Dynamic Texture Recognition[*]

Payam Saisan
UCLA
Electrical Eng.
Los Angeles, CA 90095
saisan@ee.ucla.edu

Gianfranco Doretto
UCLA
Computer Science
Los Angeles, CA 90095
doretto@cs.ucla.edu

Ying Nian Wu
UCLA
Statistics
Los Angeles, CA 90095
ywu@stat.ucla.edu

Stefano Soatto
UCLA
Computer Science
Los Angeles, CA 90095
soatto@ucla.edu

Abstract

Dynamic textures are sequences of images that exhibit some form of temporal stationarity, such as waves, steam, and foliage. We pose the problem of recognizing and classifying dynamic textures in the space of dynamical systems where each dynamic texture is uniquely represented. Since the space is non-linear, a distance between models must be defined. We examine three different distances in the space of autoregressive models and assess their power.

1. Introduction

Recognition of objects based on their images is one of the central problems in modern Computer Vision. We consider objects as being described by their geometric, photometric and dynamic properties. While a vast literature exists on recognition based on geometry and photometry, less has been said about recognizing scenes based upon their dynamics. In this paper we consider the problem of recognizing a sequence of images based upon a joint photometric-dynamic model. This allows us to recognize not just steam from foliage, but fast turbulent steam from haze, or to detect the presence of strong winds by looking at trees.

We represent images of stationary processes as the output of a stochastic dynamical model. The model is learned from the data, and recognition is performed in the space of models. The implementation of this idea, however, is not simple. First, the map from a sequence to a model is not necessarily one-to-one: very different scenes can be output of the same model. Second, even the simplest linear models learned from data represent equivalence classes of statistics: the same scene can result in very different models depending on the initial condition. Recognition in the space of models amounts to doing statistics on quotient spaces that have a non-trivial Riemannian structure.

Recognition of complex motion patterns in images is an active area of research in computer vision. Extensive work has been conducted for the case of human motion and in particular facial expressions, for instance [2, 8, 3, 16, 13]. Some methods are based on optical flow. For each frame the flow can be approximated with a small-dimensional vector in a suitable basis, as in [7], and the recognition is done with hidden Markov models (HMMs), or, as in [2], a spatio-temporal representation of the optical flow can be built. Others look at different spatio-temporal features [12].

In this paper we take a different approach: we do not choose local features, nor do we compute optical flow. Instead, we start from the assumption that the sequences of images are realizations of second-order stationary stochastic processes (the covariance is finite and shift-invariant). We set out to classify and recognize *not* individual realizations, but statistical models that generate them. This entails choosing a distance between models. This problem has been first addressed by Martin in [11], where a distance for single-input, single-output (SISO) linear Gaussian processes has been introduced. We propose and analyze three distances. The first uses principal angles between specific subspaces derived from AR[1] models. The second is the extension of the distance proposed by Martin. Both draw on recent results of De Cock and De Moor [4]. Finally, we also look at the geodesic distance.

2. From image sequences to dynamical models

We start from the assumption that a sequence of images $y(t),\ t = 1 \ldots \tau$ is a realization of a second-order stationary stochastic process. This means that the joint statistics between two time instants is shift-invariant. Although this may seem like a severely restrictive assumption, it has been shown in [14, 6] that sequences such as foliage, water, smoke, and steam are well captured by this model. These sequences are called “dynamic textures”.

It is well known that a positive-definite covariance sequence with rational spectrum corresponds to an equiva-

[*] This research is supported in part by NSF grant IIS-9876145, and ARO grant DAAD19-99-1-0139. We wish to thank Alessandro Chiuso.

[1] AR stands for autoregressive.

lence class of second-order stationary processes [10]. It is then possible to choose as a representative of each class a Gauss-Markov model – that is the output of a linear dynamical system driven by white, zero-mean Gaussian noise – with the given covariance. In other words, we can assume that there exists a positive integer n, a process $\{x(t)\}$ (the "state") with initial condition $x_0 \in \mathbb{R}^n \sim \mathcal{N}(0, P)$ and a symmetric positive semi-definite matrix $\left[\begin{array}{cc} Q & S \\ S^T & R \end{array}\right] \geq 0$ such that $\{y(t)\}$ is the output of the following Gauss-Markov "ARMA" model[2]:

$$\begin{cases} x(t+1) = Ax(t) + v(t) \quad v(t) \sim \mathcal{N}(0, Q); \;\; x(0) = x_0 \\ y(t) = Cx(t) + w(t); \quad w(t) \sim \mathcal{N}(0, R) \end{cases} \tag{1}$$

for some matrices $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{m \times n}$, where $S = E[w(t)v^T(t)]$.

The choice of matrices A, C, Q, R, S is not unique, in the sense that there are infinitely many models that give rise to exactly the same measurement covariance sequence starting from suitable initial conditions: one can substitute A with TAT^{-1}, C with CT^{-1}, Q with TQT^T, S with TS, and choose the initial condition Tx_0, where $T \in \mathcal{GL}(n)$ is any invertible $n \times n$ matrix, and obtain the same output covariance sequence. In other words, the basis of the state-space is arbitrary, and any given process has *not* a unique model, but an *equivalence class* of models. In order to be able to identify a unique model of the type (1) from a sample path $y(t)$, it is therefore necessary to choose a representative of each equivalence class (i.e. a basis of the state-space): such a representative is called a *canonical model realization*. It is canonical in the sense that it does not depend on the choice of basis in the state space (because it has been fixed).

While there are many possible choices of canonical realizations (see [9]), we are interested in one that is "tailored" to the data. Since we work with images, we will make the following assumptions about the model (1):

$$m >> n; \; \mathrm{rank}(C) = n \tag{2}$$

and choose a realization that makes the columns of C orthonormal:

$$C^T C = I_n. \tag{3}$$

This guaranties that the matrix C is an element in the Stiefel manifold $V(m, n)$ (the set of n orthonormal vectors in $\mathbb{R}^m$) and that the stochastic realization corresponding to a given dataset is uniquely determined. We shall see that the classification/recognition problem can be posed by computing statistics on such a manifold.

The problem of going from data to models can be formulated as follows: *given* measurements of a sample path of the process: $y(1), \ldots, y(\tau)$; $\tau >> n$, estimate $\hat{A}, \hat{C}, \hat{Q}$, a canonical realization of the process $\{y(t)\}$. As described in [14], the choice of $C \in V(m, n)$ results in a canonical realization. Ideally, we would want the maximum likelihood solution from the finite sample, that is the argument of

$$\max_{A,C,Q} p(y(1), \ldots, y(\tau)|A, C, Q). \tag{4}$$

Notice that we do not model the covariance of the measurement noise since that carries no information on the underlying process. In practice, for reasons of computational efficiency, we settle for a suboptimal solution described in [14]. From this point on, therefore, we will assume that we have available – for each sample sequence – a model in the form $\{A, C, Q\}$. For the purpose of recognition, we consider processes with different input noise covariance as equivalent. Therefore, the problem of recognizing dynamic textures can be posed as the problem of recognizing pairs $\{A, C\}$ estimated from data.

2.1. Geometric structure of the space of models

Models, learned from data as described in the previous section, do not live in a linear space. While the matrix A is only constrained to be stable (eigenvalues within the unit circle), the matrix C has non-trivial geometric structure for its columns form an orthogonal set. Let $C \in V(m, n)$, $m \geq n$ be a point on the Stiefel manifold of n-frames in $\mathbb{R}^m$, $C^T C = I_n$, endowed with the Euclidean metric $g_e(X, Y) \doteq \mathrm{tr}(X^T Y)$ where $X, Y \in TV(m, n)$, the tangent plane to the Stiefel manifold. It is shown in [5] that geodesic trajectories in $V(m, n)$ have the general form

$$R\exp(Xt)I_{m,n} \quad \text{where} \quad I_{m,n} = \left[\begin{array}{c} I_n \\ 0 \end{array}\right] \in \mathbb{R}^{m \times n} \tag{5}$$

and $R \in O(m)$; X is a skew-symmetric matrix having blocks

$$X \doteq \left[\begin{array}{cc} F & -G^T \\ G & 0 \end{array}\right]. \tag{6}$$

Note that X belongs to a linear space that is isomorphic to $\mathbb{R}^{mn-n(n+1)/2}$, and could therefore be used as a local coordinatization of the Stiefel manifold $V(m, n)$. We will use the structure of the geodesic in order to define a distance in $V(m, n)$ as follows: consider two points $C_1, C_2 \in V(m, n)$ and the geodesic connecting them: $C(t) \mid C(0) = [C_1 \; U]\exp(0)I_{m,n} = C_1$ and $C(t) = [C_1 \; U]\exp(Xt)I_{n,m} = C_2$ for a particular value of X, t and for any U, an orthogonal completion of C_1. Then we define

$$d : V(m, n) \times V(m, n) \longrightarrow \mathbb{R}; \quad (C_1, C_2) \mapsto \|Xt\|_F \tag{7}$$

where the subscript F indicates the Frobenius norm.

2.2 Probability distributions on Stiefel manifolds

In order to provide a simple statistical description on the space of models, we assume that A and C are independent,

[2] ARMA stands for autoregressive moving average

so that their statistical description can be addressed separately. While specifying a probability density in the space of transition matrices A is straightforward (indeed, we will adopt a Gaussian density), doing so for the output matrices C is not trivial since, as we have just seen, the space has a non-trivial curvature. In this paragraph we introduce a class of probability densities on the Stiefel manifold that can be used to model the statistics of C. Consider the following function $p : V(m,n) \longrightarrow \mathbb{R}$

$$p(C) \doteq \frac{1}{Z}\exp(\mathrm{tr}(\Sigma\mu^T C)) \tag{8}$$

where $\mu \in V(m,n)$, $\Sigma = \Sigma^T \geq 0$ and $Z = \int dP(C)$ where $dP(C) = p(C)d\mu(C)$ with $d\mu$ the base (Haar) measure on $V(m,n)$. We call this function a *Langevin (or Gibbs) density* on $V(m,n)$, owing to its similarity to Langevin distributions on the sphere. μ plays the role of the *mode* of the density, and Σ plays the role of the *dispersion*. It is easy to verify that the above density has a unique maximum for $C = \mu$. The functional expression of p can be used to compute likelihood ratios for recognition once the parameters μ, Σ, have been inferred.

In order to estimate the sufficient statistics from samples, let C_i, $i = 1 \dots N$ be a fair sample from the density $p(C)$. It follows from the law of large numbers that

$$\hat{m} \doteq \frac{1}{N}\sum_{i=1}^{N} C_i \longrightarrow \int C dP(C) \doteq m(\mu,\Sigma). \tag{9}$$

Having a closed-form expression of the integral $m(\mu,\Sigma)$, one could use samples to compute $\hat{m}$ and use the equation above to compute statistics. However, we do not pursue that avenue further here. Instead, we consider the maximum likelihood estimation of the sufficient statistics by considering the joint density of a fair sample $p(C_1,\dots,C_N|\mu,\Sigma)$, which can be written as

$$\prod_{i=1}^{N} p(C_i) = \frac{1}{\prod Z_i}\exp(\sum_{i=1}^{N}\mathrm{trace}(\Sigma\mu^T C_i)) \tag{10}$$

For example, for the case $\Sigma = I$ we can look for $\hat{\mu}$ that solves the following problem

$$\max_{\mu\in V(m,n)} p(C_1,\dots,C_N|\mu) = \max \mathrm{trace}(\mu\sum_{i=1}^{N} C_i).$$

Letting $\frac{1}{N}\sum_{i=1}^{N} C_i = U_\Sigma S_\Sigma V_\Sigma^T$ be a singular value decomposition, then

$$\hat{\mu} = U_\Sigma V_\Sigma^T.$$

This also clarifies the relationship between the sample mean $\hat{m}$ and the sample median $\hat{\mu}$: the latter consists of the orthogonal factors of the former, or in other words it is the orthogonal *projection* of $\hat{m}$ onto $V(m,n)$.

3. Recognizing dynamical models

As we have articulated in the previous section, a dynamic texture is described by a linear dynamical system and represented by the matrices A,C,Q. This space has a non-trivial curvature structure that must be taken into account when doing statistics or comparisons between models.

In this section we consider three approaches to recognition. One involves computing likelihood ratios, with an explicit form for the probability density of the models. The second involves computing angles between subspaces of the measurement span. The third only involves computing distances between models.

Suppose that two groups of points on $V(m,n)$ are given: $U_1,\dots,U_k$, fair samples from a Langevin distribution with mean μ_U and dispersion Σ_U, and $V_1,\dots,V_l$ samples from a distribution with mean μ_V and dispersion Σ_V. Given a new point C, we want to decide to which "group" it belongs. From a decision-theoretic point of view, the goal is to construct a density corresponding to each hypothesis, $p(C|U)$, $p(C|V)$, and to compute the likelihood ratio

$$\rho(C) = \frac{p(C|U)}{p(C|V)} \tag{11}$$

where the parameters Σ_U and μ_U in $p(C|U)$ are computed from the samples U_i as above, and so for Σ_V and μ_V. A decision can then be made based on whether the ratio is larger or smaller than one. This setting can be generalized to decisions among a number of hypotheses in a straightforward fashion [15].

While included in the discussion, likelihood ratios were not part of our experimental scheme. In our experiments we focused mainly on subspace angles and distances between models.

Let $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{m\times p}$ be two matrices with full column rank. The principal angles $\theta_k \in \left[0, \frac{\pi}{2}\right]$ between range(A) and range(B) are defined as

$$\cos(\theta_k) = \max_{\substack{x\in\mathbb{R}^p \\ y\in\mathbb{R}^q}} \frac{\left|x^T A^T B y\right|}{\left\|Ax\right\|_2 \left\|By\right\|_2}, \quad \text{for } k = 1,2,...,\min(p,q)$$

Subspace angles are the largest of these angles. A closed form solution is presented in [4]. We use subspace angles between dynamic texture models to compute distances.

For the sake of simplicity in our implementations we assumed to be dealing with AR models. So, we discuss and compute principal angles and Martin distances between AR models defined by $\{A, C\}$ pairs. While this assumption may seem restrictive, the results are nonetheless encouraging (see Section 4.2).

Now, let $M_1 \doteq (A_1, C_1)$ and $M_2 \doteq (A_2, C_2)$ represent two AR models with cepstrum coefficients $c_1(n)$ and $c_2(n)$ for $n = 0, \pm1, \pm2$. The Martin distance is defined for SISO

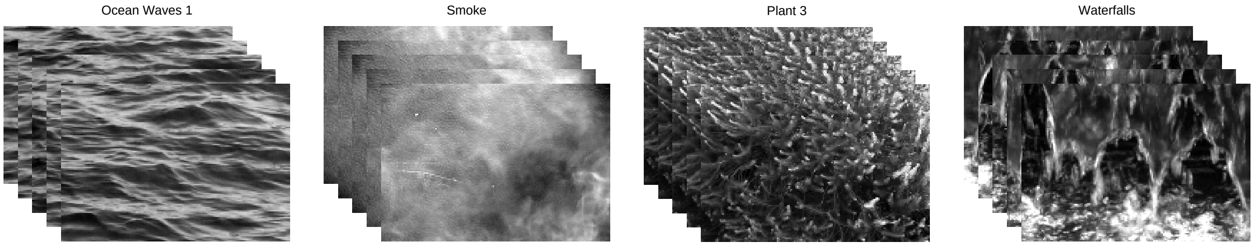


Figure 1. *Samples from the dynamic texture database. We have collected a total of more than 250 sequences, consisting of moving scenes of foliage, water, ocean waves, smoke etc. Each sequence is 150 frames long and* 220×320 *pixels. In our experiments each sequence has been divided into two subsequences of 75 frames for a total of more than 500 sequences. From these sequences 200 samples were selected at random to build the data set we used to run our experiments.*

systems as

$$d(M_1, M_2) = \sqrt{\sum_{n=0}^{\infty} n\, |c_1(n) - c_2(n)|^2}. \qquad (12)$$

As shown in [4] this distance is related to the principal angles between specific subspaces derived from the AR model, namely the range of (extended) observability spaces. The extended observability matrix for system M_i is denoted by $\mathcal{O}_\infty(M_i) \doteq [C_i^T \quad A_i^T C_i^T \quad \ldots \quad (A_i^T)^n C_i^T \quad \ldots]^T$. The distance defined by Martin can be expressed directly in terms of the principal angles between the column spaces of $[\mathcal{O}_\infty(M_1)]$ and $[\mathcal{O}_\infty(M_2)]$. If we denote by θ_i the i^{th} principal angle between these spaces, the distance defined by Martin is given by

$$d_M(M_1, M_2)^2 = \ln \prod_{i=1}^{n} \frac{1}{\cos^2 \theta_i}. \qquad (13)$$

The proof can be found in [4]. While this distance is given for scalar AR processes, one can measure the distance between multivariate AR models using the same concept.

In our implementation we approximate the extended observability matrix directly with the observability matrix and compute principal angles, θ_i, between observability spaces, $\mathcal{O}_n(M_1)$ and $\mathcal{O}_n(M_2)$.

4. Experiments

We constructed a comprehensive database of dynamic textures; sequences capturing natural phenomena such as ocean waves, steam, and plants. Included in the database are similar sequences with different dynamics. For example, we have water stream recorded from different angles, thus moving at different orientations and speeds. The database includes 76 different kinds of dynamic textures. Each of them is represented by 8 distinct instances. Each subsequence consists of 75 frames. All frames are in color where the size of an individual frame is 220×320 pixels. Figure 1 shows a few samples of the original dataset.

In our experiments, we used 50 dynamic texture categories, each with four subsequences. In order to reduce computational complexity, we first filtered and subsampled the original sequences to 110×160 pixels. We then chose a reduced patch size of 48×48 pixels by carefully cropping from each sequence regions to include key statistical and dynamical features (e.g. the flame of a candle light). Finally, we transformed color data to 256 gray levels.

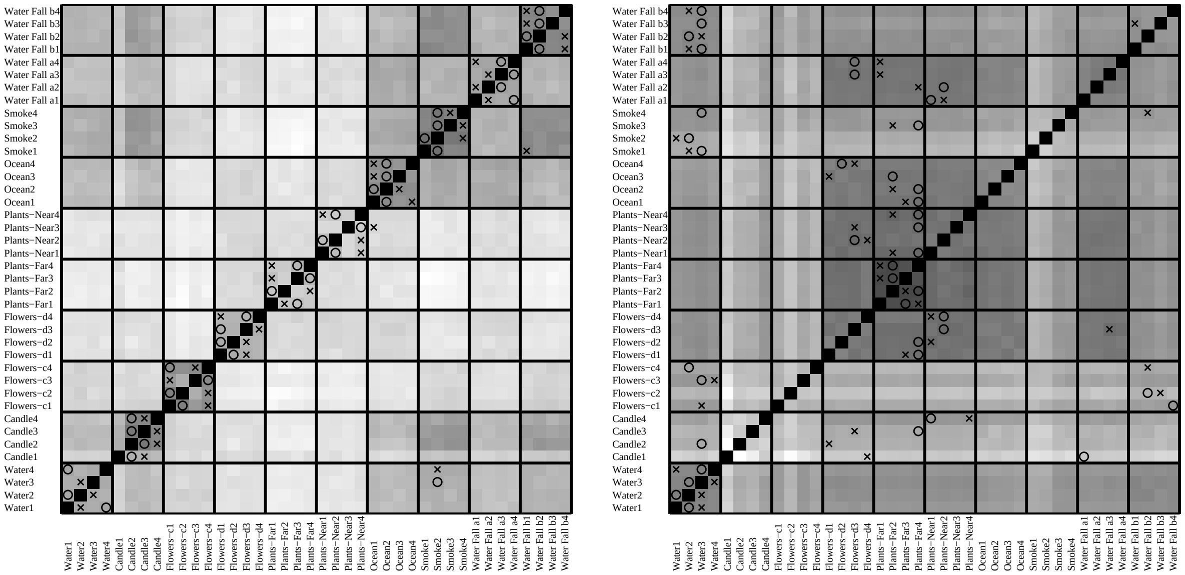
4.1. Learning Dynamic Textures

Our experimental paradigm consisted of first learning a reduced basis and then computing the dynamical system parameter, A. To perform the learning process we used the closed-form solution proposed in [14]. Therefore, we view the choice of the matrix C as a dimensionality reduction step, and we seek for a decomposition of the frames in the linear form (1) where C can be a set of principal as well as independent components. Once the system parameters are identified, we compute distances between models using the geodesic distance, subspace angles, and Martin's distance generalized to multi-input, multi-output (MIMO) models.

While principal component analysis (PCA) is based on second order statistical dependencies, independent component analysis (ICA) can be used to obtain a basis whose components are maximally independent. In our formulation of ICA we assume the following model $Y^T = X^T C^T$, where image frames flattened into column vectors make up the columns of the data matrix Y. Here Y^T is assumed to be a linear combination of an unknown set of independent sources, basis images (i.e. the columns of C). In our case, a typical 75 frame sequence was represented using a reduced basis of 20 components for both ICA and PCA. The reduced ICA representation was obtained using Bell and Sejnowski's infomax algorithm [1].

4.2. Recognition

Given the model parameters of a dynamic texture, we computed the subspace angles and distances between 200 dynamic textures. We then calculated the recognition rates for each distance and model representation (ICA/PCA), see Table 1. A correct detection for a given dynamic texture was defined as having one of the three other dynamic textures in its category as its closest neighbor.



(a) (b)

Figure 2. *Figure (a) on the left is the result of an experimental run on a small subset of the data base (*40 *dynamic textures), using Martin's distance and PCA basis. Moving along the vertical axis, we mark the first (o) and second (x) nearest neighbors. For example, the closest dynamic texture to* `Smoke1`*, along the vertical axis, is* `Smoke2` *along the horizontal axis. Similarly, the second closest dynamic texture to* `Smoke1` *is* `Water Fall b1`*. Although this subsampled dataset is small, the discrimination power of Martin's distance is already visible. Figure (b) on the right displays another run on the same subset of the data. Here the distance between models was the geodesic distance. The poor recognition rate for the geodesic distance is visible from the large number of nearest neighbors (o) falling outside of the "same family" grid lines. It should be noted that the recognition rates reported in Table 1 were obtained using 200 dynamic textures.*

	Geodesic Dist.	Subsp. Angles	Martin Dist.
PCA	5.5%	24.5%	89.5%
ICA	2%	35%	71%

Table 1. *A summary of the recognition rate percentages. We compute the probability of correct recognitions as the percentage of the number of correct nearest neighbor hits.*

Simulation results suggest significant separation among similar categories of dynamic textures using PCA as basis and Martin's distance, leading to a recognition rate of 89.5% (with ICA and Martin's distance we reached 71%). The results were not encouraging with the geodesic distance. While subspace angles were better than the geodesic distance, the recognition rate of 24.5% (35% for ICA) proved them still ineffective.

References

[1] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–59, 1995.

[2] M. G. Black. Explaining optical flow events with parameterized spatio-temporal models. In *Proc. of Conference on Computer Vision and Pattern Recognition*, volume 1, pages 326–32, 1999.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hmm for complex action recognition. In *Proc. of Conference on Computer Vision and Pattern Recognition*, pages 213–44, 1997.

[4] K. De Cock and B. De Moor. Subspace angles between linear stochastic models. In *Proc. the 39th IEEE Conference on Decision and Control*, volume 2, pages 1561–6, Sydney, NSW, Australia, Dec 2000.

[5] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1999.

[6] A. W. Fitzgibbon. Stochastic rigidity: image registration for nowhere-static scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision*, volume 1, pages 662–9, Vancouver, BC, Canada, July 2001.

[7] J. Hoey and J. J. Little. Representation and recognition of complex human motion. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, volume 1, pages 752–9, Hilton Head Island, NC, June 2000.

[8] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:852–72, Aug 2000.

[9] T. Kailath. *Linear Systems*. Prentice Hall, 1980.

[10] L. Ljung. *System Identification: theory for the user*. Prentice Hall, 1987.

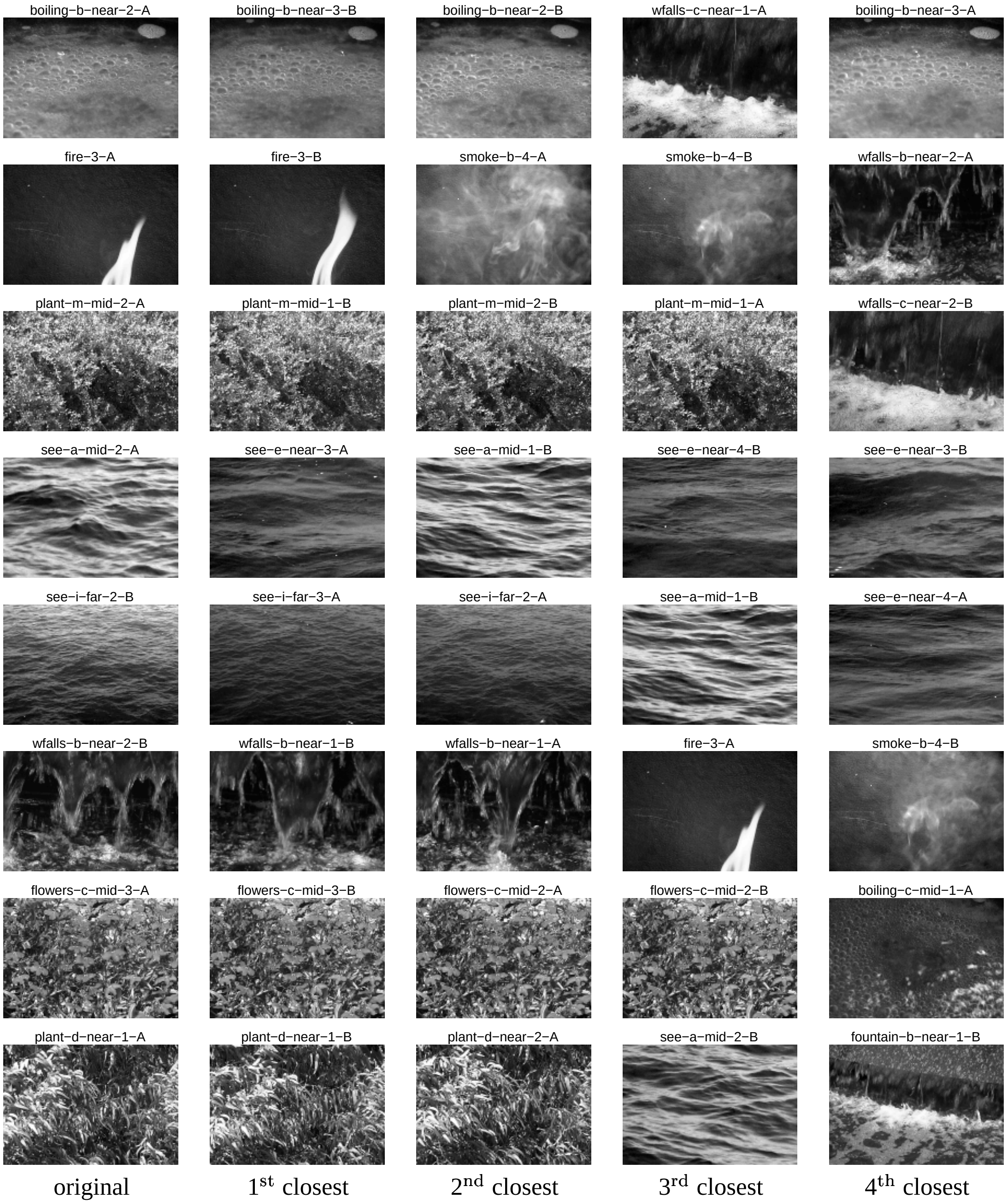


Figure 3. *Results of the nearest neighbor computation using subspace angles. The first column shows a sample from one of the original subsequences. The distance from the model of this subsequence to every other subsequence is computed, and a sample of the sequence "closest" to the test is shown in the second column. The third column shows the second closest sequence and so on.*

[11] R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48:1164–7, Apr 2000.

[12] A. A. Niyogi. Analyzing and recognizing walking figures in xyt. In *Proc. of Conf. on Comp. Vision and Pattern Recogn*, pages 469–74, Seattle, WA, June 1994.

[13] C. S. Pinhanez and A. F. Bobick. Human action detection using pnf propagation of temporal constraints. In *Proc. of Conference on Computer Vision and Pattern Recognition*, pages 898–904, 1998.

[14] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *Proc. of IEEE Intl. Conf. on Computer Vision*, volume 2, pages 439–46, Vancouver, BC, Canada, July 2001.

[15] H. Van Trees. *Detection and Estimation Theory*. Krieger, 1992.

[16] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21:884–900, Sept 1999.