Stratified Regularity Measures with Jensen-Shannon Divergence

Kazunori Okada
San Francisco State University
San Francisco, CA USA
kazokada@sfsu.edu

Senthil Periaswamy, Jinbo Bi
Siemens Medical Solutions
Malvern, PA USA
{senthil.periaswamy,jinbo.bi}@siemens.com

Abstract

This paper proposes a stratified regularity measure: a novel entropic measure to describe data regularity as a function of data domain stratification. Jensen-Shannon divergence is used to compute a set-similarity of intensity distributions derived from stratified data. We prove that derived regularity measures form a continuum as a function of the stratification's granularity and also upper-bounded by the Shannon entropy. This enables to interpret it as a generalized Shannon entropy with an intuitive spatial parameterization. This measure is applied as a novel feature extraction method for a real-world medical image analysis problem. The proposed measure is employed to describe ground-glass lung nodules whose shape and intensity distribution tend to be more irregular than typical lung nodules. Derived descriptors are then incorporated into a machine learning-based computer-aided detection system. Our ROC experiment resulted in 83% success rate with 5 false positives per patient, demonstrating an advantage of our approach toward solving this clinically significant problem.

1. Introduction

This paper introduces stratified regularity measure (SRM): a new entropic measure to describe data regularity under arbitrary spatial structures. The main concern of this work is to advocate the importance of choosing the appropriate data domain stratification/partitioning for measuring information from discrete images. Measuring regularity in images is one of the most fundamental problems in computer vision. The low-level image statistics/features, such as intensity variance and entropy, offer ubiquitous computational tools, providing a foundation to more complex vision algorithms. In their raw form, however, they are naturally constrained by specific imaging process-dependent factors of given data, such as pixel discretization and boundaries, and may not be able to capture some intuitive perceptual regularities in data.

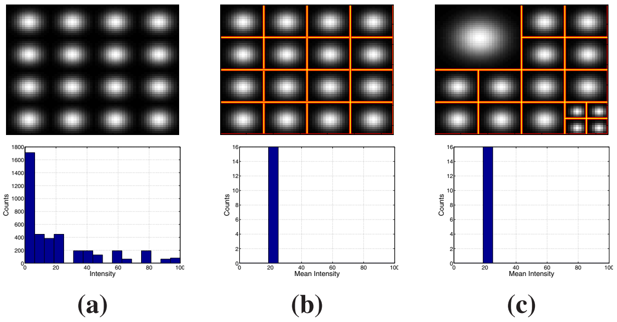


Figure 1. Shannon entropy does not reflect the *perfect* regularity of the Gaussian patches in image (a) because of the wide spread of its intensity histogram. When a well-chosen data domain stratification is given as shown in image (b), the corresponding histogram of each patch's mean intensity value becomes single-peaked, resulting in zero entropy that reflects the regularity. This also holds true even with scale variations shown in image (c), resulting in the exactly same histogram as in (b).

For example, the top-left image (a) in Figure 1 shows a *perfect* synthetic regularity of homogeneously distributed Gaussian patches. Both intensity variance and entropy of the entire image, however, result in non-zero values due to the significant spread of its intensity distribution shown below in the plot (a). This result fails to capture the perceptual regularity of the perfectly aligned patterns. Now suppose we provide an appropriate *data domain stratification* that reflects the spatial structure of the regularity as shown in the image (b). Then we proceed on computing any low-level intensity statistics of choice and taking its distribution. The plot (b) shows such a distribution of mean intensities of each stratum. Due to the stratification in image (b), it becomes a single-peak histogram thus the variance and entropy of this distribution become zero, indicating correctly the perfect regularity. Now consider an image with the same number of Gaussian patches but with different sizes as shown in the top-right image (c). By finding an appropriate stratification again as shown, both variance and entropy become zero, capturing now a regularity that is invariant against the histogram-preserving deformation of scaling. Observe that we made the simple measurement functions behave differently by choosing appropriate domain stratifications without modifying its functional form.

This paper presents a general framework of this type applied to the well-known Shannon entropy. By nature, as shown above, the entropy measure ignores spatial information because the intensity histogram does not appreciate spatial geometric structures in the original image. For this reason, the entropy cannot correctly capture obvious perceptual homogeneity of images such as texture-like regularity with a periodic occurrence of local patterns. The proposed SRM addresses this issue by measuring regularity in the following three successive steps: 1) stratifying/partitioning data domain, 2) deriving an intensity distribution for each stratum, 3) computing *set-similarity* of the distributions by using Jensen-Shannon divergence [16].

Set-similarity is a set-wise similarity measure, describing how set members are similar to one another collectively. Such a concept is useful in various vision applications (e.g., population co-registration [28, 25]) however it has not been explored fully in literature. Jensen-Shannon divergence offers an information-theoretic distributional set-similarity measure that extends naturally from its pair-wise version which is closely related to the popular Kullback-Leibler divergence.

As one of our main contributions, we prove that the proposed SRM descriptor forms a *continuum* of information-theoretic regularity measures as a function of *granularity* of the stratification. We show that the proposed measure is non-negative and upper-bounded by the Shannon entropy. And conditions for equality are given according to the level of granularity. At the finest level (each stratum contains only a single pixel), the measure is equivalent to the entropy, and the measure becomes zero at the coarsest level (entire image is a single stratum) or *when all distributions are equivalent*. These results enable an interpretation of the proposed measure as a *generalized Shannon entropy* equipped with a spatial parameterization by way of the data domain stratification. Furthermore, its continuum and bounds suggest a potential to estimate underlying spatial structure of data (i.e., segmentation) into collectively most similar/distinctive strata by minimizing/maximizing the SRM, respectively. As an illustrative example, this paper presents our initial pilot study, using a simple regular block tessellation of pseudo-regular texture pattern, for demonstrating the feasibility of this approach. Recent advancements in texture modeling [17, 9] can handle much more complex texture patterns however they do not incorporate this type of information-theoretic measures.

This paper also presents an application of the proposed measure as a novel feature extraction method for a challenging real-world medical image analysis problem: computer-aided detection of ground-glass nodules (GGN) in 3D CT scans [14, 26]. Ground-glass nodules are a class of lung tumors which is characterized by its fuzzy irregular appearance and its clinical significance with a high malignancy rate [10, 19]. To represent the target pattern more accurately, we design new SRM descriptors by stratifying each 3D volume of interest with linear-polar bins, similar to the shape context [4], and by computing distributions for both intensity and gradient magnitudes. We then incorporate these SRM descriptors into a state-of-the-art computer-aided detection (CAD) system using asymmetric cascade of sparse hyperplane classifiers [5]. The results of our quantitative ROC experiments demonstrate the effectiveness of the proposed approach with a high detection rate of 83% with five false positive per patient, which fares among other state-of-the-art GGN-CAD systems [14, 26].

The rest of this paper is organized as follows. Section 2 provides a brief review of the Shannon entropy and Jensen-Shannon divergence Section 3 then introduces the proposed SRM approach and provides its formal and experimental analyses. Sections 4 describes our experimental study for ground-glass nodule detection problem. The paper is concluded with discussing our future work in Section 5.

2. Review: Shannon Entropy and Jensen-Shannon Divergence

Shannon [20] introduced the *Shannon entropy* as a generic measure of information, providing a foundation of the information theory. Let $\mathbf{X}$ be a discrete random variable and $\mathbf{p}(x)$ be an arbitrary probability distribution of $x \in \mathbf{X}$. Then the Shannon entropy $H(\mathbf{p})$ is defined as,

$$H(\mathbf{p}) = -\sum_{x \in \mathbf{X}} \mathbf{p}(x) \log \mathbf{p}(x) \tag{1}$$

The entropy $H(\mathbf{p})$ is a non-negative continuous concave function and is maximized when $\mathbf{p}(x)$ is equiprobable over the entire discrete domain of $\mathbf{X}$ and minimized at zero when $\mathbf{p}(x)$ becomes a delta function.

The applications of the entropy especially in the information theory and communication fields are vast. Beyond this original scope, the entropy has been used to measure randomness or homogeneity in multi-dimensional image data by applying it over the image's intensity histogram. The randomness is indicated by high values of the entropy that is maximized when each intensity value is equiprobable. On the other hand, the homogeneity is indicated by its low values, minimized at zero when the image is single-colored. Such entropy-based descriptors and estimations have been applied to a wide range of vision-related problems including texture modeling [27], image registration [24, 28], independent component analysis [3], scale selection and image descriptor [13], image retrieval [23], object recognition [15], to name a few.

Jensen-Shannon divergence (JS) was proposed by Lin [16] as a new distributional similarity function. For a set of K arbitrary distributions $\{\mathbf{p}_k(x)\}$ of $\mathbf{X}$, it provides an overall *set-similarity* measure in the following form with the above entropy function,

$$JS(\mathbf{p}_1, ..., \mathbf{p}_K) = H(\sum_{k=1}^{K} \boldsymbol{\pi}(k)\mathbf{p}_k) - \sum_{k=1}^{K} \boldsymbol{\pi}(k)H(\mathbf{p}_k) \quad (2)$$

where $\boldsymbol{\pi}(k)$ denotes a normalized positive weight vector such that $\sum_{k=1}^{K} \boldsymbol{\pi}(k) = 1$. It is also straightforward to show that JS divergence can be rewritten as a weighted average of the *Kullback-Leibler* (KL) divergence between each component distribution and an average distribution,

$$\begin{aligned} JS(\mathbf{p}_1, ..., \mathbf{p}_K) &= \sum_{k=1}^{K} \boldsymbol{\pi}(k) KL(\mathbf{p}_k || \mathbf{q}) \quad (3) \\ \mathbf{q}(x) &= \sum_{k=1}^{K} \boldsymbol{\pi}(k)\mathbf{p}_k(x) \end{aligned}$$

where KL divergence is defined as

$$KL(\mathbf{p}||\mathbf{q}) = \sum_{x \in \mathbf{X}} \mathbf{p}(x) \log \frac{\mathbf{p}(x)}{\mathbf{q}(x)} \quad (4)$$

JS divergence can also be interpreted as the mutual information of two random variables X (observed image feature) and Y (latent class variable) when we set $\mathbf{p}_k$ by the likelihood $P(X|Y)$ and $\boldsymbol{\pi}(k)$ by the prior $P(Y{=}k)$ [23].

The following facts described in [16] and elsewhere are relevant to this study and noteworthy to recall.

1. **Set-Similarity:** JS in (2) naturally generalizes its corresponding pairwise version (K=2) to any finite number of distributions. It is one of the few distributional set-similarity measures found in literature.

2. **Relation to KL:** JS is an extension of KL to a well-formed one as can be seen in (3). Although its wide popularity, KL in (4) is undefined, thus uncomputable, at x where $\mathbf{p}(x) = 0$ or the distribution has zero value. The pairwise JS (K=2) resolves this problem while providing other desirable properties. For instance, it is a non-negative function due to Jensen's inequality, it provides both lower- and upper-bounds in terms of variational distance and Bayes error, its square-root is a true metric [22].

3. **Minimum:** JS equals to zero *iff* $\mathbf{p}_1$=$\mathbf{p}_2$=..=$\mathbf{p}_K$ since $H(\mathbf{p})$ is concave and the Jensen's inequality in (2).

Recently, JS divergence has been applied in a variety of vision-related tasks however most of them focused on the pairwise version while our focus for this study is its application in $K > 2$ settings similar to [6, 25]. They include object recognition with boosting [11], image segmentation with edge detection [1], automatic scale/bandwidth selection [6], active learning of training samples [18], groupwise point pattern registration [25].

3. Stratified Regularity Measures

The proposed stratified regularity measures (SRM) consist of three successive steps: 1) stratify image/data domain, 2) construct intensity distribution from each partition, 3) compute a JS divergence with a set of distributions. The following describes each step.

3.1. Data Domain Partitioning

As a specific stratification we will consider in this paper, this section introduces *mutually exclusive and exhaustive* (MEE) data domain partitioning of an image. This domain partitioning yields sets of data points/pixels/voxels in which all data points appear only once thus no sub-partitions overlap. Let $I(\mathbf{x})$ denote d-dimensional discrete image data where values in M discretized intensity levels $I \in (1, .., M) \subset \mathcal{Z}_+$ are distributed over a d-dimensional lattice $\mathbf{x} = (x_1, .., x_d) \in (1, .., D_1) \times (1, .., D_2) \times .. \times (1, .., D_d) \subset \mathcal{Z}_+^d$. The total number of data points in $I(\mathbf{x})$ is expressed by $N = \prod_{n=1}^{d} D_n$. Let Ω denote the set of all N distinctive data points from $I(\mathbf{x})$. We define the *MEE domain partition* $\mathcal{P}$ of Ω to be a set consisting of K MEE subsets of Ω such that $\mathcal{P} = \{\mathcal{Q}_1, .., \mathcal{Q}_k, .., \mathcal{Q}_K\}$, $\mathcal{Q}_i \cap \mathcal{Q}_j = \emptyset$ $\forall (i \neq j) \in (1, .., K) \times (1, .., K)$, $\bigcup_{k=1}^{K} \mathcal{Q}_k = \Omega$, and $\left(\sum_{k=1}^{K} |\mathcal{Q}_k|\right) = |\Omega| = N \geq K$, where $\emptyset$ denotes an empty set and $|\cdot|$ denote the cardinality of a set. We call a MEE subset $\mathcal{Q}_k$ a *sub-partition*, forming a stratum. Note that this partitioning is a general class of stratification since a sub-partition is simply a bag of pixels.

3.2. Intensity Histograms and Distributions

Next, a set of intensity distributions $\mathbf{p}_k$ from all the sub-partitions are derived. From each sub-partition $\mathcal{Q}_k$, we first construct an intensity histogram $\mathbf{h}_k(i)$ over a finite discrete domain $i \in (1, ..., M)$. The empirical intensity distribution is derived by normalizing the histogram,

$$\mathbf{p}_k(i) = \frac{\mathbf{h}_k(i)}{N_k} \quad (5)$$

where the total count of the sub-partition N_k is given by

$$N_k = |\mathcal{Q}_k| = \sum_{i=1}^{M} \mathbf{h}_k(i) \quad (6)$$

Suppose now that we construct an intensity histogram $\mathbf{h}(i)$ over the same domain i from entirety of the original image

$I(\mathbf{x})$. The corresponding distribution is given by,

$$\mathbf{p}(i) \quad = \quad \frac{\mathbf{h}(i)}{N} \tag{7}$$

It is straightforward to see that histograms $\{\mathbf{h}_k(i)\}$ derived from the sub-partitions are a linear decomposition of the histogram $\mathbf{h}(i)$ due to the MEE data domain partitioning,

$$\mathbf{h}(i) \quad = \quad \sum_{k=1}^{K} \mathbf{h}_k(i) \tag{8}$$

where $N = \sum_{k=1}^{K} N_k$. Note also that this linear decomposition property holds true even when we consider continuous distributions in a form of standard kernel density estimator,

$$N\mathbf{f}(i) \quad = \quad \sum_{k=1}^{K} N_k \mathbf{f}_k(i) \tag{9}$$

where $\mathbf{f}(x) = \frac{1}{n}\sum_{j=1}^{n} K(\frac{x-x_j}{h})$ with a positive constant bandwidth h.

3.3. Set-Similarity by Jensen-Shannon Divergence

Finally, the set of intensity distributions $\{\mathbf{p}_k(i)\}$ are subjected to the JS divergence for deriving the proposed regularity measures. The original JS formulation in (2) contains arbitrary weights. We set these weights according to the counts or probability mass of each sub-partition,

$$\boldsymbol{\pi}(k) = \frac{N_k}{N} \tag{10}$$

Substituting (5) - (10) to the JS divergence (2) yields the following form of JS divergence as a new image regularity measure,

$$JS_p(\mathbf{p}_1, ..., \mathbf{p}_K) = H(\mathbf{p}) - \frac{1}{N}\sum_{k=1}^{K} N_k H(\mathbf{p}_k) \tag{11}$$

with (8) for discrete histograms and with (9) for continuous distributions estimated by the density estimator $f(x)$. We call this measure *partitioned Jensen-Shannon divergence* (JS_p),

As demonstrated in Section 3.7, JS_p exhibits a strong dependency to the number of components K; Its magnitude rapidly decreases by decreasing K. Since K can be treated as a variable to be estimated as discussed later, it is of benefit that we devise a divergence measure normalized over K. We define such a normalized JS divergence as follows and call it *normalized Jensen-Shannon divergence* (JS_n),

$$JS_n(\mathbf{p}_1, ..., \mathbf{p}_K) \quad = \quad \frac{N}{K} JS_p(\mathbf{p}_1, ..., \mathbf{p}_K) \tag{12}$$

This normalization can be related to the well-known power law of natural scene power spectral statistics [2, 21] in that K equals the squared spatial frequency f^2 when a regular block partitioning is assumed.

3.4. SRM as Generalized Shannon Entropy

This section proves some formal properties of the new SRMs JS_p and JS_n and their relation to the Shannon entropy. As discussed earlier, the entropy $H(\mathbf{p})$ represents regularity (i.e., homogeneity, randomness and/or information) of the image data from which the distribution $\mathbf{p}(i)$ was constructed. On the other hand, given an arbitrary partitioning, the proposed measures ask how similar the sub-partitions are by computing a set-similarity of distributions using the JS divergence. Intuitively, this approach introduces a spatial parameterization into the entropic regularity measurement as motivated in the introduction. Due to the MEE partitioning, both $H(\mathbf{p})$ and $JS(\mathbf{p}_1, ..., \mathbf{p}_K)$ utilize the same amount of data but their formal relationship must be further studied.

The form of JS_p contains the Shannon entropy $H(\mathbf{p})$ of the original image as one of the pair of additive terms. Since JS_p is a non-negative function, it immediately yields that JS_p is upper-bounded by the Shannon entropy

$$0 \leq JS_p(\mathbf{p}_1, ..., \mathbf{p}_K) \leq H(\mathbf{p}) \tag{13}$$

Furthermore, conditions for equalities in (13) can be studied with *granularity* of the partitioning/stratification. Finer granularity yields a higher number of sub-partitions K. When the partitioning is given at the *finest* granularity or K=N, N_k=1 $\forall k$, each data point forms a sub-partition thus every $\mathbf{p}_k$ becomes a delta function. This vanishes the second term of RHS in (11) thus the partitioned JS divergence becomes equivalent to the Shannon entropy under this condition. Using the above arguments, we can prove a proposition about the upper-bound equality saying that both JS_p and JS_n are equivalent to H only at the finest granularity,

$$H(\mathbf{p}) = JS_p(\mathbf{p}_1, ..., \mathbf{p}_K) = JS_n(\mathbf{p}_1, ..., \mathbf{p}_K) \text{ iff } K = N \tag{14}$$

On the other hand, when the partitioning is given at the *coarsest* granularity or K=1, the entire image forms a single sub-partition. This condition is definable using the form (2) although it no longer serves as a divergence measure in any sense. (KL divergence, on the other hand, is not definable in this condition.) Under this condition, the second term of RHS in JS_p becomes $H(\mathbf{p})$ thus it cancels with the first term, resulting zero. Together with the original condition for the zero equality described in Section 2, the above arguments prove another proposition about the lower-bound equality,

$$JS_p = JS_n = 0 \text{ iff } K = 1 \text{ or } \mathbf{p}_1 = \mathbf{p}_2 =, .., = \mathbf{p}_K \tag{15}$$

These two propositions suggest that the proposed regularity measures using JS_p can be interpreted as a *generalized Shannon entropy*. This is in a sense that they provide a *continuum* of measures of regularity parameterized flexibly

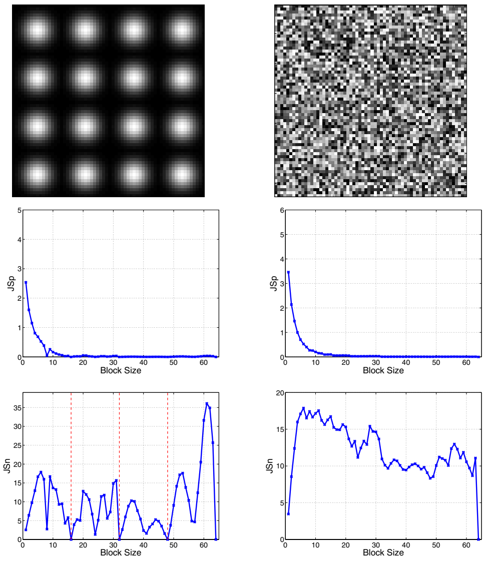


Figure 2. Comparison of JS_p and JS_n.

by arbitrary image partitioning $\{\mathcal{Q}_k\}$. At the finest level of the partitioning with $K = N$, they become equivalent to the Shannon entropy. At the coarsest level with $K = 1$, they assume zero. For any other partitioning, they are bounded by the entropy and minimized when the component distributions are equivalent. Furthermore, they inherit more desirable properties of the original divergence function than the popular KL divergence as discussed in Section 2.

3.5. SRM Descriptor

Using the above results, the following scheme for SRM descriptor is proposed.

1. partitioning Ω to an arbitrary $\mathcal{P}$ as defined in Section 3.1,

2. deriving empirical distributions $\mathbf{p}$, $\{\mathbf{p}_k\}$ from $\mathcal{P}$ by using (5)-(7),

3. computing JS_p and/or JS_n with $\mathbf{p}, \{\mathbf{p}_k\}$ by using (11)-(12).

Given a specific partition $\mathcal{P}$, this provides a scalar entropic regularity measure for the specified underlying stratification/partitioning. Given a set of pre-determined partitions $\{\mathcal{P}_l\}$, the above procedure can be applied to each member of the set, resulting in a vector of such measures. Note that the image domain Ω can be arbitrary chosen (e.g., a local image patch) without loss of generality. When computing a set of SRM descriptors for a single image with various partitions, the first entropy term in (11) can be computed once and stored for subsequent computations.

3.6. Partition Estimation with SRM

The continuum and bounds of the proposed SRM described in Section 3.4 allow us to consider a continuous estimation of data domain partitioning using the proposed regularity measures. Exploiting the property described in (15), a generic estimation problem is formulated by *minimizing* JS_n in order to find the MEE partition $\mathcal{P}_{min}$ of $I(\mathbf{x})$ which makes intensity distributions of all sub-partitions as *similar* as they can be,

$$\mathcal{P}_{min} = \text{argmin}_{\mathcal{P}} JS_n(I(\mathbf{x}), \mathcal{P}) \tag{16}$$

where the argument notation of JS_n introduced in (12) is slightly abused to indicate its dependency to both data and a partition. On the other hand, exploiting the property described in (14), another generic problem can be formulated by *maximizing* JS_n in order to find the MEE partition $\mathcal{P}_{max}$ of $I(\mathbf{x})$ which makes intensity distributions of all sub-partitions as *dissimilar* or *distinctive* as they can be,

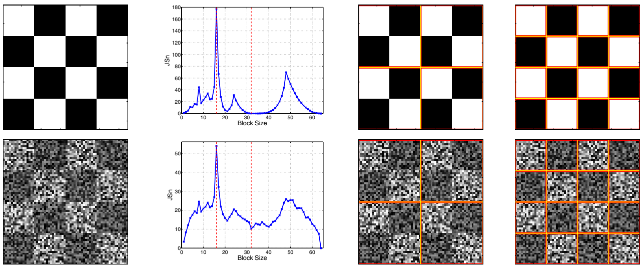
$$\mathcal{P}_{max} = \text{argmax}_{\mathcal{P}} JS_n(I(\mathbf{x}), \mathcal{P}) \tag{17}$$

This estimation framework can be understood as a latent classification problem through the mutual information interpretation discussed in Section 2. The minimization/maximization leads to partition an image into classes which minimizes/maximizes the information about the latent class variable conveyed by the measurements. Such minimization/maximization of information is known to yield classes with most/least similar densities.

3.7. Illustrative Examples: Synthetic Data

For illustration purpose, in the following, we study with a simplified data partitioning of regular image tessellations with blocks of various sizes. This simple partitioning is parameterized by the scalar block size, offering a clear way to visualize the parameter space. Handling the boundary condition, we allow blocks at the right and bottom borders to have a different size than others. Given a 2D square image domain $\mathbf{x} \in (1, .., D) \times (1, .., D)$, the block size ranges from $(1, 1)$ to (D, D) with D total choices.

Figure 2 compares the properties of JS_p (middle row) and JS_n (bottom) computed for two 64-by-64 synthetic images (top). The plots show the divergences computed for incrementally increasing block sizes from 1 to 64. The left image is the same one used in Figure 1, consisting of sixteen Gaussian patches of 16-by-16 size. The right one is an image with random intensity values. The Shannon entropy $H(\mathbf{p})$ of the left and right images are 2.5380 and 3.4615, respectively. For both images, plots demonstrate the JS_p's bounds as in (13)-(15). Also JS_n are much more informative in comparison with JS_p. For the Gaussian images, JS_n displays more clearly the recursive regularity of the patterns



(a) Input **(b)** JS_n **(c)** $\mathcal{P}_{min}$ **(d)** $\mathcal{P}_{max}$

Figure 3. Robustness of the minimum and maximum generalized entropy estimation against noise. The red dash lines in (b) indicate the selected block size by (16) and (17).

with $JS_n = 0$ at block sizes of 16, 32, and 48 indicated by red dash lines. On the other hand, JS_n for the random image are less variant over block sizes than the Gaussian one, indicating the lack of regularity. It is difficult to distinguish this qualitative difference between the two JS_p curves.

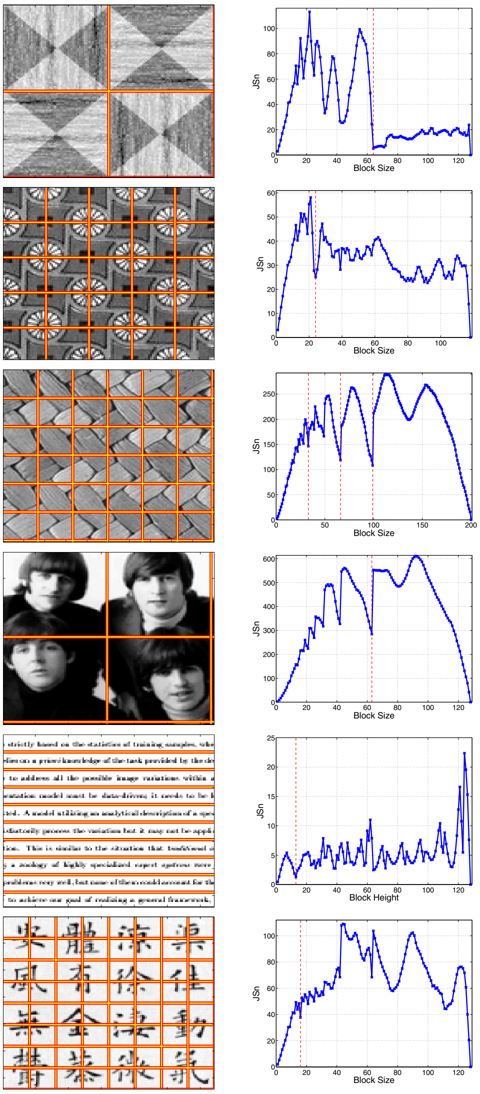
Figure 3 illustrates the aforementioned estimation framework and its robustness against data noises. The first row shows that $\mathcal{P}_{min}$ and $\mathcal{P}_{max}$ of the 64-by-64 checker-flag image corresponds to the intuitive block partitions at 32 and 16, respectively. This property remains true even when we apply very strong random intensity noise as shown in the second row. Similar robustness was observed when we vary the histogram's bin size.

3.8. Illustrative Examples: Pseudo-Regular Texture Patterns

To further study the robustness of our approach, we apply our estimation framework to a simple texture analysis problem. Figure 4 shows six illustrative examples of the results $\mathcal{P}_{min}$. The minimum entropy principle with (16) was applied to regularly tessellated texture-like patterns of wood tile, textile, and basket, as well as people and texts in various languages. We employed the same regular block partitioning except for the fifth row's example where we consider partitioning in only vertical direction. The results show successful regular partitioning even for cases with some geometrical transformations (1-3 rows) and significant appearance variations across sub-partitions (4,6 rows).

4. Ground-Glass Nodule Detection in CT Scans

This section describes our application of the proposed SRM descriptor to an automatic ground-glass nodule (GGN) detection system using 3D CT scans. Our study addresses the clinical demand raised by a recent clinical study [10] has shown that the lung tumors characterized as GGNs have a higher chance to develop into malignant cancers. As illustrated in Figure 5, GGNs exhibits fuzzy and irregular appearances, making it difficult to detect and characterize as well. We meet this technical challenge by exploiting the proposed SRM descriptor as a feature extraction method within a machine learning-based computer-aided diagnosis (CAD) framework. The choice of features is a key for developing a successful CAD system. Our target GGNs can appear not only in fuzzy intensity distribution (non-solid) but also with a number of solid cores among non-solid background (part-solid). Characterizing intensity pattern regularity of these targets requires a measure that can be flexible in terms of underlying spatial structure (i.e., solid core distributions) such as our SRM descriptors.



(a) $\mathcal{P}_{min}$ **(b)** JS_n

Figure 4. Examples of block size selection results with 2D texture images. From top to bottom: (1) wooden tiles with rotations, (2) patterns in a textile, (3) interwoven patterns of a basket, (4) picture of four people, (5) English text, (6) Chinese text. For all images, we used intensity values range within 8-bit grayscale and histogram of 32 bins. Image sizes varied from 120 to 200 $pixels^2$. The selected block sizes are displayed by red dash lines in (b).

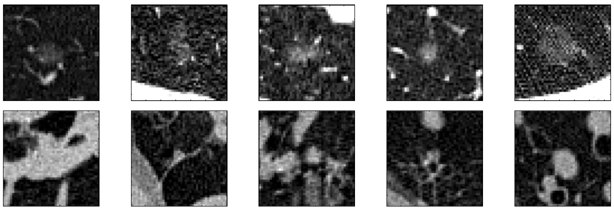


Figure 5. Examples of the ground-glass nodules (first row) and false-positive cases (second row) in 2D cross sections.

Our solution utilizes JS_p with 3D linear-polar partitions to compute candidate features chosen by the feature selection process described below. Each candidate is first approximately segmented by robustly fitting an ellipsoid. Then the local volume around the candidate is affine-warped so that the fitted ellipsoid is aligned to a sphere whose radius is 21 voxels. We let the voxels lying within the sphere form the domain Ω then impose a linear-polar partition to Ω similar to the 3D shape context [4]. We considered the SRM descriptors computed for distributions of the raw intensities and the gradient magnitudes. We also used two linear-polar partitions: 1) concentric spheres with various radii (onion) and 2) cones by regularly tessellating spherical coordinates (cake). The granularity of the partition is set to 7 onion and 12 cake partitions (i.e., 3 voxel layers in each onion slice and 30 degrees in each cake slice) by the same strategy described in Section 3.7. After fixing the partition, the normalization factor in JS_n becomes constant, thus we used JS_p in this experiment opting for its computational simplicity.

Together with four SRM descriptors (i.e., the raw intensity or gradient values versus the cake or onion partitions), we compute other 55 local intensity and geometric features such as intensity moments, tumor size, and boundary curvature and isotropy. These features are first subjected to a data-driven feature selection to identify the most descriptive features that distinguish GGN patterns from any non-GGN structures. Selected features are then used to train a detector with a binary classifier using asymmetric cascade of sparse hyperplane classifiers as proposed in [5].

For feature selection, the linear discriminant analysis with a greedy search was employed. Given a subset of features S, our greedy approach finds a new single feature f from the feature pool that improves classification performance when considering the expanded subset of features $S \bigcup \{f\}$. The search process starts with an empty set of features S, and stops when no feature in the feature pool improves classification performance significantly when added to S. The procedure selected nine features including a SRM descriptor with the gradient input and cake partition, as well as the original Shannon entropy computed for the gradient magnitudes in Ω.

Figure 6 display a scatter plot of a pair of the selected descriptors: the gradient-cake SRM descriptor against the local mean intensities. Circles denote GGNs while dots denote all other non-GGN candidates. The plot shows that the two distributions are fairly separated however there are some GGNs that may appear similar to other candidates, demonstrating the difficulty of the problem.

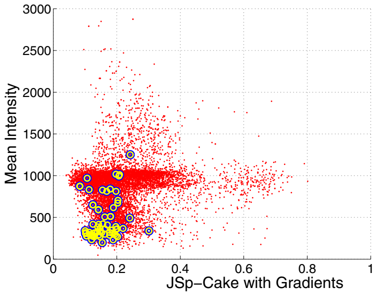


Figure 6. A scatter diagram of the randomness feature JS_p with gradient magnitudes plotted against the mean intensity feature. The circles show the feature values for the GGN cases while the red small dots show for all detected candidates.

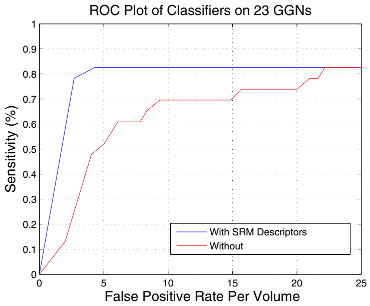


Figure 7. A ROC analysis of the ground-glass nodule CAD system with (blue) and without (red) the proposed SRM descriptors.

For studying the classification performance, we compare the detection performance of the CAD systems trained using the same scheme with and without the proposed SRM descriptors. CT volumes of 59 patients were used and 23 GGNs were identified and labeled by radiologists. Then leave-one-patient-out (LOPO) cross validation [8] performance was reported on the system's detection rates and false positive rates per volume. Figure 7 summarizes our ROC analysis. Note that the horizontal axis of the figure denotes the number of expected false positives per patient, which is a clinically more meaningful parameter than the standard false positive rate. The detection rate achieved 78% and 83% at false positive rate of 3 and 5 per patient, respectively. The results clearly show the advantage of our approach against the system without the SRM descriptors.

5. Discussion and Future Work

This paper introduced new measures to describe regularity of data flexibly in terms of spatial structure. We prove that the proposed SRM can be interpreted as a continuum of regularity measures upper-bounded by the Shannon entropy and parameterized by domain stratification/partitioning. Exploiting these properties, we successfully applied the measures for computer-aided detection of ground-glass nodules in CT scans.

Deriving robust solutions for the estimation problems in (16) and (17) in a full partition space is out of this paper's scope but one of the important future works to be addressed. Such a solution is required to apply the proposed methods to a wider range of applications. It is obvious that the simplistic exhaustive search for regular tessellation used in this paper does not extend to solve this combinatoric search problem. We plan to explore sampling-based solutions such as the MCMC method for this purpose.

Our theoretical contributions are generic and not constrained to the specific choice of the vision-based applications explored in this paper. The entropy is one of the most ubiquitous tools in the domain of machine learning, computer vision, and image understanding. The proposed frameworks can be applied to a vast range of computational theories and applications using the information theoretic measures, some of which are overviewed in Section 2. Other future work include analyzing the proposed framework formally in the context of other information theoretic learning [7, 3] and statistical modeling [12, 27] frameworks.

References

[1] C. Atae-Allah, J. F. Gomez-Lopera, J. Martinez-Aroza, R. Roman-Roldan, and P. Luque-Escamilla. Image segmentation by Jensen-Shannon divergence. In *ICPR*, 2000.

[2] J. J. Atick. Entropy minimization: A design principle for sensory perception? *International Journal of Neural Systems*, 3:81–90, 1992.

[3] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7(6):1129–59, 1995.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape context. *IEEE Trans. Pattern Anal. Machine Intell.*, 24:509–522, 2002.

[5] J. Bi, S. Periaswamy, K. Okada, T. Kubota, G. Fung, M. Salganicoff, and R. Rao. Computer aided detection via asymmetric cascade of sparse hyperplane classifiers. In *ACM SIGKDD*, pages 837–844, 2006.

[6] D. Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(2):281–288, 2003.

[7] D. Donoho. On minimum entropy deconvolution. *Applied Time Series Analysis II*, page 565609, 1981.

[8] M. Dundar, G. Fung, L. Bogoni, M. Macari, A. Megibow, and B. Rao. A methodology for training and validating a CAD system and potential pitfalls. In *CARS*, 2004.

[9] J. Hays, A. Leordeanu, M. Efros, and Y. Liu. Discovering texture regularity as higher-order correspondence problem. In *ECCV*, 2006.

[10] C. I. Henschke, D. F. Yankelevitz, R. Mirtcheva, G. McGuinness, D. McCauley, and O. S. Miettinen. CT screening for lung cancer: frequency and significance of part-solid and non-solid nodules. *AJR Am. J. Roentgenol.*, 178(5):1053–1057, 2002.

[11] X. Huang, S. Z. Li, and Y. Wang. Jensen-Shannon boosting learning for object recognition. In *CVPR*, 2005.

[12] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[13] T. Kadir and M. Brady. Scale, saliency, image description. *Int. J. Comput. Vision*, 45:83–105, 2001.

[14] K. Kim, J. Goo, J. Kim, B. Min, K. Bae, and J. Im. Computer-aided diagnosis of localized ground-glass opacity in the lung at ct: Initial experience. *Radiology*, 237:657–661, 2005.

[15] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *ICCV*, 2005.

[16] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Info. Theory*, 37(1):145–151, 1991.

[17] Y. Liu, W. Lin, and J. Hays. Near regular texture analysis and manipulation. In *Siggraph*, pages 368–376, 2004.

[18] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. Mooney. Active learning for probability estimation using Jensen-Shannon divergence. In *Euro. Conf. Machine Learning*, 2005.

[19] K. Okada, D. Comaniciu, and A. Krishnan. Robust anisotropic Gaussian fitting for volumetric characterization of pulmonary nodules in multislice CT. *IEEE Trans. Med. Imaging*, 24:409–423, 2005.

[20] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[21] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[22] F. Topsoe. Inequalities for the Jensen-Shannon divergence. Draft available at http://www.math.ku.dk/topsoe/.

[23] N. Vasconcelos and M. Vasconcelos. Scalable discriminant feature selection for image retrieval and recognition. In *CVPR*, 2004.

[24] W. M. Viola, P. Wells III. Alignment by maximization of mutual information. *Int. J. Comput. Vision*, 24(2):137154, 1997.

[25] F. Wang, B. Vemuri, and A. Rangarajan. Groupwise point pattern registration using a novel CDF-based Jensen-Shannon divergence. In *CVPR*, 2006.

[26] J. Zhou, S. Chang, D. Metaxas, B. Zhao, M. Ginsberg, and L. Schwartz. An automatic method for ground glass opacity nodule detection and segmentation from ct studies. In *EMBS*, pages 3062–3065, 2006.

[27] S. C. Zhu, Y. N. Wu, and D. Mumford. Minmax entropy principle and its application to texture modeling. *Neural Comput.*, 9(8):1627–1660, 1997.

[28] L. Zollei, E. Learned-Miller, W. Grimson, and W. Wells III. Efficient population registration of 3D data. In *ICCV WS on CVBIA*, pages 291–301, 2005.