Modeling Dynamic Scenes with Active Appearance

Gianfranco Doretto
GE Global Research
One Research Circle, Niskayuna, NY 12309
doretto@research.ge.com

Abstract

In this work we propose a model for video scenes that contain temporal variability in shape and appearance. We propose a conditionally linear model akin to a dynamic extension of active appearance models. We formulate the problem variationally, and propose a framework where a model complexity cost dictates the "modeling responsibility" of each of the factors: appearance, shape and motion. We render the learning problem well-posed by reverting to a physical and a dynamic prior, and use the finite element method to compute a numerical solution. We illustrate our model to learn and simulate the shape, appearance, and motion of scenes that exhibit some form of temporal regularity, intended in a statistical sense.

1. Introduction

In modeling complex visual phenomena one can employ rich models that characterize the global statistics of the image, or choose simple classes of models to represent the local statistics of a spatio-temporal "segment," together with the partition of the data into such segments. Each segment could be characterized by certain statistical regularity properties in time and/or space. The former approach is often pursued in computer graphics, where a global model is necessary to capture effects such as mutual illumination or cast shadows. However, such models are not suitable for inference, since their parameters (often infinite-dimensional) cannot be uniquely inferred from the data. For instance, the complex appearance of sea waves can be attributed to a scene with simple reflectance and complex geometry, such as the surface of the sea, or with simple geometry and simple reflectance, for instance a mirror reflecting the radiance of a complex illumination pattern.

Since a "physically correct" model of the shape, motion and appearance of complex scenes cannot be inferred, one can resort to modeling visual complexity in terms of statistical variability from a nominal model – best if such a model contains all and only the parameters that can be identified. The simplest instance of this program is to use linear statistical analysis to model the variability of a data set as an affine variety; the "mean" is the nominal model, and a Gaussian density represents linear variability. This is done, for instance, in Eigenfaces [21] where appearance variation is modeled by a Gaussian process, in Active Shape Models [10] where shape variation is represented by a Gaussian Procrustean density [7], and in Dynamic Textures [20, 12], where motion is modeled by a Gauss-Markov process. Active Appearance Models (AAM's) [10], or linear morphable models [22], go one step beyond in combining the representation of appearance and shape variation into a conditionally linear model, in the sense that if the shape is known then appearance variation is represented by a Gaussian process, and vice versa. Naturally, one could make the entire program more general and non-linear by "kernelizing" each step of the representation [19] in a straightforward way.

In this paper we seek to expand this program and *model the statistics of data segments that exhibit spatio-temporal stationarity using conditionally linear processes for appearance, shape and motion.* In other words, rather than modeling only appearance (eigenfaces), only shape (active shape models) or only motion (dynamic textures), using linear statistical techniques, we model all three simultaneously.[1] Therefore, our work could be thought of as extending AAM's to the temporal domain, or extending dynamic textures to temporal variations of the domain.

Note that, as we have suggested, there is ambiguity in how these three factors interact: We could attribute all the responsibility for the variation of a data set to changes in appearance (i.e. the range of the image), or – under suitable conditions[2] – to changes in geometry (i.e. transformations

This research was conducted at the University of California, Los Angeles, with support by NSF ECS-0200511, AFOSR F49620-03-1-0095/E-16-V91-G2, and ONR N00014-03-1-0850/N00014-02-1-0720.

[1] Eventually this will have to be integrated into a higher-level spatio-temporal segmentation scheme, but such a high-level model is beyond our scope, and here we concentrate in modeling and learning each segment in isolation.

[2] When the data set can be represented as the transitive action of a group of deformations of the domain of the image.

of the domain of the image). *We are interested in developing a modeling framework where a complexity cost dictates the "modeling responsibility" of each factor.*

Unlike traditional AAM's, we do not use "landmarks," and our work follows the lines of the more recent efforts in AAM's, such as the work of Baker et al. [5] and Cootes et al. [9]. This work also relates to other approaches for modeling the temporal statistics of video segments, such as [18, 13, 23, 6].

In Section 2 we propose a variational formulation of the modeling framework. In Section 3 we set up the learning problem and propose an automatic model selection scheme that arbitrates the modeling responsibility between appearance, shape, and motion. This provides us with a principled framework where physical, and dynamic priors can be easily imposed to render the learning problem well-posed. Finally, we compute a numerical solution using the finite element method (FEM) [15].

Our models can be used to support detection (segmentation), classification (recognition) as well as simulation (synthesis) tasks. We illustrate the power of the models using the latter criterion. We compare our results with existing models, and show significant improvement in both fidelity (RMS error) and complexity (model order).

2. Dynamic Active Appearance

We propose to jointly model the variability in geometry (shape), radiometry (appearance) and dynamics (motion) of a scene as a conditionally linear model. Before doing that, however, we must define a nominal, or "mean," model. This has to be identifiable, in the sense that all of its parameters have to be uniquely determinable given sufficiently exciting data. We could start with an approximation of a physical model at time t, for instance a family of deforming surfaces $S_t : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$, viewed from a moving viewpoint $g_t \in SE(3)$, reflecting energy via a bi-directional reflection distribution function (BRDF), under a certain illumination. Unfortunately, it is trivial to show that this model cannot be identified, as the counterexamples in the previous section illustrate. Indeed, even if we assume that the scene is Lambertian, so the BRDF can be represented by an albedo $\rho_t : \mathbb{R}^3 \to \mathbb{R}_+$, but allow arbitrary illumination, we cannot infer the model uniquely [8]. More in general, illumination and reflectance play exchangeable roles (a consequence of the Helmholtz reciprocity principle [24]), and therefore any physical model with an explicit illumination other than ambient (constant) would be an overkill for our purpose. Therefore, we will start by a simple Lambertian scene in ambient illumination as our nominal model. Deviations from this model in terms of shape, appearance and motion will be represented statistically, as we describe in the next section.

2.1. Derivation of the nominal model

Under the assumptions discussed above, the intensity of the image at position x_t and time t can be written as $I_t(x_t) = \rho_t(p), p \in S_t$ where $x_t = \pi(g_t p)$ and $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ is the canonical perspective projection. To simplify this model we can parameterize $S_t : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3;\ x \mapsto S_t(x)$. This already highlights the ambiguity between shape S_t and rigid motion g_t, since we only measure their composition, and we could attribute the variability in the image to either factor. Therefore, we lump them into the warp $w_t \doteq \pi(g_t S_t) : \Omega \to \Omega$. Then, the reader will notice the ambiguity in shape S_t and appearance ρ_t in $I_t(x_t) = \rho_t(S_t(x))$, so again we could attribute image variation to either factor. With a bad abuse of notation, we rename $\rho_t \doteq \rho_t \circ S_t$ as the template (note that the domain of ρ_t is now $\Omega \subset \mathbb{R}^2$, rather than $S_t \subset \mathbb{R}^3$). We can then rewrite the model as

$$\boxed{\begin{cases} I_t(x_t) = \rho_t(x), & x \in \Omega \subset \mathbb{R}^2 \\ x_t = w_t(x), & t = 1, 2, \ldots, T \end{cases}} \tag{1}$$

which is reminiscent of deformable templates, except that here we do not know the template ρ_t. If we think of an image as a function with a domain Ω and a range $\mathbb{R}_+$, we have that shape and motion are warped together in the domain deformation w_t, and shape and appearance are merged in the range deformation ρ_t. Naturally, there is ambiguity even between these two factors, as one can easily see by substituting $I_t(x_t) = \rho_t(w_t^{-1}(x_t)),\ x_t \in w_t(\Omega)$, assuming the domain deformation to be homeomorphic, from which one can see that all the modeling responsibility could be delegated to w_t, yielding the notion of deformable templates, or to ρ_t. Recently Miller and Younes have proposed various joint models [17], and so have Fitzgibbon and Zisserman in their work on the joint manifold distance [14]. We will seek for model complexity to dictate the assignment of modeling responsibility to ρ_t and w_t, as we explain in the next section.

2.2. Variability from the nominal model

Rather than representing the deviation from Lambertian reflection with a BRDF, the deviation from rigid motion with some physical deformation model, we use a statistical model, indeed the simplest possible one, which corresponds to assuming that the variability of appearance, shape and motion is conditionally linear. This means that shape is modeled as a Gaussian shape space; given shape, appearance variation is modeled by a Gaussian distribution, and given shape and appearance, motion is modeled by a Gauss-Markov process in the joint representation:

$$\boxed{w_t(x) = w_0(x) + W(x)s_t, \quad x \in \Omega} \tag{2}$$

where the mean warp $w_0 : \mathbb{R}^2 \to \mathbb{R}^2$ and the principal warps $W : \mathbb{R}^2 \to \mathbb{R}^{2\times k}$ are vector- and matrix-valued functions respectively, $s_t \in \mathbb{R}^k$ is the shape parameter, and k is the shape state dimension. Similarly, we assume that

$$\boxed{\rho_t(x) = \rho_0(x) + P(x)\alpha_t, \quad x \in \Omega} \tag{3}$$

where $\rho_0 : \mathbb{R}^2 \to \mathbb{R}_+$, and $P : \mathbb{R}^2 \to \mathbb{R}^{1\times l}$, are the mean template and principal templates respectively, $\alpha_t \in \mathbb{R}^l$ is the appearance parameter, and l is the appearance state dimension. They, in turn, can be modeled by a dynamical system, so we assume that there exist suitably sized matrices A, B, and a covariance matrix C and a Gaussian process $\{\xi_t\}$ with initial condition ξ_0 such that

$$\boxed{\begin{cases} \xi_{t+1} = A\xi_t + n_t \quad n_t \sim \mathcal{N}(0, Q) \\ \begin{bmatrix} s_t \\ \alpha_t \end{bmatrix} = C\xi_t \end{cases}} \tag{4}$$

where n_t is a white and zero-mean Gaussian process with a covariance Q. Note that traditional AAM's assume that $x \in \{x_1, \ldots, x_N\}$, a set of "landmark points" in (2), and then extend it by interpolation to Ω in order to perform linear statistical analysis in (3). Baker et al. [5] have proposed an extension where (2) is performed on Ω, and we will comment on the differences in the next section.

3. Learning

Given $I_t(x_t),\ x_t \in w_t(\Omega)$, for $t = 1, \ldots, \tau$, learning the model amounts to determining the functions $w_0(\cdot)$ (mean warp), $W(\cdot)$ (principal warps), $\rho_0(\cdot)$ (mean template), $P(\cdot)$ (principal templates), the dynamic parameters A, C and Q that minimize a discrepancy measure between the data and the model. In formulas, we are looking for

$$\begin{cases} \arg\min_{w_0,W,\rho_0,P,A,C,Q} \int_\Omega \sum_t (I_t(w_t(x)) - \rho_t(x))^2 \, \mathrm{d}x \\ \text{subject to (1), (2), (3), (4) and} \\ \int_\Omega P_{.i}(x) P_{.j}(x) \, \mathrm{d}x = \delta_{ij} = \int_\Omega W_{.i}(x) W_{.j}(x) \, \mathrm{d}x \end{cases} \tag{5}$$

in addition to minimizing additional regularizing terms for the functions $w_0(\cdot)$, $W_{.i}(\cdot)$, $\rho_0(\cdot)$, $P_{.i}(\cdot)$, and the process $\left\{ \begin{bmatrix} s_t \\ \alpha_t \end{bmatrix} \right\}$ to guarantee that the problem is well-posed (in the above notation δ_{ij} is the Kronecker delta, while $W_{.i}$, and $P_{.i}$ represent the i-th column of W, and P respectively). The last set of constraints impose orthogonality of the shape and appearance bases, and could be relaxed under suitable conditions. Needless to say, this is a tall order. In the rest of this section we show how to reduce this problem to finite dimensions using the finite element method (FEM), which provides with a straightforward way to regularize the unknowns.

3.1. Solving the learning problem

Solving problem (5) entails performing a minimization in an infinite dimensional space. In order to avoid this, we reduce the problem to a minimization in a finite dimensional space, and describe an alternating minimization procedure. At the i-th step of the iteration we assume that $w_0(x)$, $W(x)$, and $W_1^\tau = [w_1(x), \cdots, w_\tau(x)]$ are known, and want to solve the following problem

$$\arg\min_{\rho_0,P,\alpha} \int_\Omega \sum_{t=1}^{\tau} (I_t(w_t(x)) - \rho_0(x) - P(x)\alpha_t)^2 \, \mathrm{d}x \ . \tag{6}$$

In the first iteration we assume $W(x) = 0, \forall x \in \Omega$, which means that all the modeling responsibility is delegated to the appearance, represented by $\rho_t(x)$. Also, without loss of generality, one can assume that $x = w_0(x), \forall x \in \Omega$.

The minimization problem (6) is linear, and can be solved in closed form once a model selection criterion has been chosen. More precisely, the mean template ρ_0 can be computed as the sample mean of the images warped to the domain Ω:

$$\rho_0(x) = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t(w_t(x)) \ , \quad x \in \Omega \ . \tag{7}$$

Then, after removing the mean template, one can compute the principal components of the data set $[I_1(w_t(x)), \cdots, I_\tau(w_t(x))]$, that we indicate with $U_\rho(x) : \mathbb{R}^2 \to \mathbb{R}^{1\times\tau}$ (computationally, this can be done by performing an SVD of the data set), so that one can write $[I_1(w_t(x)), \cdots, I_\tau(w_t(x))] = U_\rho(x)\Sigma_\rho V_\rho^T$, where $V_\rho \in \mathbb{R}^{\tau\times\tau}$ is a unitary matrix, and $\Sigma_\rho \in \mathbb{R}^{\tau\times\tau}$ contains the singular values $\sigma_{\rho,1} \geq \sigma_{\rho,2} \geq \cdots \geq \sigma_{\rho,\tau}$, in its diagonal.

At this point, in order to estimate $P(\cdot)$, one needs to select the dimensionality of α_t. Since we are interested in setting up a procedure that automatically attributes the percentage of modeling responsibility of the appearance and the shape, this is a delicate step. We propose to perform automatic model selection of the appearance by looking at the energy of the principal components $U_\rho(\cdot)$. In more detail, we compute the normalized appearance energy $\tilde{\sigma}_{\rho,i} = \sigma_{\rho,i} / \sum_{j=1}^{\tau} \sigma_{\rho,j}$, and set $P(\cdot)$ to be the collection of principal components with normalized appearance energy above a certain threshold γ_ρ:

$$\begin{cases} l = \max_i \{i | \tilde{\sigma}_{\rho,i} \geq \gamma_\rho\} \ , \\ P(x) = [U_{\rho,1}(x), \cdots, U_{\rho,l}(x)], \quad x \in \Omega \ . \end{cases} \tag{8}$$

This approach to model selection is very similar to model selection techniques that have been used for long time within the system identification community [1].

Once the number of principal components l is known, one can estimate the appearance state $\alpha_1^\tau = [\alpha_1, \cdots, \alpha_\tau]$

by simply computing the following matrix product

$$\alpha_1^\tau = \Sigma_{\rho,1:l,1:l} V_{\rho,:,1:l}^T \ , \tag{9}$$

where here we have made use of Matlab notation to indicate the selection of the first l columns and rows of Σ_ρ, and the first l columns of V_ρ.

In the next step of the i-th iteration we assume that $\rho_0(x)$, $P(x)$, and α_1^τ are known, and want to solve the following problem

$$\arg \min_{w_0,W,s} \int_\Omega \sum_{t=1}^{\tau} (I_t(w_0(x) + W(x)s_t) - \rho_t(x))^2 \ \mathrm{d}x \ . \tag{10}$$

To simplify this complex minimization, we split it in two steps. In the first one we solve the following subproblem

$$\arg \min_{w} \int_\Omega \sum_{t=1}^{\tau} (I_t(w_t(x)) - \rho_t(x))^2 \ \mathrm{d}x \ , \tag{11}$$

allowing us to estimate $w_0(x)$ and $W(x)$, while in the second one we estimate the shape state $s_1^\tau = [s_1, \cdots, s_\tau]$ by solving

$$\arg \min_{s} \int_\Omega \sum_{t=1}^{\tau} (I_t(w_0(x) + W(x)s_t) - \rho_t(x))^2 \ \mathrm{d}x \ . \tag{12}$$

The minimizations (11) and (12) are discussed in the following two sections respectively.

3.1.1 Estimation of the nominal and principal warps

In solving problem (11) we seek to guarantee that: (a) $w_t(x)$ is a homeomorphism in x, since in (1) we require the warp to be invertible as it cannot handle occlusions; (b) $w_t(x)$ varies smoothly in time, as we expect two adjacent images of a video sequence to be similar (in accordance with (4)); (c) $w_t(x)$ is Gaussian distributed around the mean warp $w_0(x)$ (to meet (2), and partly (4)). Obviously, none of the conditions (a), (b), and (c) are guaranteed to be satisfied if we minimize the functional in (11) as it is, from which the need for a regularization.

To regularize the functional in (11) we view the problem of estimating the warp $w_t(x)$ as a three-dimensional stress analysis problem, where we consider the set of points $\{(x,t) \mid x \in \Omega, \ t \in [1,\tau] \subset \mathbb{R}\}$ as a three-dimensional Euclidean space filled with a homogeneous isotropic linear elastic material. In absence of external loads, each particle located at a point $(w_0(x),t)$ is in equilibrium. Under this condition we have $(w_t(x),t) = (w_0(x),t), \ \forall \ (x,t)$. In presence of external loads with a potential energy given by the functional in (11), a particle located in $(w_0(x),t)$ moves to $(w_t(x),t)$, and is subject to a displacement $u(x,t) \doteq [w_t^T(x) - w_0^T(x) \quad 0]^T$. This displacement generates a strain in the structure that is given by $\varepsilon = [\partial u_1/\partial x_1 \ \ \partial u_2/\partial x_2 \ \ 0 \ \ \partial u_1/\partial x_2 + \partial u_2/\partial x_1 \ \ \partial u_1/\partial t \ \ \partial u_2/\partial t]^T$, which increases the total potential energy of the system by an amount of $1/2 \int_\Omega \int_1^\tau \varepsilon^T D \varepsilon \ \mathrm{d}t\mathrm{d}x$, where D is an elasticity matrix containing the appropriate material properties [15].

The problem of estimating the warp $w_t(x)$ can therefore be solved by minimizing the total potential energy given by the following functional

$$\int_\Omega \int_1^\tau (I_t(w_t(x)) - \rho_t(x))^2 \ \mathrm{d}t\mathrm{d}x + \frac{1}{2} \int_\Omega \int_1^\tau \varepsilon^T D \varepsilon \ \mathrm{d}t\mathrm{d}x \ , \tag{13}$$

where, for notational consistency, the summation in t of (11) has been replaced with an integral. The first part of the functional obviously represents the data fidelity term, while the second part is a regularization, or model prior term.

Unlike problem (6), minimizing (13) entails a non-linear minimization in an infinite dimensional space. To reduce the problem to a non-linear minimization in a finite dimensional space, we use the finite element method [15]. By using this approach we discretize the structure under consideration into a collection of finite elements connected to each other at several nodes. Then, the displacement $u_e(x,t)$, inside each element e, is approximated by a function of the nodal displacements $v_e - v_{0e}$, where v_{0e} is the nodal position at equilibrium and v_e is the new nodal position. More precisely $u_e(x,t) = N_e(x,t)(v_e - v_{0e})$, where $N_e(x,t)$ is the so called shape function matrix, and the elemental strain vector ε_e can be expressed in terms of the nodal displacements as $\varepsilon_e = B_e(v_e - v_{0e})$, where B_e contains appropriate derivatives of the shape functions. Finally, the strain energy stored in the element can be written as $1/2 \int_{vol} \varepsilon_e^T D \varepsilon_e \ \mathrm{d}t\mathrm{d}x = 1/2 (v_e - v_{0e})^T K_e (v_e - v_{0e})$, where $K_e = \int_{vol} B_e^T D B_e \ \mathrm{d}t\mathrm{d}x$ is the so called element stiffness matrix, and the total strain energy becomes

$$\begin{aligned} \frac{1}{2} \int_\Omega \int_1^\tau \varepsilon^T D \varepsilon \ \mathrm{d}t\mathrm{d}x &= \frac{1}{2} \sum_e (v_e - v_{0e})^T K_e (v_e - v_{0e}) \\ &= \frac{1}{2} (v - \bar{v}_0)^T K (v - \bar{v}_0) \ , \end{aligned} \tag{14}$$

where $K = \sum_e K_e^{(aug)}$ is the global stiffness matrix obtained by summing the appropriately augmented elemental stiffness matrices and which is proved to be positive definite. The vector $v^T = [v_1^T, \cdots, v_\tau^T]$ is the global nodal position vector, where we ordered the nodes in such a way that v_t contains the positions of all the nodes that are present in the time slice t, since it is there where eventually we will allow the nodal points to lie. Finally, we have $\bar{v}_0^T = [v_0^T, \cdots, v_0^T]$, since at equilibrium we require every time slice to have the nodal positions to lie in the same place, and this is required to satisfy condition (c).

With this framework in place it becomes natural to choose a triangular prism as shape of the basic element e. In

this way, v_t in the time slice t, or image plane at time t, will represent the vertex locations of a triangulated mesh that uniquely identify a piecewise affine warp. With $w_0(x; v_0)$ we indicate the warping from Ω to a nominal domain identified by v_0, and $w_t(x; v_t)$ indicates the warping from Ω to a domain identified by v_t (see [16] for an accurate description of the implementation of a piecewise affine warping map).

After this discretization, in lieu of minimizing the functional (13), we will concentrate on the following non-linear optimization problem

$$\arg\min_{v} \int_{\Omega} \sum_{t=1}^{\tau} (I_t(w_t(x; v_t)) - \rho_t(x))^2 \, \mathrm{d}x + \frac{1}{2}(v - \bar{v}_0)^T K (v - \bar{v}_0) \; , \quad (15)$$

that can be efficiently solved iteratively by using the so called inverse compositional image alignment algorithm described in [3, 4], while its modifications for the case that handles priors can be found in [2].

During the iteration of the inverse compositional algorithm, there is the need to invert an Hessian matrix, which one can show having the following expression:

$$\begin{bmatrix} H_1 & & 0 \\ & \ddots & \\ 0 & & H_\tau \end{bmatrix} + \begin{bmatrix} K_1 & K_2 & \cdots & K_\tau \\ K_2^T & K_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & K_2 \\ K_\tau^T & \cdots & K_2^T & K_1 \end{bmatrix} . \quad (16)$$

The first term is a block diagonal matrix, and each block, given by the following expression $H_t = \int_\Omega (\partial w_t / \partial v_t)^T \nabla \rho_t{}^T \nabla \rho_t \partial w_t / \partial v_t \; \mathrm{d}x$, is symmetric and positive definite. The second term is a symmetric positive definite block Toeplitz matrix. Given the lack of an efficient algorithm for the computation of the inverse of the Hessian, if the number of finite elements in the structure is too big, computing the inverse of (16) could become a problem. On the other hand, it is possible to reduce the complexity of the prior (14), and assume no dependency between deformations of the meshes at different time instants. This corresponds to imposing $K_2 = \cdots = K_\tau = 0$. The main advantage of using this reduced prior is computational efficiency, because one can run the inverse compositional algorithm image by image, and not on the entire image sequence. More importantly, the Hessian that needs to be inverted to process the image at time t becomes[3] $H_t + K_1$. The main drawback, instead, is the fact that the prior does not impose a smooth variability in time of the warp $w_t(x)$. However, this issue is addressed in Section 3.1.2.

[3]The matrix K_1 is the stiffness matrix for the case of an elastic plane subject to stress, and its computation, based on a decomposition in triangular elements, can be found in many standard books [15].

Once we obtain an estimate for v, we proceed by updating the mean warp:

$$w_0(x) = \frac{1}{\tau} \sum_{t=1}^{\tau} w_t(x; v_t) \; . \quad (17)$$

Once we remove the mean warp from the data set W_1^τ, the principal warps $W(\cdot)$ can be computed through the SVD, $W_1^\tau \doteq U_w(x) \Sigma_w V_w^T$, where $V_w \in \mathbb{R}^{\tau \times \tau}$ is a unitary matrix, and $\Sigma_w \in \mathbb{R}^{\tau \times \tau}$ contains the singular values $\sigma_{w,1} \geq \sigma_{w,2} \geq \cdots \geq \sigma_{w,\tau}$, in its diagonal. In order to estimate $W(\cdot)$, one needs to select the dimensionality of s_t. As we did for the estimation of $P(\cdot)$, we perform automatic model selection of the shape by looking at the normalized energy of the principal components $U_w(\cdot)$. If $\tilde{\sigma}_{w,i} = \sigma_{w,i} / \sum_{j=1}^{\tau} \sigma_{w,j}$ is the normalized shape energy, we set $W(\cdot)$ to be the collection of principal components with normalized shape energy above a certain threshold γ_w:

$$\begin{cases} k = \max_i \{ i | \tilde{\sigma}_{w,i} \geq \gamma_w \} \; , \\ W(x) = [U_{w,1}(x), \cdots, U_{w,k}(x)], \quad x \in \Omega \; . \end{cases} \quad (18)$$

3.1.2 Estimation of the shape state and dynamic parameters

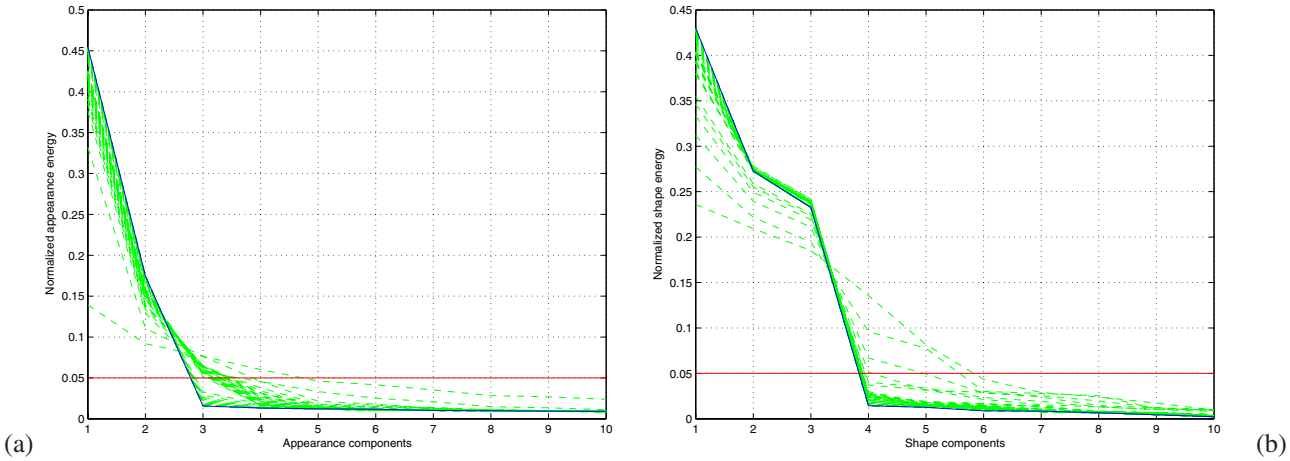
Once the number of principal warps k is known, one can obtain a first estimate of the shape state s_1^τ by computing the product (in Matlab notation)

$$s_1^\tau = \Sigma_{w,1:k,1:k} V_{w,:,1:k}^T \; . \quad (19)$$

This estimate needs to be updated as a consequence of the fact that $w_0(x)$ and $W(\cdot)$ are known. Moreover, the temporal statistics of $\left\{ \begin{bmatrix} s_t \\ \alpha_t \end{bmatrix} \right\}$ is supposed to be second-order stationary, as we plan to model it with a linear dynamic system. In other words, the minimization (12) should be done subject to the prior model (4). To this end, we estimate s_1^τ by solving another optimization problem

$$\arg\min_{s} \int_{\Omega} \sum_{t=1}^{\tau} (I_t(w_0(x) + W(x) s_t) - \rho_t(x))^2 \, \mathrm{d}x + \sum_{t=1}^{\tau-1} \| \xi_{t+1} - A \xi_t \|_F^2 \; , \quad (20)$$

where the prior aims at minimizing the Frobenius norm of the residuals in (4). Again, this minimization can be performed by using the inverse compositional algorithm [3]. Note that at each iteration of the algorithm, the matrix A needs to be estimated as it appears in the derivative of the prior. We omit the derivation of the prior derivative, which is tedious as it involves tensor algebra computations (for more details, the interested reader can consult [11]). Suffice it to say that the matrix A is estimated via least squares,



(a) (b)

Figure 1. *Test sequence. Evolution of the normalized appearance energy (a) ($\gamma_\rho = 0.05$), and the normalized shape energy (b) ($\gamma_w = 0.05$).*

which means that, if $\xi_{t_1}^{t_2} = [\xi_{t_1}, \cdot, \xi_{t_2}]$, then A is computed by the following expression

$$A = \xi_2^\tau {\xi_1^{\tau-1}}^T (\xi_1^{\tau-1} {\xi_1^{\tau-1}}^T)^{-1} , \tag{21}$$

whereas the covariance of the driving noise n_t is

$$Q = \frac{1}{\tau - 1} (\xi_2^\tau - A\xi_1^{\tau-1})(\xi_2^\tau - A\xi_1^{\tau-1})^T . \tag{22}$$

Here we have made the unspoken assumption that C is the identity matrix. In practice this may not be true, and C appears in the prior derivative and it has to be recomputed at every iteration as well.

4. Results

To validate the learning procedure described in Section 3 we produced a synthetic Test sequence with mean template ρ_0, and two principal templates $P(\cdot)$, as depicted in the top left section of Figure 2, and we applied a random stationary process α_t to simulate appearance variation, and a random action process $g_t \in SE(2)$ to simulate shape variation. The convergence of the learning procedure is better illustrated by Figure 1. The plot (a) shows the evolution of the normalized appearance energy $\{\tilde{\sigma}_{\rho,i}\}$, while the plot (b) shows the evolution of the normalized shape energy $\{\tilde{\sigma}_{w,i}\}$. At the end of the first iteration we have an appearance state dimension $l = 4$, and a shape state dimension $k = 5$. Note that the thresholds where set to $\gamma_\rho = 0.05$, and $\gamma_w = 0.05$. As the learning procedure approaches convergence, l and k decrease until they rich the values of 2 and 3 respectively. Since for building the Test sequence we used two principal templates, and a planar rigid motion, which has three degrees of freedom, the inferred appearance and shape state dimensions are correct. Figure 2 also shows the estimated mean template $\hat{\rho}_0$, and the estimated principal templates $\hat{P}(\cdot)$ (top center section); the mean, and the first two principal components of the original sequence (top right section); three training images (bottom left section); three synthesized images using the dynamic texture (DT) model [12] (bottom center section); and three synthesized images using the proposed active appearance dynamic texture (AADT) model (bottom right section). Figure 3 depicts a representation of the estimated mean warp $\hat{w}_0$, and the three principal warps $\hat{W}(\cdot)$.

Besides the Test sequence we collected other four sequences that we call Flowers, Candle, Duck, and Flag. For each of these sequences, for the same given thresholds γ_ρ and γ_w, we learnt both the DT model, and the AADT model. In Table 1, l_{DT} indicates the dimension of the state of the DT model, whereas l_{AADT} and k_{AADT} indicate the appearance and shape state dimensions of the AADT model. Since the majority of the model parameters is used to encode either the principal components of the DT model, or the principal templates of the AADT model, comparing l_{DT} and l_{AADT} is informative of the reduction of the complexity of the AADT model. As expected, at this reduction corresponds an increase of the shape state dimension, going from zero to k_{AADT}. In particular, Table 1 "suggests" that $l_{\mathrm{DT}} \approx l_{\mathrm{AADT}} + k_{\mathrm{AADT}}$.

The last three columns of the table report the root mean square errors (RMSE) per pixel in the reconstruction. $\mathrm{RMSE}_{\mathrm{DT}}$ and $\mathrm{RMSE}_{\mathrm{DT2}}$ are the errors for the DT model with state dimension l_{DT} and l_{AADT} respectively, whereas $\mathrm{RMSE}_{\mathrm{AADT}}$ is the error for the AADT model. One may notice that $\mathrm{RMSE}_{\mathrm{AADT}}$ and $\mathrm{RMSE}_{\mathrm{DT}}$ are fairly similar. This was expected since the models are inferred while retaining principal components and templates that are above the same threshold. The synthetic Test sequence is an exception because it was produced with exactly two principal templates containing sharp edges that the DT model can hardly capture. On the other hand, the comparison between $\mathrm{RMSE}_{\mathrm{DT}}$

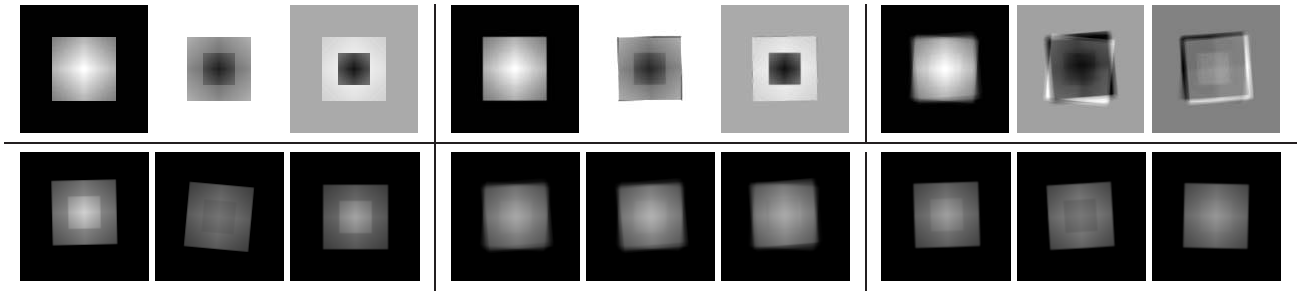


Figure 2. *Test sequence. Top row: ground truth of the mean template and principal templates (left), estimated mean template and principal templates (center), mean image and first two principal components of the training sequence (right). Bottom row: frames of the training sequence (left), synthesis using the DT model (center), synthesis using the AADT model (right).*

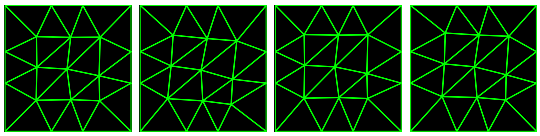


Figure 3. *Test sequence. Left to right: mean warp and principal warps of the estimated model.*

and RMSE_{DT2} highlights the degradation of the reconstruction error when the DT model is forced to have the same model complexity of the AADT model (because the state dimension of the DT model is equal to the appearance state dimension of the AADT model).

Finally, we validate the ability of the AADT model, and learning procedure to capture the spatio-temporal properties of a video sequence by using the model to extrapolate new video clips. For the Candle, Flowers, Duck, and Flag sequences, Figure 4 and 5 show one frame of the original sequence, the same frame with the triangulated mesh defined by v_t, one frame synthesized with the DT model, and one frame synthesized with the AADT model. Even if the reconstruction errors of the two models are comparable, their simulation reveals that the AADT model sensibly prevails the simpler DT model. This is true especially when a video sequence contains moving objects with defined structure and sharp edges, suggesting that the AADT model can capture a statistical description of the temporal statistical properties of a video sequence that is higher than the second-order.

5. Conclusions

We have presented a model for portions of image sequences where appearance, shape, and motion can be represented by conditionally linear models. These capture segments that exhibit certain statistical stationarity properties in time and/or space, and can be found by a spatio-temporal segmentation procedure.

Our approach can be though of as extending the work on Active Appearance Models [5, 10] to the temporal domain, or extending dynamic textures [12, 20] to the spatial domain.

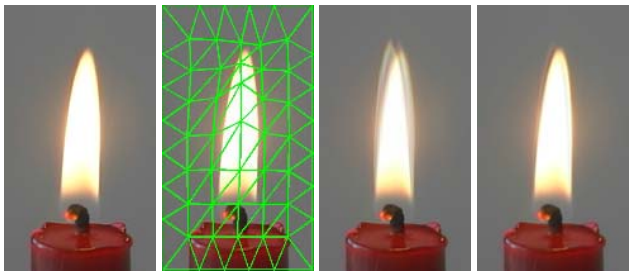


Figure 4. *Candle sequence. Left to right: original frame, original frame with mesh v_t, synthesis using the DT model, synthesis using the AADT model.*

We have presented a variational formulation of the problem, and an efficient computational solution that uses standard numerical approaches (finite element methods). We have validated our modeling and learning frameworks with a ground truth sequence, and have illustrated the superior modeling power of this approach, with respect to dynamic textures, in terms of model complexity, reconstruction error, and extrapolation power (prediction-error).

References

[1] K. S. Arun and S. Y. Kung. Balanced approximation of stochastic systems. *SIAM J. on Mat. An. and App.*, 11(1):42–68, 1990.

[2] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: a unifying framework: part 4. Technical report, CMU Robotics Institute, Pittsburgh, PA, February 2004.

[3] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proc. of CVPR*, volume 1, pages 1090–1097, 2001.

[4] S. Baker and I. Matthews. Lucas-Kanade 20 years on: a unifying framework. *IJCV*, 56(3):221–255, 2004.

[5] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE Trans. PAMI*, 26(10):1380–1384, 2004.

Sequence	l_{DT}	l_{AADT}	k_{AADT}	$\mathrm{RMSE}_{\mathrm{DT}}$	$\mathrm{RMSE}_{\mathrm{AADT}}$	$\mathrm{RMSE}_{\mathrm{DT2}}$
Test	4	2	3	7.3626	2.3870	9.9951
Flowers	22	19	6	4.0272	4.1290	4.4273
Candle	11	7	7	2.1146	2.1484	2.9134
Duck	16	11	6	1.6801	1.6001	1.8580
Flag	18	10	8	2.9926	3.2491	3.6392

Table 1. *For every sequence: l_{DT} is the state space dimension of the DT model, l_{AADT} and k_{AADT} are the appearance and shape state dimensions of the AADT model, $\mathrm{RMSE}_{\mathrm{DT}}$ and $\mathrm{RMSE}_{\mathrm{AADT}}$ are the root mean square reconstruction errors per pixel using the DT and AADT models respectively.*

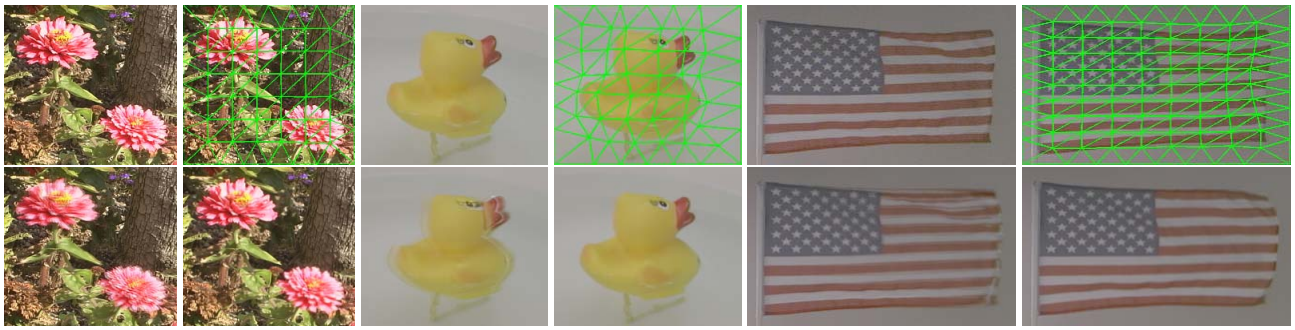


Figure 5. *Flowers, Duck, and Flag sequences. For each sequence: original frame (top left), original frame with mesh v_t (top right), synthesis using the DT model (bottom left), synthesis using the AADT model (bottom right).*

[6] N. Campbell, C. Dalton, D. Gibson, and B. Thomas. Practical generation of video textures using the autoregressive process. In *Proc. of BMVC*, pages 434–443, 2002.

[7] T. K. Carne. The geometry of shape spaces. *Proceedings of the London Mathematical Society*, 3(61):407–432, 1990.

[8] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs. In search of illumination invariants. Technical report, Yale University, 2004. Submitted to IJCV.

[9] T. Cootes, S. Marsland, C. Twining, K. Smith, and C. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *Proc. of ECCV*, pages 316–327, 2004.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. PAMI*, 23(6):681–685, 2001.

[11] G. Doretto. *DYNAMIC TEXTURES: modeling, learning, synthesis, animation, segmentation, and recognition*. PhD thesis, UCLA, March 2005.

[12] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, 2003.

[13] A. Fitzgibbon. Stochastic rigidity: image registration for nowhere-static scenes. In *Proc. of ICCV*, volume 1, pages 662–669, 2001.

[14] A. W. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *Proc. of CVPR*, volume 1, pages 26–33, 2003.

[15] T. J. R. Hughes. *The Finite Element Method - linear static and dynamic finite element analysis*. Dover Publications, Inc., 2000.

[16] I. Matthews and S. Baker. Active appearance models revised. *IJCV*, 60(2):135–164, 2004.

[17] M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: a general framework. *IJCV*, 41(1/2):61–84, 2001.

[18] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa. Video textures. In *Proceedings of SIGGRAPH*, pages 489–498, 2000.

[19] B. Schölkopf and A. Smola. *Learning with kernels: SVM, regularization, optimization, and beyond*. The MIT press, 2002.

[20] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. In *Proc. of ICCV*, volume 2, pages 439–446, 2001.

[21] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cog. Neur.*, 3(1):71–86, 1991.

[22] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. PAMI*, 19(7):733–742, 1997.

[23] Y. Z. Wang and S. C. Zhu. A generative method for textured motion: analysis and synthesis. In *Proc. of ECCV*, pages 583–598, 2002.

[24] T. E. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: exploiting reciprocity for surface reconstruction. *IJCV*, 49(2/3):215–227, 2002.