THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 111:357–367, 1997 August

GALAXY MORPHOLOGY WITHOUT CLASSIFICATION: SELF-ORGANIZING MAPS

AVI NAIM,[1] KAVAN U. RATNATUNGA,[1] AND RICHARD E. GRIFFITHS[1]
Johns Hopkins University, Department of Physics and Astronomy, Baltimore, MD 21218
Received 1996 August 16; accepted 1997 March 25

ABSTRACT

We examine a general framework for visualizing data sets of high (greater than 2) dimensionality and demonstrate the framework by taking the morphology of galaxies at moderate redshifts as an example. The distributions of various populations of such galaxies are examined in a space spanned by four purely morphological parameters. Galaxy images are taken from the *Hubble Space Telescope* Wide Field Planetary Camera 2 in the *I* band (using the F814W filter). Since we have little prior knowledge on how galaxies are distributed in morphology space, we use an unsupervised learning method (a variant of Kohonen's self-organizing maps, or SOMs). This method allows the data to organize themselves onto a two-dimensional space while conserving most of the topology of the original space. It thus enables us to visualize the distribution of galaxies and study it more easily. The process is fully automated, does not rely on any kind of "eyeball" classification and is readily applicable to large numbers of images. We apply it to a sample of 2934 galaxies and find that the morphology correlates well with the apparent magnitude distribution and, to a lesser extent, with color and bulge dominance. The resulting map traces a morphological sequence similar to the Hubble sequence, albeit two-dimensional. We use the SOM as a diagnostic tool and rediscover a population of bulge-dominated galaxies with morphologies characteristic of peculiar galaxies. This result is achieved *without* recourse to classification by eye. We also examine the effect of noise on the resulting SOM, and conclude that our results are reliable down to an *I* magnitude of 24. We propose using this method as a framework into which more physical data can be incorporated as they become available. We hope that this method will lead to a deeper understanding of galaxy evolution.

Subject headings: galaxies: evolution — galaxies: fundamental parameters — galaxies: structure

1. INTRODUCTION

Morphological classification of galaxies was originally envisaged as a tool for studying the evolution of galaxies (e.g., Hubble 1936). Much as in other fields of science, as the amount of data grew the classifications were revised and became more and more refined (Sandage 1961; de Vaucouleurs 1959; van den Bergh 1960, 1976). At some point the question arose as to how well these refinements correlate with physical quantities and processes within galaxies. In an excellent review, Roberts & Haynes (1994) showed that morphological types in the local universe do correlate with color, H I mass, and other quantities *in the mean*, but that there is a large scatter about the mean. This implies that morphological classification has become overly refined, at least as far as its relation to physical properties is concerned.

A major limitation of most classification schemes for galaxies is that the schemes were devised solely using samples of nearby galaxies because of the lack of imaging capabilities at higher redshifts. This situation has changed with the advent of the *Hubble Space Telescope* (*HST*) and very large ground-based telescopes. The morphology of large numbers of galaxies at moderate redshifts ($z < 1$) is now available, and preliminary results (Griffiths et al. 1994; Glazebrook et al. 1995; Driver, Windhorst, & Griffiths 1995; Abraham et al. 1996) indicate that many galaxies at moderate redshifts do not fit comfortably on the Hubble sequence. It is an obvious challenge to try to incorporate galaxies at different redshifts into one coherent scheme.

A great deal of work has been done recently on morphological classification of faint galaxy images. Most of it, however, relies on "eyeball" classifications: e.g., Cowie, Hu, & Songaila (1995) presented deep *I*-band Wide Field Planetary Camera 2 (WFPC2) images of a *K*-selected sample. They gave a qualitative eyeball account of the change they saw in the morphology of galaxies around $K = 19.5$ mag. Driver et al. (1995) divided galaxies in a deep WFPC2 field into three eyeball classes and analyzed the number counts as a function of type. van den Bergh et al. (1996) produced a morphological catalog of galaxies in the Hubble Deep Field (HDF) that was again based on eyeball classifications. In addition, they supplied two quantitative parameters for those galaxies (light concentration and asymmetry) that allow for a more objective analysis. Odewahn et al. (1996) used both eyeball classifications and trained artificial neural networks to obtain classifications for galaxies in deep *HST* fields. Their network utilized parameters derived from surface brightness profiles in *U*, *B*, *V*, and *I* filters. The move from pure eyeball classification to automated classification using objective parameters has been inevitable because of the large quantities of images that have become available over the past few years. The parameters used by van den Bergh et al. (1996) proved to be a useful first step in this direction, although they gave a very crude separation of eyeball types. Using light-profile parameters, Odewahn et al. (1996) discussed the possible makeup of the population of blue galaxies. Both these papers tied their quantitative parameters to classifications on the existing Hubble sequence, which is apparently insufficient for the full range of morphologies detected with *HST*.

Since the Hubble sequence appears too refined on the one hand, and not general enough on the other, we suggest a

[1] Current address: Department of Physics, Wean Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.

more general approach. In recognition of the fact that morphology is a continuous quantity, we abandon any attempt to tag each galaxy with a specific type. Instead, we use a space spanned by four morphological parameters, previously introduced elsewhere (Naim, Ratnatunga, & Griffiths 1997), and examine the distribution of various populations of galaxies in it. We start with a large, complete, magnitude-limited sample of *HST* WFPC2 images described in § 2. We have little prior knowledge of the distributions of galaxies in this space. For this reason, we use a variant of an *unsupervised learning* technique called "self-organizing maps" (SOMs). It allows data taken from a space of high dimensionality to organize themselves into a two-dimensional "histogram," while retaining most of the original topology. The resulting map can then be plotted and analyzed. SOMs, explained in detail in § 3, therefore combine nonlinear clustering with a dimension-reduction technique. SOMs have been little used in astronomy to date (the one example we are aware of is Mähönen & Hakala 1995) and, as we show below, prove a valuable tool for unsupervised data analysis. However, one important point has to be stressed from the outset: we are using a nonparametric method here, in the sense that the results are not described in terms of functional dependencies between the parameters we use. Consequently, SOMs are primarily a *diagnostic tool* that should be used only as a first step toward forming a model that explains the observations. The method's most important feature is the ability to identify special populations that merit closer examination. We first demonstrate the application of SOMs to a synthetic data set (§ 4) and then apply them to the sample of *HST* galaxies (§ 5). A discussion follows in § 6.

2. SAMPLE SELECTION AND MORPHOLOGICAL PARAMETERS

2.1. *Sample Selection*

It is easiest to select a suitably large sample from data that were collected uniformly. The 27 contiguous fields of the Groth-Westphal strip (Groth et al. 1994) make an excellent such collection. *I*-band (F814W) images were preferred over *V*-band (F606W) images (which are also available for the same fields) for two reasons: first, exposures in *I* were about 50% longer and typically resulted in images with higher signal-to-noise ratios; second, at the expected redshifts of these galaxies the *I* filter corresponds roughly to the rest-frame *B* band, in which most existing morphological schemes were defined, while the *V* filter corresponds to a much bluer rest-frame band in which images appear much more broken up.

Our indications from previous work (Naim et al. 1997) are that down to an isophotal magnitude of $I = 24.0$ mag a distinction between morphologically "normal" and "peculiar" galaxies is still possible, although it suffers increasingly from effects of noise toward the faint end. We decided to attempt the same limit here and to then examine a subset of the sample with higher signal-to-noise ratio to see what effect the noise had on our results. There were 3391 images brighter than $I = 24$ mag in the Groth-Westphal strip. The Medium Deep Survey (MDS) pipeline, using a maximum likelihood method (K. U. Ratnatunga, R. E. Griffiths, & E. J. Ostrander 1997, in preparation), fits simple photometric models ($r^{1/4}$ law, exponential disk, and combinations of the two) to galaxy images. It was found that the fitted half-light radius parameter is very useful in separating stars and compact objects from galaxies, and the limiting value was empirically set at 0.″1 (1 image pixel). It is clear that some distant galaxies, as well as closer compact objects, have half-light radii smaller than this limit. Therefore, not all of the 421 images that were removed from the sample because of failing this test are indeed stars. However, images whose half-light radius is smaller than 0.″1 are typically no more than 3–4 pixels across, thus containing almost no morphological information. Consequently, we use this cutoff not only as a safeguard against contamination by stars but also as a practical lower limit for the derivation of our parameters. As well as the 421 images mentioned above, fewer than 20 other images were rejected by the program that calculates the morphological parameters because of low quality (e.g., too high a fraction of missing pixels). During classifications by eye (see below), several more images (less than 20) were rejected because of other problems (e.g., a nearby star overlapped the galaxy). The final sample contains 2934 entries.

Isophotal magnitudes are tightly correlated with the *integrated* signal-to-noise index, ν, which is calculated by summing the individual signal-to-noise ratios that are greater than 1 over image pixels (see Ratnatunga et al. 1997 for details). Note that since we are using the *integrated* signal-to-noise index, the values we are dealing with are typically of order 100. At the limiting magnitude of $I = 24.0$ mag all but six galaxies in the sample have $\nu > 100$, which is, incidentally, the limit below which no disk-plus-bulge photometric model fit was attempted by the maximum likelihood software (although pure bulge and pure disk models were attempted down to much lower values).

2.2. *Morphological Parameters*

A full description of the four parameters we use is given in Naim et al. (1997). We therefore give only a brief description of the parameters here. In designing these parameters we attempted to give as full a description as possible of the features that stand out in galaxy images, while remaining neutral with respect to quantities such as the underlying photometric model or the color of the image. Our parameters are the following:

1. *Blobbiness.*—The degree to which bright pixels stand out, accentuating bright localized structure. This parameter may be correlated with regions of intense star formation. Briefly, this parameter is calculated for each bright image pixel as the ratio of brighter pixels to the total number of pixels in a semi-circular environment around the central bright pixel.

2. *Isophotal center displacement.*—The displacement of geometric centers of various isophotes from each other, as a measure of overall asymmetry. This parameter, by detecting tidal tails, may be related to merging history.

3. *Isophotal filling factor.*—The fraction of pixels belonging to a certain isophote out of the number of pixels in the ellipse enclosing that isophote. This is a measure of overall structure: in featureless images this fraction is expected to be higher than in images exhibiting a great deal of structure, because in the latter bright pixels will be found at higher radii, making the enveloping ellipse much bigger. This expectation is verified for an eyeballed subset of our sample (Naim et al. 1997, see also below), in which late spiral and peculiar galaxies average a value of less than 0.2 for this

parameter, early spirals average close to 0.3, and elliptical/lenticular galaxies average over 0.35.

4. *Skeleton ratio* of detected structures, indicating how elongated the structures are. Briefly, for detected structures in the galaxy image, this parameter is the ratio of pixels making up the "backbone" of the structures to the total number of pixels in the structures.

The first three parameters are evaluated from the raw I-band image, while the fourth is derived from the residual image, which is the image left over after subtraction of the best-fit photometric model given by the maximum likelihood software.

3. SELF-ORGANIZING MAPS

The motivation behind SOMs derives from our inability to plot data in more than three dimensions. Kohonen (1989) suggested a nonlinear mapping from a given M-dimensional space ($M > 2$) onto a two-dimensional map in a way that maintains as much as possible of the topology of the higher dimensional space. SOMs are therefore one implementation of *unsupervised learning*, a generic name referring to methods for describing data without any prior knowledge of how they cluster. Self-organization takes place in an iterative manner with little user intervention. The role played by the user is reduced to defining the organizing criterion (i.e., the criterion determining which vector is mapped to which node in the SOM). The resulting map can be regarded as a two-dimensional histogram, although its axes do not carry the usual parametric meaning. The numbers on the x- and y-axes represent positions in the map, not values of the M parameters making up the space of the data.

Let a given data set contain N vectors of dimension M, each describing a single object (e.g., a galaxy). In the case of our galaxy sample, $N = 2934$ and $M = 4$. The "data space" is therefore M-dimensional. Define the map as a two-dimensional array of discrete nodes. Throughout this paper we use square maps of size 16×16 nodes. The nodes occupy positions in what we refer to as the (two-dimensional) "map space." The link between the two spaces is realized by assigning each node of the map an M-dimensional "characteristic" vector from the data space. Note that this assignment is done in an automated way, with no input from the user, i.e., it is truly an unsupervised operation. The key measure in the process of self-organization is distance. Distances are calculated independently for each space. For greater clarity we will refer to the distances in the following as data distance and map distance, respectively. The user's role is confined to choosing a certain distance measure (e.g., the L^2 norm, also known as the Euclidean distance), which serves as the organizing criterion. Each object in the data set is mapped to the node whose characteristic vector is closest to it in the sense of that distance measure (the "winning" node). In each iteration of the training process the entire data set is mapped to the SOM, and then the characteristic vectors of the nodes are updated according to the objects mapped onto them. Topology is preserved by allowing nodes in the vicinity of the winning node to be updated as well. Over many iterations this will cause nodes that lie close to each other to develop similar characteristic vectors, and therefore, eventually, whole regions in the SOM will correspond to specific populations in the data set. While nearby nodes will represent finer details within each population, nodes far away from each other will represent significantly different populations. The iterations are stopped once some convergence criterion (see below) is met.

Normally one initializes the map nodes to have random characteristic vectors at first. However, this procedure could assign very different vectors to adjacent nodes, while similar vectors could be found far from each other. This could result in two different populations of galaxies overlapping in the resulting SOM or in a single population being artificially split between two or more regions in the map. It has been suggested that the first problem could be overcome by running the SOM several times, each time starting with a different set of random characteristic vectors, and choosing only the "best" run, e.g., in the sense of minimizing the χ^2 difference between all objects and the characteristic vectors of the nodes to which they were mapped. However, since this learning process is unsupervised, there may be many very different minima of such a measure, each corresponding to a different topology, with little to choose between them. In addition, this solution does not answer the second problem we raise. Furthermore, randomizing the initial characteristic vectors makes the entire process unrepeatable.

In order to avoid these difficulties, we first run a simple clustering algorithm (SCA) on the data and use the emerging crude clusters to decide how to initialize the map vectors. Our version of the SOM algorithm consists of two stages: In the startup phase we employ the SCA to get a rough idea of how the objects cluster. The SCA initially defines each object in the data set as an independent "group" in data space. The L^2 norm (Euclidean distance) is adopted as the data-distance measure, and a search radius is defined that increases linearly with the number of iterations. In each iteration, groups whose centers of mass lie within a search radius of each other are merged, and so the number of groups decreases monotonically with time. The stopping criterion for the SCA is met once the three largest groups contain, between them, more than half these vectors. The critical number was set to three because three vectors define a plane and can therefore be mapped in a topologically faithful manner onto our two-dimensional map. Note, however, that the three largest groups need not represent the most diverse combinations of the morphological parameters. For this reason we examine all groups containing more than 1% of the data when the SCA is stopped (typically of order 10 groups). Out of all the vectors representing the "centers of mass" of these groups, we select those that contain a maximum or a minimum value of at least one of the parameters. Since we are using four parameters, the number of such selected vectors (N_v) is in the range two to eight, but is expected to be closer to eight in most cases. The results of running the SCA (and any other crude clustering algorithm) over a given data set are expected to be quite independent of the exact details of the algorithm. Different distance measures may result in somewhat different results, but since we are using the SCA only as the first stage in our analysis, such differences are not important for the final outcome.

In the second stage we iterate through all possibilities of selecting three so-called "anchors," or "key vectors," out of these N_v vectors to initialize and train the SOM. The selected vectors are assigned to three nodes in the map in a way that conserves their relative data distances. All other nodes *within the triangle enclosed by these anchors* are then assign-

ed characteristic vectors that are weighted averages of these three key vectors, the weight being the inverse of the map distance from each anchor node:

$$\boldsymbol{C}^{(i,j)} = \frac{\sum_{k=1}^{3} (\boldsymbol{C}_k / d_k^{(i,j)})}{\sum_{k=1}^{3} (1/d_k^{(i,j)})} , \tag{1}$$

where, for $k \in \{1, 2, 3\}$, $\boldsymbol{C}_k$ is one of the three key characteristic vectors and $d_k^{(i,j)}$ is the map distance between node (i, j) and the node in which the anchor vector resides. Only the region inside the triangle is used. This procedure allows the map nodes to span much of the variance in the data from the outset and guarantees the repeatability of the results. We select only three anchors because, again, three points define a plane and can therefore be mapped in a topologically faithful manner onto the two-dimensional map. Choosing all possible combinations of three vectors for the role of anchors allows us to search for the combination that best represents the data in an unsupervised way. Repeatability is guaranteed because the entire process is deterministic and does not require input from the user.

Next comes self-organization. We again adopt the L^2 norm as our data-distance measure. For each data vector $\boldsymbol{V}$ (describing one galaxy), the winning node is node (i, j) for which the data distance between its characteristic vector $\boldsymbol{C}^{(i,j)}$ and the data vector $\boldsymbol{V}$ is minimal. This distance is given by

$$d(i, j) = \left\{ \sum_{l=1}^{M} [C^{(i,j)}(l) - V(l)]^2 \right\}^{1/2} , \tag{2}$$

where the argument l denotes that the lth component of the vector is being taken. Once the entire data set has been mapped to the SOM, the characteristic vectors of each node is updated. There are two possible sources of alteration at a given node, namely, the objects that were mapped directly onto that node and the objects mapped to nearby nodes, which affect that node by virtue of the attempt to conserve topology. Let $\langle \boldsymbol{V}^{(i,j)} \rangle$ be the average of all the vectors mapped onto node (i, j). Then the first contribution is of the form

$$\Delta \boldsymbol{C}_1^{(i,j)} = \langle \boldsymbol{V}^{(i,j)} \rangle , \tag{3}$$

and the contributions of the second kind will come from nodes (i_1, j_1) around node (i, j) and will each have the form

$$\Delta \boldsymbol{C}_2^{(i,j)\,|\,(i_1,j_1)} = \exp\left[\frac{-(d_m^{(i_1,j_1)})^2}{2\sigma^2} \right] \langle \boldsymbol{V}^{(i_1,j_1)} \rangle , \tag{4}$$

where $d_m^{(i_1,j_1)}$ is the map distance between (i, j) and (i_1, j_1). The "environment kernel" chosen here is a Gaussian whose width, σ, is a decreasing function of the number of iterations n_i:

$$\sigma(n_i) = 1/n_i . \tag{5}$$

The reason for the dependence of σ on the number of iterations is that as the structure of the map becomes more organized it is desirable to limit the effect of the environment. If the other nodes were always allowed to contribute at the same level, the process of self-organization might never converge and finer details in the map could be washed away. For practical purposes of reducing the number of calculations, the environment of node (i, j) from which the nodes (i_1, j_1) are taken is limited to a square of side 7 (i.e., vertical/horizontal map-distance of no more than 3) centered on node (i, j). There is no need to go any farther, because even when σ is maximal at 1 (during the first iteration), the coefficient d_m drops to about 0.01 at a map distance of 3, and therefore nodes further away from (i, j) are unlikely to contribute to (i, j) significantly. The updated value of the characteristic vector of node (i, j) is therefore given by

$$\boldsymbol{C}_{\text{new}}^{(i,j)} = (1 - \eta)\boldsymbol{C}_{\text{old}}^{(i,j)} + \eta \frac{\Delta \boldsymbol{C}_1^{(i,j)} + \sum_{(i_1,j_1) \neq (i,j)} \Delta \boldsymbol{C}_2^{(i,j)\,|\,(i_1,j_1)}}{1 + \sum_{(i_1,j_1) \neq (i,j)} \exp\left[-(d_m^{(i_1,j_1)})^2 / 2\sigma^2 \right]} , \tag{6}$$

where the denominator in the second term is the normalization factor for all the weighted contributions and η is a parameter that describes the "learning rate" of the SOM. We set η to 0.02. It is not advisable to make η large because then the changes in the characteristic vectors can become erratic.

At the end of each iteration, we monitor the rms difference between the current and previous characteristic vector of each node. We stop training the map when the largest of these differences has dropped below 0.1% of its maximum possible value. Typically this criterion leads to convergence within several thousand iterations.

Self-organization is repeated for all selections of three anchors. For each such selection all the vectors in the data set are mapped onto the trained SOM, and the χ^2 difference between the data vectors and the characteristic vectors of the nodes to which they were mapped is monitored to find the best triplet. The SOM resulting from the best triplet is then chosen as the best overall SOM.

4. AN EXAMPLE: NONLINEAR MAPPING IN FOUR DIMENSIONS

We test the ability of the SOM to handle nonlinear mapping in several dimensions by first defining a curve in a space of the same dimensionality as our galaxy data set. In order to demonstrate the ability of the SOM to retain topological information, the curve is specified in parametric form, which conveys a clear notion of the order of points along the curve. The curve is given by

$$\boldsymbol{F}(\theta) = (\sin\theta, \cos\theta, \sin\theta\cos\theta, \sin^2\theta) , \tag{7}$$

where θ is the free parameter. We choose five points along the curve, corresponding to θ-values of $\pi/12$, $\pi/6$, $\pi/4$, $\pi/3$, and $5\pi/12$. Around each of these points we randomly scatter 400 other points. There is little overlap between these five clouds of points, and the relations between any two components of $\boldsymbol{F}$ are nonlinear. The SOM software is trained on a data set containing all 2000 points; the results are shown in Figure 1. The top left panel shows the mapping of the full data set; there appear to be three to five distinct concentrations. The other five panels each depict one group of points (denoted by the corresponding value of θ). It is plain to see that the SOM *maintains the order of the groups along the curve*, although some mixing between adjacent groups takes place. The SOM is therefore capable of mapping nonlinear data sets while conserving much of the topology. Note also that although the initial organization of the SOM has the form of a triangle, in this case it forms an obtuse triangle, closely resembling the true shape of the distribution of points in the original space—that of a one-dimensional curve in a four-dimensional space. One possible drawback of this representation is that groups tend to be more con-

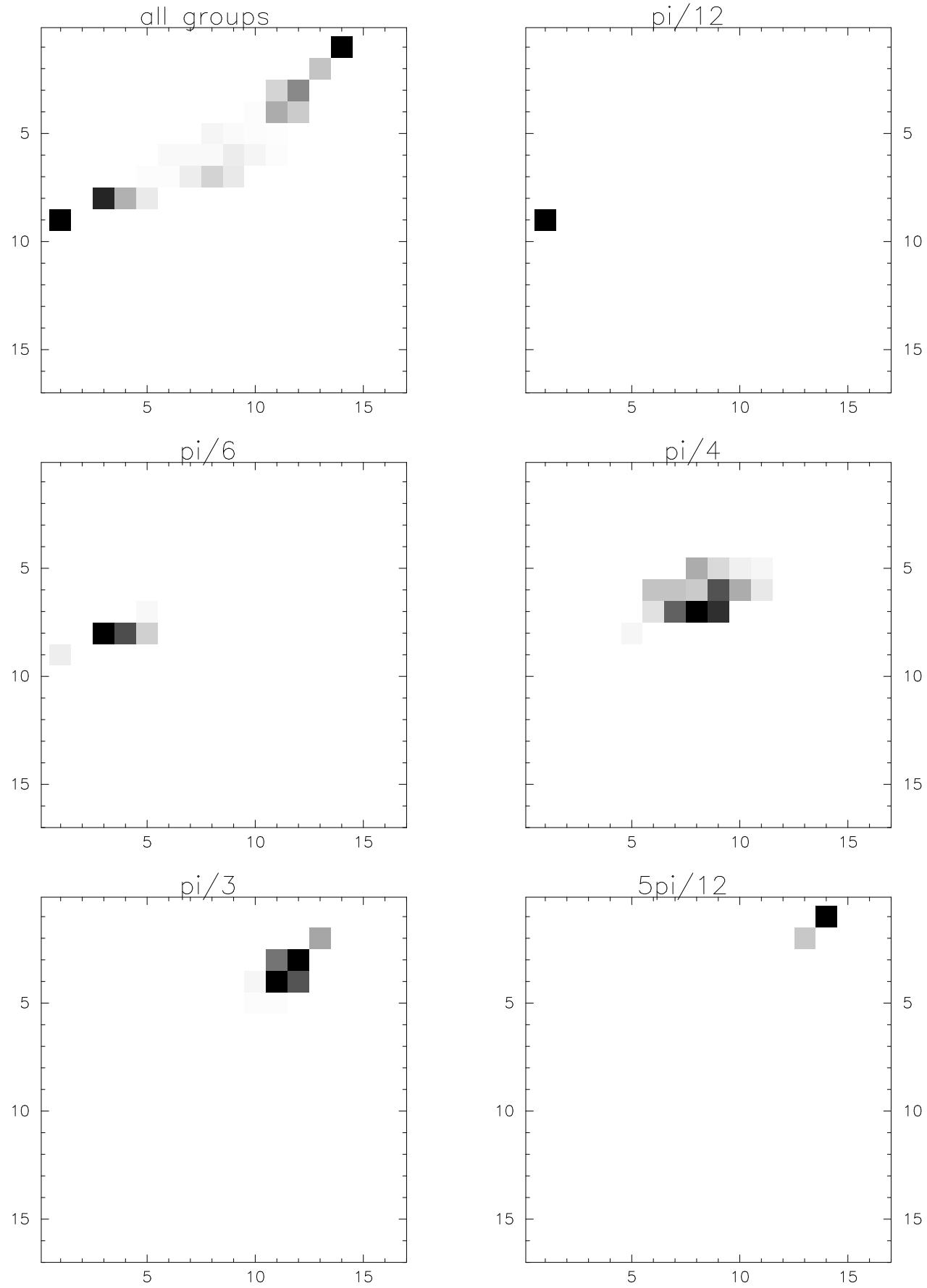

FIG. 1.—Mapping of the synthetic data set onto its SOM. Numbers along the axes represent position in the map, not values of any of the four dimensions of the data.

centrated close to the vertices of the triangular region, implying a steeper gradient in the parameters associated with vectors near the vertices. The mapping is therefore not completely topologically faithful. The numbers along the axes represent positions on the map, not values of θ or any other parameters.

5. GALAXY DISTRIBUTIONS IN MORPHOLOGY SPACE

As a preliminary step, principal components analysis (PCA) of the data set was performed in order to represent as much of the variance in the data as possible by replacing the original axes by *linear* combinations of the original axes. However, the first principal component (PC) only spans 48% of the variance, and the first two PCs span only 71% of it. PCA is therefore inadequate for mapping these data in two dimensions, and a nonlinear method is indeed required.

5.1. *Mapping Galaxy Populations*

We next proceeded to analyze the sample of 2934 galaxies with the SOM software. The best resulting map (in terms of the χ^2 between the data and the nodes to which they were mapped) is shown in Figure 2. Shading progresses from light for low population levels to dark for highly populated regions. Although only the vertices of the triangular map were initialized with vectors corresponding to actual clusters of data points, the final map is well populated in all nodes. This shows that the SOM training process refines the crude results of the clustering algorithm and brings out finer structure. However, mapping the full data set like this is not very informative without examining the characteristic vectors associated with each node. In Figure 3 we show four panels, each depicting the distribution of values of a single morphological parameter in the SOM. There are apparent trends in the parameter distributions: blobbiness is lowest around the left vertex and grows as one moves right, especially toward the upper right. Center displacement is highest in the top right vertex and decreases toward both of the other vertices. The filling factor generally grows toward the left vertex and somewhat toward the bottom right vertex but then decreases again. The skeleton ratio has the clearest trend, growing strongly as one moves away from the bottom right vertex.

With the help of Figure 3 one can now identify the morphologies associated with the map of Figure 2. The area of the left vertex is populated by smooth, symmetric galaxies with a high filling factor. This description corresponds to the appearance of elliptical galaxies. As one moves right toward the center of the map, two trends become apparent: toward the top right vertex galaxies are much more blobby and increase in asymmetry (higher center displacement). The filling factor drops but the skeleton ratio is high, so this region should correspond to images with a great deal of elongated structure, such as spiral galaxies or galaxies with tidal tails. Toward the bottom right the skeleton ratio drops sharply, while the values of the filling factor and the center displacement do not have clear trends. These results imply galaxies of generally "knotty" appearance, some of which are very asymmetric with a lot of apparent structure, while others are less asymmetric and exhibit less structure. These morphologies largely correspond to peculiar galaxies.

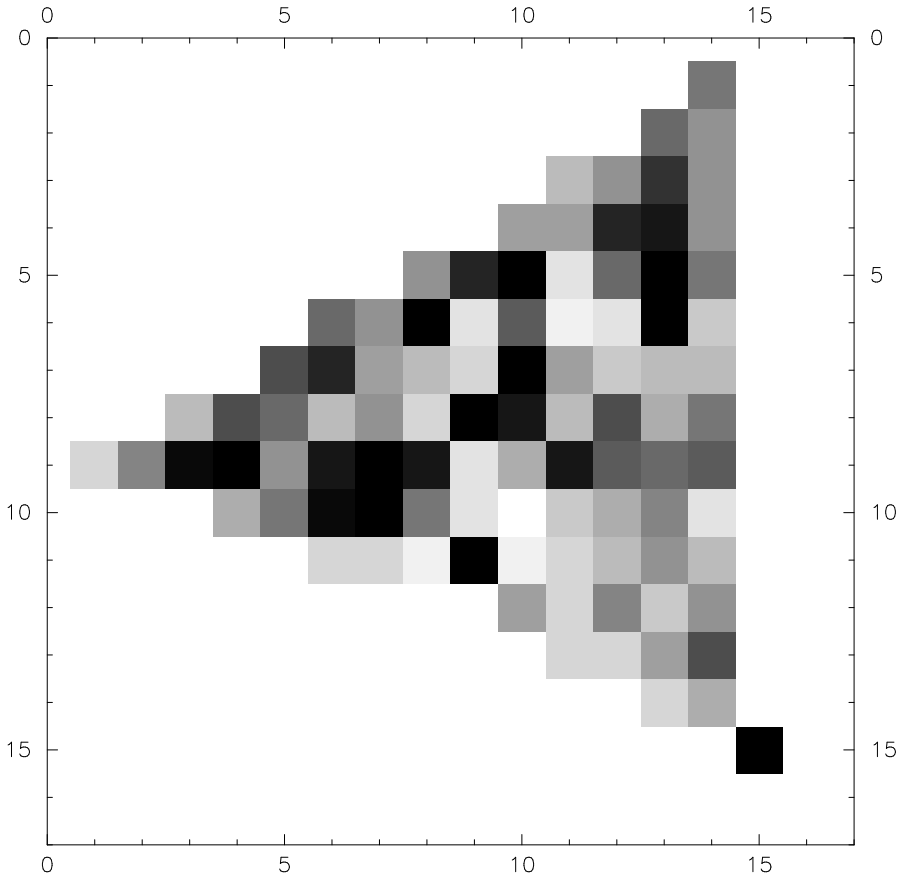

FIG. 2.—Mapping of the full sample of 2934 galaxies onto its trained SOM. Darker color indicates a more populated node. Numbers along the axes denote position in the map, not values of morphological parameters.

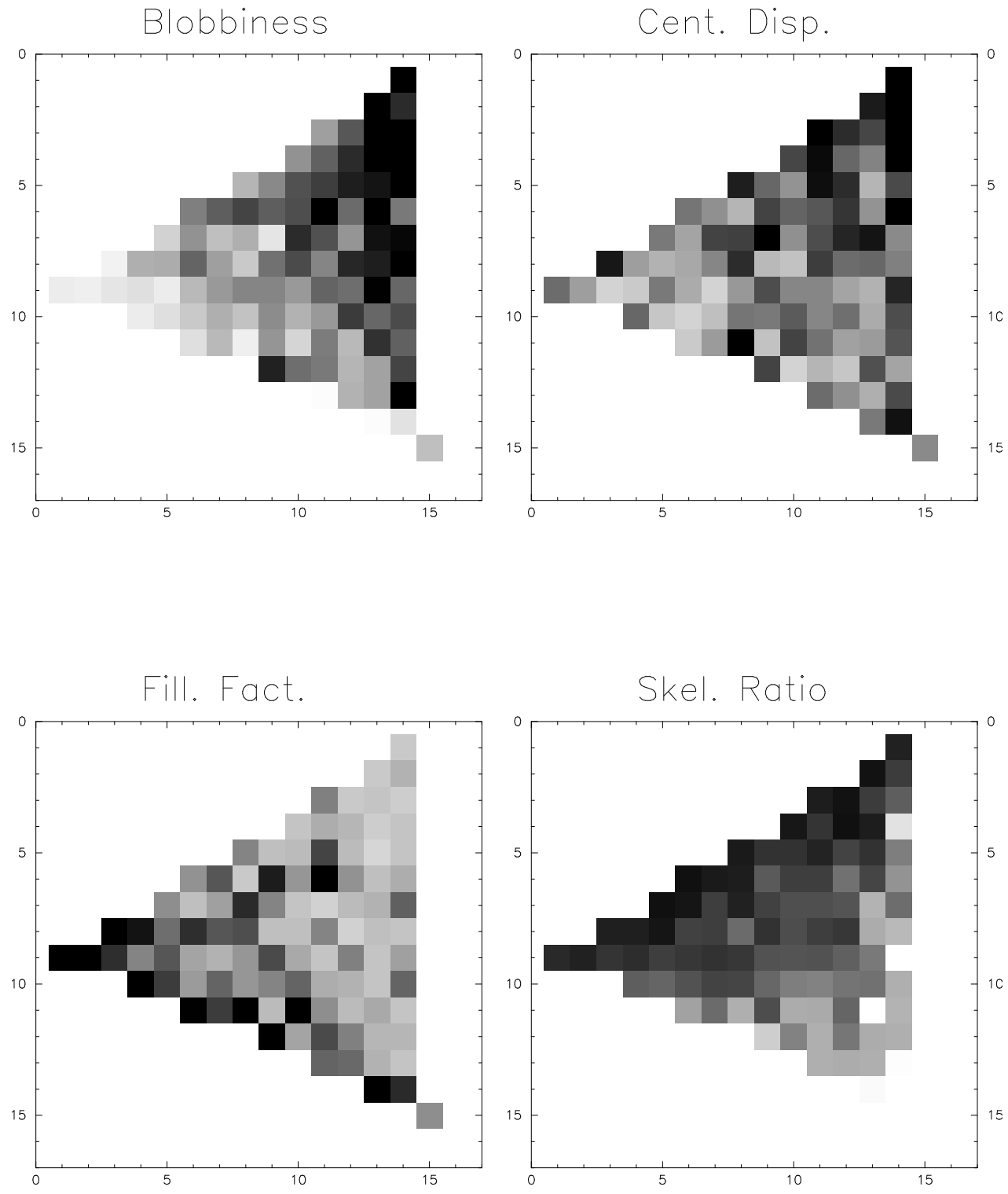

FIG. 3.—Distributions of parameter values in the trained SOM. *Top left*, blobbiness; *bottom left*, isophotal filling factor; *top right*, isophotal center displacement; *bottom right*, skeleton ratio. The darker the shade, the higher the value of the parameter.

In order to verify the above interpretations and to study how different properties of galaxies correspond to morphology, we defined subsets of our sample according to several criteria and mapped these subsets onto the trained SOM. In Figure 4 each panel shows the mapping of one subset, *normalized to the total size of that subset*. This means that the intensities are relative within each panel and should not be directly compared between panels. The panels in the bottom row depict populations selected by eyeball classification. Such classifications were made by one of us (A. N.; see Naim et al. 1997) for roughly one-third of the entire data set as a preparation to using *supervised learning* for these galaxies. Elliptical and S0 galaxies are depicted in the left panel, spirals in the middle panel, and peculiars in the right panel. The locations of these subsets on the map match what we expect from analysis of the characteristic vectors

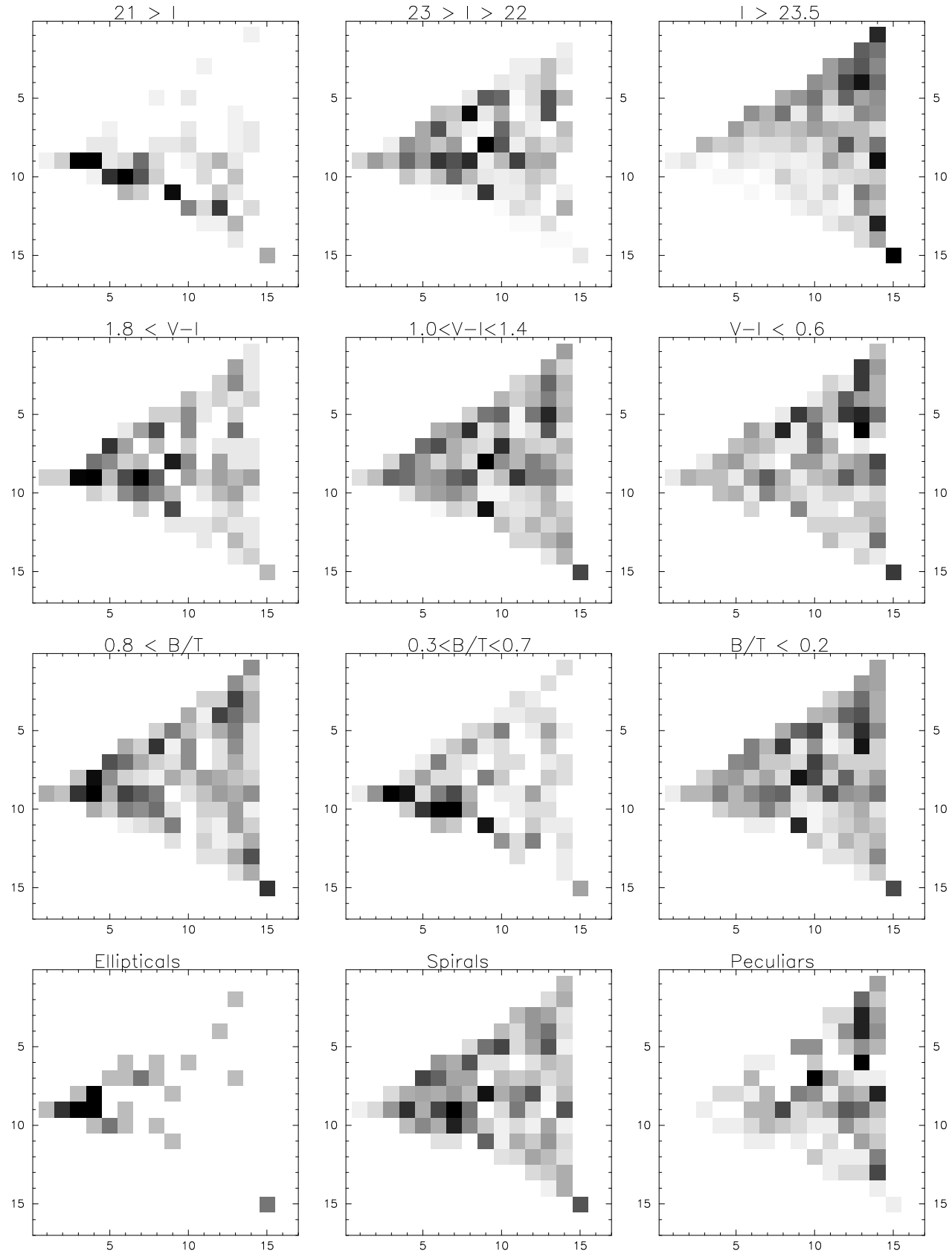

FIG. 4.—Mapping of subsets of the sample onto the trained SOM. *Top row*, subsets selected by apparent *I* magnitude; *second row*, those selected by color; *third row*, by bulge dominance; *bottom row*, by eyeball classification. Refer to Fig. 3 for the changes in each of the four parameters as a function of position in the map.

above: elliptical galaxies are mostly concentrated around the left vertex, spirals are well spread out but appear more concentrated toward the center of the map, and peculiars are mostly found in the right-hand side of the map. This morphological sequence generalizes the one-dimensional Hubble sequence into a two-dimensional map. Roughly speaking, the horizontal axis depicts mostly the change in overall smoothness and symmetry of images, while the vertical axis describes the nature and frequency of structure in the images.

The top three panels of Figure 4 depict the distributions of galaxies in three subsets selected by apparent isophotal magnitude. The left panel depicts galaxies brighter than $I = 21$ mag, the middle panel depicts galaxies in the range 22 mag $< I <$ 23 mag, and the right panel contains galaxies fainter than $I = 23.5$ mag. The gaps in the ranges of apparent magnitude shown in these panels are intended to reduce the overlap and accentuate trends, since the distributions form a continuum. The same applies to the panels describing color and bulge dominance below. However, these magnitude ranges were chosen a priori. The magnitude limits represent a compromise between representing the full range of magnitudes and ensuring that no single bin is underpopulated. There are 278 galaxies brighter than $I = 21$ mag, 722 in the range 22 mag $< I <$ 23 mag, and 946 fainter than $I = 23.5$ mag. The shift in concentration of galaxies with apparent magnitude is evident. Since the redshift distribution of galaxies is a function of apparent magnitude, these three panels may describe, in a statistical way, the evolution of galaxy morphologies with redshift. Verifying this would require many spectroscopic redshifts, though, and work is in progress along these lines (A. Naim et al. 1997, in preparation). The trend we see here is clear: at the bright end the smooth, symmetric galaxies are much more prominent than at the faint end.

The panels in the second row depict subsets selected according to the only available color, $V-I$. The left panel contains red galaxies with (isophotal) $V-I > 1.8$. The middle panel contains intermediate-color galaxies ($1.0 < V-I < 1.4$), and the right panel depicts blue galaxies ($V-I < 0.6$). Color appears to follow morphology, albeit with significant scatter. There is a trend that blue galaxies occupy the upper half of the right side of the map. The panels in the third row describe subsets selected by bulge-to-total ratio, defined as the light contribution of the bulge component divided by the combined contributions of the bulge and disk components. This ratio is calculated from the maximum likelihood fits of bulge and/or disk models to the galaxy image (Ratnatunga et al. 1997, in preparation, contains many details about the subtleties of these fits). The left panel describes bulge-dominated galaxies (B/T > 0.8), the middle panel describes intermediate cases ($0.3 <$ B/T < 0.7) and the right panel depicts disk-dominated galaxies (B/T < 0.2). Interestingly, the bulge-dominated galaxies appear less concentrated in the right-hand side than the intermediate cases. We verify this impression in Figure 5, where we show the mean positions of five subsets, selected by B/T ratios, on the trained SOM. The scatter around these means is considerable, but there is nevertheless a general trend of leftward movement with increasing B/T ratio, which is reversed by the last subset. This is an indication of a change in morphology among bulge-dominated galaxies. Closer examination of Figure 4 confirms that galaxies with B/T > 0.8 cluster in two regions,

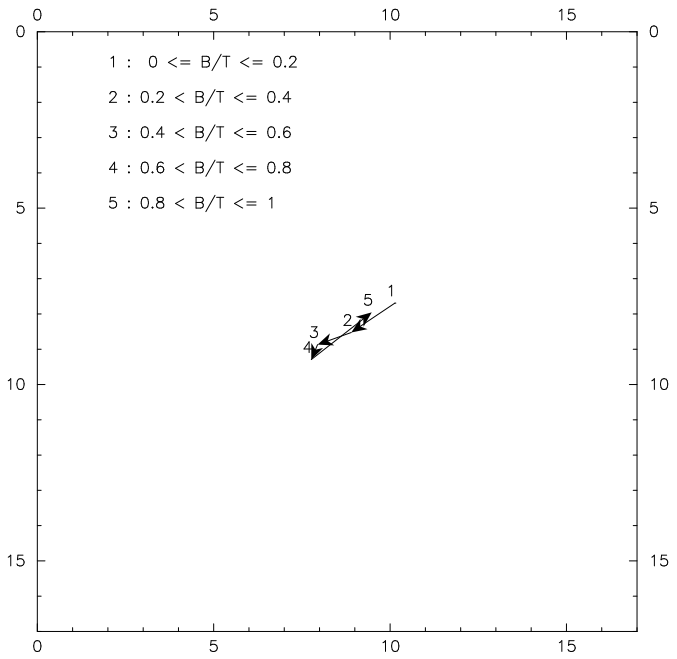

FIG. 5.—Mean positions of subsets selected by B/T ratio in the trained SOM. The scale is the same as in Fig. 2. The trend set by the subsets up to B/T of 0.8 is reversed by the $0.8 <$ B/T ≤ 1 subset, indicating the existence of bulge-dominated galaxies with blobby, asymmetric morphologies. See Fig. 3 for the changes in each of the four parameters as a function of position in the map.

one corresponding to smooth, symmetric morphologies, and one corresponding to blobby and asymmetric morphologies. This latter population has already been noted (Naim et al. 1997). It may correspond to the "blue nucleated galaxies" found in the Canada-France Redshift Survey (Schade et al. 1995), although verifying this point would require further work.

5.2. *The Effect of Noise*

One possible source of the apparent correlation between blobbiness and asymmetry of images on one hand, and apparent magnitude on the other, is that, as one looks at fainter and fainter images, noise sets in and changes the appearance of the images. To investigate this possibility, we show, in Figure 6, how galaxies of high integrated signal-to-noise index are mapped on the same trained SOM used previously. While the full sample contains galaxies that virtually all have $\nu > 100$, the subset shown in Figure 6 was selected to have $\nu > 500$. The fraction of bright galaxies in this subset is naturally higher than in the full sample, so it is difficult to completely decouple the effect of reducing the noise from that of selecting brighter galaxies. Nevertheless, while the concentration of blobby, asymmetric galaxies appears less prominent in Figure 6, it still denotes a significant population. Had that population appeared only because of noise, it should have disappeared in this figure completely. We thus conclude that blobby, asymmetric galaxies indeed exist and that their numbers do increase as one looks at fainter and fainter magnitudes.

We turn back to Figure 3 now, in order to see how noise might have affected the evaluation of our parameters. The panels describing the distributions of blobbiness, isophotal filling factor, and isophotal center displacement show the trends one would expect, although finer details are also

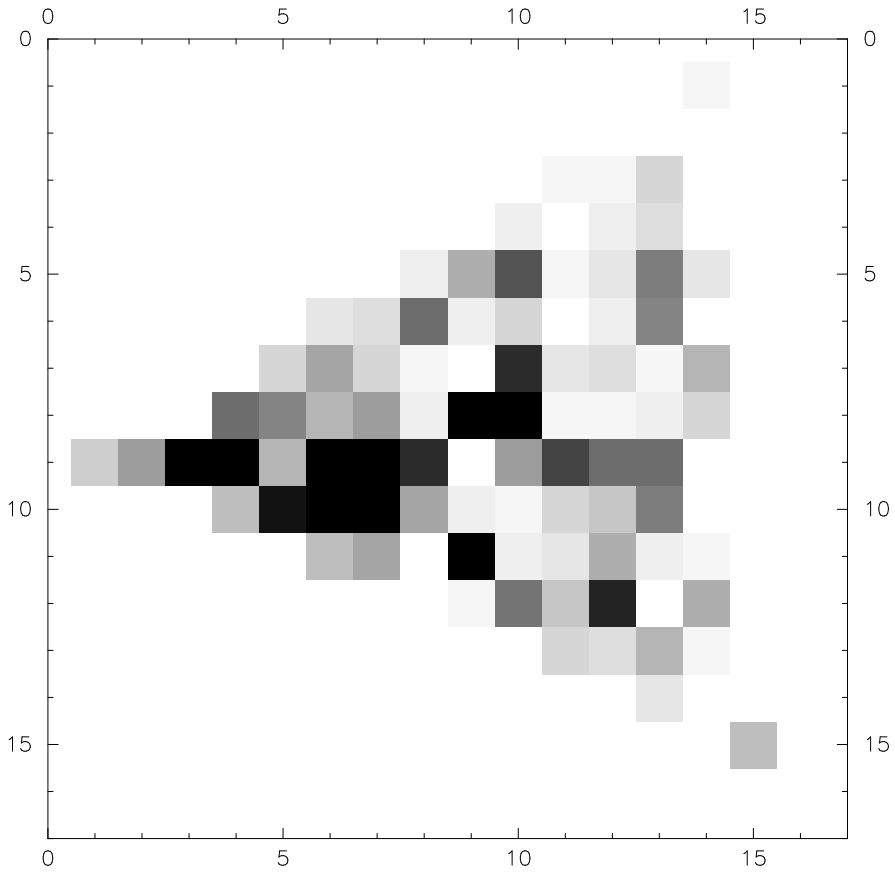

FIG. 6.—Mapping of images with high signal-to-noise ratio ($\nu > 500$) onto the trained SOM. See also Fig. 2.

evident, allowing one or more parameters to vary slightly from one node to the next. The one problematic parameter is the skeleton ratio: while the map shows the expected small values in the region of the peculiar galaxies (because of nearly round star-forming regions) and higher values in the region occupied by the spirals (because of elongated arms), the values are disturbingly high for ellipticals, for which one would expect no features at all (and consequently a value of zero for the skeleton ratio). We note that unlike the other three parameters, which were evaluated from the raw images of the galaxies, the skeleton ratio is measured from the residual images, left after the best-fit photometric model had been subtracted. The skeleton ratio is measured for features that stand out relative to the residual image, and when the residual contains no real features (e.g., in an elliptical galaxy), noise may result in the "detection" of spurious structure. We suspect that this is the source of the relatively high skeleton ratio that characterizes nodes in the region occupied by ellipticals, but further work is needed in order to verify that these features are not real. Luckily, this effect appears to influence most of the bulge-dominated, featureless galaxies in the same way, thus not disturbing their clustering properties. On the other hand, the skeleton ratio is very useful in distinguishing spirals from peculiars, and should not be discarded.

6. DISCUSSION

It has always been important to examine individual galaxies in detail and study the processes dictating their appearance. However, for the fuller picture of galaxy evolution one must employ statistical analysis. One must find quantitative parameters that capture the diversity of galaxy morphologies, while not becoming too specialized or numerous. Here we continue to use the set of four parameters introduced in a previous paper (Naim et al. 1997). However, unlike in that work, our aim here is to analyze the data in an unsupervised way in order to learn new things. One serious difficulty that arises with even a modest number of parameters is that of visualizing data in more than three dimensions. We therefore make use of our variant of the Kohonen SOM, which allows one to cast a distribution in several dimensions into a two-dimensional map. Our algorithm is not necessarily the best for this purpose, and other algorithms exist. Using SOMs allows us to visualize the distributions of galaxies and point out interesting populations for further study. In this respect the SOM is a diagnostic tool, facilitating the first step that needs to be taken with any kind of data: looking at it.

We examine the SOM on a synthetic data set and confirm its ability to perform nonlinear mapping while maintaining the correct topological order of the higher dimension space. We then apply it to our *HST* galaxy sample. In the resulting SOM, galaxies cluster in several groups in morphology space. We confirm the picture that emerged from previous work (Glazebrook et al. 1995; Driver et al. 1995; Abraham et al. 1996), according to which the galaxy population becomes more and more dominated by blobby, asymmetric morphologies as one examines fainter and fainter galaxies. Further, we show that the colors of galaxies at moderate redshifts become significantly

bluer. This could be partly due to the shift in rest-frame band as one goes to higher redshifts, but actual measured redshifts are needed in order to evaluate how much this effect contributes to the trend we see in the SOM. Bulge dominance also appears related to morphology, the blobby galaxies being more disk dominated. However, a population of bulge-dominated galaxies with high blobbiness and asymmetry that was observed by supervised classification (Naim et al. 1997) is rediscovered here in an independent way. Note that this result is achieved without any recourse to eyeball classification, i.e., the existence of the population can be inferred without suspecting it from the outset. Bulge-dominated galaxies with blue colors were also found by Koo et al. (1996), and some of them exhibit peculiar morphologies (e.g., "knots"). That study was limited to a small number of galaxies, and therefore no statistical conclusions can be drawn regarding the bulge-dominated peculiar galaxies. Pascarelle et al. (1996) reported the discovery of compact (half-light radius $\sim 0''.1$) blue objects that are apparently subgalactic clumps. It is possible that these clumps, once assembled closer together, give rise to the bulge-dominated peculiars that we identify in our sample, although this is by no means certain. Alternatively, bulge-dominated peculiars may be older galaxies caught in the process of merging with dwarf companions. We have no way of telling with current data.

Noise becomes progressively more important as one considers fainter images, but our analysis shows that it cannot fully account for the trends we detect. The skeleton-ratio parameter is most affected by noise in smooth, symmetric images, but does not significantly bias the clustering properties of that population as a whole and is still very useful in separating two other populations (corresponding to the eyeball classes of spirals and peculiars). An improved version of this parameter may give better results, though. K-corrections are also of great importance at redshifts of order unity, as discussed, e.g., by Odewahn et al. (1996). However, we have not studied their effect on our parameters in this paper, because we only have two filters for the data presented here (I and V). A study into the effect of the filters used on the measured morphological parameters is currently under way, using MDS fields that were taken in three filters (BVI). Any effect the K-corrections may have on our parameters will, of course, influence the resulting SOM.

To summarize, since morphological classification has become too refined, we adopt an approach that utilizes morphology without any classification. The SOM succeeds in mapping different morphologies to different regions of the map, and we are encouraged by the apparent correlation of morphology with other quantities, such as color and bulge dominance. These correlations allow us to use morphology as a selection criterion for further studies of specific populations (e.g., mergers). However, understanding galactic evolution requires the addition of more physical information, such as rotation curves and full spectral analysis. In this paper we propose a framework into which such information could be incorporated once it becomes available. Our hope is that this modest first step will eventually lead to the emergence of an overall scheme incorporating most aspects of galaxy formation and evolution.

We would like to thank Ofer Lahav, Jens Hjorth, Bob Abraham, and Richard Ellis for raising important points regarding SOMs and morphology and its implications. As always, Eric Ostrander's contribution to the MDS pipeline was invaluable. We also thank the referee for a thorough report. This research was supported by funding from the *HST* Medium Deep Survey under GO grants p2684 and following.

REFERENCES

Abraham, R. G., van den Bergh, S., Glazebrook, K., Ellis, R. S., Santiago, B. X., Surma, P., & Griffiths, R. E. 1996, ApJS, 107, 1
Cowie, L. L., Hu, E. M., & Songaila, A. 1995, AJ, 110, 1576
de Vaucouleurs, G. 1959, in Handbuch der Physik, Vol. 53, ed. S. Flügge (Berlin: Springer), 275
Driver, S. P., Windhorst, R. A., & Griffiths, R. E. 1995, ApJ, 453, 48
Glazebrook, K., Ellis, R. S., Santiago, B. X., & Griffiths, R. E. 1995, MNRAS, 275, L19
Griffiths, R. E., et al. 1994, ApJ, 435, L19
Groth, E. J., Kristian, J. A., Lynds, R., O'Neil, E. J., Balsano, R., & Rhodes, J. 1994, BAAS, 26, 1403
Hubble, E. 1936, The Realm of Nebulae (New Haven: Yale Univ. Press)
Kohonen, T. 1989, Self-Organization and Associative Memory (3d ed.; Berlin: Springer)
Koo, D. C., et al. 1996, ApJ, 469, 535
Mähönen, P. H., & Hakala, P. J. 1995, ApJ, 452, L77
Naim, A., Ratnatunga, K. U., & Griffiths, R. E. 1997, ApJ, 476, 510
Odewahn, S. C., Windhorst, R. A., Driver, S. P., & Keel, W. C. 1996, ApJ, 472, L13
Pascarelle, S. M., Windhorst, R. A., Driver, S. P., Ostrander, E. J., & Keel, W. C. 1996, ApJ, 456, L21
Roberts, M. S., & Haynes, M. P. 1994, ARA&A, 32, 115
Sandage, A. R. 1961, The Hubble Atlas of Galaxies (Washington: Carnegie Inst. Washington)
Schade, D., Lilly, S. J., Crampton, D., Hammer, F., Le Fèvre, O., & Tresse, L. 1995, ApJ, 451, L1
van den Bergh, S. 1960, ApJ, 131, 215
———. 1976, ApJ, 206, 883
van den Bergh, S., Abraham, R. G., Ellis, R. S., Tanvir, N. R., Santiago, B. X., & Glazebrook, K. 1996, AJ, 112, 359