Big)p(u_{:,j})\mathrm{d}U\mathrm{d}F. \quad (8)$$

Note that integration over X is unfeasible since X is an input, in a rather complex non-linear manner, of $p(f_{:,j}|u_{:,j}, X)$, which contains the kernel matrix K_{ff}.

To see that, the specific form of $p(f_{:,j}|u_{:,j}, X)$ is derived based on Eq. (2) and Eq. (5):

$$\begin{aligned} p(F|U, X) &= \prod_{j=1}^{p} p(f_{:,j}|u_{:,j}, X) \\ &= \prod_{j=1}^{p} p(f_{:,j}|a_j, \Sigma_f), \end{aligned} \quad (9)$$

where $a_j = K_{fu}K_{uu}^{-1}u_{:,j}$, $\Sigma_f = K_{ff} - K_{fu}K_{uu}^{-1}K_{uf}$.

Thus, we use variational distribution $q(F, U, X)$ to approximate the true posterior $P(F, U, X|Y)$ with the form:

$$q(F, U, X) = \left(\prod_{j=1}^{p} p(f_{:,j}|u_{:,j}, X)q(u_{:,j})\right)q(X). \quad (10)$$

Here, $q(X)$ is a variational distribution that follows:

$$q(X) = \prod_{i=1}^{n} N(x_{i,:}|\mu_{i,:}, S_i), \quad (11)$$

where each covariance matrix S_i is diagonal. Another variational distribution $q(U)$ is arbitrary and will be explained later. In terms of Jensen's inequality, the lower bound $F\big(q(X), q(U)\big)$ of the objective function could be derived by:

$$F\big(q(X), q(U)\big) = \int q(F, U, X) \log \frac{|K_{uu}|^{\lambda} p(Y)}{q(F, U, X)} \mathrm{d}X \mathrm{d}F \mathrm{d}U. \tag{12}$$

After inserting Eq. (10) into Eq. (12), we have:

$$F\big(q(X), q(U)\big) = \int \prod_{j=1}^{p} p(f_{:,j}|u_{:,j}, X) q(u_{:,j}) q(X) \log \frac{|K_{uu}|^{\lambda} \prod_{j=1}^{p} p(y_{:,j}|f_{:,j}) p(u_{:,j}) p(X)}{\prod_{j=1}^{p} q(u_{:,j}) q(X)} \mathrm{d}X \mathrm{d}F \mathrm{d}U. \tag{13}$$

Let $\langle \cdot \rangle_p$ be a shorthand for expectation with respect to the distribution p, then:

$$F(q(X), q(U)) = \sum_{j=1}^{p} \left(\int q(u_{:,j}) q(X) \langle \log p(y_{:,j}|f_{:,j}) \rangle_{p(f_{:,j}|u_{:,j}, X)} \mathrm{d}X \mathrm{d}u_{:,j} + \langle \log \frac{p(u_{:,j})}{q(u_{:,j})} \rangle_{q(u_{:,j})} \right) - \int q(X) \log \frac{p(X)}{q(X)} \mathrm{d}X + \lambda \log |K_{uu}|. \tag{14}$$

Since $K_{uu} = k(z_{i,:}, z_{k,:})$, the term $\lambda \log |K_{uu}|$ can be placed outside the integral. To simplify the sum term at the beginning of Eq. (14), let the formulation within the sum notation be denoted by:

$$\hat{F}_j\big(q(X), q(U)\big) = \langle \log \frac{p(u_{:,j})}{q(u_{:,j})} \rangle_{q(u_{:,j})} + \int q(u_{:,j}) q(X) \langle \log p(y_{:,i}|f_{:,j}) \rangle_{p(f_{:,j}|u_{:,j}, X)} \mathrm{d}X \mathrm{d}u_{:,j}. \tag{15}$$

Thus, the lower bound $F\big(q(X), q(U)\big)$ may be re-expressed in the form:

$$F\big(q(X), q(U)\big) = \sum_{j=1}^{p} \hat{F}_j\big(q(X), q(U)\big) - KL\big(q(X)||p(X)\big) + \lambda \log |K_{uu}|. \tag{16}$$

Now, the lower bound $F\big(q(X), q(U)\big)$ consists of three parts: $\hat{F}_j\big(q(X), q(U)\big)$, $KL\big(q(X)||p(X)\big)$ and the log likelihood of the diversity encouraging prior $\lambda \log |K_{uu}|$. Since both of $q(X)$ and $p(X)$ are Gaussian distributions, the KL term can easily be calculated:

$$KL\big(q(X)||p(X)\big) = \frac{1}{2} \sum_{i=1}^{n} tr\big(\mu_{i,:} \mu_{i,:}^T + S_i - log S_i\big) - \frac{nq}{2} \tag{17}$$

According to the definitions of $q(X)$ and distributions in Eq. (5) , $\hat{F}_j\big(q(X), q(U)\big)$ can be expressed as:

$$\hat{F}_j(q(X), q(U)) = \frac{1}{2\sigma^2} tr\big(\langle K_{ff} \rangle_{q(X)}\big) - \frac{1}{2\sigma^2} tr\big(K_{uu}^{-1} \langle K_{uf} K_{fu} \rangle_{q(X)}\big) + \int q(u_{:,j}) \log \frac{e^{\langle \log N(y_{:,j}|a_j, \sigma^2 I_p) \rangle_{q(X)}} p(u_{:,j})}{q(u_{:,j})} \mathrm{d}u_{:,j} \tag{18}$$

Note that there is a KL-like quantity in the Eq. (18), such that the optimal $q(u_{:,j})$ is supposed to be proportional to:

$$q(u_{:,j}) \propto e^{\langle \log N(y_{:,j}|a_j, \sigma^2 I_p) \rangle_{q(X)}} p(u_{:,j}). \tag{19}$$

$\hat{F}_j(q(X), q(U))$ can be upper bounded by $\hat{F}_j(q(X))$ after applying the *reversing Jensen's inequality* (King and Lawrence 2006) to the KL-like quantity containing $q(u_{:,j})$:

$$\hat{F}_j(q(X)) = \frac{1}{2\sigma^2} tr\big(\langle K_{ff} \rangle_{q(X)}\big) - \frac{1}{2\sigma^2} tr\big(K_{uu}^{-1} \langle K_{uf} K_{fu} \rangle_{q(X)}\big) + \log \int e^{< \log N(y_{:,j}|a_j, \sigma^2 I_p) >_{q(X)}} p(u_{:,j}) \mathrm{d}u_{:,j}. \tag{20}$$

Now $q(U)$ is optimally eliminated, $\hat{F}_j(q(X))$ can be calculted as follows:

$$\hat{F}_j(q(X)) = \left[\log \frac{\sigma^{-n} |K_{uu}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}} |\sigma^{-2} \psi_2 + K_{uu}|^{\frac{1}{2}}} e^{-\frac{1}{2} y_{:,j}^T W y_{:,j}} \right] - \frac{\psi_0}{2\sigma^2} + \frac{1}{2\sigma^2} tr(K_{uu}^{-1} \psi_2), \tag{21}$$

where $\psi_0 = tr\big(\langle K_{ff} \rangle_{q(X)}\big)$, $\psi_1 = \langle K_{fu} \rangle_{q(X)}$, $\psi_2 = \langle K_{uf} K_{fu} \rangle_{q(X)}$ and $W = \sigma^{-2} I_n - \sigma^{-4} \psi_1 (\sigma^{-2} \psi_2 + K_{uu}^{-1}) \psi_1^T$.

With $\hat{F}_j\big(q(X), q(U)\big)$ and $KL\big(q(X)||p(X)\big)$ in hand, we can optimize the parameters θ in our model using a gradient-based algorithm.

Repairing Damaged Video

A set of partially observed frames $Y_* = \{Y_*^u, Y_*^o\}$ in a video sequence is given in the prediction stage. Here, Y_*^o denotes the observed part in the video frames and Y_*^u refers to the missing part. Our task is to calculate the following predictive density:

$$P(Y_*|Y) \approx \int P(Y_*|F_*) q(F_*|X_*) q(X_*) \mathrm{d}X_* \mathrm{d}F_*. \tag{22}$$

Like F and X, F_* and X_* are namely the latent variables of new testing data Y_*. To compute above, we need to optimise with respect to the parameters (u_*, S_*) of the Gaussian variational distribution $q(X_*)$. The standard GP prediction is employed here to obtain $q(X_*)$ which can be further factorized as follows:

$$q(X_*) = \int p(X_*|X) q(X) \mathrm{d}X, \tag{23}$$

where q(X) is already obtained in the training stage and $p(X_*|X)$ can be derived from the *conditional GP prior* (Rasmussen and Williams 2006).

According to the predictive density, we need to compute $q(F_*|X_*)$ now. By following Eq. (10), $q(F_*|X_*)$ is:

$$q(F_*|X_*) = \prod_{j=1}^{p} \int p(f_{:,j}|u_{:,j}, X) q(u_{:,j}) \mathrm{d}u_{:,j}. \quad (24)$$

Note that $p(f_{:,j}|u_{:,j}, X)$ is already known in training stage and the specific expression of variational distribution $q(u_{:,j})$ is given in Eq. (19).

So, to predict Y_*, we need to first predict its latent function F_* according to:

$$q(F_*) = \int q(F_*|X_*) q(X_*) \mathrm{d}X_*. \quad (25)$$

With $q(F_*|X_*)$ and $q(X_*)$ in hand now, the mean is $E[F_*]$=$B^T \Psi_1^*$ and the covariance is:

$$Cov(F_*) = \sigma^2 I + B^T \Big(\Psi_2^* - \Psi_1^* (\Psi_1^*)^T \Big) B + \Psi_0^* I - tr\Big(\Big(K_{uu}^{-1} - (K_{uu} + \sigma^{-2} \Psi_2)^{-1} \Big) \Psi_2^* \Big) I, \quad (26)$$

where $B = \sigma^{-2}(K_{uu} + \sigma^{-2}\Psi_2)^{-1}\Psi_1^T Y$, $\psi_0^* = tr(K_{uu})$, $\Psi_1^* = \langle K_{u*} \rangle$ and $\Psi_2^* = \langle K_{u*} K_{u*}^T \rangle$. Notice that the Ψ statistics involving the test latent variable x_* appear naturally in these expressions. Using the above expressions, the predicted mean of Y_* is equal to $E(F_*)$ and the predicted covariance is equal to $Cov(F_*) + \sigma^{-1} I$.

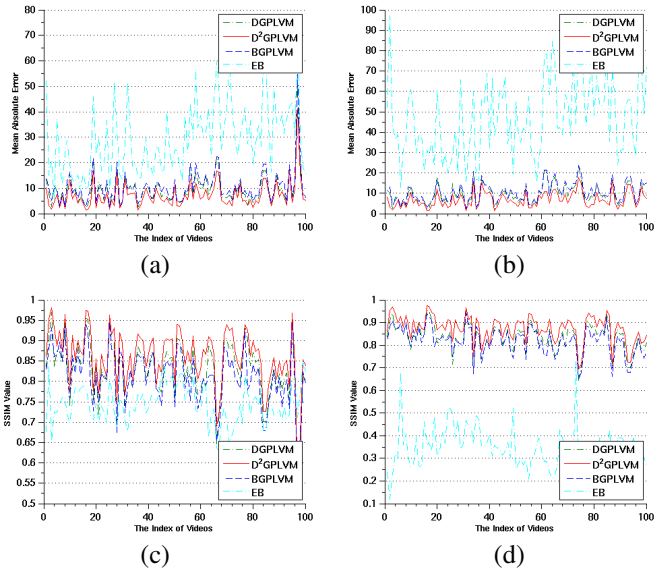

Figure 2: (a)-(c) MAEs and SSIMs of block damage repair in videos. (b)-(d) MAEs and SSIMs of repairing video with scattered damage repair in videos. The red line is our method. The smaller the MAE value is, the more similar the video frames between repaired video and ground truth, and vice versa for SSIM.

Table 1: Average MAE and SSIM of Plots (a), (b), (c), (d) in Fig. 2 respectively.

	BGPLVM	DGPLVM	EB	**D²GPLVM**
MAE of (a)	10.68	9.53	29.26	**7.33**
MAE of (b)	11.84	10.73	52.87	**7.91**
SSIM of (c)	0.83	0.84	0.76	**0.88**
SSIM of (d)	0.89	0.90	0.37	**0.95**

Results

In this section, we conduct experiments on movie clips from the Hollywood dataset [1]. Since D²GPLVM requires a certain number of frames for training, the clips used for testing were at least seven seconds in length. For each sequence, 40 percent of the frames were randomly selected to generate artificial damage. To be precise, we assumed that half the pixels in one frame were missing, and providing two kinds of damage for testing: block damage and scattered damage. More specifically, block damage was generated by cutting off either the left/right or top/bottom half part of one frame. Furthermore, a number of missing pixels were distributed across the whole frame to produce scattered damage. Our method was performed on 100 movie clips in the dataset and compared with three other methods: BGPLVM (Titsias and Lawrence 2010), DGPLVM and exemplar based inpainting (EB) (Criminisi, Perez, and Toyama 2004).

To demonstrate the robustness of our approach, we tested both black-and-white and color films. Since the test sequence resolution was not fixed, it was not easy to obtain specific times used for training and repair. However, in general, it took about one minute for training and around two seconds for repair. The code was run on Matlab 2014a on a computer configured with a 3.2GHz CPU and 8GB memory. Of note, our approach maintained a similar execution time for repair as DGPLVM and BGPLVM, but was much faster than exemplar-based inpainting, which took over 15 minutes for entire video inpainting. Note that only a single parameter λ in our model controls the diversity of the optimized inducing points, which is manually set as 0.01 for all 100 videos here. The results can be further improved if the cross-validation method is employed.

First, mean absolute error (MAE) and structural similarity (SSIM) were calculated to assess the fidelity of recovered frames on all clips. The MAE and SSIM values for one clip were subsequently obtained by averaging the sum of the MAEs and SSIMs for all frames. Exemplar based inpainting consistently showed the worst performance (Fig. 2). DGPLVM incorporates the dynamical sequential prior into BGPLVM, rendering it more competent than BGPLVM for time sequence based video repair. By taking the diversity encouraging prior into account, our method substantially enhanced damaged video recovery (Fig. 2). Moreover, so as to further clearly demonstrate the superiority of our method, we also computed the average value of MAEs and SSIMs in the plots of (a), (b), (c), (d). This can be seen in Table 1 that

[1] http://www.di.ens.fr/~laptev/actions/hollywood2/

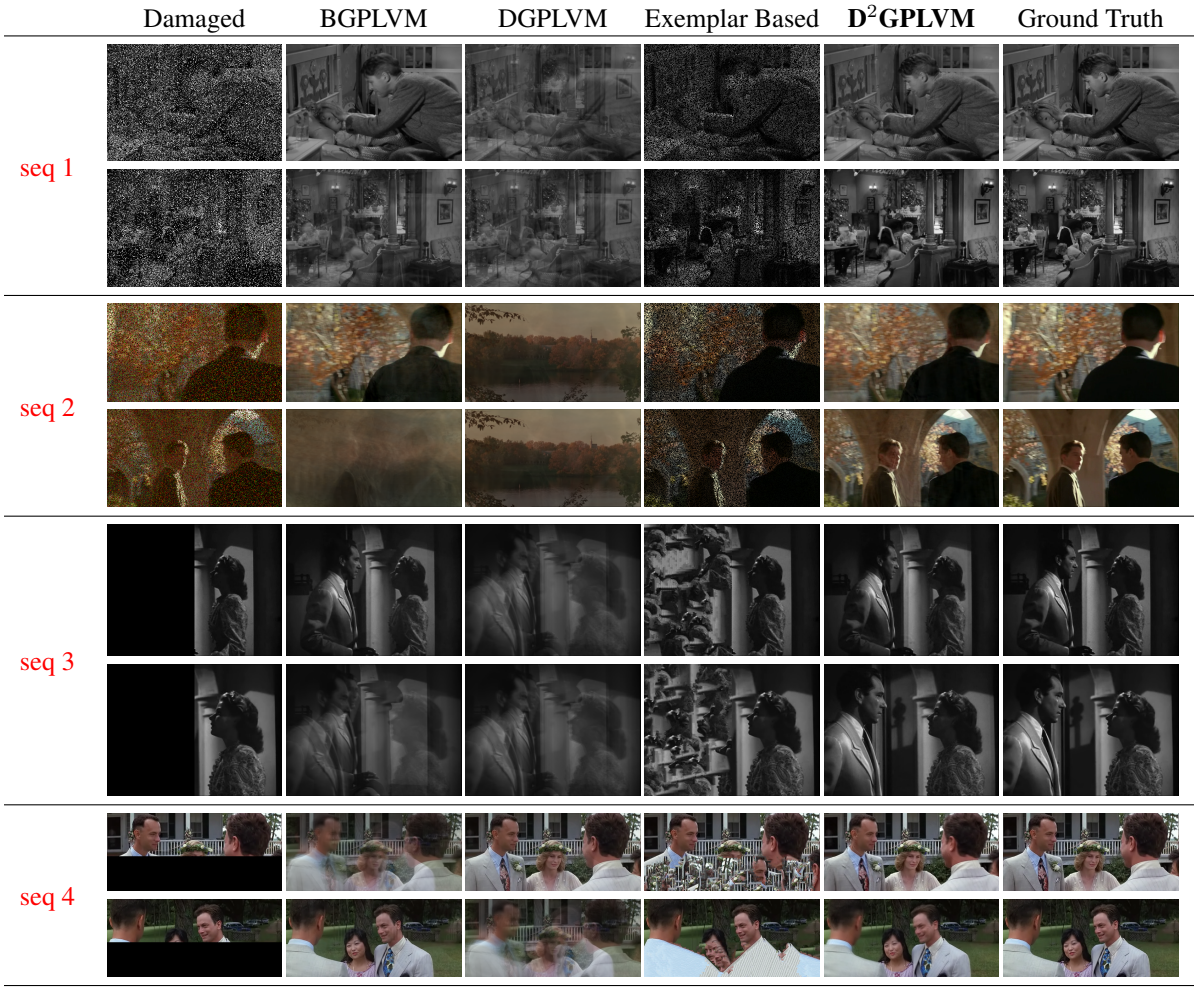

Figure 3: A schematic of video repair. Two recovered frames of the four video sequences are randomly displayed. Our method consistently outperforms the others and provides smooth and clearly recovered results.

our method unexceptionally significantly outperforms other methods in respect of the MAE and SSIM measurements.

Examples of video reapir in four video sequences by four different methods are shown in Fig. 3. The original frames with half the pixels missing are shown in the first column, while the next four columns are a comparison with the other methods and the last column is the ground truth. As mentioned above, since exemplar based inpainting only exploits local image features for texture synthesis, it is likely to introduce errors during repair. Without special constraints, BGPLVM and DGPLVM tend to generate ghost effects in the recovered frames because the inducing points created by BGPLVM and DGPLVM are usually more similar than diverse. Consequently, these methods fail to extract comprehensive distinct scenes to better recover damaged videos, especially those with constantly changing shots, a dynamic background and/or moving objects. Our method offers stable repair and outperforms the other approaches in different video scenarios.

Conclusion

This paper presents D^{2}GPLVM for video repair by simultaneously exploring dynamic and diversity properties under the GPLVM framework. Compared to DGPLVM, the diversity encouraging prior is essential in our model to extract more distinct features from the observed training frames. Meanwhile, our method has the inherent advantage of recovering incomplete frames with more complex sceneries since these types of video usually have more diverse characteristics that cannot be captured by traditional DGPLVM. Instead of using local neighboring texture synthesis, our probabilistic model utilizes global temporal information to recover smoother frames with large missing holes. Finally, our model reformulates the objective function of traditional DGPLVM and introduces a lower bound of the objective function by variational inference. Therefore, the objective function of our model is analytically tracktable by maximizing its variational lower bound. We illustrate the effectiveness of our method by comparing it with a number of other methods on a movie dataset of 100 various video clips.

Acknowledgments

This research is supported by Australian Research Council Projects (No: DP-140102164 & No: FT-130101457).

References

Amano, T., and Sato, Y. 2002. Image interpolation using bplp method on the eigenspace. *IEICE Transactions on Information and Systems* J85-D-II(3):457–465.

Amano, T. 2004. Image interpolation by high dimensional projection based on subspace method. In *International Conference on Pattern Recognition*, 665–668.

Bertalmio, M., and Sapiro, G. 2000. Image inpainting. In *ACM SIGGRAPH*.

Bertalmio, M.; Vese, L.; Sapiro, G.; and Osher, S. 2009. Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing* 12(8):882–889.

Candela, J. Q., and Rasmussen, C. E. 2005. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research* 1939–1959.

Criminisi, A.; Perez, P.; and Toyama, K. 2004. Region filling and object removal by exemplar-based inpainting. *IEEE Transactions on Image Processing* 13(9):1200–1212.

Criminisi, A.; Prez, P.; and Toyama, K. 2003. Object removal by exemplar-based inpainting. In *International Conference on Computer Vision and Pattern Recognition*, 721–728.

Csato, L., and Opper, M. 2002. Sparse on-line gaussian processes. *Neural Computation* 14(3):641–668.

Csato, L. 2002. Gaussian processes iterative sparse approximations. *PhD thesis, Aston University*.

Cutler, R., and Davis, L. S. 2009. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):781–796.

Damianou, A. C.; Titsias, M. K.; and Lawrence, N. D. 2014. Variational inference for uncertainty on the inputs of gaussian process models. *ArXiv:1409.2287*.

Drori, I.; Cohen-Or, D.; and Yeshurun, H. 2003. Fragment-based image completion. In *ACM SIGGRAPH*.

Efros, A., and Leung, T. K. 1999. Texture synthesis by non-parametric sampling. In *International Conference on Computer Vision*, 1033–1038.

Hosoi, T.; Kobayashi, K.; Ito, K.; and Aoki, T. 2011. Fragment-based image completion. In *IEEE International Conference on Image Processing*, 3441–3444.

Jia, J., and Tang, C. K. 2004. Inference of segmented color and texture description by tensor voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6):771–786.

Jia, J.; Wu, T. P.; Tai, Y. W.; and Tang, C. K. 2004. Video repairing: Inference of foreground and background under severe occlusion. In *International coference on Computer Vision and Pattern Recognition*, 364–371.

Jia, J.; Tai, Y. W.; Wu, T. P.; and Tang, C. K. 2006. Video repairing under variable illumination using cyclic motions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(5):832–839.

King, N. J., and Lawrence, N. D. 2006. Fast variational inference for gaussian process models through kl-correction. In *European Conference on Machine Learning*.

Medioni, G.; Lee, M. S.; and Tang, C. K. 2000. A computational framework for feature extraction and segmentation. *Amsderstam: Elsevier Science*.

Min, R., and Dugelay, J. 2013. Inpainting of sparse occlusion in face recognition. In *IEEE International Conference on Image Processing*, 1425–1428.

Mo, Z.; Lewis, J. P.; and Neumann, U. 2004. Face inpainting with local linear representations. In *British Machine Vision Conference*, 347–356.

Rasmussen, C. E., and Williams, C. 2006. Gaussian processes for machine learning. *MIT Press*.

Seeger, M.; Williams, C.; and Lawrence, N. D. 2003. Fast forward selection to speed up sparse gaussian process regression. In *Ninth International Workshop on Artificial Intelligence and Statistics*.

Seitz, S. M., and Dyer, C. R. 1997. View invariant analysis of cyclic motion. *International Journal of Computer Vision* 25(3):231–251.

Snelson, E., and Ghahramani, Z. 2006. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, 1257–1264.

Sobiecki, A.; Telea, A.; Giraldi, G.; Neves, L.; and Thomas, C. E. 2013. Low-cost automatic inpainting for artifact suppression in facial images. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

Sun, J.; Yuan, L.; Jia, J.; and Shum, H.-Y. 2005. Image completion with structure propagation. In *ACM SIGGRAPH*.

Titsias, M. K., and Lawrence, N. D. 2010. Bayesian gaussian process latent variable model. In *Thirteenth International Workshop on Artificial Intelligence and Statistics*, 844–851.

Titsias, M. K. 2009. Variational learning of inducing variables in sparsegaussian processes. In *Twelfth International Conference on Artificial Intelligence and Statistics*, 567–574.

Wexler, Y.; Shechtman, E.; and Irani, M. 2004. Space-time video completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1120–1127.