A Model for Dynamic Shape and Its Applications

Che-Bin Liu and Narendra Ahuja
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{cbliu, ahuja}@vision.ai.uiuc.edu

Abstract

Variation in object shape is an important visual cue for deformable object recognition and classification. In this paper, we present an approach to model gradual changes in the 2-D shape of an object. We represent 2-D region shape in terms of the spatial frequency content of the region contour using Fourier coefficients. The temporal changes in these coefficients are used as the temporal signatures of the shape changes. Specifically, we use autoregressive model of the coefficient series. We demonstrate the efficacy of the model on several applications. First, we use the model parameters as discriminating features for object recognition and classification. Second, we show the use of the model for synthesis of dynamic shape using the model learned from a given image sequence. Third, we show that, with its capability of predicting shape, the model can be used to predict contours of moving regions which can be used as initial estimates for the contour based tracking methods.

1. Introduction

Changes in the shape of a dynamic object offer important cues for object recognition. In this paper, we are concerned with models of gradual changes in the shape of a 2-D region. We present a simple model of shape variation which was seen limited use in the past work. This model models the changes in the 2-D shape of a region in terms of the changes in its contour representation. Specifically, an autoregressive time series model of the changes in the Fourier coefficients of the region contour is used. We use it to model, recognize, and synthesize 2-D dynamic shape. We present applications to (i) modeling fire motion and detecting fire in video sequences, (ii) classification of objects based on motion patterns, (iii) synthesis of novel image sequences of evolving shapes, and (iv) object boundary prediction for use by contour tracking methods.

The 2-D shape representation and its use has received much attention in computer vision. A survey of shape analysis methods can be found in [7]. Pavlidis [11] proposed the following three classifications for shape based methods using different criteria. (i) *Boundary* (or *External*) or *Global* (or *Internal*): Algorithms that use region contour are classified as external and boundary, such as Fourier transforms based approaches; Those that use interior region for the analysis are classified as internal and global, such as moment based methods. (ii) *Numeric* or *Non-numeric*: This classification is based on the result of the analysis. For instance, medial axis transform generates a new image with a symmetric axis, and is categorized as non-numeric. In contrast, Fourier and moment based methods produce scalar numbers, and thus are in numeric category. (iii) *Information Preserving* or *Non-preserving*: Approaches that allow users to reconstruct shapes from their shape descriptors are classified as information preserving. Otherwise, they are information non-preserving.

We propose a dynamic shape model that describes shape at any given time using Fourier transform coefficients and an autoregressive (AR) model to capture the temporal changes in these coefficients. The Fourier description possesses boundary, numeric, and information preserving properties. The autoregressive model is a simple probabilistic model that has shown remarkable effectiveness in the mapping and prediction of signals. As Srivastava [19] points out, the temporal change of Fourier representation may not be linear. However, a linear model is more manageable to approximate such a process, and requires a small number of observations to estimate parameters.

The remainder of this paper is organized as follows. In Section 2, we present our dynamic shape model and its parameter estimation. In Section 3, we apply the proposed approach to modeling and detection of fire in video sequences. In Section 4, we classify several objects and visual phenomena based on their evolving region contours. In Section 5, we apply the model to synthesis of evolving shape sequences. In Section 6, we use our model to predict object shape in a video sequence for object contour tracking. Section 7 discusses the limitations of the proposed model. Section 8 summarizes the contribution of this work.

2. Dynamic Shape Model

Our dynamic shape model includes two parts: a spatial representation of 2-D shape and a temporal representation of shape variation. The detailed model and its parameter estimation are described in the following sections. We also compare our model to other relevant models.

2.1. Spatial Representation of Shape

Fourier Descriptors (FD), the Fourier Transform coefficients of the shape boundary, represents a 2-D shape using an 1-D function. There are several variations of Fourier based 1-D boundary representation in literature [9]. In this paper, we use Persoon and Fu's method [13] for its simplicity.

Given an extracted region in an image, we first retrieve its boundary using eight-connected chain code. Assume that we have N points from the chain code representation of the boundary. We express these points in complex form: $\{z_i | z_i = x_i + jy_i\}_{i=1}^{N}$ where (x_i, y_i) are the image coordinates of boundary points as the boundary is traversed clockwise. The coefficients of the Discrete Fourier Transform (DFT) of $\{z_i\}_{i=1}^{N}$ are

$$a_k = \frac{1}{N}\sum_{i=1}^{N} z_i \exp(-j\frac{2\pi}{N}ik), k = -\lfloor\frac{N-1}{2}\rfloor, \ldots, \lfloor\frac{N}{2}\rfloor. \quad (1)$$

If M harmonics are used ($M \leq \lfloor\frac{N-1}{2}\rfloor$), the coefficients $\{a_m\}_{m=-M}^{M}$ are the Fourier Descriptors used to characterize the shape. To reconstruct L boundary points $\{\tilde{z}_l\}_{l=1}^{L}$ using M harmonics, we perform inverse DFT as:

$$\tilde{z}_l = \sum_{m=-M}^{M} a_m \exp(j\frac{2\pi}{L}ml) \qquad l = 1, \ldots, L. \quad (2)$$

Note that $a_0 = \frac{1}{N}\sum_{i=1}^{N} z_i$ represents the center of gravity of the 1-D boundary, which does not carry shape information. We neglect this term to achieve translation invariance for recognition and classification. We keep this term for synthesis and shape prediction because it accounts for scale changes.

Most related works in Fourier based shape description discuss about similarity measures that make FD invariant to relevant transformations, e.g., rotation, translation and scaling. The requirement for each invariance depends on the applications. In this paper, we do not consider rotation invariance because we need to reconstruct the boundary of shape. Since rotation invariance is not relevant, we can always choose the starting point as the topmost boundary pixel along the vertical axis through the center of gravity of the entire shape. Our representation approximates scale invariance (if we drop a_0 term) since we have dense sampling of points along region boundary using chain code. Chain code expression discretizes the arc and Equation (1) normalizes the arc length [1].

2.2. Temporal Representation of Shape Variation

The stochastic characteristics of boundary motion are estimated by an autoregressive model of changes in Fourier coefficients of the region boundary. The autoregressive (AR) model, also known as a linear dynamical system (LDS), is used based on the assumption that each term in the time series depends linearly on several previous terms along with a noise term [8]. In this work, the AR model is used to capture different levels of temporal variation in FDs.

Suppose v_k are the m-dimensional random vectors observed at equal time intervals. The m-variate AR model of order p (denoted as AR(p) model) is defined as

$$v_k = w + \sum_{i=1}^{p} A_i v_{k-i} + n_k. \quad (3)$$

The matrices $A_i \in R^{m \times m}$ are the coefficient matrices of the AR(p) model, and the m-dimensional vectors n_k are uncorrelated random vectors with zero mean. The m-dimensional parameter vector w is a vector of intercept terms that is included to allow for a nonzero mean of the time series.

Our dynamic shape model uses FDs to represent shape, so the random vector v_k is in a form of FD at time k. To select the optimum order of the AR model, we adopt Schwarz's Bayesian Criterion [16] which chooses the order of the model so as to minimize the forecast mean-squared error. We estimate the parameters of our AR model using Neumaier and Schneider's algorithm [10] which ensures the uniqueness of estimated AR parameters using a set of normalization conditions.

2.3. Comparison to Other Models

Models of active contour tracking that predict contour motion and deformation [1, 3, 23] have been proposed to account for dynamic object shape. For example, Terzopoulos and Szeliski [23] incorporate Kalman filtering with the original snake model [4]. Blake et al. [1] propose a contour tracking method that works particularly well for affine deformation of object shape. Snake based methods process the contour directly in the spatial domain and consider local deformations [4, 12]. In contrast, in our representation, shape information is distributed in each coefficient of FD. Thus, we consider global deformations. Only a few methods, such as [1, 22], consider both local and global deformations. Local deformations of all contour points comprise too large a data set to be convenient for shape recognition and

[1] Note: scale invariance is achieved if the distances between a pixel and its eight neighbors are considered as equal.

classification. In addition, models of active contour tracking predict motion and deformation for one image frame. In contrast, we model global temporal characteristics of a whole image sequence. Most importantly, most work on deformable shape modeling is aimed at region contour identification by using a deformable, evolving snake to converge on the desired contour. Instead, in our work, the evolving shape description is aimed at describing a temporal changing shape.

There is also some work using level sets to represent dynamic shape such as [26]. The advantage of the level set method is its ability to handle topology changes. However, as will be shown later, our model requires significantly less computation.

3. Application I: Recognition

In this section, we will show that using the temporal information of shape variation improves recognition results that use shape only. We choose the problem of fire recognition in video sequences as an example.

Fire has diverse, multispectral signatures, several of which have been utilized to devise different methods for its detection. Most of the methods can be categorized into smoke, heat, or radiation detection. However, there are only a few papers about fire detection in computer vision literature. Healey et al. [2] use a purely color based model. Phillips et al. [14] use pixel color and its temporal variation, which does not capture the temporal property of fire which is more complex and requires a region level representation.

3.1. Fire Detection Algorithms

Our fire detection algorithms include two main steps: (i) Extract potential fire regions in each image; and (ii) Represent each extracted region using FD and AR parameters, and then use a classifier to recognize fire regions.

To extract potential fire regions in each image, we use algorithms described in [6]. For each potential fire region, we represent it independently by taking the magnitude of its FDs. We then find matching regions in previous images of the sequence, and estimate parameters of the AR model for the corresponding fire regions. The FD and estimated AR parameters are both used as features of current region. We use a two-class Support Vector Machine (SVM) classifier [24] with RBF kernel for fire region recognition.

3.2. Experimental Results

The video clips used in our experiments are taken from a random selection of commercial/training video tapes. They include different types of fires such as residential fire, warehouse fire, and wildland fire. We use images captured at day time, dusk or night time to evaluate system performance under different lighting conditions. We also use other image sequences containing objects with fire-like appearances such as sun and light bulbs as negative examples. The video clips that we tested our algorithm on contain a total of 3956 image frames in 36 sequences. Figure 1 shows some selected fire images used in our experiments. The (red) contours depicted in the images are the detected fire region contours.

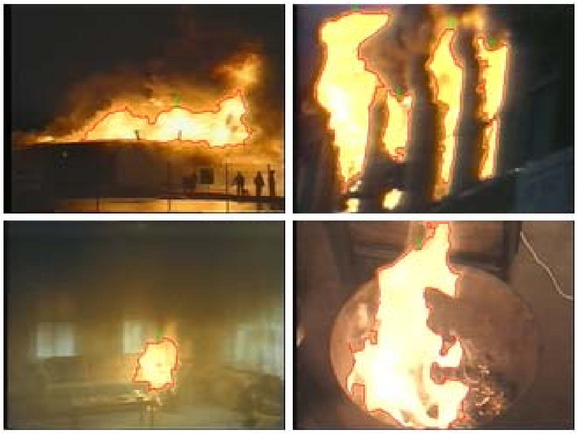

Figure 1: Selected fire images used in experiments.

In our test data, the potential region extraction algorithm extracted a total of 1319 fire-like region contours, 1089 of which were true fire region contours. For shape representation in terms of Fourier Descriptors, we find that using 40 coefficients (i.e. $M = 20$) is sufficient to approximate the relevant properties of the fire region contours. In this experiment, we assume that different FDs at any given time k are independent of each other, so we have diagonal coefficient matrices in our AR model, where $A_i(m, n) = 0$ if $m \neq n$. Thus it can be viewed as modeling $2M$ independent time series. We also find that the AR(1) model yields the minimal forecast mean-squared error. Therefore, we use 40 AR coefficients to represent the stochastic characteristics of the temporal changes in FDs.

Table 1: Recognition rate of fire and non-fire contour recognition.

Experiments	Fire	Non-Fire
Use shape only (FD)	0.996	0.904
Use shape + evolution (FD + AR)	0.999	1.0

We tested our algorithms in two ways: The first set of experiments with only spatial information of region contours (FD only as the feature vector), and the second set of experiments with spatial and temporal information of region contour evolution (FD and AR parameters as the feature vector). In the second set of experiments, we required that a fire contour be seen in at least previous four frames. Note that three frames are the minimum requirement to estimate parameters of our AR(1) model. For each set of experiments, we repeated the test ten times using one-tenth of fire and non-fire region contours to train the SVM classifier, and the other region contours for test. In this way, we used many more fire examples than counter examples on training. This was intended to tilt the detector in favor of false positives vs false negatives as corroborated by the experimental results. The average recognition rate is shown in Table 1. It is clear that temporal information of shape evolution indeed improved the detection performance and reduced false alarm rate significantly.

4. Application II: Classification

In this section, we demonstrate that the temporal information of shape variation alone is a good discriminant for classifying several objects and visual phenomena. Under our proposed framework, we show that object shape variation is indeed an important visual cue for object classification.

Follow the model presented in Section 2. Assume that M harmonics in the FDs are used to represent the region boundary of an object in each image of the sequence, and AR(1) model is used to describe boundary dynamics. We then have $2M$ AR coefficients to represent the temporal characteristics of the evolving object shape in an image sequence. Let $\{a_n\}$ and $\{b_n\}$ be AR coefficients modeling a dynamic shape α and a dynamic shape β, respectively. We define the distance between the two dynamic shape sequences as

$$d(\alpha,\beta) = (\sum_{n=1}^{2M} |a_n - b_n|^2)^{1/2}. \tag{4}$$

A simple nearest-neighbor classifier using metric (4) is used for classification.

4.1. Experimental Results

The image sequences used in the experiments include two running human sequences, three waving flag sequences, and two fire sequences. The fire contours are extracted as described in [6]; The region boundaries of flags and running human are semi-automatically extracted using active contour method [4] for each image frame. We use forty FDs to approximate each object boundary. The AR parameters are estimated using each whole sequence. Therefore, the estimated AR parameters represent the global dynamics of the object boundary in a sequence. The experiments are done using the cross-validation method. Only one out of seven image sequences is misclassified, where a running human sequence is classified as a waving flag sequence.

5. Application III: Synthesis

In this section, we apply our model to synthesis of dynamic shape. In particular, we synthesize fire boundary sequences, where the dynamic shape model is obtained from a fire image sequence in as described Section 3. We choose fire as an example because fire region can be modeled as nested subregions, where each subregion shows temporal variation (see Figure 2, leftmost image).

Synthesis of dynamic shape is a novel topic in computer vision. The most relevant work are those of image based dynamic/temporal texture synthesis. Some of them use only local image structures and ignore the underlying dynamics [25]. Some other works that learn the underlying dynamics in pixel level [21] or in image subspace [18] do not use region level image structures. Instead, they learn the global dynamics of the whole image. In our method, we learn the dynamics of regions using region boundaries.

Many physics based methods have been proposed to produce visual phenomena such as fire [15, 17, 20]. However, since these methods do not learn dynamics from images, they are not capable of generating subsequent images based on a given image. Image based method, such as [18], generates an image sequence if given an initial image and the learned image dynamics. But the resulting images will show significant artifacts if the region of motion is not fixed. Our approach is image based, and it directly deals with temporal variation of regions.

5.1. Synthesis Results

Our synthesis of new sequence is based on Equation (3), after AR parameters have been estimated from the given image sequence. For a given initial image, we retrieve the object boundary in the image and represent it using FDs. We perform desired number of iterations of the AR model to estimate FDs for the entire synthesis sequence. The shape sequence are reconstructed using the estimated FDs (2).

In this experiment, we use a fire sequence as a training example. A fire region is modeled as a nested ring structure where each ring is associated with a color spectrum. Although the changes in color is continuous, we threshold the fire region (by grayscale intensity) into three subregions. Each region boundary in the given image sequence are independently modeled by our approach. The color spectra of each region are modeled as a mixture of Gaussian. Once the parameters of three AR models have been estimated, we use the mean boundaries in the given sequence as initial boundaries, and simulate the AR models to generate subsequent boundaries. An inner region boundary is confined to its outer region boundary so that we maintain the nested ring structure. To avoid spin-up effects, the first thousand time steps of the AR models are discarded. The pixel colors of each region are drawn from respective color models. Figure 2 shows the nested ring model, an example fire image of the input video and some selected synthesized fire image frames.

Our method is capable of solving the following two problems: Given a fire image sequence, (i) generate a new sequence of fire shapes, where both shapes and dynamics are similar to the given image sequence; (ii) also given an initial fire shape, generate a new sequence of fire shapes, where the dynamics is similar to the given image sequence. To achieve photo-realistic fire rendering, since we can solve problem (i), we need only a more sophisticated model that enforces spectral gradient to fill colors in the synthesized fire region. For non-photo-realistic fire rendering, such as cartoon drawing, we ask artists to draw fire regions as nested rings and assign a color for each subregion. Our approach will automatically generate subsequent images based on the learned dynamical model. The synthesized sequence can then be overlaid into other image sequences.

6. Application IV: Shape Prediction

The capability of predicting shape comes naturally in our dynamic shape model. In this section, we apply our method to tracking deformable objects. The contour based tracking methods consist two parts: obtaining an initial contour and conforming the initial contour to object boundary. A good initial contour estimate provides a predicted contour closer to true object boundary in both geometry and position.

Most works on contour tracking are based on the active contour model (or snake model) proposed by Kass et al. [4]. Some works assume that the motion of the object is slow and its deformation is small [5]. So the optimal contour estimate in the previous image frame is used as the initial contour in the current frame. When the changes in shape are large, these methods are very likely to fail. Other works that estimate motion and deformation are compared to our method in Section 2.3.

Using our proposed framework, the contours are again represented by FDs. To account for large changes in shape, we estimate our AR model locally using a small number of previous image frames. A first-order AR model is estimated. Then the initial contour is predicted by Equation 3 with $n_k = 0$. Note that the zeroth term of FDs has positional information. So our dynamical model simultaneously predicts the position and shape for the current image frame. Any contour based tracking methods can then be used to conform the contour to object boundary.

6.1. Experimental Results

We test our algorithms using a Bream sequence, where a fish initially swims to the right, makes a sharp turn, and then swims to the left. We choose this image sequence because there are large changes in shape when the fish makes a sharp turn, which makes the tracking challenging. We compare our method to the method that predicts only shape translation but not shape deformation.

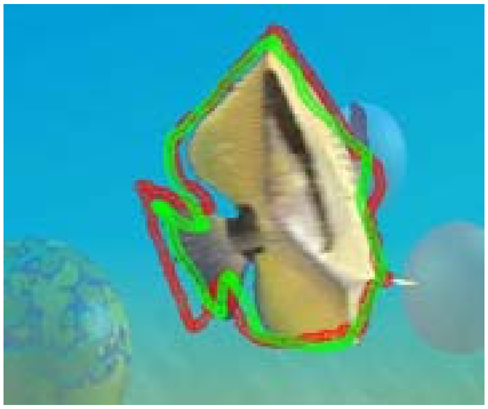

Figure 3: The green contour is predicted by our dynamic shape model, and the red contour is the optimal contour of the previous image frame with predicted translation.

Figure 3 shows the estimated initial contours of both methods. It is clear that our method accounts for scale change in horizontal dimension, but the other method does not. The fin on the upper right side of the fish is partially occluded in the previous image frame. Both methods do not predict this discontinuous change in shape. But our method does move the fin upward according to its appearance in previous image frames. The quality of the converged contour by any snake model will benefit from a better initial shape prediction.

7. Limitations

In Section 2.1, we approximate scale invariance for FD by densely sampling along the boundary to obtain the chaincode. However, for small regions, the spatial quantization is likely to introduce considerable noise to the FD. To avoid this problem, we eliminate regions smaller than a certain size. Consequently, our model does not detect small or far away fires. Small regions are expected to increase misclassification rate and synthesis results are better for larger regions.

The AR model is a linear dynamical system. There may be cases where linear model is not sufficient. In such cases, nonlinear dynamical model can be adopted under the proposed framework. Similarly, any other shape description method with boundary, numeric, and information preserving properties may be used in place of FD.

8. Conclusion

In this paper, we have proposed a novel model for dynamic shape. Although both FD and AR model have been

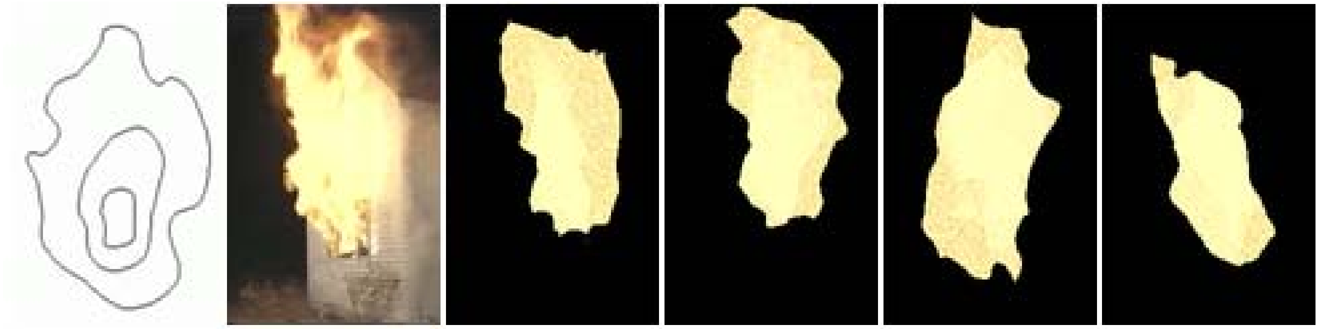

Figure 2: Leftmost image: A nested ring structure models the fire region. Second image: An example fire image from the given video sequence. Others: Selected frames of the synthesized fire image sequence.

well established, using them together to analyze temporal shape variation is not discussed in literature. Traditional shape analysis focuses on spatial similarity, but not temporal similarity. The autoregressive model has been applied mainly to model 1-D signals [8] and 2-D pixel interdependences [18, 21]. We are not aware of any work on AR modeling of region shape changes.

References

[1] A. Blake, R. Curwen, and A. Zisserman. Affine-invariant contour tracking with automatic control of spatiotemporal scale. In *ICCV*, pages 66–75, 1993.

[2] G. Healey, D. Slater, T. Lin, B. Drda, and D. Goedeke. A system for real-time fire detection. In *Computer Vision and Pattern Recognition*, pages 605–606, 1993.

[3] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision*, volume 1, pages 343–356, 1996.

[4] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1987.

[5] F. Leymarie and M. Levine. Tracking deformable objects in the plane using an active contour model. *IEEE Trans. on PAMI*, 15(6):617–634, 1993.

[6] C.-B. Liu and N. Ahuja. Vision based fire detection. In *17th International Conference on Pattern Recognition*, 2004.

[7] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[8] H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer-Verlag, 1991.

[9] S. Mori, H. Nishida, and H. Yamada. *Optical Character Recognition*. John Wiley & Sons, 1999.

[10] A. Neumaier and T. Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software*, 27(1):27–57, 2001.

[11] T. Pavlidis. A review of algorithms for shape analysis. *Computer Graphics and Image Procesing*, 7:243–258, 1978.

[12] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Trans. on PAMI*, 13(7):715–729, 1991.

[13] E. Persoon and K. Fu. Shape discrimination using fourier descriptors. *IEEE Transactions on Systems, Man and Cybernetics*, 7(3):170–179, March 1977.

[14] W. Phillips, III, M. Shah, and N. da Vitoria Lobo. Flame recognition in video. In *Fifth IEEE Workshop on Applications of Computer Vision*, pages 224–229, December 2000.

[15] W. T. Reeves. Particle systems – a technique for modeling a class of fuzzy objects. *ACM Transactions on Graphics*, 2:91–108, April 1983.

[16] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[17] K. Sims. Particle animation and rendering using data parallel computation. *ACM Computer Graphics (SIGGRAPH '90)*, 24(4):405–413, 1990.

[18] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *IEEE International Conference on Computer Vision*, pages 439–446, 2001.

[19] A. Srivastava, W. Mio, E. Klassen, and X. Liu. Geometric analysis of constrained curves for image understanding. In *Proc. Second IEEE Workshop on Variational, Geometric and Level Set Methods in Computer Vision*, 2003.

[20] J. Stam and E. Fiume. Depicting fire and other gaseous phenomena using diffusion processes. *Proceedings of ACM SIGGRAPH 1995*, pages 129–136, 1995.

[21] M. Szummer and R. W. Picard. Temporal texture modeling. In *IEEE International Conference on Image Processing*, volume 3, pages 823–826, 1996.

[22] D. Terzopoulos and D. Metaxas. Dynamic 3d models with local and global deformations: deformable superquadrics. *IEEE Trans. on PAMI*, 13(7):703–714, 1991.

[23] D. Terzopoulos and R. Szeliski. Tracking with kalman snakes. In A. Blake and A. Yuille, editors, *Active Vision*, pages 3–20. MIT Press, Cambridge, MA, 1992.

[24] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 1999.

[25] L.-Y. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of ACM SIGGRAPH 2000*, pages 479–488, 2000.

[26] A. J. Yezzi and S. Soatto. Deformotion: Deforming motion, shape average and the joint registration and approximation of structures in images. *International Journal Comput. Vision*, 53(2):153–167, 2003.