Modeling Dynamic Scenes with Active Appearance

Technical Report UCLA-CSD-TR040053
December 2, 2004

Gianfranco Doretto

Department of Computer Science
UCLA
Los Angeles, CA 90095

Stefano Soatto

Department of Computer Science
UCLA
Los Angeles, CA 90095

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 use finite-element methods 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 in space and/or time. 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 dataset as an affine variety; the "mean" is the nominal model, and a Gaussian density represents linear variability. This is done, for instance, in Eigenfaces [17] where appearance variation is modeled by a linear Gaussian process, in Active Shape Models [9, 7] where shape variation is represented by a Gaussian Procrustean

density, and in Dynamic Textures [11], where motion is modeled by a Gaussian density. Active Appearance Models (AAM) [9], or linear morphable models [18], go one step beyond in combining the representation of appearance and shape variation into a conditionally linear Gaussian process. Naturally, one could make the entire program more general and non-linear by "kernelizing" each step of the representation [16] 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 shape, motion and appearance.* 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 dataset to changes in appearance (i.e. the range of the image), or – under suitable conditions[2] – to changes in geometry (i.e. transformations 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, such as the work of Baker et al. [6] and Cootes et al. [10].

In Section 3 we formulate the problem in a variational framework, and in Section 3.1 we propose a numerical solution using finite-element methods. This provides us with a principled modeling framework where physical priors (stiffness) can be easily imposed, yielding an automatic technique for model selection. We also impose photometric priors implicitly by the use of an explicit dynamical model of appearance, akin to dynamic textures [11].

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, measured using prediction error, together with the overall complexity of the model. 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), photometry (appearance) and dynamics (motion) of a scene as a conditionally linear process. 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, 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 (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 $\rho_t : S_t \to \mathbb{R}_+$, but allow arbitrary illumination, we cannot infer the model uniquely [8]. More in general, illumination

[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 dataset can be represented as the transitive action of a group of deformations of the domain of the image.

and reflectance play exchangeable roles (a consequence of Helmholtz reciprocity principle [19]), 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 in shape S_t and 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 $w_t \doteq g_t S_t : \Omega \to \mathbb{R}^2$. 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 $\rho_t \doteq \rho_t \circ S_t$ (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 [15], and so have Fitzgibbon and Zisserman in their work on the joint manifold distance [12]. We will seek for 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 shape, motion and appearance 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 Gaussian distribution in the joint representation:

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

where $w_0 : \mathbb{R}^2 \to \mathbb{R}^2$ and $W : \mathbb{R}^2 \to \mathbb{R}^k$ are vector- and matrix-valued functions respectively. 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 $P : \mathbb{R}^2 \to \mathbb{R}^l$. Here $s_t \in \mathbb{R}^k$ and $\alpha_t \in \mathbb{R}^l$ are the shape and appearance parameters respectively. They, in turn, can be modeled by a dynamical system, so we assume that there exist suitably sized matrices A, B, 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. [6] have proposed an extension where (2) is performed on Ω, and we will comment on the differences in the next section.

3. Learning

Given $I_t(x_t),\ x_t \in w_t(\Omega)$, learning the model amounts to determining the functions $w_0(\cdot)$ (mean deformation), $W(\cdot)$ (principal warps), $\rho_0(\cdot)$ (mean template), $P(\cdot)$ (principal components), the dynamical model parameters A, C and covariance Q that minimize a discrepancy measure between the data and the model. In formulas, we are looking for

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

in addition to minimizing additional regularizing terms for the functions $w_0(\cdot), W_{\cdot i}(\cdot), \rho_0(\cdot), P_{\cdot i}(\cdot)$ to guarantee that the problem is well-posed. 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 finite-element methods (FEM), which provides with a straightforward way to regularize the unknowns. Once the learning part is done, modeling is straightforward since the spatio-temporal statistics of a data segment are now captured by the finite-dimensional parameters α_t, s_t which are easy to be determined by using a chain of singular value decompositions (SVD's).

3.1. Solving the learning problem

Solving problem (5) entails performing a minimization in an infinite dimensional space. In order to avoid dealing with such a complicated problem, in this section we describe an alternating minimization procedure, together with a reduction of the problem to a minimization in a finite dimensional space.

The first step of the optimization starts by assuming that $w_0(x)$, $W(x)$, and $W_1^\tau = [w_1(x), \cdots, w_\tau(x)]$ are known, and we are interested in solving the following problem

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

Note that during the first iteration all the modeling responsibility is delegated to the appearance $\rho_t(x)$, which means that $W(x) = 0, \forall x \in \Omega$, also note that one can assume, without loss of generality, that $x = w_0(x), \forall x \in \Omega$.

The minimization problem (6) is linear, and can readily be solved in closed form, once a model selection criterion has been chosen. More precisely, the nominal appearance 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 nominal appearance, one can compute the principal components of the dataset $[I_1(w_t(x)), \cdots, I_\tau(w_t(x))]$, that we indicate with $U_\rho(x) : \mathbb{R}^2 \rightarrow \mathbb{R}^T$ (computationally, this can easily be done by performing a SVD of the dtaset), 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 the appearance state 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 energy $\tilde{\sigma}_{\rho,i} = \sigma_{\rho,i} / \sum_{j=1}^{\tau} \sigma_{\rho,j}$, and define $P(\cdot)$ to be the collection of principal components with normalized energy higher than 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 way of doing 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.

The second step of the optimization starts by assuming that $\rho_0(x)$, $P(x)$, and α_1^τ are known, and we are interested in solving 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 decide to split it in two steps. In the first one we will solve the following problem:

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

which will allow us to estimate $w_0(x)$ and $W(x)$, while in the second one we will estimate the warping state $s_1^\tau = [s_1, \cdots, s_t]$ by solving

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

We are going to discuss the minimization (11) and (12) in the following two sections respectively.

3.1.1 Estimation of the nominal and principal warps

In solving problem (11) we are interested in guaranteing that: (a) $w_t(x)$ is a homeomorphism in x, and this is because in (1) we require the warping to be invertible as it cannot handle occlusions; (b) $w_t(x)$ varies smoothly in time, as we expect that two adjacent images of a video sequence have not changed much (and this is in accordance with (4)); (c) $w_t(x)$ is Guaussian distributed around the nominal warp $w_0(x)$, and this is to satisfy (2), and partly (4). It is obvious that none of the three 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 $u(x,t) \doteq [w_t^T(x) - w_0^T(x) \;\; 0]^T$. 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 [13].

The problem of estimating the warping $w_t(x)$ 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 finite-element methods [13]. 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 $\bar{v}_0^T = [v_0^T, \cdots, v_0^T]$ 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 [14] 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

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

that can efficiently be solved iteratively by using the so called inverse compositional image alignment algorithm described in [4, 5], while its modifications for the case that handles priors can be found in [3].

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

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

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 assumption 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, 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, since the regularization that we are going to describe in the next section addresses exactly this issue, not imposing time smoothness at this stage is a problem that can be overcome in most of the cases.

Once we obtain an estimate for v, we proceed by updating the nominal warp, and this can be done by computing the following sample mean

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

To compute the principal warps $W(\cdot)$, once we remove the nominal warp from the dataset W_1^τ, one can compute its singular value decomposition $U_w(x)\Sigma_w V_w^T$, 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 the shape state 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 energy, we define $W(\cdot)$ to be the collection of principal components in $U_w(\cdot)$ with normalized energy higher than a certain threshold γ_w:

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

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

$$s_1^\tau = \Sigma_{w,1:k,1:k} V_{w,:,1:k}^T \ . \tag{19}$$

3.1.2 Estimation of the shape state and dynamic parameters

At this stage the estimate of the shape state s_1^τ needs to be updated as a consequence of the fact that $w_0(x)$ and $W(\cdot)$ are known. Moreover, we are interested in learning a state in such a way that its temporal statistics is second-order stationary as we plan to model it with a linear dynamical system. In other words, the minimization of (12) has to be done subject to the prior model (4). To this end, we estimate s_1^τ by solving the following problem

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

Where the prior aims at minimizing the Frobenius norm of the residuals. Again, this minimization can be performed by using the inverse compositional algorithm [4]. Note that at each step of the algorithm, the matrix A needs to be estimated as it appears in the derivative of the prior. Given the limited space, we omit the derivation of the prior derivative, which is easy but tedious as it involves tensor algebra computations. Suffices to say that the matrix A is estimated via least squares, which means that, if $\xi_{t_1}^{t_2} = [\xi_{t_1}, \cdot, \xi_{t_2}]$, then A is computed by the following expression

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

[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 [13].

Sequence	l: PCA	l: AAM	k: AAM	RMSE$_1$: PCA	RMSE$_2$: PCA	RMSE$_3$: AAM
Test	4	2	3	7.3626	9.9951	2.3870
Flowers	22	19	6	4.0272	4.4273	4.1290
Candle	11	7	7	2.1146	2.9134	2.1484
Duck	16	11	6	1.6801	1.8580	1.6001
Flag	18	10	8	2.9926	3.6392	3.2491

Table 1: First three columns: dimensions of the state space by using PCA (dynamic textures), dimensions of the appearance state, dimensions of the shape estate. Last three columns: root mean square error per pixel for two PCA models and the active appearance model.

while the covariance of the driving noise n_t can be computed as

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

Note that in this simple exposition we have implicitly assumed that the matrix C is the identity matrix.

4. Results

In this section we will briefly describe the experiments we did to validate both the learning procedure and the correctness of the proposed model.

To validate the learning procedure, figures (1) and (2) show the results on a synthetic sequence. This sequence has been created by the nominal appearance and appearance components depicted in the row (a) of figure (1), then a random stationary motion has been applied. The convergence of the learning procedure is better illustrated by figure (2) (a) and (b). The first plot shows the evolution of the appearance normalized energy $\{\tilde{\sigma}_{\rho,i}\}$, while the second shows the evolution of the shape normalized energy $\{\tilde{\sigma}_{w,i}\}$. As the learning approaches convergence, the dimensions of the appearance state l and the shape state k, converge to the right values, namely 2 for the appearance, and 3 for the shape, since the square is subject to a planar rigid motion (see movie uploaded with the submission). Figure (1) also shows the estimated nominal appearance, appearance components (row (b)), and shape components (middle row).

Table (1) shows how well the model can represent the original dataset for a given dimensionality of the models, and for five video sequences. The root mean square errors RMSE$_1$ is the reconstruction error per pixel when the simple PCA is used, while RMSE$_3$ is the reconstruction error per pixel when our active appearance model is used. The relationship between the two models is established by the fact that the dimensionality of both of them was automatically computed by choosing the same appearance energy threshold γ_ρ. On the other hand, the error RMSE$_2$ was computed by retaining the same number of principal components that are retained in the active appearance model.

Finally, we validate the ability of the model, and learning procedure to capture the spatio-temporal information carried by a video sequence by using the model to extrapolate new images. For a sequence of flowers, a candle, a toy duck, and a flag, figure (3) shows one frame of the original sequence, the same frame with the superimposed grid identified by v_t, one frame synthesized by using PCA (that is a sample synthesized by using dynamic textures [11]), and a frame synthesized with the new active appearance model (see movie uploaded with the submission).

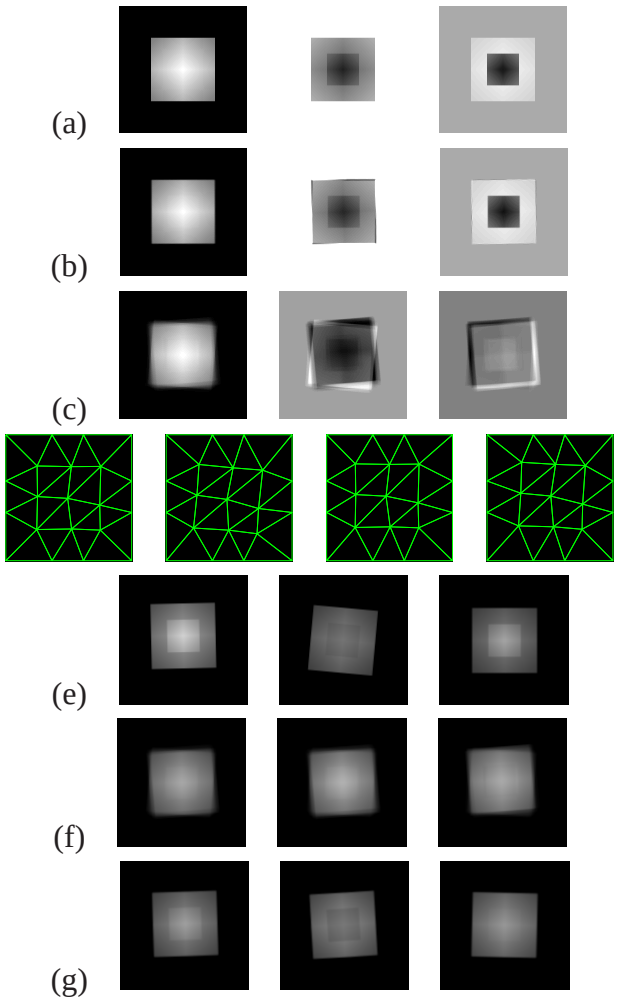

Figure 1: Top to bottom: (a) nominal appearance and appearance components (ground truth); (b) nominal appearance and appearance components (estimated); (c) mean and first two principal components; (d) shape components (estimated); (e) original sequence; (f) synthesis with PCA (dynamic textures); (g) synthesis with the active appearance model.

5. Conclusions

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

Our approach can be though of as extending the work on Active Appearance Models [6, 9] to the temporal domain, or extending dynamic textures [11] to the spatial domain.

We have presented a variational formulation of the problem, and an efficient computational solution that uses standard numerical approaches (finite-element methods).

We have illustrated the modeling power of our approach in terms of extrapolation power (prediction-error) and uploaded numerous movies to qualitatively display the behavior of these models during synthesis.

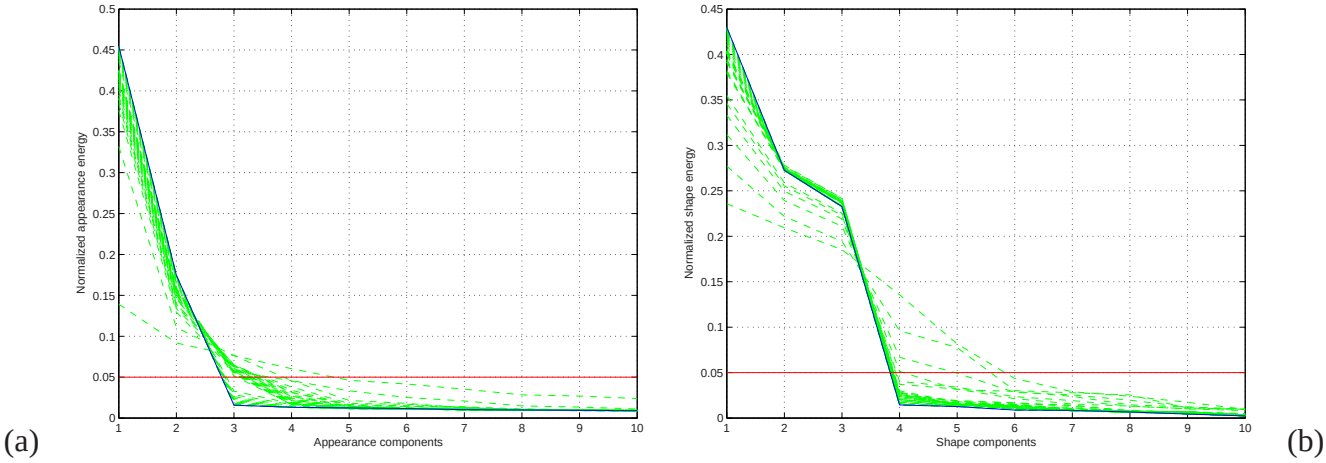

Figure 2: (a) evolution of the appearance energy, $\gamma_\rho = 0.05$ (solid line represents convergence); (b) evolution of the shape energy, $\gamma_w = 0.05$ (solid line represents convergence)

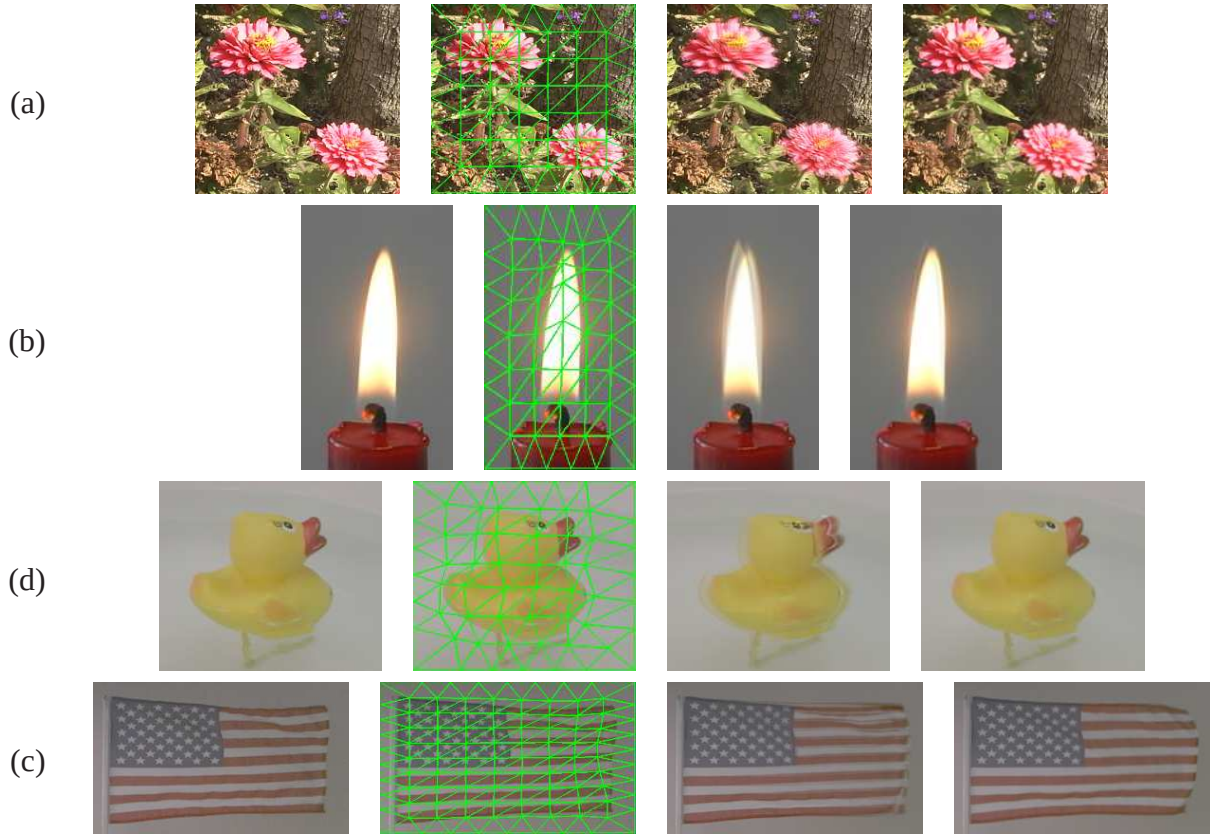

Figure 3: Top to bottom: (a) flowers sequence, (b) candle sequence, (c) toy duck sequence, (d) flag sequence. Left to right: original frame, original frame with mesh v_t, synthesized frame with PCA, synthesized frame with the active appearance model. (See the movie uploaded with the submission.)

References

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

[2] The Authors. Technical report.

[3] S. Baker, R. Gross, and Matthews I. Lucas-kanade 20 years on: A unifying framework: Part 4. Technical Report CMU-RI-TR-04-14, CMU, Pittsburgh, PA, February 2004.

[4] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proc. of CVPR*, volume 1, December.

[5] S. Baker and I. Matthews. Lucas-kanade 20 years on: a unifying framework. *Int. J. Computer Vision*, 56(3):221–255, 2004.

[6] S. Baker, I. Matthews, and J. Schneider. Image coding with active appearance models. Technical report, Carnegie Mellon University, The Robotics Institute.

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

[8] H. Chen, P. Belhumeur, and D. Jacobs. In search of illumination invariants. Technical report, Yale University and Princeton, Department of Physics, CS, EE and NEC Research Institute.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. of the Eur. Conf. on Comp. Vis.*, pages 484–496, 1998.

[10] T.F. Cootes, S. Marsland, C.J. Twining, K. Smith, and C.J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *Proc. ECCV*, May.

[11] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *Int. J. Computer Vision*, 51:91–109, 2003.

[12] A. W. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2003.

[13] T. J. R. Hughes. *The Finite Element Method - Linear Static and Dynamic Finite Element Analysis*. Dover Publishers, New York, 2000.

[14] I. Matthews and S. Baker. Active appearance models revised. *Int. J. Computer Vision*, 60(2):135–164, 2004.

[15] M. I. Miller and L. Younes. Group action, diffeomorphism and matching: a general framework. In *Proc. of SCTV*, 1999.

[16] B. Schölkopf and A. Smola. *Learning with kernels: SVM, regularization, optimizationand beyond*. MIT press, Cambridge, 2002.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neurosci.*, 3:71–86, 1991.

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

[19] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: exploiting reciprocity for surface reconstruction. In *Proc. of the ECCV*, pages 869–884, 2002.