The Astrophysical Journal, 568:396–407, 2002 March 20

STATISTICAL PATTERN RECOGNITION FOR LABELING SOLAR ACTIVE REGIONS: APPLICATION TO *SOHO*/MDI IMAGERY

M. Turmon
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109

J. M. Pap[1]
Department of Physics and Astronomy, 405 Hilgard Avenue, University of California, Los Angeles, CA 90095

and

S. Mukhtar
Computation and Neuroscience Program, California Institute of Technology, Pasadena, CA 91125
Received 2001 May 22; accepted 2001 November 14

ABSTRACT

This paper presents a new application of statistical methods for identifying the various surface structures on the Sun that may contribute to observed changes in total and spectral solar irradiance. These structures are divided for our purposes into three types: quiet Sun, faculae, and sunspots (umbra and penumbra). Each region type is characterized by the observed data present at pixels of that type. Statistical models characterizing these observables are found from expert identification of a sample set of regions or unsupervised clustering. Information about the spatial continuity of regions is incorporated into the model via a prior distribution on the label image; the contribution of the prior can be interpreted as a regularizing term. Once the parameters defining the models are fixed, the inference procedure becomes to maximize the probability of an image labeling given the observed data. This allows objective and automated classification of a large set of images. We describe the application of these procedures to computing labelings from synchronized full-disk high-resolution magnetic-field and light-intensity maps from the Michelson Doppler Imager experiment on the *Solar and Heliospheric Observatory*.

Subject headings: methods: analytical — methods: statistical — Sun: activity — Sun: faculae, plages — sunspots

1. INTRODUCTION

Study of the Sun's variability has been of high importance for both astrophysics and solar-terrestrial physics. The Sun, a fairly typical star, has the special advantage of proximity, which allows the detailed study of a variety of phenomena important for stellar physics. High-precision photometric observations of solar-type stars clearly show that year-to-year brightness variations connected with magnetic activity are a widespread phenomenon among such stars (e.g., Radick 1994). Space-borne irradiance observations over the last 2 decades have demonstrated that solar irradiance, both bolometric and at various wavelengths, varies during the course of the 11 year solar cycle (Fröhlich 1998). Since solar energy sustains the life on Earth and is the ultimate driving force for terrestrial climate, it is inescapable that we must understand why, how, and on what timescale the solar irradiance varies to better understand the role of solar variability in climatic changes.

Analyses based on 2 decade long space irradiance measurements have revealed that the surface manifestations of solar magnetic activity play an important role in solar irradiance changes (Lean et al. 1998). On the other hand, several studies have shown that current irradiance models, solely based on the effect of surface magnetic activity, cannot explain all the aspects of irradiance changes (Kuhn 1996; Fröhlich et al. 1997; Pap 1997). Unfortunately, identification of the cause of this residual variability is a difficult problem. Hints from helioseismology (Kuhn et al. 1998) and from precise photometry by the Variability of Irradiance and Gravity Oscillations (VIRGO) and Michelson Doppler Interferometer (MDI) experiments on the *Solar and Heliospheric Observatory* (*SOHO*) indicate that global effects—changes in the photospheric temperature, large-scale mixing flows or convective cells, and radius fluctuations, can all produce changes in solar irradiance. Nonetheless, it is essential to clarify to what degree the observed irradiance changes are related to surface and global effects, respectively. While studying the role of global effects in irradiance changes is not an easy and straightforward task, there are thousands of solar images whose analysis enables us to study the contribution of surface magnetic activity to irradiance changes in great detail. To understand the physical causes of the changes observed in total and spectral solar irradiances, it is necessary to study the spatial characteristics and temporal evolution of the solar magnetic fields and related thermal structures in the various layers of the solar atmosphere.

One of the largest obstacles to the use of this body of solar imagery is the amount of time and effort required to analyze it carefully. Simple manual cataloging techniques do not scale to years-long time intervals: for example, the study reported here covers the interval from 1996 July to 1997 September and uses about 10^4 solar images. In order to speed up the scientific investigation process, we have developed a system for the automated processing and analysis of various images available from space and the ground. Its core is a Bayesian image-segmentation technique driven by statistical models trained from scientist-provided image

[1] Also at Goddard Earth Sciences and Technology Center, NASA Goddard Space Flight Center, Greenbelt, MD 20771.

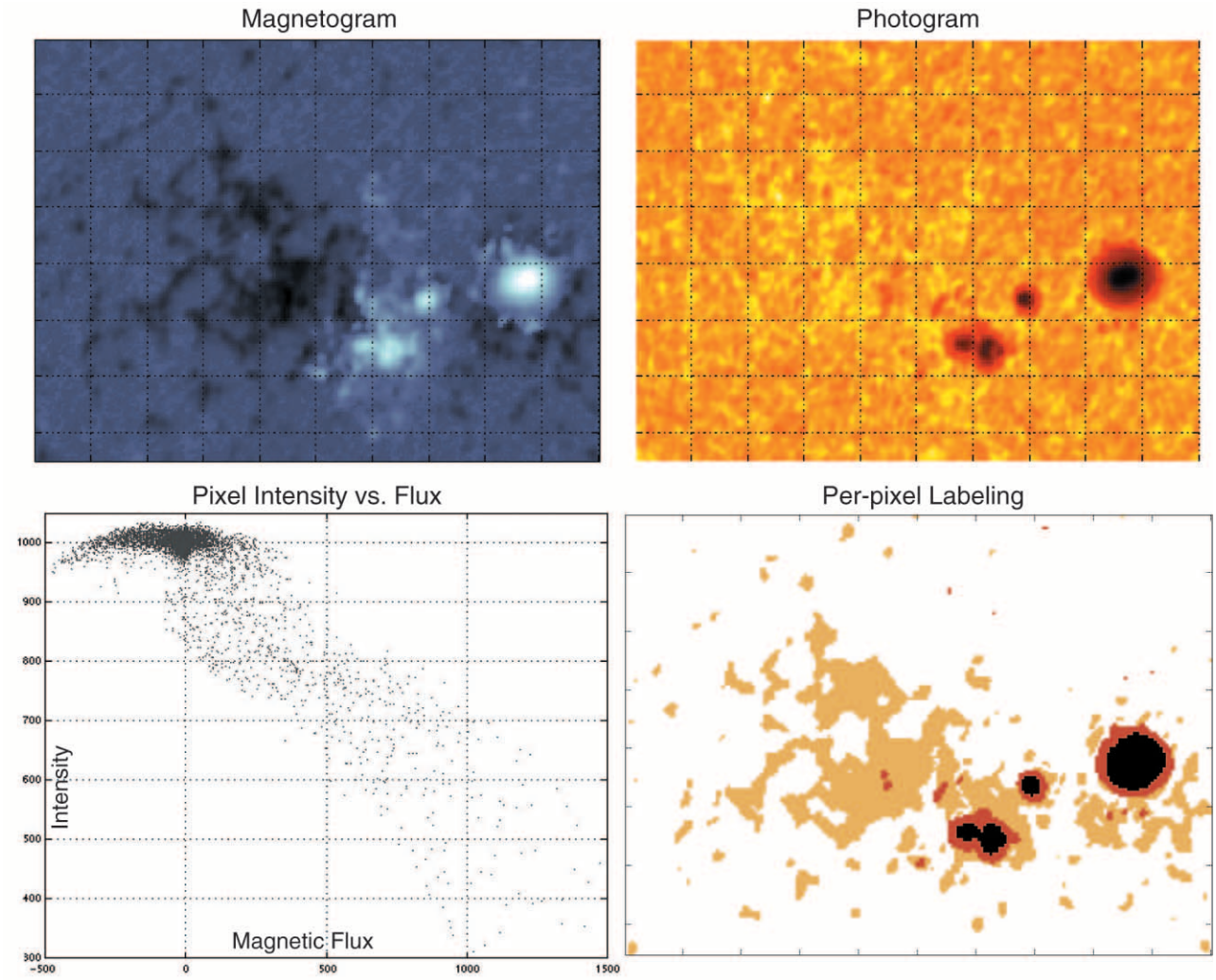

FIG. 1.—*Top*: Near-simultaneous magnetogram and photogram, with some contrast enhancement. *Bottom*: Corresponding scatter plot and rough per-pixel labeling.

labelings. Once these models are selected, labeling proceeds automatically. Models, written down in a portable definition format, can be refined over time or exchanged between observation programs. These models allow the controlled introduction of physical knowledge of the characteristics of the activity types of interest (Turmon & Pap 1997). While we have used the *SOHO*/MDI images as a testing ground for these methods, the fundamental approach described here accommodates other data sources (e.g., Ca II K images) and, in particular, allows the use of more wavelengths without modification.

Several groups are conducting work along similar lines. Bratsolis & Sigelle (1998) use similar computational machinery for smoothing image labelings but do not directly link image labels to physical classes. A major practical benefit offered by the methods we advocate is the ability to integrate the information from several images; much region identification work to date has proceeded from just one image

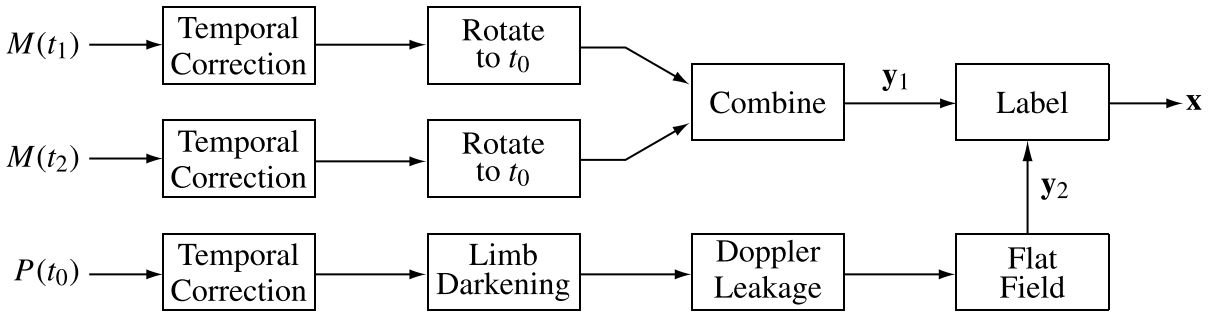

FIG. 2.—Data flow used to preprocess photograms and magnetograms in preparation for feature identification

source. Worden, White, & Woods (1998) produce labelings from Ca II K spectroheliograms by taking account of spatial structures such as contiguity and filling factors; the latter is similar to our idea of a "dominant process" (see § 4). Harvey & White (1999) find spatial structures in magnetograms, and their analysis links magnetic and intensity features, although their labelings do not do so explicitly. Fligge, Solanki, & Unruh (2000) also consider region identification based on MDI magnetograms. Several observatories have well-calibrated and regular image-analysis operations (see Walton et al. 1998; Steinegger, Bonet, & Vasquez 1997; Steinegger et al. 1998).

We note that whereas our earlier work (Pap et al. 1997) used only magnetic-field observations from MDI, the present study uses both full-disk (1024^2 pixels) magnetograms, taken approximately every 96 minutes, and quasi white-light photograms, which are taken roughly once every 360 minutes. For our purposes, it is important to distinguish further between the 1 and 5 minute magnetogram integration times used by MDI, as these produce images having different noise characteristics. All these images are taken by MDI with a CCD camera near the Ni I 676.8 nm absorption line originating in the midphotosphere (Scherrer et al. 1995). We use the level 1.5 MDI imagery throughout; magnetic strengths are given in units of MDI nominal Gauss, while photograms are given in arbitrary units referenced to a quiet-Sun level of 1000.

We first describe this image processing technique and its application to the MDI photograms and magnetograms in detail. This must, of necessity, include the preprocessing steps that remove certain persistent instrumental artifacts from the magnetograms and photograms. The centerpiece of this paper is the description of the statistical machinery needed to find these image structures. We conclude by setting out how this machinery is tuned for the MDI imagery and giving sample results illustrating the characteristics of the derived labelings. In a companion paper, we compare the time series of the various activity components with the *SOHO*/VIRGO total and spectral irradiances in the near-UV at 402 nm, visible at 500 nm, and near-IR at 862 nm. In addition to the VIRGO data, we will also study the effect of various activity components on the EUV and XUV measured by *SOHO*'s Charges, Elements and Isotopes Analysis System (CELIAS) and Solar EUV Monitor (SEM), and also on the Mg II h and k core-to-wing ratio from the Solar Ultraviolet Spectral Irradiance Monitor (SUSIM) aboard the *Upper Atmosphere Research Satellite* (*UARS*).

2. BASIS FOR IMAGE SEGMENTATION

Figure 1 illustrates the potential of the MDI data for extracting active regions from solar images. The top panels show details from a 1 minute magnetogram and a flat-fielded photogram that happened to be taken 6.0 minutes apart by MDI at 4:15 UT on 1996 August 1. (A separation of 6 minutes implies a relative displacement of less than half a pixel at disk center.) The activity shows clearly on the corresponding scatter plot (*bottom left panel*); this plot is of fundamental interest because the statistical models we build will model the probability density function in this *feature vector* representation. In fact, as demonstrated in Figure 1, the umbra and penumbra can be separated at the knee of the two-dimensional scatter plot (see especially the bottom right panel). We also note that magnetic fields of about ± 200 G may produce either sunspots or faculae; a two-parameter observation is therefore crucial to reliably separate these features. The crude labeling in the bottom right panel corresponds to a simple thresholding of these bivariate data, ignoring all spatial relationships. (We use per-pixel thresholding for inspection here but do not advocate it as a general labeling technique.) The spatial coherence of the resolved structures is apparent.

Generally speaking, these images show a wide variety of solar structures: active regions (sunspots and faculae), remnants of active regions, and the active network and the relatively quiet network that are distributed as cell-like structures over the solar disk. For the purposes of this analysis, we have concentrated on three specific structures: sunspot umbra and penumbra, faculae, and background (the so-called quiet Sun/network). These structures are relatively easy to identify directly in sample images, which facilitates model selection. Decomposition into more classes could be more informative about solar processes, but only to the extent that the extra classes indeed represent distinct physical processes.

Inferring structural properties from these observations requires a uniform, automated technique whose parameters have been determined objectively to the greatest extent possible. We can formalize the procedure illustrated above—in particular, generalizing to the case of nonsynchronous magnetograms and photograms—as in Figure 2:

1. Temporally correct and flat field a photogram.
2. Interpolate a magnetogram to the photogram observation time.
3. Infer the labeling from the magnetogram and photogram.

Before going on to detail these procedures, we introduce some notation. Images are collected by sampling in a time t and in a spatial variable $s = [s_1, s_2]$ indexing the image plane; the set of all such discrete spatial sites is S. A single pixel's observable feature vector is $y = [y_1, \ldots, y_d] \in R^d$; such vectors are grouped into an image $\boldsymbol{y}$ indexed by s. A decomposition of an image into K classes is captured by defining a labeling $\boldsymbol{x}$ of integers in $\{1, \ldots, K\}$ for each image pixel $s \in S$. When discussing the MDI imagery, the magnetic field observable is generically denoted y_1, while the intensity is y_2; these are seen as full images (magnetograms and photograms) $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. We use an observer-centered frame in which the s_1 coordinate increases toward solar east and s_2 increases toward solar north.

3. PREPROCESSING AND SYNCHRONIZATION

We begin with a photogram at a given time t_0. Ensuring the immunity of all preprocessing steps to changing solar activity is paramount due to their use in irradiance study. This temporal stability is crucial to eliminating side effects from the final region maps. Where possible, we have therefore tabulated smoothly varying correction factors across the interval under study (1996 July–1997 September) rather than using a fully image-by-image system.

First, we note that instrument throughput drops linearly with time and exhibits jumps due to instrument focus, orientation, and configuration (Bogart, Bush, & Wolfson 1998). These effects are partially removed from the level 1.5 data that we use; the remainder are taken out by dividing the whole image by a nominal throughput. This factor is pre-

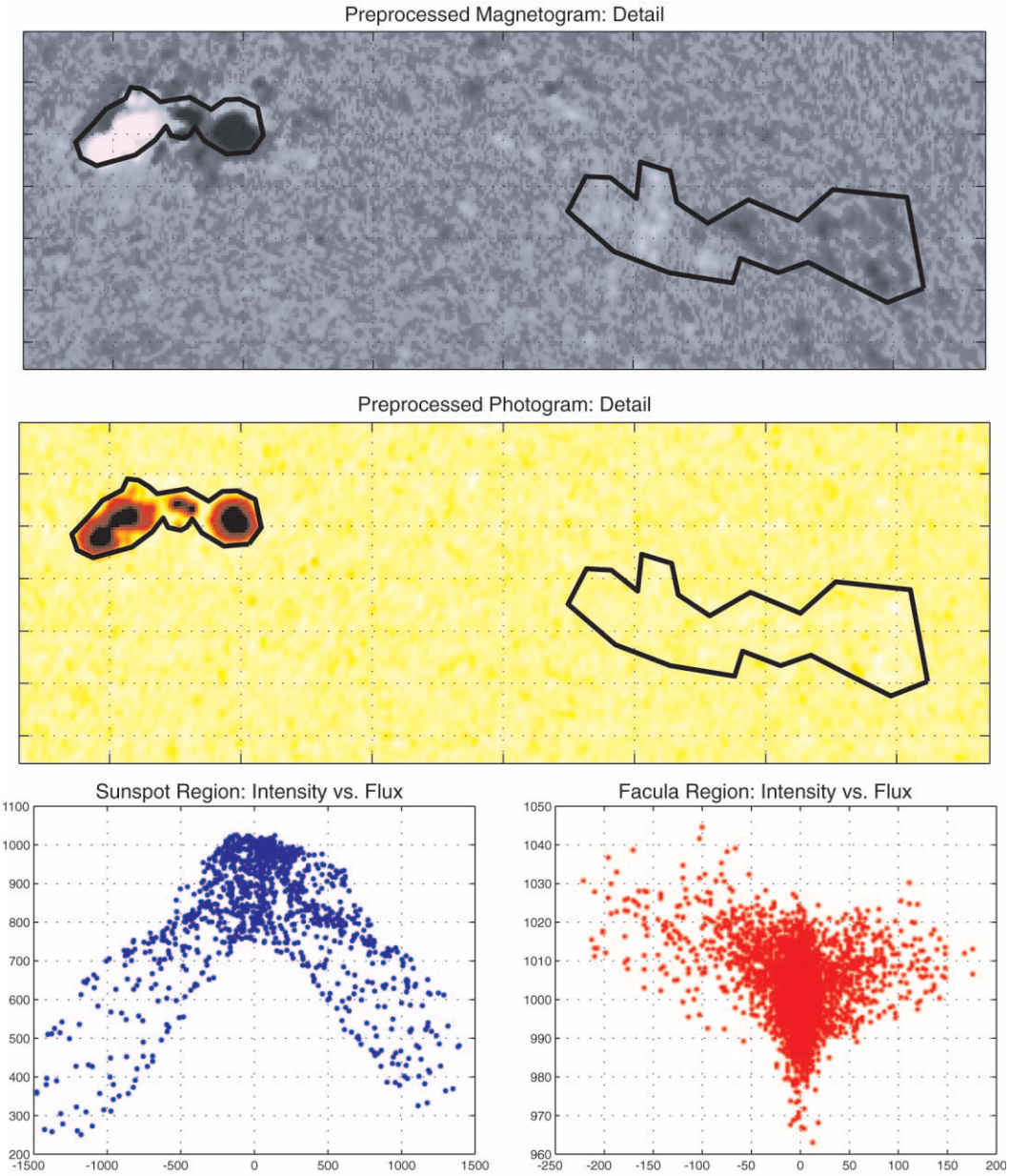

FIG. 3.—Preprocessed magnetogram and photogram from 1997 September 7, with scatter plots of sites within the two indicated regions

dominantly linear in time, with breaks at several points when the instrument was recalibrated. The factor is determined separately for each image by taking the median of 448 medians within 32×32 pixel blocks; this level is defined as the quiet Sun.

In removing center-to-limb variation, it becomes important to take into account the slight ellipticity of the MDI photograms due to instrumental effects (e.g., Kuhn et al. 1998). We do this by defining an effective radius from the image center $s_\odot$ via $r_{\rm eff}^2 = (s - s_\odot)^T R_\theta^T D R_\theta (s - s_\odot)$, where

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

$$D = \begin{bmatrix} 0.9981 & 0 \\ 0 & 1.0000 \end{bmatrix},$$

and $\theta = 30\overset{\circ}{.}5$. The ellipticity correction remains constant over the time interval studied. It corresponds to a stretching of the image along a major axis tilted $30\overset{\circ}{.}5$ south of solar west and a peak-to-peak scale variation of $1\overset{''}{.}0$ around the solar limb; this agrees with Figure 1 in the paper of Kuhn et al. (1998). This effective radius, at most $R_\odot$, is used to find $\mu = [1 - (r_{\rm eff}/R_\odot)^2]^{1/2}$. Center-to-limb variation is then removed via an adaptation of the median-based procedure of Brandt & Steinegger (1998), yielding a correction factor

$$\mathrm{LD}(\mu) = \begin{cases} 1 + \sum_1^4 \gamma_k (\log\mu)^k & \text{if } \mu \geq 0.05, \\ 0.2 - \mu[0.2 - \mathrm{LD}(0.05)]/0.05 & \text{otherwise,} \end{cases}$$

where

$$\gamma_1 = 0.44814, \quad \gamma_2 = 0.13719,$$
$$\gamma_3 = 0.02119, \quad \gamma_4 = 0.00027.$$

The correction for low μ overrides irregularities in the polynomial there with a linear drop to $\mathrm{LD}(0) = 0.2$. The function agrees with that of Bogart et al. (1998) to within 0.7% for $\mu \geq 0.2$. Both the ellipticity and $\mathrm{LD}(\cdots)$ are stable across the time interval studied here.

Another form of distortion present in these images is the leakage of the Doppler signal into the intensity. This is removed by dividing by a factor $1 + \kappa v_s$ depending on the (spatially varying) relative motion v_s of the Sun and the MDI. The Doppler correction coefficient κ varies slowly over time due to instrument configuration changes; the correct value for a given time is tabulated. Finally, a small residual flat field (accounting for roughly 1.0% spatial variation) is applied. This flat field also changes slowly over time; however, it is found separately for each image using a blocked median filter. The result is a spatially and temporally uniform photogram $\boldsymbol{y}_2$.

This photogram $\boldsymbol{y}_2$ is used with a magnetogram interpolated between two existing magnetograms as follows. First, the nearest bracketing magnetograms are found. (There are two exposure intervals, 1 and 5 minutes; both magnetograms must be of the same type.) These are preprocessed to remove temporal effects: since magnetograms are computed as a difference of channels, they are mostly immune to them. However, temporal normalization is still needed to remove an offset of about -0.4 G that affects 5 minute magnetograms taken in 1997 April–November (J. T. Hoeksema 1999, private communication). Next, standard formulas (Zappalá & Zuccarello 1991) for the latitude-dependent angular velocity of active regions in the photosphere,

$$\Omega = 14.643 - 2.2407 \sin^2 \theta \,\mathrm{deg\,day^{-1}},$$

are used to synthesize a magnetogram at t_0 from each bracketing magnetogram. Sunspot age, which we do not account for, changes Ω by about $\pm 0.^\circ 3\ \mathrm{day}^{-1}$. A 240 minute rotation to find a nearby image (needed by less than 1% of the MDI photograms we labeled) corresponds to a movement of about 18 MDI pixels at disk center. An error in Ω of $0.^\circ 5\ \mathrm{day}^{-1}$ over 240 minutes corresponds to a displacement of about half a pixel at disk center, which is negligible.

The two magnetograms are merged by taking as many pixels as possible from the nearer image and the remaining ones from the more distant one. (Averaging, or weighted averaging, would change the image statistics by lowering noise and is therefore not used in the merge.) The synthesized magnetogram $\boldsymbol{y}_1$ and the flattened photogram $\boldsymbol{y}_2$ are now combined into $\boldsymbol{y}$ and used to infer a labeling. Figure 3 shows results from the interpolation process. The top two panels are details from a magnetogram and a photogram; the photogram was taken at 17:58 UT on 1997 September 7 and preprocessed as outlined above. The magnetogram has been synthesized from bracketing 1 minute magnetograms taken at 6:24 and 19:11 UT on that day. The regions are $255'' \times 725''$. These observation times imply a minimum rotation of 73 minutes; over the time interval studied, 80% of photograms had usable magnetograms closer than this. The scatter plots of y_s for the two indicated subregions are in the bottom panels. (Quantization of the MDI magnetogram is visible in the right panel.) The bipolar spot (*left panel*) shows clearly as having lower intensity, and the penumbra is visible at the knee of the plot as in Figure 1. Most encouraging, the brightness enhancement of the decayed facula (*right panel*), far from the limb at $\mu = 0.81$, is clearly visible, increasing with magnetic flux strength. The maximum brightness enhancement amounts to about 3%–4% of the nominal intensity level. Such results are consistent with a well-calibrated limb-compensation and flat-field mechanism.

4. IMAGE DECOMPOSITION

The final step, inferring the labeling, is more complex because doing so involves scientific judgment. In the system we propose, this judgment is isolated in a falsifiable (Popper 1959) statistical model whose parameters are chosen according to a scientist-provided labeling. This practice isolates the problem-specific elements of the decision procedure into a *concise and testable model* that naturally accounts for noise in the observables and uncertainty in their relation to class labels. This distinguishes the approach described below from strictly "algorithm-based" procedures (e.g., thresholding or region growing), in which the only way to describe the labeling method is via the computer code that is ultimately used and for which reckoning with the uncertainty inherent in noisy data and uncertain labelings takes place away from its natural probabilistic setting. We also note that the framework outlined here extends naturally to any number of image observables.

Because of the abundance of prior information, we adopt a Bayesian view that allows us to use expert knowledge of observed data in the consistent framework referred to above. Following well-established statistical practice (Geman & Geman 1984; Ripley 1988; Turmon & Pap 1997), we treat the labeling step in this Bayesian framework as inference of the underlying pixel classes (symbolic variables represented by small integers) based on the observed (vector-valued) pixel characteristics. The viewpoint is that there is a family of K physical processes, and at any site $s \in S$, exactly one process is dominant, say x_s. In general, the dominance of a given process has a spatial coherence, so that labels tend to form clumps in the image plane. The observable feature vector y_s then depends only on the dominant process at s, not on the process that is active at neighboring sites. This scheme was originally borrowed from statistical physics to model images with occasional sharp transitions in the observable y_s. Direct models based on (for example) a spatially correlated Gaussian process on the observables themselves cannot account for such transitions. Introduction of the hidden label variable x_s, which controls the observable, explains the dramatic shifts by a change of the dominant process at that site.

The posterior probability of labels given the observed data is central in the Bayesian framework. Here, by Bayes rule, this probability is

$$P(\boldsymbol{x}|\boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})/P(\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x}). \qquad (1)$$

The constant of proportionality is unimportant because we are only interested in the behavior of the posterior as the labeling $\boldsymbol{x}$ is varied. We remark that the Bayesian framework allows the computation of many relevant quantities besides just $P(\boldsymbol{x}|\boldsymbol{y})$. We can easily find, for example, the K numbers comprising the posterior probability distribution $P(x_s|\boldsymbol{y})$: a concentrated posterior indicates confidence in the assigned labeling. For now, our goal is to find the optimal labeling. If we are given a reward every time we recover the correct labeling $\boldsymbol{x}$, but nothing for incorrect labelings, it can

be shown that the best strategy is to choose the labeling maximizing equation (1), or equivalently, its logarithm:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} [\log P(\boldsymbol{y}|\boldsymbol{x}) + \log P(\boldsymbol{x})]. \quad (2)$$

This is the well-known maximum a posteriori (MAP) decision rule.

To use the MAP rule, we must specify $P(\boldsymbol{x})$ and $P(\boldsymbol{y}|\boldsymbol{x})$. The assumption made above that the hidden labels drive the observables means that given the controlling variable x_s, y_s is independent of $y_{s'}$ for any $s' \neq s$. Alternatively,

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{s \in S} P(y_s|x_s), \quad (3)$$

i.e., the observables are not coupled when given the labeling.

Prior models $P(\boldsymbol{x})$ can be specified in many ways. We have used the Markov field smoothness priors

$$P(\boldsymbol{x}) = Z^{-1} \exp\left[-\sum_{s' \sim s} \beta(s, s') 1(x_s \neq x_{s'}) \right]. \quad (4)$$

Here the indicator $1(\cdots)$ is 1 if the contained proposition is true and zero otherwise. The site coupling $\beta(\cdots, \cdots) \geq 0$ can account for differing distances between pixels. The constant Z is chosen to normalize the probability mass function, and the sum extends over "neighboring" sites in S. On our rectangular grid, sites are neighbors if they adjoin vertically, horizontally, or diagonally, so each site s has eight neighbors, denoted $N(s)$. Such distributions were originally introduced to model ferromagnets, and this particular formulation is known as the "Potts model" in statistical physics (Wu 1982).

For illustration, consider first the following simplified example: the smoothness $\beta(s, s') = \beta$, a constant, and within each label type $1 \leq k \leq K$, data are conditionally Gaussian, distributed with per-class mean vectors $\mu(k)$ and RMS energy σ. Data in each class are therefore scattered isotropically about the class mean, so that

$$\begin{aligned} P(y_s|x_s) &= N(y_s; \mu(x_s), \sigma^2 I) \\ &= \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left[-\frac{1}{2} \left\| \frac{y_s - \mu(x_s)}{\sigma} \right\|^2 \right], \end{aligned} \quad (5)$$

where $N(y; \mu, \Sigma)$ is the d-dimensional normal density function with the indicated mean and covariance matrix, and $\| \cdots \|$ is the Euclidean norm, so $\|v\|^2 = \sum_{i=1}^{d} v_i^2$. Combining with equation (4), taking logarithms as in equation (2), and discarding terms not involving $\boldsymbol{x}$ yields the objective function

$$-\frac{1}{2} \sum_{s \in S} \|[y_s - \mu(x_s)]/\sigma\|^2 - \beta \sum_{s' \sim s} 1(x_{s'} \neq x_s). \quad (6)$$

The first term is the familiar likelihood function, forcing each label x_s to be such that $\mu(x_s)$ is close to the data y_s. The second, arising from the prior probability of a labeling, penalizes feature maps with many disagreements among neighboring sites, such as those with speckled patterns of activity. Together, they can be interpreted as a Lagrangian form that dictates that we maximize fidelity to the data, subject to a constraint on the physical reasonableness of the labeling. As β drops, rougher labelings are penalized less. We emphasize that any prior model (eq. [4]) with $\beta > 0$ couples the labels at neighboring pixels, and so the Bayesian inference procedure does *not* correspond to a per-pixel decision rule. With $\beta = 0$, labels are spatially uncoupled, but the decision region may still be complex depending on the per-class distributions.

In fact, to accurately describe the characteristics of the features of interest to us, a more flexible distribution than a simple Gaussian must be used. For example, in the bottom left panel of Figure 1, the sunspot class is clearly not well fitted by a normal distribution. We have employed the wider class of finite normal mixture distributions (McLachlan & Peel 2000) of the form

$$P(y) = \sum_{j=1}^{J} \lambda_j N\left(y; \mu_j, \Sigma_j\right), \quad (7)$$

where $\sum_{j=1}^{J} \lambda_j = 1$, the constituent mean vectors μ_j are arbitrary, and the covariance matrices Σ_j are symmetric positive-definite. The function $N(x; \mu, \Sigma)$ is the distribution of a Gaussian bump in d dimensions, centered at μ and with covariance Σ:

$$\begin{aligned} N(y; \mu, \Sigma) = &\frac{1}{(2\pi)^{d/2} (\det \Sigma)^{1/2}} \\ &\times \exp\left[-\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right]. \end{aligned} \quad (8)$$

By letting J increase, the underlying distribution can be fitted more exactly. The free parameters λ_j, μ_j, and Σ_j are chosen by maximum likelihood. When $J > 1$, there is no longer a closed-form solution for these parameters, so they must be estimated by numerical optimization of the likelihood function, equations (3) and (7). We have used the well-known expectation-maximization (EM) algorithm (§ 3.2, McLachlan & Peel 2000), although other numerical methods would work (Redner & Walker 1984). For use with MDI, we have modified the EM maximization to account for the symmetry constraint on the magnetic field: all distributions must be invariant with respect to a reversal in the polarity of the field. To choose J, we have used cross-validated likelihood (Smyth, Ide, & Ghil 1999; Smyth 2000).

The distributional model for the quiet-Sun class ($x_s = 1$) is easy to determine by extracting samples of quiet Sun from a selection of images in the time studied here. This results in a set of quiet-Sun parameters $\{\lambda_j, \mu_j, \Sigma_j\}_{j=1}^{J_1}$. We extracted millions of pixels of quiet Sun, but used only a randomized selection of 20,000 feature vectors to determine the 30 or so free parameters. The corresponding distribution $P(y_s|x_s = 1)$ is shown in the top row of Figure 4. The left panel shows the six components of the distribution, and the right panel shows them superimposed on a subset of the quiet-Sun data. The outer ellipses represent a value of 2.5 standard deviations in any direction from the class center, and the lines intersecting the ellipses are its major and minor axes. (The graph is not printed with axis scales equal, so they do not cross at right angles.) For the numerical parameter values, see Table 1. Some components have a direct interpretation. For example, QS-1, with a center very close to (0, 1000), is the dominant component of the quiet-Sun class. However, QS-2 and QS-3 are also significant, and are due to supergranulation cells. The QS-4 pair, with a nonzero value of magnetic field, appears due to the quiet network.

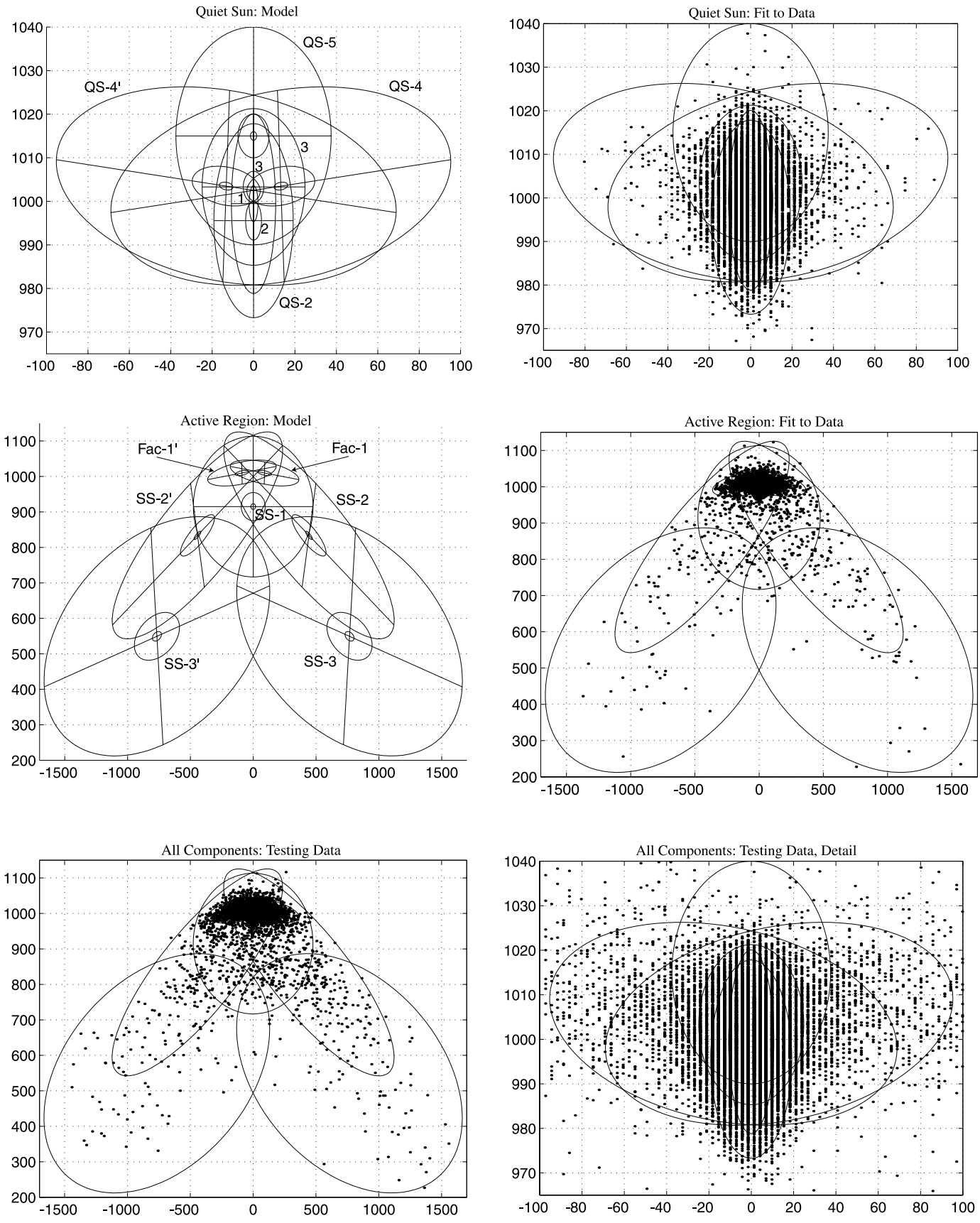

FIG. 4.—Models and their fit to data. *Top*: Quiet-Sun model components (*left*), superposed on training data (*right*). *Middle*: Same plots for the union of sunspots and faculae. *Bottom*: All model components plotted with test data (*left*) and quiet-Sun detail (*right*).

TABLE 1
MODEL COMPONENTS FOR QUIET SUN, FACULAE, AND SUNSPOT

Tag	λ_j	$\mu_{j,1}$	$\mu_{j,2}$	$(\Sigma_{j,11})^{1/2}$	$(\Sigma_{j,22})^{1/2}$	ρ_j[a]
QS-1.....	0.4308	0.00	999.46	4.28	8.25	0.00
QS-2.....	0.2085	0.00	995.55	7.66	8.91	0.00
QS-3.....	0.2024	0.00	1003.30	9.93	7.19	0.00
QS-4.....	0.0553	13.19	1003.49	32.84	9.12	0.25
QS-4′....	0.0553	−13.19	1003.49	32.84	9.12	−0.25
QS-5.....	0.0476	0.00	1015.04	15.08	9.92	0.00
Fac-1 ...	0.5000	90.10	1009.63	109.33	14.60	−0.51
Fac-1′...	0.5000	−90.10	1009.63	109.33	14.60	0.51
SS-1	0.5245	0.00	915.03	190.58	79.30	0.00
SS-2	0.1828	443.93	833.88	271.25	116.65	−0.82
SS-2′	0.1828	−443.93	833.88	271.25	116.65	0.82
SS-3	0.0550	766.36	549.42	359.29	134.87	−0.37
SS-3′	0.0550	−766.36	549.42	359.29	134.87	0.37

[a] The correlation coefficient $\rho = \Sigma_{12}/(\Sigma_{11}\Sigma_{22})^{1/2}$.

The distributions for the other region types are determined similarly, as indicated in Figure 3. Areas in images that represent sunspots and faculae are outlined, and a distribution is fitted to the resulting pooled scatter plot containing both objects. We used pooled data rather than attempting to separate the facula and spot components because we found it too difficult to ensure that the classes were not mixed; we prefer to allow the clustering algorithm to extract the classes in an unsupervised mode. Once a distribution $P(y_s|x_s \neq 1)$ (i.e., faculae and sunspot) has been found, it is decomposed manually into its component clusters. This is shown in the middle panels of Figure 4. The faculae appear as two components with slightly enhanced intensity, and the sunspot components are the lower five, typically with much greater magnetic field. We used 5000 feature vectors sampled at random from a collection of active regions to fit, again, roughly 30 parameters. The fit of the model to the pooled data used to determine it is shown in the middle right panel.

At this point, $P(\boldsymbol{y}|\boldsymbol{x})$ is fully specified via equation (3). Because we have constructed not just a decision rule but a full model for the observed data, we can check its validity in a falsification experiment (Popper 1959) using independently drawn *test data* sampled at random from the entire set of image pixels. The bottom panels in Figure 4 show this experiment. The left panel shows the full range of feature vector values, and the right panel is a detail plot of the central portion. This shows a good fit of the class models to independently sampled data.

The decision regions corresponding to these class models are shown in Figure 5; this is a map across the feature space of the class having the largest density. This map has some novel features. For example, we see that these distributional models do not support the practice of "axis-parallel" threshold rules for class membership. At the quiet-Sun boundary, for example, the presence of the supergranulation stretches the decision region along the intensity coordinate.

The other ingredient is the prior $P(\boldsymbol{x})$. Its contribution to the maximization will be small because the labeling should be predominantly determined by the observed data and the model $P(y|x)$. For now, we have let $\beta(s, s') = \beta$, a constant. Straightforward considerations (Ripley 1988, p. 97) show that $0.45 < \beta < 1.10$ for a neighborhood size of eight is needed to preserve corner configurations. We have used a value $\beta = 0.5$ for the labelings reported here to ensure that

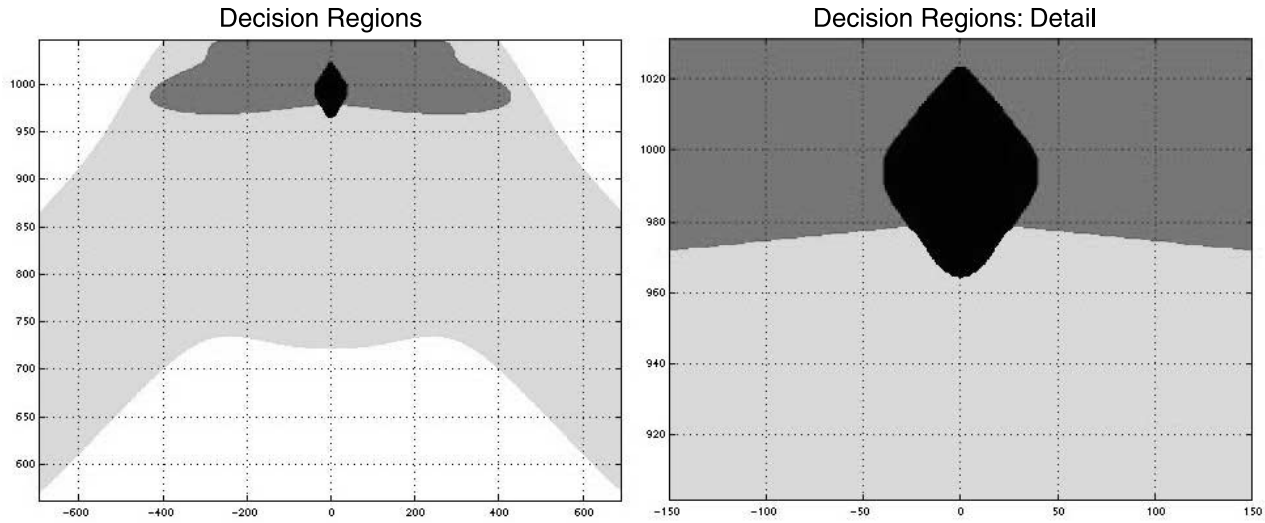

FIG. 5.—Decision regions across feature space. The abscissa is magnetic field and the ordinate is intensity; the most probable class is plotted. The left plot shows the full range of feature space and the right details the quiet Sun.

the prior term will be moderate compared to the likelihood term. Model accuracy might be improved by allowing the label coupling $\beta(s, s')$ to depend on the three-dimensional distance between sites. As sites become farther apart, they would be less coupled; in particular, this would allow substantial label fluctuation in the radial direction near the limb, while providing for moderate label smoothing at disk center.

5. ALGORITHMS FOR LABELING

The labeling is implicitly defined via the maximization of equation (2). The objective function can be written, similarly to equation (6), as

$$\log P(\boldsymbol{x}|\boldsymbol{y}) = \sum_{s \in S} \log P(y_s|x_s) - \sum_{s' \sim s} \beta(s, s') 1(x_{s'} \neq x_s) + C \qquad (9)$$

for some constant C depending only on $\beta(\cdots, \cdots)$. The key functions of the data that must be computed in order to evaluate this probability are the *probability maps* $P(y_s|x_s = k)$ for each pixel $s \in S$ and class $1 \leq k \leq K$; for the MDI images, we of course use the mixture probability of Table 1.

Direct approaches to finding an exact maximizer are difficult because the space of labelings is discrete and huge: of size $K^{|S|}$, where $|S|$ is the number of image sites. In practice, numerical methods adapted from statistical physics (Metropolis et al. 1953; Geman & Geman 1984) are employed to draw a *sample* from $P(\boldsymbol{x}|\boldsymbol{y})$; this sampling scheme is bootstrapped to find the maximizer. The stochastic sampling methods, originally developed for simulations of Ising models of magnetism, work by iteratively refining a labeling as follows. Starting from some initial labeling, update each site's label by drawing from the distribution $P(x_s|\boldsymbol{x}_{N(s)}, \boldsymbol{y})$. One scans repeatedly through all $s \in S$, updating labels in turn. Since most labels rarely change, it is useful to maintain a table of neighbor counts for each site. Also, since each label-update amounts to rolling a die with the given probabilities, one can further speed the computation (Ripley 1988, p. 100) by noting that a series of die rolls is equivalent to a single draw of a geometric random variable representing the waiting time until the die outcome changes. This yields a method for drawing a full labeling $\boldsymbol{x}$ from the distribution $P(\boldsymbol{x}|\boldsymbol{y})$.

To find its maximizer, introduce a nonnegative temperature parameter T that controls the sharpness of the *annealed* posterior

$$P_T(\boldsymbol{x}|\boldsymbol{y}) = 1/Z_T \exp\left[T^{-1} \log P(\boldsymbol{x}|\boldsymbol{y})\right], \qquad (10)$$

where the constant Z_T normalizes the distribution to sum to unity. This has the effect of multiplying the log-posterior (eq. [9]) by T^{-1}. As $T \to 0$, it is easy to see that P_T concentrates on the most likely elements in $P(\boldsymbol{x}|\boldsymbol{y})$; sampling from P_T thus becomes equivalent to choosing a maximizing labeling. It can be shown (Geman & Geman 1984) that if temperature is decreased slowly enough, a maximizer is indeed obtained. While seemingly indirect, these methods are well established in the statistics literature for this class of problems. Computations for $|S| = 1024^2$ take on the order of 2 minutes on a modern workstation; memory requirements are for K floating-point probability maps, one table of neighbor counts for each site, and one integer labeling.

6. DISCUSSION

Determining structural information about solar phenomena is one way to understand and refine the mechanisms of solar irradiance variability. We have demonstrated that the MDI photograms and magnetograms, used together, have the potential to identify these solar structures accurately. The first step in using these image sources together is a well-calibrated, temporally stationary, preprocessing scheme for putting flattened photograms and magnetograms in the same reference frame. Once this is done, a formalism for automated image segmentation based on contemporary image segmentation techniques is applied to the normalized data. The parameters in this model are fitted from scientist-provided image labelings, and their accuracy can be checked by standard statistical methods.

As an example, we show in Figure 6 the data and labeling for 1996 August 30, 07:35 UT. The preprocessed magnetogram was interpolated from 1 minute magnetograms taken 263 minutes earlier and 6 minutes later. The photogram has been preprocessed using the procedure outlined in § 3. The

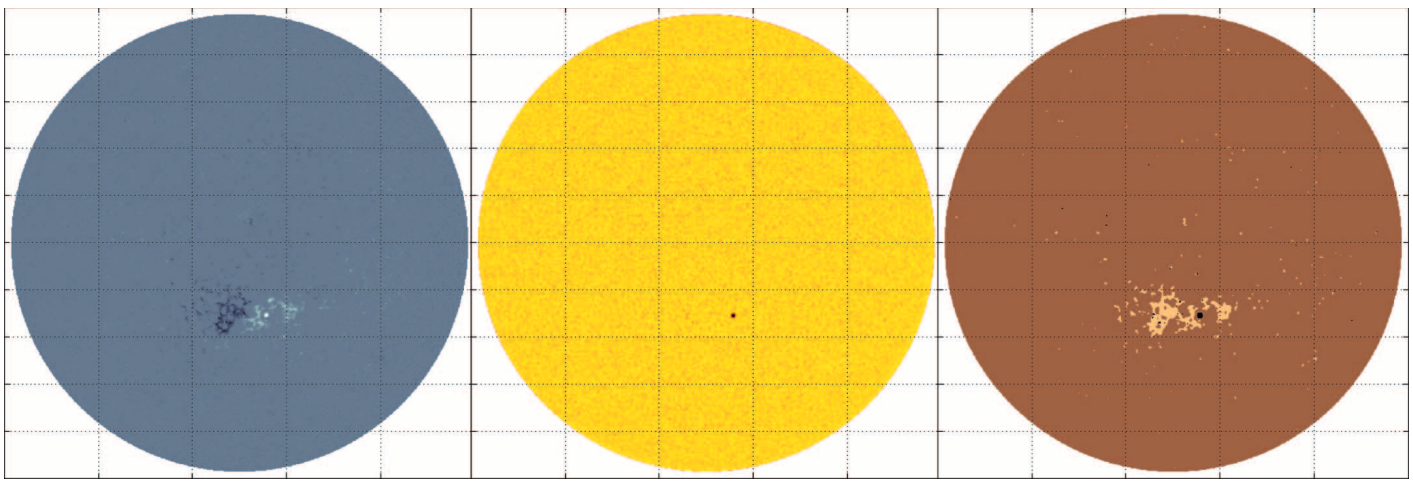

FIG. 6.—Preprocessed magnetogram (*left*), photogram (*middle*), and labeling (*right*) for 1996 August 30, 07:35 UT

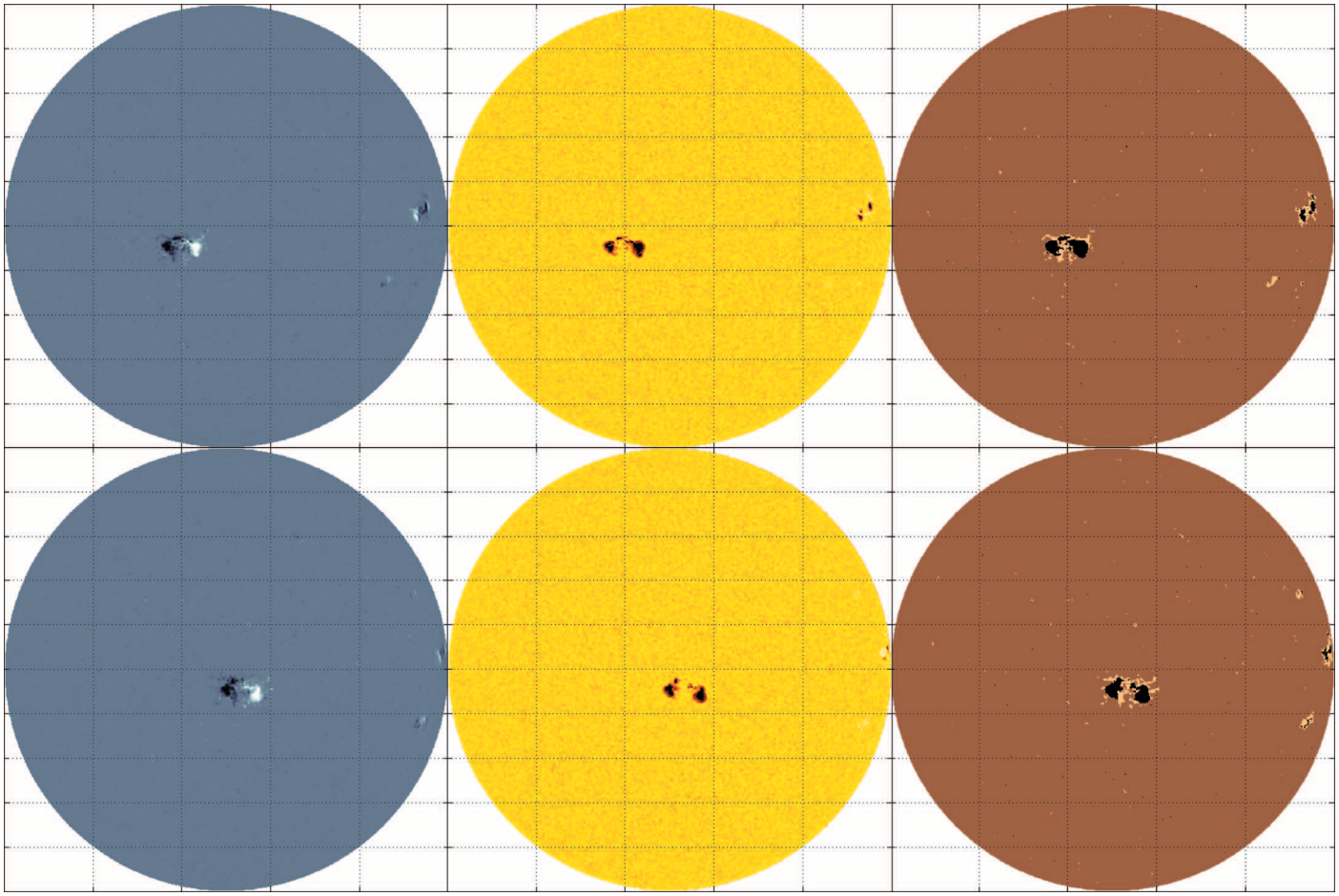

FIG. 7.—Preprocessed magnetograms (*left*), photograms (*middle*), and labelings (*right*) for 1996 November 25, 11:51 UT (*top row*) and November 26, 15:57 UT (*bottom row*).

labeling shows a small, concentrated sunspot (NOAA 7986) surrounded by a much larger facula. To the left of the principal sunspot is another small sunspot near the center of the negative-polarity (dark-colored) magnetic region. Inspection of the photogram shows that there is indeed a small area of significant intensity decrease (about 10% over several pixels) that results in this classification. The NOAA Solar Geophysical Data Bulletin (SGD) also identifies this feature as a sunspot. This labeling is in close agreement with the one of Fligge et al. (2000) on the same day (their Fig. 2). In particular, three small facular regions near the equator and other small faculae below the principal facula are visible in both sets of labelings.

Figure 7 shows results from 1996 November 25 and 26; the top panel is chosen to match the day shown in Figure 3 of Fligge et al. (2000). In this case, the 5 minute magnetograms are used. Interpolation in the top panels used magnetograms 36 minutes earlier and 443 minutes later; bottom panels, 90 minutes earlier and 485 minutes later. The active regions here are more complex and evolve rather quickly over time, so exact agreement between the two labelings is not expected. However, both the labelings of Fligge et al. (2000) and that in the top panel here identify two active regions, each with two main components. The leftmost region, NOAA 7999, has in turn several small components between the two main spots that roughly correspond in both labelings and in the SGD. The lower series shows images from the next day, November 26, and demonstrates the ability to consistently track active regions to within 2–3 pixels of the limb. Both sunspot components of the right-hand active region (NOAA 7997) are resolved at the edge of the limb. The new sunspot region that has appeared between the two larger groups is NOAA 8000 and also appears for the first time in the SGD for November 26. This spot is related to the facula identified in the previous day's labeling. These examples show a high degree of spatial coherency in the resulting labelings, as well as good agreement with other labelings.

The population of active regions included in this study was limited. Only three spots (NOAA 7981, 7999, and 8083, present in many images) were strong enough to cause more than about 10 pixels per image to drop below 250 in intensity. The model of Table 1, therefore, is rather unlikely to generate any intensities below that level. Of course, a more comprehensive survey that extended up to solar maximum uncovered many more such spots (about 40 in the MDI data up to 2000 December), and their inclusion would change the models somewhat. For example, Figure 8 shows thumbnail images and corresponding feature vector plots for 2 days near solar maximum. The top images (2001 April 7, 6:24 UT) show NOAA 9415, which is a complex bipolar spot with clear umbra and penumbra having intensity well below 200 in several places. None of the spots observed in the 14 month window treated here had such a complex magnetic configuration or such low intensities. The bottom images (2000 November 25, 11:11 UT) highlight an artifact

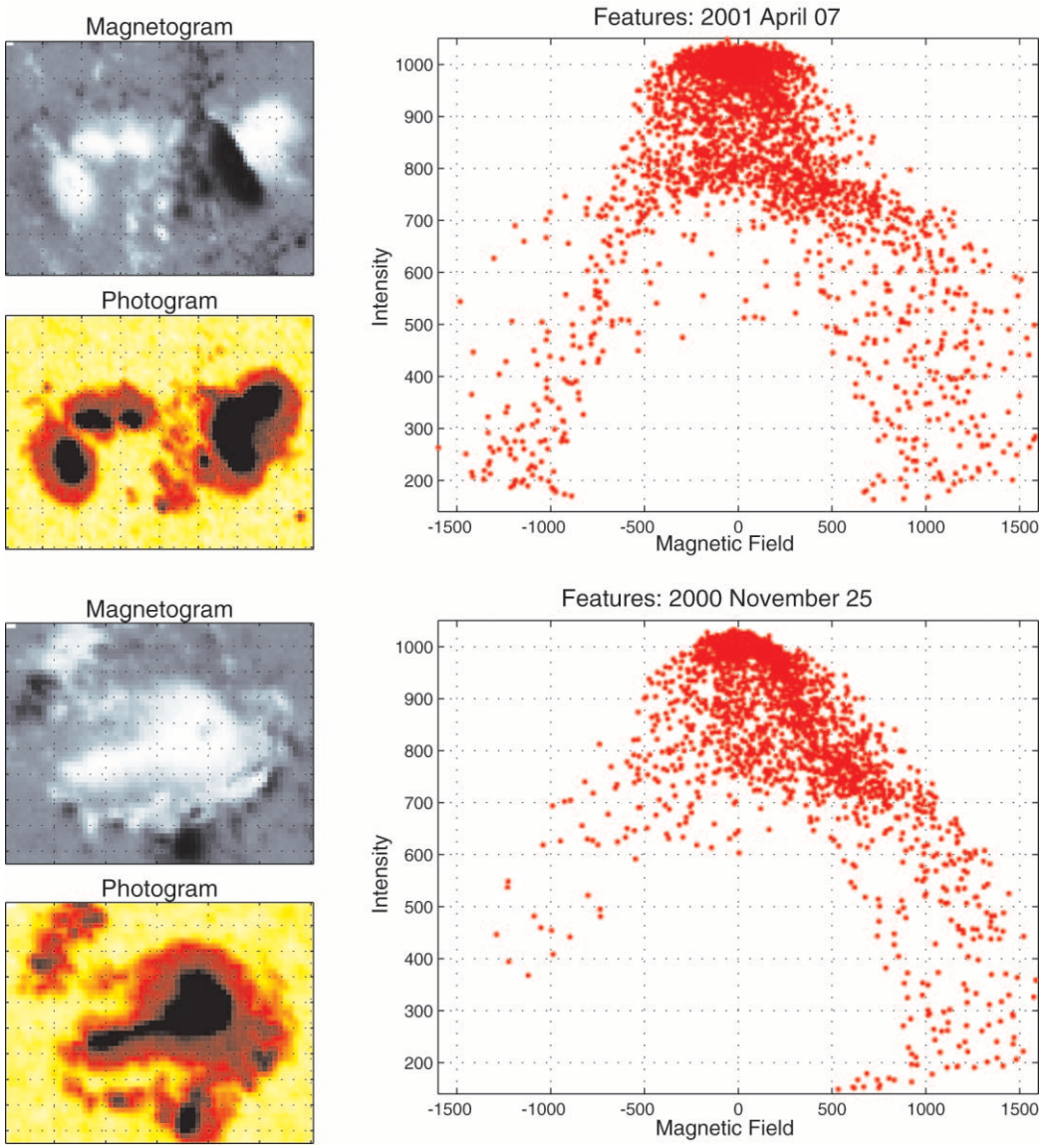

FIG. 8.—Preprocessed magnetograms and photograms (*left*) and corresponding feature vectors (*right*) for two strong active regions (*upper*: NOAA 9415; *lower*: NOAA 9236).

of the MDI instrument at very low intensity, where the computed magnetic field is erroneously low due to an onboard lookup table limitation (Liu & Norton 2001).[2] The effects of this distortion will rightly only appear once active-region models are fitted using data near solar maximum.

In assessing the impact of this temporal nonstationarity on labelings, we note that a model may have excellent discriminative capacity, even though it does not have complete generative capacity (Mjolsness & DeCoste 2001). The model for feature vectors (eq. [7]) generates observable pixels mimicking MDI at midcycle (Fig. 4, *bottom*) but fails in some parts of feature space to generate plausible peak-cycle sunspot pixels. Its generative capacity is therefore incomplete, and it will not pass a falsification experiment such as that in Figure 4. However, to identify active regions, we need only the class maximizing equation (9) to be unchanged. So, far from class boundaries, the class probabilities could be inaccurate and still lead to the correct labeling. The previously unseen pixels in Figure 8 will thus be explained as arising from the sunspot class because the other two classes are even less probable.

We have also applied these techniques to terrestrial imagery from Mount Wilson. For those data, the atmospheric point-spread function (PSF) blurs all features, rendering spatial regularization less important: we have taken $\beta = 0$ for that imagery. However, even at lower spatial resolutions, it remains useful to be able to deal coherently with multiwavelength observables and to let the decision regions be determined by explicit and checkable models. With the advent of active optics and better telescopes, the effective resolution of ground measurements is increasing. For example, the Precision Solar Photometric Telescopes, already operating in Rome, Hawaii, and Sacramento Peak, provide high photometric precision and high spatial resolution (2048×2048 pixels) full-disk images at two continuum wavelengths (402 and 609 nm) and also in the Ca II K line. Forthcoming ground-based observations by SOLIS (Synoptic Optical Long-term Investigations of the Sun) and

[2] Available at http://soi.stanford.edu.

ATST (Advanced Technology Solar Telescope) will further advance our understanding of the dynamics and evolution of active regions by producing high-resolution images. In particular, SOLIS will provide full-disk vector magnetograms in about 15 minutes and with 1″ pixels. ATST will produce visible and infrared images from 0.3 to 35 μm with a resolution of 0.″1 or better. Even with these high-resolution observations, the atmospheric PSF will remain a problem if it varies significantly in time. The region models, since they are built from pooled data, will reflect the diversity of the PSF, but features in relatively more blurred images may be oversmoothed. These seeing effects could be modeled if side information about the time-varying PSF is obtained.

One item not stressed here is the model flexibility allowed in defining $P(y_s|x_s)$. This distribution, relating the region types to the observables, can be let to depend on the spatial location of a site. This may be an advantage for specifying the characteristics of faculae, which have a spatially varying contrast. We have already made attempts to establish an empirical relation between the intensity and magnetic flux of different features as a function of their position on the solar disk. Spatially varying sensor noise can be modeled in a similar way. Whether or not all such second-order adjustments are incorporated in the statistical model, identification of the various regions will facilitate finding the so-called empirical calibration curves between the MDI magnetic field values and the continuum intensity images as a function of the evolution of the active region involved. It is anticipated that having labelings readily at hand will allow other such "per-region" quantities to be computed.

While we have not done so here, in a full "pattern-theoretic" approach (Grenander & Miller 1994), the label information can be linked into larger, more abstract structures describing individual active regions and facular groups. Such structural models have been used, for example, to infer galactic shapes from digitized images (Ripley & Sutherland 1990) and to describe chromospheric plages (Turmon & Mukhtar 1997). The plage models in particular can be readily adapted to describe photospheric faculae via an automatically determined bounding polygon.

SOHO is a mission of international cooperation between ESA and NASA. The authors gratefully acknowledge the effort of the MDI and VIRGO teams. In particular, Rick Bogart helped us to understand the physical origin of several features of the MDI images. Jeneen Sommers explained the fine points of the MDI FITS data and metadata. The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, by the University of California, Los Angeles, and by the Goddard Earth Science and Technology Center, University of Maryland, Baltimore County, under a contract with NASA. The research was supported by a grant of the *SOHO* Guest Investigator Program.

REFERENCES

Bogart, R. S., Bush, R. I., & Wolfson, C. J. 1998, in Proc. *SOHO* 6/GONG 98 Workshop (ESA-SP 418; Noordwijk: ESA), 127
Brandt, P. N., & Steinegger, M. 1998, Sol. Phys., 177, 287
Bratsolis, E., & Sigelle, M. 1998, A&AS, 131, 371
Fligge, M., Solanki, S. K., & Unruh, Y. C. 2000, A&A, 353, 380
Fröhlich, C. 1998, in Solar Electromagnetic Radiation Study for Solar Cycle 22, ed. J. M. Pap, C. Fröhlich, & R. K. Ulrich (Dordrecht: Kluwer), 391
Fröhlich, C., et al. 1997, Sol. Phys., 170, 1
Geman, S., & Geman, D. 1984, IEEE Trans. Pattern Anal. & Machine Intell., 6, 721
Grenander, U., & Miller, M. I. 1994, J. R. Stat. Soc. Ser. B, 56, 549
Harvey, K. L., & White, O. R. 1999, ApJ, 515, 812
Kuhn, J. 1996, in Global Changes in the Sun, ed. T. Roca-Cortes (New York: Cambridge Univ. Press), 231
Kuhn, J. R., Bush, R. I., Scheick, X., & Scherrer, P. 1998, Nature, 392, 155
Lean, J. L., Cook, J., Marquette, W. H., Johannesson, A., & Willson, R. C. 1998, ApJ, 492, 390
Liu, Y., & Norton, A. A. 2001, MDI Measurement Errors: The Magnetic Perspective (Tech Rep. SOI-TN-01-144; Stanford: SOI)
McLachlan, G., & Peel, D. 2000, Finite Mixture Models (New York: Wiley)
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087
Mjolsness, E., & DeCoste, D. 2001, Science, 293, 2051
Pap, J. 1997, in Past and Present Variability of the Solar-Terrestrial System: Measurements, Data Analysis, and Theoretical Models, ed. G. C. Castagnoli & A. Provenzale (Washington, DC: IOS), 1
Pap, J., Turmon, M., Mukhtar, S., Bogart, R., Ulrich, R., Fröhlich, C., & Wehrli, Ch. 1997, in Proc. 31st ESLAB Symp. (ESA-SP 415; Noordwijk: ESA), 477
Popper, K. R. 1959, The Logic of Scientific Discovery (New York: Basic Books)
Radick, R. 1994, in The Sun as a Variable Star: Solar and Stellar Irradiance Variations, ed. J. Pap, C. Fröhlich, H. S. Hudson, & S. K. Solanki (New York: Cambridge Univ. Press), 109
Redner, R. A., & Walker, H. F. 1984, SIAM Rev., 26, 195
Ripley, B. D. 1988, Statistical Inference for Spatial Processes (New York: Cambridge Univ. Press)
Ripley, B. D., & Sutherland, A. I. 1990, Philos. Trans. R. Soc. London A, 332, 477
Scherrer, P. H., et al. 1995, Sol. Phys., 162, 129
Smyth, P. 2000, Stat. & Comput., 10, 63
Smyth, P., Ide, K., & Ghil, M. 1999, J. Atmos. Sci., 56, 3704
Steinegger, M., Bonet, J. A., & Vasquez, M. 1997, Sol. Phys., 171, 303
Steinegger, M., Bonet, J. A., Vasquez, M., & Jimenez, A. 1998, Sol. Phys., 177, 279
Turmon, M., & Mukhtar, S. 1997, in Proc. 1997 Int. Conf. Image Proc., Vol. III (New York: IEEE), 320
Turmon, M. J., & Pap, J. M. 1997, in Proc. 2d Conf. on Statistical Challenges in Modern Astronomy, ed. G. J. Babu & E. D. Feigelson (New York: Springer), 408
Walton, S. R., Chapman, G. A., Cookson, A. M., Dobias, J. J., & Preminger, D. G. 1998, Sol. Phys., 179, 31
Worden, J. R., White, O. R., & Woods, T. N. 1998, ApJ, 496, 998
Wu, F. Y. 1982, Rev. Mod. Phys., 54, 235
Zappalá, R. A., & Zuccarello, F. 1991, A&A, 242, 480