System Theoretic Methods in Computer Vision and Image Processing

Octavia Camps

Mario Sznaier

***Abstract*— Dynamic vision and imaging systems have the potential to substantially improve our quality of life. However, key issues that must be addressed in order to deploy these systems in unstructured environments are their potential fragility and the need to process vast amounts of information in real time. As we show in this paper, these issues can be addressed by appealing to a common systems theoretic substrate that allows for recasting a wide range of problems into a tractable convex optimization form. These ideas are illustrated with several applications including multiframe tracking, motion segmentation, texture analysis/synthesis and video reconstruction and inpainting.**

I. Introduction

Dynamic vision and imaging – the confluence of dynamics, computer vision, image processing and control – is uniquely positioned to enhance the quality of life for large segments of the general public. Aware sensors endowed with tracking and scene analysis capabilities can prevent crime, reduce time response to emergency scenes and allow elderly people to continue living independently. Enhanced imaging methods can substantially reduce the amount of radiation required in medical imaging procedures and in cancer therapy. Moreover, the investment required to accomplish these goals is relatively modest, since a large number of imaging sensors are already deployed and networked. For instance, the number of outdoor surveillance cameras in public spaces is already large (10,000 in Manhattan alone), and will increase exponentially with the introduction of camera cell phones capable of broadcasting and sharing live video feeds in real time. The challenge now is to develop a theoretical framework that allows for *robustly* processing this vast amount of information, within the constraints imposed by the need for real time operation in dynamic, partially stochastic scenarios. The goal of this paper is to illustrate the central role that dynamic models and their associated predictions can play in developing a comprehensive, computationally tractable robust dynamic vision and imaging framework. Establishing a connection with a rich set of robust systems theory tools allows for recasting a wide spectrum of problems arising in this context – robustly tracking an object in a sequence of frames, modelling appearance changes, recovering structure from motion, recognizing classes of activities, and classifying textured images –into a tractable, finite dimensional convex optimization.

II. Notation

$\overline{\sigma}(\mathsf{A})$	maximum singular value of A.
$\mathcal{H}_{\infty,\rho}$	space of transfer functions analytic in $\lvert z\rvert \leq \rho$, equipped with the norm $\|G\|_{\infty,\rho} \doteq ess\sup_{\lvert z\rvert<\rho} \overline{\sigma}(G(z))$. The case $\rho = 1$ will be denoted as usual simply by $\mathcal{H}_\infty$
$\mathcal{BH}_\infty(K)$	open K–ball in $\mathcal{H}_\infty$
ℓ_p	space of vector valued sequences equipped with the norm: $\|x\|_p^p \doteq \sum_{i=0}^{\infty} \|\mathbf{x}_i\|_p^p$, $p \in [1,\infty]$ and $\|x\|_\infty \doteq \sup_i \|\mathbf{x}_i\|_\infty$.

Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, {camps,msznaier}@ece.neu.edu

III. Interpolation Problems in Dynamic Vision

In this section we show that many dynamic vision problems such as robustly tracking an object in a sequence of frames, obtaining structure from motion and motion segmentation can be reduced to a convex optimization problem, through the use of well established system–theoretic tools.

A. Multiframe Tracking

A requirement common to most dynamic vision applications is the *ability to track* objects in a sequence of frames. Current approaches integrate correspondences between individual frames over time, through a combination of target dynamics, empirically learned noise distributions and past position observations [4, 8]. However, while successful in many scenarios, these approaches still remain vulnerable to model uncertainty, occlusion and appearance changes, as illustrated in Figure 1.

As we show next, this difficulty can be solved by modelling the motion of the target as the output of a dynamical system, to be indentified from the available data. To this effect, start by modelling y_k, the present position of a given target feature as:

$$y(z) = \mathcal{F}(z)e(z) + \eta(z) \tag{1}$$

where $\mathbf{e}$ and $\eta_k \in \mathcal{N}$ represent a suitable input and measurement noise, respectively, and where the operator $\mathcal{F}$ is not necessarily ℓ_2 stable. Further, we will assume that the following *a priori* information is available:

(a) Set membership descriptions $\eta_k \in \mathcal{N}$ and $e_k \in \mathcal{E}$. These can be used to provide deterministic models of the stochastic signals e, η.

(b) $\mathcal{F}$ admits an expansion of the form $\mathcal{F} = \overbrace{\sum_{j=1}^{N_p} p_j \mathcal{F}^j}^{\mathcal{F}_p} + \mathcal{F}_{np}$. Here $\mathcal{F}^j$ are known, given, not neces-

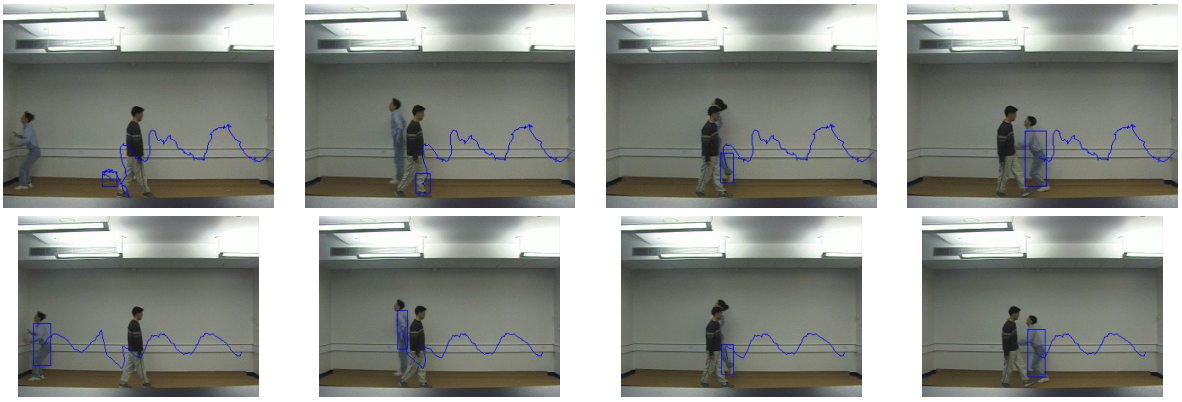


Fig. 1. Tracking in the presence of occlusion. Top: Unscented Particle Filter based tracker loses the target due to occlusion. Bottom: Combination Identified Dynamics/Kalman Filter tracks through the occlusion.

sarily ℓ_2 stable operators that contain all the information available about possible modes of motion of the target.

(c) $\mathcal{F}_{np} \in \mathcal{BH}_{\infty,\rho}(K)$ for some known $\rho \leq 1$, e.g. a bound on the divergence rate of the approximation error of the expansion $\mathcal{F}_p$ to $\mathcal{F}$ is available.

In this context, the next location of the target feature y_k can be predicted by first identifying the relevant dynamics $\mathcal{F}$ and then using it to propagate its past values. In turn, identifying the dynamics entails finding an operator $\mathcal{F}(z) \in \mathcal{S} \doteq \{\mathcal{F}(z)\colon \mathcal{F} = \mathcal{F}_p + \mathcal{F}_{np}\}$ such that $y - \eta = \mathcal{F}e$, precisely the class of interpolation problem addressed in [10]. As shown there, such an operator exists if and only if the following set of equations in the variables $\mathbf{p}, \mathbf{h}$ and K is feasible:

$$\mathsf{M}_R(\mathbf{h}) = \begin{bmatrix} \mathsf{R}_\rho^2 & \mathsf{T}_h^T \\ \mathsf{T}_h & K^2\mathsf{R}_\rho^{-2} \end{bmatrix} \geq 0 \tag{2}$$

$$\mathbf{y} - \mathsf{T}_e\mathsf{P}\mathbf{p} - \mathsf{T}_e\mathbf{h} \in \mathcal{N} \tag{3}$$

where T_x denotes the Toeplitz matrix associated with a given sequence $\mathbf{x} = [x_1, \ldots, x_n]$, $\mathsf{R}_\rho \doteq \mathrm{diag}\,[1\ \rho\ \cdots\ \rho^n]$, $\mathsf{P} \doteq [f^1\ f^2\ \cdots\ f^{N_p}]$,where f^i is a column vector containing the first n Markov parameters of the i-th transfer function $\mathcal{F}^i(z)$ and $\mathbf{h}$ contains the first n Markov parameters of $\mathcal{F}_{np}(z)$

A Simple Tracking Example: Consider again the problem illustrated in Figure 1. The experimental information consists of centroid position measurements from the first 20 frames, where the target is not occluded. The *a priori* information, estimated from the non–occluded portion of the trajectory is:

1) 5% noise level
2) $\mathcal{E} = \delta(0)$, i.e. motion of the target was modelled as the impulse response of the unknown operator F[1].
3) $\mathcal{F}_p \in \mathrm{span}[\frac{1}{z-1},\ \frac{z}{z-a},\ \frac{z}{(z-1)^2},\ \frac{z^2}{(z-1)^2},\ \frac{z^2-\cos\omega z}{z^2-2\cos\omega z+1},\ \frac{\sin\omega z^2}{z^2-2\cos\omega z+1}]$ where $a \in \{0.9, 1, 1.2, 1.3, 2\}$ and $\omega \in \{0.2, 0.45\}$
4) $F_{np} \in \mathcal{BH}_{\infty,\rho}(K)$, with $\rho = 0.99$

As shown in Figure 1, a tracker that uses the identified dynamics for prediction is now able to track the target past the occlusion. It is worth emphasizing that the combination identified dynamics/Kalman filter significantly outperforms a tracker based solely on an unscented particle filter [4], even though the latter has substantially higher computational complexity. Hence, exploiting dynamical information through the use of control–motivated tools, leads to *both* robustness improvement and substantial computational complexity reduction.

Subsampling and data gating: A salient feature of the framework described above is its ability to furnish *deterministic, worst–case bounds* on the prediction error that can be used to disambiguate among targets with a low computational cost. Specifically, given a sequence $\{y_k\}_{k=0}^{N-1}$ of measurements of the location f_k of the feature, define the consistency set as:

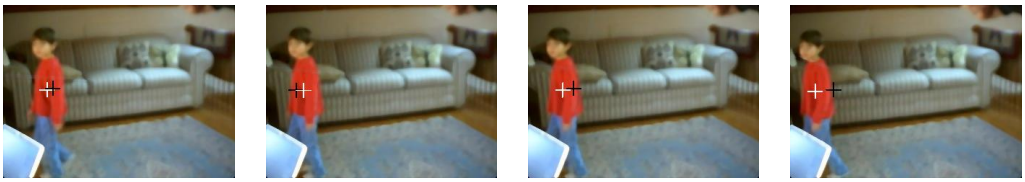
$$\mathcal{T}(\mathbf{y}) \doteq \left\{\mathcal{F} \in \mathcal{S}\colon \{y_k - (\mathcal{F} * e)_k\}_{k=0}^{N-1} \in \mathcal{N}\right\} \tag{4}$$

i.e, the set of all models consistent with both the *a priori* information and the experimental data. Since both, the "true" operator $\mathcal{F}_o$ that maps the input $\mathbf{e}$ to the feature locations $\mathbf{f}$ and the identified one belong to the consistency set, it follows that, given the first N measurements $\{y_k\}_{k=0}^{N-1}$, a bound on the worst case prediction error over the horizon $[0, M-1]$, $M > N$, is given by:

$$\begin{aligned} \|\hat{\mathbf{f}} - \mathbf{f}\|_{\ell_\infty[0,M-1]} &\leq \sup_{\mathcal{F}_i \in \mathcal{T}(\mathbf{y})} \|\mathcal{F}_1 e - \mathcal{F}_2 e\|_{\ell_\infty[0,M-1]} \\ &\leq 2 \sup_{\mathcal{F} \in \mathcal{T}(\mathbf{0})} \|\mathcal{F}\|_{\ell_\infty[0,M-1]} \end{aligned}$$

where the last inequality follows from standard information based complexity arguments (see for instance Lemma 10.3 in [12]). When $\mathcal{N}$ is convex, computing this bound reduces to

[1]This is equivalent to lumping together the dynamics of the plant and the input signal.



Frame 14 Frame 16 Frame 18 Frame 20

Fig. 2. Prediction (black) versus Ground Truth (white))

a convex optimization problem. In particular, the case $\mathcal{N} = \{\eta\colon \|\eta\|_\infty \leq \eta_{\max}\}$ leads to a Linear Programming problem.

Frame	13	15	17	20
Actual error	8.87	10.04	10.31	26.05
Worst case bound	13.00	17	21	27

Fig. 3. Propagation of the Prediction error. Target width is 30 pixels.

Figure 3 compares the actual and upper bound of the error for the sequence partially shown in Figure 2. In this experiment, the position measured in frame 12 was propagated forward using the identified dynamics and the bounds computed by solving a single LP problem. Note that this procedure extends trivially to the case of several targets, each with its own dynamics, providing an effective tool for disambiguating targets with neighboring tracks, since candidate features that fall outside these bounds can be discarded. In addition, these bounds provide a mechanism to balance computational requirements and data obsolescence, by establishing *a priori* that no new data is required from Frame 12 until Frame 20, where the error becomes comparable with the width of the target.

B. Dynamic Appearance Modelling and Computational Complexity Issues.

Arguably, one of the hardest challenges in tracking is to overcome changes to its appearance, due to factors such as target motion, self-occlusion, target articulations, and changes in illumination. In principle, this difficulty can be solved by using *dynamic* appearance models that incorporate time–evolution information and have better predictive capabilities. In turn, as argued in [5], these models can be obtained using the same robust identification approaches employed to identify the motion dynamics. However, moving beyond a few simple descriptors requires addressing the issues of high computational costs, due to the poor scaling properties of LMI based identification algorithms[2].

A possible way of addressing the challenge noted above is through the use of recently introduced nonlinear dimensionality reduction techniques to map the data to a lower dimensional manifold where the identification/tracking is performed. Since the projection onto the lower dimensional manifold can be modelled as a static nonlinearity, this approach leads naturally to a Wiener system structure of the form illustrated in Figure 4, consisting of the interconnection of a LTI system $H(z)$ and a memoryless nonlinearity $f(.)$. The effectiveness of this approach in substantially reducing the computational complexity entailed in tracking complex targets is illustrated next using the problem of human motion modelling and tracking. The experimental data, partially shown in Figure 5(a) consists of the first 20 frames of a human walking on a treadmill, each having 1728 pixels. Thus, modelling pixel evolution become infeasible even when using just a few frames. On the other hand using the risk–adjusted approach proposed in [6] (recall that the computational complexity of this approach scales polynomially with the number of data points) and the following *a priori* information

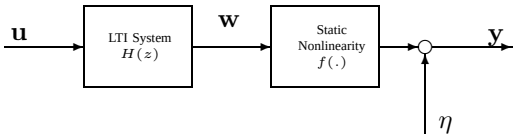


Fig. 4. Wiener System Structure

1.- $\omega \in R^3$ (this hypothesis is motivated by the physics of the problem, where ω is related to the coordinates of the centroid of the target).

2.- The static nonlinearity $\mathbf{f}(.)$ has the form[3]: $\mathbf{f}(\mathbf{x}) = \mathbf{B}\mathbf{\Psi}(\mathbf{x})$ where $\mathbf{B} \in R^{1726\times 6}$ is an unknown matrix and the bases $\mathbf{\Psi}(\mathbf{x})\colon R^3 \to R^6$ are given by:

$$\mathbf{\Psi}(\mathbf{x}) = [\exp(-0.8\|\mathbf{x}-\mathbf{t_1}\|_2^2), \exp(-0.8\|\mathbf{x}-\mathbf{t_2}\|_2^2), 1, \mathbf{x}^T]^T$$

where

$$\mathbf{t_1} = \begin{bmatrix} 0.6833 & -0.4521 & -0.0033 \end{bmatrix}$$
$$\mathbf{t_2} = \begin{bmatrix} -0.7552 & 0.4997 & 0.0036 \end{bmatrix}$$

led to model with a fifth order linear portion that interpolates the data within 10%. The predictive power of this model is shown in the bottom portion of Figure 5(a). Finally, Figure 5(b) shows close agreement between the temporal

[2]Recall that the computational complexity of conventional LMI solvers scales as (number of decision variables)^{10}[9].

[3]This hypothesis is motivated by the bases proposed in [2] to map human silhouettes to lower dimensional spaces.

evolution of the points on the manifold and the positions predicted using the linear dynamic model. This substantiates the conjecture, originally posed in [7], that human motion tracking can be decoupled into two problems: (a) a linear tracking problem in a low dimensional manifold, accounting for the *dynamics* of the motion, and (b) a nonlinear, static mapping that accounts for the changes in appearance of the target.

C. Structure and Motion Recovery from Dynamics:

When tracking an unknown number N_o of moving objects, it is of interest to identify (i) the number of objects, (ii) the individual dynamics and, (iii) assign points in the image to each. To illustrate the issues involved, start by considering P features from a single rigid object, tracked over F frames with image coordinates $\{(u_t^p, v_t^p)\}$, $p = 1, \ldots, P$, $t = 1, \ldots, F$. Define the measurement matrix $W_{1:F}$, by:

$$\mathcal{W}_{1:F} = \begin{bmatrix} u_t^p - u_t \\ v_t^p - v_t \end{bmatrix} \in R^{2P \times F} \tag{5}$$

where (u_t, v_t) denote coordinates of the centroid of the features. Under the assumptions of affine projection it can be shown [14] that $\mathcal{W}_{1:F}$ has at most rank 3 and can be decomposed into a "rotation" matrix $R_{1:F}$ and a "structure" matrix S

$$\mathcal{W}_{1:F} = \begin{bmatrix} R_{1:F}^u \\ R_{1:F}^v \end{bmatrix} S = R_{1:F} S \tag{6}$$

In the case of multiple objects, the number of objects and the corresponding geometry can be obtained by factoring $\mathcal{W}$ into rank 3 submatrices. This basic idea lies at the core of factorization based approaches (see for instance [16, 15]), leading to computationally efficient solutions. However, its success hinges upon identifying the correct point-correspondences across frames. Thus, it is sensitive to noise, partial occlusion or large affine warping of feature templates due to large inter frame rotations. In such cases, predictions provided by estimating the dynamics of the moving objects can play a critical role in connecting previous measurements with current data. The main idea is to parametrize R_{t+1} the rotation matrix between frames 1 and $t+1$ as $R_{t+1} = e^{j\omega_t} R_t$, and *identify* the dynamics governing the time evolution of ω from past data. This leads to a hybrid, *bootstrap*–type approach, where, at any given instant, a factorization of $\mathcal{W}$, is used to learn the dynamics of the time–varying motion of the object(s). In turn, these dynamics are used to predict future feature positions that can be used to disambiguate tracks, or even fill in for partially missing data, avoiding the need for dropping frames where not all features are present. In addition, the associated error bounds can be used to limit the size of search windows. The potential of this approach is illustrated in Figure 6, comparing a purely factorization–based approach against the proposed hybrid one while reconstructing a stuffed teddy bear. As shown in Figure 6(b), bootstrapping SfM with identification of the motion dynamics resulted in significantly smaller ratios between the fourth and third singular values of $\mathcal{W}$, indicating a significant improvement of the tracking data (recall that for a single object, ideally we should have $\text{rank}(\mathcal{W}) = 3$).

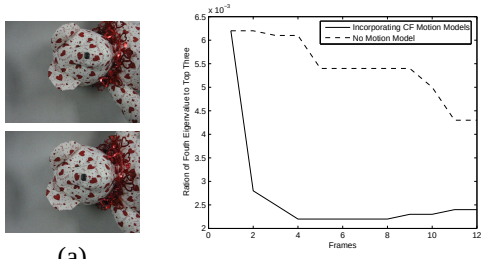


Fig. 6. (a) First and last frames. (b) Ratio of the fourth to the third singular values.

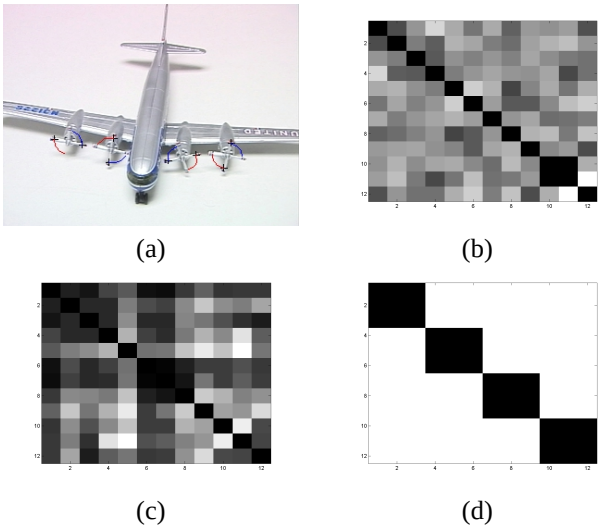
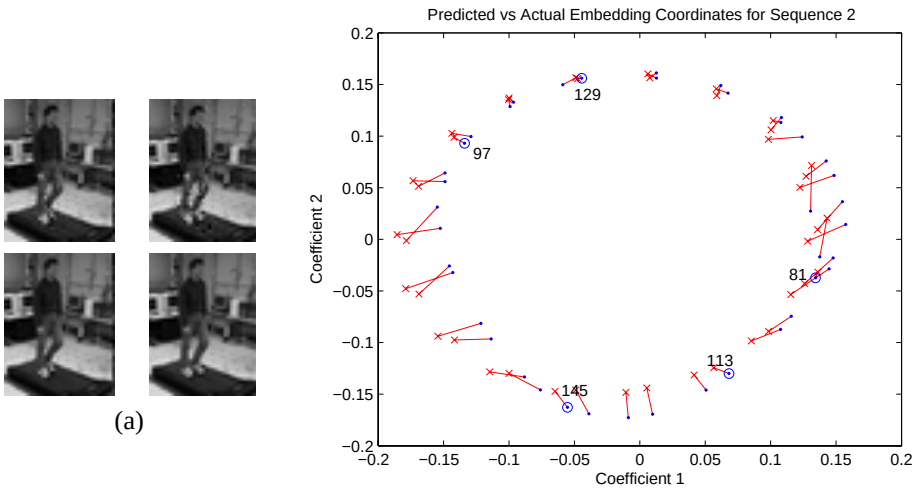


Fig. 7. (a) All propellers move at the same speed. Right wing propellers move counterclockwise, while left wing propellers move clockwise. (b) Costeira-Kanade motion segmentation. (c) Zelnik-Manor-Irani motion segmentation using six eigenvectors. (d) Dynamics based motion segmentation.

A second source of fragility in currently existing approaches stems from the difficulty in disambiguating objects that partially share motion modes, such as the same–wing propellers of the airplane shown in Figure 7(a). It can be easily shown that in this case $\text{rank}(\mathcal{W}) = 6$. Hence, as shown in Figure 7 (b)–(c), any motion segmentation approach based solely on finding linearly independent subspaces of the column space of $\mathcal{W}$ will fail, since it cannot distinguish this case from the case of two independently moving propellers. Intuitively, the main difficulty here is that any approach based on properties of $\mathcal{W}$ that are invariant under column permutations, *take into account only geometrical constraints, but not dynamical ones*.

As we show next, robustness can be substantially improved by grouping points according to the complexity of the model



(a) (b)

Fig. 5. Learning the dynamics of complex appearances using a Wiener system. (a) Top: Walking sequence (from CMU MoBo database), Bottom: impulse response of the identified Wiener system. (b) Evolution on a 2D projection of the 3D manifold: predicted(red) and actual(blue).

required to explain their relative motion. Intuitively, this formalizes the idea that points on the same rigid share more modes of motion than points on different objects, leading to less complex models. Specifically, begin by associating to the j^{th} object, its centroid $\mathbf{O}^{(j)}$ and an affine basis $b^{(j)}$, centered at $\mathbf{O}^{(j)}$, defined by three no coplanar vectors $\mathbf{V}_i^{(j)}$. Finally, denote by $o^{(j)}(k)$, $v_i^{(j)}(k)$ the coordinates of the image of $\mathbf{O}^{(j)}(k)$ and the projections of $\mathbf{V}_i^{(j)}(k)$ onto the image plane, respectively. Given any point $\mathbf{P}_i^{(j)}$ belonging to the j^{th} object, the coordinates at time k of its image $\mathbf{p}^{(i)}(k)$ are given by:

$$\mathbf{p}_i^{(j)}(k) = \mathbf{o}^{(j)}(k) + \alpha_i^{(j)}\mathbf{v}_1^{(j)}(k) + \beta_i^{(j)}\mathbf{v}_2^{(j)}k + \gamma_i^{(j)}\mathbf{v}_3^{(j)}(k)$$

where $\alpha_i^{(j)}, \beta_i^{(j)}$ and $\gamma_i^{(j)}$ are the *affine invariant* coordinates of $\mathbf{P}_i^{(j)}$ with respect to the basis $b^{(j)}$. Note that, for any two points $\mathbf{P}_r^{(j)}, \mathbf{P}_s^{(j)}$ in the same object, the dynamics of $\mathbf{o}^{(j)}$ are *unobservable* from $\delta_{r,s}(k) \doteq \mathbf{p}_r^{(j)}(k) - \mathbf{p}_s^{(j)}(k)$. Thus, the underlying subsystem is rank deficient when compared to a subsystem describing difference between points on different objects. Roughly speaking, the relative motion of points in a given object, carries no information about the motion of other objects. It follows that points can be clustered in objects according to the complexity of the model required to explain their relative motion. In turn, the order of this model can be estimated by simply computing the rank of the Hankel matrix constructed from the pair-wise differences $\delta_{rs}(k)$, leading to a simple segmentation algorithm, computationally no more expensive than a sequence of SVDs. The effectiveness of this approach is illustrated in Figure 7(d), showing that it correctly identified the presence of four independently moving objects.

IV. Interpolation Problems in Textured Image Processing

Texture analysis, classification and synthesis has been the subject of research in image processing, computer graphics and computer vision for over three decades, with applications ranging from medical diagnosis to entertainment to human computer interfaces. During the past few years, significant advances have been made in addressing multiple aspects of the problem, ranging from inpainting and synthesis to classification. However, at present, each sub-problem is addressed using a specific set of tailored tools, only loosely connected to those used to solve other subproblems. For instance, most texture recognition schemes rely on representations in terms of statistics of the responses to a collection of filters [3]. On the other hand, most synthesis approaches look at texture as the samples of some probabilistic distribution [3]. The objective of this section is to briefly illustrate how the use of system theoretic tools can lead to a unified framework capable of exploiting the synergism between different aspects of the problem to improve robustness and reduce the computational burden.

A. Texture Modelling and Synthesis

Compact models of textured images can be obtained by treating the intensity values $\mathcal{I}(k,l)$ at the (k,l) pixel of the image as the output of a second order stationary stochastic process. Equivalently, $\mathcal{I}(k,l)$ can be modelled as the output of a *two-dimensional*, discrete linear shift-invariant system driven by white noise, reducing the problem to an identification one: obtaining a model G from image data, possibly corrupted by noise. Note however that while most currently existing identification techniques deal with causal, one-dimensional systems, texture modelling requires

considering two–dimensional, *non–causal* systems, since the intensity value at a pixel is likely to depend on the values of all pixels in its neighborhood, not just on those preceding it in some ordering of the image pixels. This difficulty can be circumvented by considering a given $n \times m$ image as one period of an infinite 2D signal with period (n, m). Thus, at any given location (i, j) in the image, the intensity values $\mathcal{I}(r, s)$ at other pixels are available also at position $(r - qn, s - qm)$, and the integer q can always be chosen so that $r - qn < i, s - qm < j$. From this observation, it follows that the unknown system G admits a state space representation of form:

$$\begin{aligned} x'(i,j) &= Ax(i,j) + Bu(i,j) \\ \mathcal{I}(i,j) &= Cx(i,j) + Du(i,j) \end{aligned} \tag{7}$$

where

$$x'(i,j) = \begin{bmatrix} x^v(i+1,j) \\ x^h(i,j+1) \end{bmatrix}, x(i,j) = \begin{bmatrix} x^v(i,j) \\ x^h(i,j) \end{bmatrix}$$

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$$

subject to an additional constraint of the form

$$\begin{aligned} g(i+N,j) &= g(i,j) \\ g(i,j+M) &= g(i,j) \end{aligned} \quad \text{for some finite } N, M > 0$$

where $g(.,.)$ denotes the impulse response of G. With these assumptions, the problem becomes one of identifying a state–space realization from experimental data, subject to a periodicity constraint, precisely the type of problems solved in [1]. The potential of this approach is illustrated in Fig. 8, where it was used to expand partial images by first identifying the underlying model and then simply computing its impulse response.

Fig. 8. Using 2-D Models to Expand Images

B. Texture Classification

In this section we briefly indicate how the models obtained above can be used for texture classification. Proceeding as in [13], we will recast the problem into a robust semi-blind model (in)validation form. To this effect, we will postulate that all images corresponding to realizations of a given texture $\mathcal{T}$ can be obtained as the output of a 2-D operator S to an unknown input signal e with unit spectral density, applied in $(-\infty, 0] \times (-\infty, 0]$. This leads to the set-up shown in Figure 9, where $T(z_1, z_2)$ represents a nominal model of a particular texture, $h(i, j)$ and $y(i, j)$ denote the intensity value of the ideal and actual images, respectively, and where

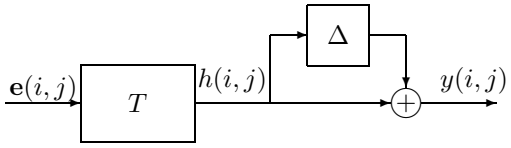


Fig. 9. The Texture Recognition Set-up

the (unknown) operator $\Delta(z_1, z_2)$ describes the mismatch between these two images.

In this context, given a set of texture families, each represented by a model T_i, an unknown specimen can be classified by (i) performing a sequence of invalidation models to find the lowest uncertainty value $\|\Delta_i\|$ required to explain the specimen in terms of the model T_i, and (ii) assigning the unknown texture to the family corresponding to smallest uncertainty norm. By using the proposed identification technique to obtain a (separable) model of the nominal texture, the corresponding 2-D model invalidation problem can be reduced to two decoupled 1-D semi–blind validation problems that can be solved using the LMI–based technique developed in [13].

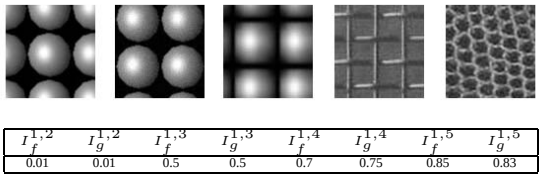


Fig. 10. Top: Sample Textures. Bottom: Optimal γ

Figure 10 shows the results of applying the technique outlined above to classify several images. Here $I_f^{1,j}$ and $I_g^{1,j}$ denote the results obtained when comparing the decompositions corresponding to the first image against the models obtained from the j^{th} texture. As shown in the table, the proposed technique correctly indicates that the first three images belong to the same family[4].

C. Video Inpainting as a Rank Minimization Problem

Video inpainting, that is the process of seamlessly restoring or altering portions of a video clip, has been the subject of considerable attention in the past few years (see for instance [11] and references therein), but the problem is far from solved. Existing algorithms are limited in the types of sequences that can handle and have relatively high computational complexity. In this section we briefly outline how the use of Systems Theory ideas can lead to simple, computationally efficient algorithms that exploit (global)

[4]The higher values of $I_f^{1,3}$ and $I_v^{1,3}$ are due to the use of a lower quality image for the third texture.

spatio–temporal information. The main idea is to (i) find a set of descriptors that encapsulate the information necessary to reconstruct a frame, (ii) find an optimal estimate of the value of these descriptors for the missing/corrupted frames, and (iii) use the estimated values to reconstruct the frames. In turn, the optimal descriptor estimates can be efficiently obtained postulating that the correct values of the missing descriptors are such that the resulting inpainted sequence is described by the simplest possible (eg. lowest order) dynamical model[5]. Since the order of the underlying model can be estimated from the Hankel matrix of the data, this idea leads to a rank minimization problem, which in turn can be relaxed to an LMI optimization, resulting in the following algorithm:

1.- Given the observed values of the descriptors f^o, form the following (Hankel) matrix:

$$H_f \doteq \begin{bmatrix} f_1 & f_2 & \cdots & f_{n/2} \\ f_2 & f_3 & \cdots & f_{n/2+1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n/2} & f_{n/2+1} & \cdots & f_{n-1} \end{bmatrix} \quad (8)$$

Here f denotes either the observed data f_k^o, if the k frame is present, or the unknown value f_k^m, if the frame needs to be inpainted, and n denotes the total number of frames.

2.- Estimate the values f^m which are maximally consistent with f^o by solving the following Linear Matrix Inequality (LMI) optimization problem,

$$\begin{array}{lll} \text{minimize w.r.t} & f^m & Tr(Y) + Tr(Z) \\ \text{subject to} & & \begin{bmatrix} Y & H_f \\ (H_f)^T & Z \end{bmatrix} \geq 0 \end{array}$$

where $Y^T = Y \in \mathcal{R}^{n\times n}$, $Z^T = Z \in \mathcal{R}^{n\times n}$ and $H_f \in \mathcal{R}^{n\times n}$.

The potential of this approach is illustrated in Fig. 11, where it was used to restore the occluded person. In this particular example, the positions $f_k = (x_k^i, y_k^i)$ of the 6 feature points indicated in the figure were chosen as descriptors. The video has 36 frames, and occlusion occurs in frames 17 through 19. Using the algorithm outlined above implemented in MATLAB to inpaint the missing descriptors required approximately 20 seconds on a P-III 1.2G PC.

V. Conclusions

Dynamic vision and imaging is arguably one of the few areas where both further advances and widespread field deployment are being held up not by the lack of a supporting infrastructure, but the lack of *supporting theory*. In this paper paper we illustrated the central role that systems theory can play in developing a comprehensive framework leading to provably robust dynamic vision and imaging systems. In turn, these fields can provide a rich environment both draw inspiration from and to test new developments in systems theory. For instance, the applications addressed in this paper point out, among others, to the need for further research into low complexity nonlinear identification methods, the development of worst-case identification methods for switched systems that are not necessarily ℓ^2 stable (to allow for parsing video sequences into different activities), and to extend currently available 1-D identification methods to the 2-D case.

[5]It can be analitycally shown that this is indeed the case for periodic sequences, but empirical results show that this hypothesis works well also for non–periodic textures.

Acknowledgements

Support from NSF under grants ECS-0221562, IIS-0312558 and ECS-0501166, and AFOSR under grant FA9550-05-1-0437 is gratefully acknowledged.

References

[1] T. Ding, M. Sznaier, and O. Camps. Robust identification of 2-d periodic systems with applications to texture synthesis and classification. In *45thCDC*, pages 3678–3683, San Diego, CA, 2006.

[2] Ahmed Elgammal and Chan-Su Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Computer Vision and Pattern Recognition*, pages 681–688, 2004.

[3] D. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2003.

[4] M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[5] Hwasup Lim, Octavia I. Camps, and Mario Sznaier. A caratheodory-fejer approach to appearance modelling. In *IEEE Computer Vision and Pattern Recognition*, pages 301–307, 2005.

[6] W. Ma, H. Lim, M. Sznaier, and O. Camps. Risk adjusted idenitification of wiener systems. In *45thCDC*, pages 2512–2515, San Diego, CA, 2006.

[7] V. Morariu and O. Camps. Modelling correspondences for multi camera tracking using nonlinear mani- fold learning and target dynamics. In *IEEE Computer Vision and Pattern Recognition*, pages 545–552, 2006.

[8] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(9):1016–1034, September 2000.

[9] F. Paganini and E. Feron. LMI methods for robust $\mathcal{H}_2$ analysis: A survey with comparisons. In L. El Ghaoui and S. Niculescu, editors, *Recent Advances on LMI methods in Control*. SIAM press, 1999.

[10] P. A. Parrilo, R. S. Sanchez Pena, and M. Sznaier. A parametric extension of mixed time/frequency domain based robust identification. *IEEE Trans. Autom. Contr.*, 44(2):364–369, 1999.

[11] K. A. Patwardhan, G. Sapiro, and M. Bertalmio. Video inpainting of occluding and occluded objects. In *Proceedings of ICIP 2005*, volume 2, pages 69–72. IEEE, 2005.

[12] R. Sánchez Peña and M. Sznaier. *Robust Systems Theory and Applications*. Wiley & Sons, Inc., 1998.

[13] M. Sznaier, M. C. Mazzaro, and O. Camps. Semi-blind model (in)validation with applications to texture classification. In *44thCDC–ECC'05*, pages 6065–6070, Seville, Spain, 2005.

[14] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Interna-*

Fig. 11. Top: original sequence. Middle: observed (red) and estimated (green) descriptors. Bottom: inpainted sequence.

tional Journal of Computer Vision, 9(2):137–154, November 1992.

[15] Jing Xiao, Jinxiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. In *The 8th European Conference on Computer Vision (ECCV 2004)*, May 2004.

[16] L. Zelnik-Manor and M. Irani. Degeneracies, dependencies and their implications in multi-body and multi-sequence factorization. In *IEEE Computer Vision and Pattern Recognition*, pages 287–293, 2003.