Learning Repetitive Patterns for Classifying Non-Rigidly Deforming Texture Surfaces

Roman Filipovych and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
{rfilipov,eribeiro}@fit.edu

Abstract

In this paper, we address the relatively unexplored problem of classifying texture surfaces undergoing significant levels of non-rigid deformation. State-of-the-art texture classification methods have demonstrated to be very effective for classifying fronto-parallel texture fields. Recently, affine-invariant descriptors have been proposed as an effective way to model local perspective distortion in textures. However, if the effects of local surface curvature distortion are large, affine-invariant descriptors become unreliable. Our contribution in this paper is twofold. First, we propose a method for learning representative basic elements of non-fronto-parallel texture fields undergoing non-rigid deformations. Secondly, we demonstrate the effectiveness of our texture learning method for the classification of non-rigid deforming texture surfaces. We test our method on a set of images obtained from man-made texture surfaces.

1. Introduction

In this paper, we address the problem of classifying images of non-rigid texture surfaces. In particular, we focus ourselves on the case where the surface is covered with a repetitive and sometimes sparse texture pattern. Additionally, the observed surface is assumed to undergo significant amounts of random curvature-induced deformation. This distortion will cause the appearance of local texture to vary in unpredictable ways. Perceiving and modeling the appearance of repetitive textures are important visual tasks with a number of applications including surface tracking, texture classification, and texture synthesis. However, obtaining accurate descriptions from non fronto-parallel texture fields is not a trivial problem as the observed pattern appearance can vary significantly with both the viewing geometry and the surface orientation [17, 7, 12, 9]. This appearance variation can be problematic for most standard texture classification methods. Indeed, changes in local curvature produce nonlinear warping of some image regions. Consequently, texture descriptors evaluated on these warped image regions are likely to be unreliable.

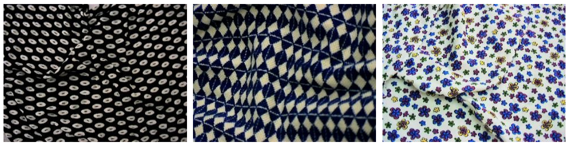

Figure 1. Deforming texture surfaces.

Our goal in this paper is twofold. First, we describe a method for learning basic undistorted affine-invariant texture primitives from videos of a deforming surface. Our texture modeling approach is similar to other methods based on learning *textons* (i.e., primitive texture elements) [10, 9]. However, in our method, we propose to learn the appearance of basic texture elements using distances calculated on an isometric mapping representation [19]. This mapping allows us to remove elements containing high levels of curvature distortion from the learned models. A more detailed description of our texture learning method can be found in [6]. Secondly, we show how our texture model can be used for classifying a set of novel videos of the same surface pattern undergoing varying levels of free-form deformations. Figure 1 shows samples of the textured surfaces used in the study presented in this paper. To accomplish these goals, we commence by investigating the distribution of extracted affine-invariant texture descriptors on a nonlinear manifold embedding. Here, we assume that the population of affine-invariant descriptors lies on a lower dimensional manifold describing mainly variations in surface orientation and curvature. The learned manifold seems to describe the departure from local planarity of affine-invariant descriptors. Under the low dimensional manifold assumption, we describe a learning procedure that allows us to group repetitive texture elements while selecting the best set of candidates to represent the actual undistorted repetitive texture components [6]. Finally, we compare our approach with a K-Means-based texton learning classification method for the task of classifying videos of deforming texture surfaces.

The remainder of this paper is organized as follows. Section 2 provides a survey of the related literature. Section 3 describes the details of our texture primitive selection method. The preliminary results of our study are shown in Section 4. Finally, in Section 5, we present our conclusions and directions for future investigations.

2. Related literature

Finding general representations for texture is a challenging problem. In fact, despite extensive research efforts by the computer vision community, there is no currently widely accepted method to model the complexity encountered in all available textures. State-of-the-art texture classification algorithms have successfully approached the texture representation problem by means of statistical descriptors. These descriptors can be built from the response of convolution filters [10, 20], image regions and pixel distributions [9], and frequency-domain measurements [5, 2].

In this paper, we focus ourselves on the problem of classifying texture surfaces undergoing non-rigid deformations. This a relatively unexplored computer vision problem. Indeed, most texture classification methods are based on measurements obtained from planar fronto-parallel texture fields [10, 20, 5]. For example, Leung and Malik [10] introduced a filter bank-based descriptive model for textures that is capable of encoding the local appearance of both natural and synthetic textures. This method achieves impressive classification rates of natural textures due to their ability to learn representative statistical histogram-based models of each texture. However, it is unclear how they would perform on non-rigid deforming surfaces.

There has been some recent attempts to address the classification problem from images of non-rigid, non fronto-parallel textured surfaces [17, 7, 12, 9]. For example, Chetverikov and Foldvari [3] use a frequency-domain affine-invariant representation for local texture regions. Recent work by Lazebnik *et al.* [9] describe an effective algorithm for retrieval and classification of non-rigid and non fronto-parallel textures. They propose a texture classification method based on learned texture primitives (i.e., *textons*) using affine-invariant descriptors. Once the texture primitives are at hand, Lazebnik *et al.* perform classification by using the Earth Movers Distance similarity measure [16] between learned model descriptors and descriptors of new images. This signature-based approach is indeed very effective under the assumptions of orthographic viewing geometry and low-curvature surfaces. However, for surfaces presenting high levels of curvature deformation, the folds and bends of the surface will significantly reduce the ability of affine-invariant descriptors to capture accurate local texture representation. As a result, the learned appearance of basic texture components is likely to be less representative of the actual texture.

Depending on the curvature of the surfaces, the deformation of texture elements can present a significant degree of nonlinearity. This inherent non-linear behavior may cause the clustering metrics to be incorrect. Nonlinear manifold learning techniques such as Isomap [19], Local Linear Embedding (LLE) [15], and Laplacian Eigenmaps [1] are suitable candidates for the analysis of such deformations. For example, Souvenir and Pless [18] characterize deformations in magnetic resonance imaging. Nonlinear manifold learning is also a useful technique for the synthesis of dynamic textures. Liu *et al.* [11] approach the dynamic texture synthesis problem using nonlinear manifold learning and traversing. In [6], Filipovych and Ribeiro proposed the use of non-linear manifold learning methods for modeling the texture deformations caused by surface curvature.

3. Our method

The method proposed in this paper extends our previous work in [6] by applying the texture learning method to the problem of classifying non-rigid deforming surfaces. For completeness, the main steps of the learning method are also summarized in this section. Our classification method is divided into two main stages. In the first stage, a model of the basic repetitive texture primitives is learned from a set of video frames from a training dataset. This stage does not require the availability of fronto-parallel views of any of the textures to be learned. The learning step aims at creating a dictionary containing the most representative primitives while removing texture components that are highly distorted by surface curvature. This learning stage contrasts with the method proposed by Lezebnik *et al.* [9] in two main points. First, we express the appearance variation in the population of local texture affine-invariant descriptors using nonlinear manifold distances rather than the standard Euclidean distance. Secondly, we propose a selection process that removes learned components highly distorted by surface curvature. The second stage of the method is a classification step that measures the similarity between texture models. Here, the model having the maximum similarity is considered to represent the class of the novel texture.

3.1. Learning stage.

Step 1 - Extraction of affine-invariant regions. The first step of our algorithm consists of extracting a large number of image subregions from a set of video frames of the observed surface. This step is subdivided into two main parts. First, a large set of affine-invariant interest points is detected on each image or video-frame. Here, we use the Kadir and Brady's salient feature detector [8] as it provides information about the affine scale of the detected image features. This detector outputs elliptical image subregions centered at each feature of interest. The extracted subregions are subsequently normalized to a common scale-invariant shape (e.g., circle). The remaining rotation ambiguity can be removed by representing the normalized subregions using the spin-image affine-invariant descriptor proposed in [9]. This descriptor is essentially a pixel gray-level intensity histogram calculated on a scale-invariant polar representation of an image subregion. It represents the radial frequency of normalized pixel intensities. The polar mapping of the pixel intensity transforms rotations into translations. The spin-image histogram representation that follows is translation invariant. Figure 2 illustrates the spin-image construction process. The representation allows for full affine-invariance. Let $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N)$ be the set of affine-invariant descriptors obtained in this step, where N is the total number of subregions.

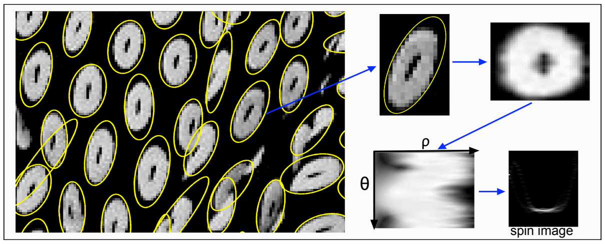

Figure 2. Spin-image construction.

Step 2 - Nonlinear manifold mapping of descriptors. This step aims at obtaining a compact representation of the most significant repetitive patterns on the image. However, the nonlinear nature of the distortion in the set of extracted affine-invariant descriptors does not always allow for correct distance measurements in the original feature space. Additionally, the spin-image descriptor itself carries a significant level of information redundancy. To obtain a better description of the variation in the dataset, we assume that basic texture elements along with their nonlinear deformations lie on a low-dimensional nonlinear manifold in which the two intrinsic dimensions of variability describe mainly local surface orientation and curvature distortion. Based on this assumption, we perform Isomap [19] on the original distribution $\mathbf{S}$. Isomap allows for data dimensionality reduction while preserving the nonlinear manifold's intrinsic geometry. The reduced dimension set of subregions produced by this step is given by $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.

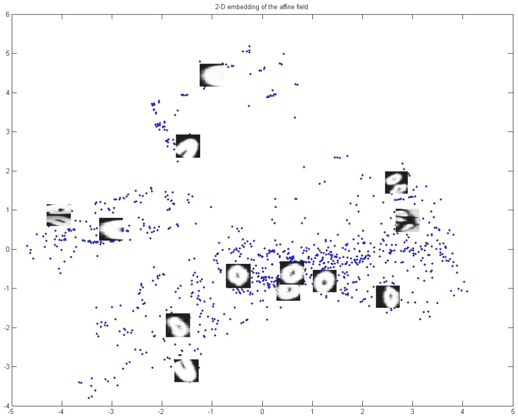

Figure 3. 2-D Isomap embedding of texture spin-images for a deforming surface.

Figure 3 illustrates two dimensions of the learned Isomap manifold for the local affine-invariant patch distribution. For clarity, only two dimensions of the manifold are shown. Superimposed samples of embedded images are also shown. We extracted a large number of affine-invariant descriptors from a video of the patterned surface shown in the first column of Figure 1. The plot shows a dense cluster near the center containing mostly locally planar patches. On the other hand, nonlinearly deformed patches tend to group themselves into clusters with respect to deformation similarity. Finally, occluded or distorted elements form relatively sparse groups with large within-class variation.

Step 3 - Learning representative texture components. The goal of this step is to determine the most representative classes of texture elements in $\mathbf{X}$ given by the previous step. We model the distribution of affine-invariant descriptors as a mixture of K Gaussian densities given by $p(\mathbf{x}|\Theta) = \sum_{i=1}^{K} \alpha_i p_i(\mathbf{x}|\theta_i)$, where $\mathbf{x}$ is an affine-invariant descriptor in the Isomap manifold space, α_i represent the mixing weights such that $\sum_{i=1}^{K} \alpha_i = 1$, Θ represents the collection of parameters $(\alpha_1, \ldots, \alpha_K, \mu_1, \Sigma_1, \ldots, \mu_K, \Sigma_K)$, and p_i is a multivariate Gaussian density function parameterized by μ and Σ (i.e., the mean vector and covariance matrix, respectively). Each mixture component represents a set of texture descriptors of similar appearance on the image. The model parameters can be estimated by using the Expectation-Maximization (EM) algorithm [4].

In order to obtain sharper representations of the learned texture components, we select a single descriptor from each cluster to represent a basic texture component in the image. In other words, a set of texture components is selected as:

$$\boldsymbol{\tau}_j = \arg\max_{\mathbf{x}_i} p_j(\mathbf{x}_i|j) \qquad j = 1, \ldots, K \tag{1}$$

A set of basic components is obtained by this process and is used to create a dictionary representation $\mathbf{d} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K\}$. However, the nonlinear nature of the surface distortion will compromise the representativeness of some of the learned mixture components. As a result, the learned clusters might not represent actual texture components but a geometrically warped version of them. Next, we propose a way to remove these non-representative elements from our dataset of learned texture primitives.

Step 4 - Texture component model. Our main goal in this step is to distinguish between distributions of affine transformed basic texture elements and their nonlinearly deformed counterparts. The nature of the nonlinear texture deformations is mostly anisotropic (i.e., directional appearance). Consequently, we expect the clusters of non-affine distorted elements to have a relatively small number of data points. Here, distributions with low prior probability will most likely correspond to regions distorted by nonlinear transformations. Elements falling within such distributions can be safely discarded and therefore removed from the dictionary. The remaining distributions may represent two cases. The first case corresponds to classes of affine transformed elements that are representative of the texture. The second case represents classes consisting of nonlinearly distorted regions. Our experiments have shown that the distribution of nonlinearly distorted elements have high within-class variation. Based on this observation, we rank the remaining dictionary elements based on the decreasing order of the within-class variation of their classes (i.e., we use value of the determinant of the covariance matrix for the class, $|\Sigma_j|$). The top-ranked elements are selected as the ones that represent classes of locally planar surface regions:

$$\mathbf{d} = \{\mathbf{s}_j\} \quad \text{such that} \quad |\Sigma_j| \geq |\Sigma_{j+1}| \tag{2}$$

where $\mathbf{s}_j \in \mathbf{S}$. The above procedure is performed for each texture class and it is similar to signature generation [16].

We represent the appearance of each texture using the learned set of basic components. Figure 4 shows the results for a patterned fabric used in our experiments. Figure 4(a) shows (from top to down) the ranked sequence of learned texture components obtained by our method. Figure 4(b) shows three frame images with the learned locally planar texture components superimposed on the fabric's surface.

Representing the local texture information using spin-images permits to achieve a certain level of robustness against non-rigid deformations [9]. The improvement proposed in our method is to use nonlinear manifold embedding to unwrap the texture nonlinear distortions and only then eliminate classes representing the distorted regions.

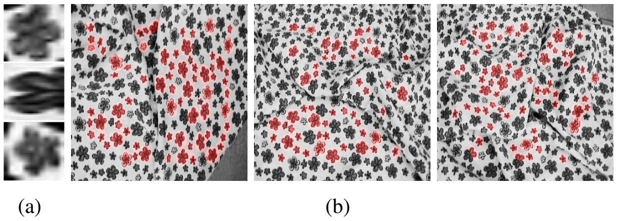

Figure 4. Learned locally planar regions.

3.2. Classifying novel texture sequences.

Once the appearance models for each texture in the training database are at hand, the classification stage consists simply of performing the four steps of the texture model learning for a novel texture sequence and measuring the overall dissimilarity between the components in the appearance model of the novel texture and the previously learned models for each texture class. We use the following dissimilarity measurement:

$$\alpha(\mathbf{d}_1, \mathbf{d}_2) = \sum_i \min_j \|\mathbf{s}_{1i} - \mathbf{s}_{2j}\| \tag{3}$$

The above dissimilarity measurement is a non-symmetric variant of the Hausdorff measure [13].

4. Experimental results

Our experiments are divided into two main parts. First, we evaluated our *texton* learning algorithm on video sequences of a number of texture surfaces for increasing levels of deformation. The surfaces used in our experiments consisted of patterned fabrics bought from a local shop. To produce the deformations, we have deformed the fabric manually while recording the video sequences. Three of these patterns are shown in Figure 6. Secondly, we compare the classification results between our method and the standard K-Means learning method.

We commence by extracting a large set of subregions from a sample of frames of video sequences for each texture class. Our current method does not use any temporal information and a sparse set of frames is usually sufficient for the algorithm to work. We extracted approximately 2,000 local affine invariant descriptors from the set of images. The feature extraction stage was followed by a ten-dimensional

		Blobs			Diamonds			Flowers		
		Low	Medium	High	Low	Medium	High	Low	Medium	High
K-Means	Blobs	0.76	0.81	0.65	1.50	1.31	1.31	1.91	1.81	1.61
	Diamonds	1.40	1.21	1.87	1.48	1.42	1.39	0.68	0.36	0.72
	Flowers	1.97	1.77	1.63	1.50	1.38	1.34	0.67	0.36	0.72
Our Method	Blobs	3.12	3.11	3.42	4.25	4.13	3.95	3.92	3.95	4.18
	Diamonds	4.06	4.11	4.05	3.11	3.64	3.75	4.11	4.11	4.05
	Flowers	4.13	4.16	3.99	4.05	4.00	3.99	2.88	3.35	3.48

Figure 6. Dissimilarity matrix between K-Means learning and our method.

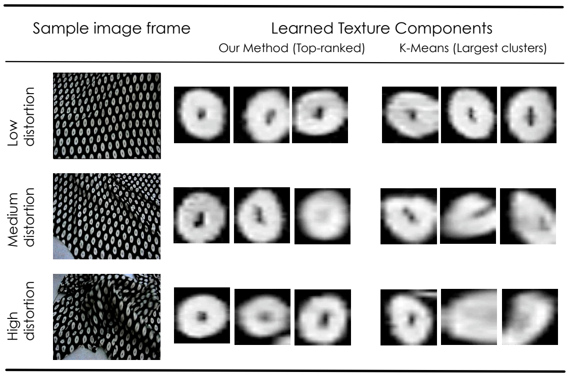

Figure 5. Learned textons.

Isomap embedding of the corresponding affine invariant descriptors. After the EM learning step was performed, we selected classes with the highest prior probabilities such that the sum of priors formed 60% of the total population. The number of mixture components was experimentally determined, and was set to 10 (See [6] for details). For every such class, the basic texture element was selected using Equation 1. Finally, the ranking stage was performed to remove noisy components and rank the most representative ones. Here, we selected the 75% top ranked elements. A single patch descriptor was selected by the algorithm whenever the calculations produced no resulting components. Our results were consistent for different numbers of mixture components. It is common that among the top-ranked texture primitives there will be several exemplars of a single element. This might be caused by region detection errors or illumination changes.

The learning step performed by our method determined the repetitive texture elements consistently for various levels of curvature distortion. Figure 5 shows a qualitative comparison between the primitives learned by our algorithm and a standard K-Means-based method. In the figure, each row shows a sample video frame of each levels of distortion (i.e., low, medium, and high, respectively). The figure also displays three of the estimated basic texture primitives besides each corresponding video frame. The appearance of the primitives obtained by our method was quite consistent for different levels of distortion. This contrasts with the high-variance results obtained by the K-Means learning method. It is worth noting that the textures shown in our experiments have a relatively low complexity in terms of the variety in both number and shape of textons. Despite this low appearance complexity, the K-Means-based method was unable to extract correct representations of the basic texture elements.

In the second part of our experiments, we provide a quantitative comparison between the classification results obtained by our algorithm and typical results obtained by clustering the affine-invariant feature space using the K-Means algorithm. Figure 6 shows an one-against-all classification dissimilarity matrix for a sample of textures in our experiments. The columns of the matrix correspond to the training dataset while the rows correspond to the dataset of novel videos. From the results, our method consistently obtained maximum dissimilarity values (i.e., minimum distance) for all the correct texture classes. In contrast, the K-Means-based classification did not show any level of consistency for the videos used in this experiment.

It is possible that some surfaces may have no prominent textons. In general, there are two possible cases to consider. In the first case, the texture does not actually have textons and the proposed algorithm cannot be applied. In the second case, the texture distortion is so high that the number of actual locally planar regions is very small. As a result, the current method would not be able to correctly extract undistorted textons. A possible solution might be to make use of shading information whenever available.

The manifold learning stage is a crucial step of our algorithm. Our experiments show that, for high level of surface distortions, the use of spin-images without the non-linear manifold embedding did not provide satisfactory results.

Overall, the results show that the proposed method is able to distinguish between locally planar texture elements and their nonlinearly distorted versions. Additionally, we are able to accomplish promising classification results even when significant levels of curvature distortion are present.

5. Conclusions and future work

In this paper, we proposed a classification method that learns basic texture primitives of patterned surfaces distorted by non-rigid motion. The algorithm uses nonlinear manifold learning to capture the intrinsic dimensionality of the distortion of non-rigid deforming texture surfaces. A selection procedure for finding the most representative local texture components was presented. The learned primitives were used to create appearance models of the texture in videos of deforming surfaces. We applied our texture learning method to the problem of classifying deforming texture surfaces. Our experiments showed the effectiveness of the method on a set of images obtained from patterned fabric surfaces undergoing a range of non-rigid deformations.

There are several avenues for future work. First, experiments demonstrating the limits of the proposed method as a function of the amount of surface distortion should be performed. To accomplish this, we need to devise a numerical measure that reflects the amount of curvature-induced distortion present on an image. We have not encountered such a measure documented in the literature. Additional interesting future directions include further investigation of the effects of curvature on local texture measurements as well as the introduction of both spatio-temporal information and inherent texture repetitiveness into the nonlinear manifold learning stage [14]. Studies aimed at developing these ideas are in hand and will be reported in due course.

5.0.1 Acknowledgments

This research was supported by U.S. Office of Naval Research under contract: N00014-05-1-0764.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Neural Information Processing Systems*, pages 585–591, 2001.

[2] A. Bhalerao and R. Wilson. Affine invariant image segmentation. In *British Machine Vision Conference*, 2004.

[3] D. Chetverikov and Z. Foldvari. Affine-invariant texture classification. In *IEEE International Conference on Pattern Recognition*, page 3901, Barcelona, Spain, 2000.

[4] A. Dempster, N.M.Laird, and D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Stat. Soc., Series B*, 39(1):1–38, 1977.

[5] M. N. Do and M. Vetterli. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Trans. Image Process.*, 11(2):146–158, 2002.

[6] R. Filipovych and E. Ribeiro. Learning basic patterns from repetitive texture surfaces under non-rigid deformations. In *International Conference on Image Analysis and Recognition (ICIAR)*, Montreal, Canada, 2007.

[7] J. Hays, M. Leordeanu, A. A. Efros, and Y. Liu. Discovering texture regularity as a higher-order correspondence problem. In *European Conference on Computer Vision (2)*, pages 522–535, 2006.

[8] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *European Conference on Computer Vision*, pages Vol I: 228–241, 2004.

[9] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[10] T. Leung and J. Malik. Recognising surfaces using three-dimensional textons. In *ICCV99*, pages 1010–1017, 1999.

[11] C. B. Liu, R. S. Lin, N. Ahuja, and M. H. Yang. Dynamic textures synthesis as nonlinear manifold learning and traversing. In *British Machine Vision Conference*, pages 859–868, 2006.

[12] Y. Liu, W.-C. Lin, and J. H. Hays. Near regular texture analysis and manipulation. *ACM Transactions on Graphics (SIGGRAPH)*, 23(3):368 – 376, August 2004.

[13] O. Masoud and N. Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743(15), 2003.

[14] A. Rahimi, B. Recht, and T. Darrell. Learning appearance manifolds from video. In *IEEE Conference on Computer Vision and Pattern Recognition - Vol. 1*, pages 868–875, San Diego, California, USA, 2005.

[15] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, December 2000.

[16] Y. Rubner and C. Tomasi. Texture-based image retrieval without segmentation. In *International Conference on Computer Vision - Vol. 2*, page 1018, 1999.

[17] F. Schaffalitzky and A. Zisserman. Geometric grouping of repeated elements within images. In *Shape, Contour and Grouping in Computer Vision*, pages 165–181, 1999.

[18] R. Souvenir and R. Pless. Isomap and nonparametric models of image deformation. In *IEEE Workshop on Motion and Video Computing (WACV/MOTION'05) - Vol. 2*, pages 195–200, Colorado, USA, 2005.

[19] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[20] S. C. Zhu, C. en Guo, Y. Wang, and Z. Xu. What are textons? *Int'l Journal of Computer Vision*, 62(1-2):121–143, 2005.