TWO DIMENSIONAL HISTOGRAM ANALYSIS USING THE HELMHOLTZ PRINCIPLE

ARJUNA FLENNER AND GARY A. HEWER
NAWCWD - Physics and Computational Sciences
China Lake, CA 93555, USA

CHARLES S. KENNEY
University of California - ECE Department
Santa Barbara, CA 93555, USA

(Communicated by Luminita Vese)

ABSTRACT. An algorithm for two dimensional histogram modal analysis is presented. A major challenge in two dimensional histogram analysis is to provide an accurate location and description of the extended modal shape. The approach presented in this paper combines the Fast Level Set Transform of the histogram and the Helmholtz principle to find the location and shape of the modes. Furthermore, the algorithm is devoid of any a priori assumptions about the underlying density or the number of modes. At the core, this approach is a new way to manage and search the number of regions that must be examined to identify meaningful sets. Computational issues required a new tail sum bound on the multinomial distribution to be stated and proven. This bound reduces to the Hoeffding inequality for the binomial distribution. The histogram segmentation procedure was applied to the two problems of image color segmentation and correlation pattern recognition. With no a priori knowledge about the color image assumed, the two dimensional modal analysis is applied to the CIELAB color space to find perceptually uniform dominant colors. The modal analysis is also extended to correlation pattern recognition to find multiple targets in a single correlation plane.

1. Introduction. Histograms modal analysis is a concise and flexible technique to identify the global properties of large data sets. Despite a vast aggregate of research, a definitive answer on how to locate the modes of a histogram has been elusive. Segmentation of a histogram into its separate modes is a difficult task for many reasons. These reasons include: the number of modes is unknown, the number of candidate regions that must be examined to localize the modes is large, and the extent of the mode is often irregular. Delon et al. [9, 10, 11], successfully circumvented these limitations by developing a nonparametric approach for one dimensional histogram analysis based on the Helmholtz principle. This paper will extend their ideas to create a robust two dimensional histogram analysis procedure that combines the Helmholtz principle with the topographic map of the FLST tree structure.

The modes of a multinomial 2-D histogram are complex. The complexity exists since the location and number of local maxima are unknown. Furthermore, their

2000 Mathematics Subject Classification. Primary: 68T45, 62H35.
Key words and phrases. Helmholtz Principle, Histogram analysis.

This research was supported by the Office of Naval Research under Office of Naval Research Grant Numbers N00014-02-1-0318, N00014-07-WX20685, and N00014-07-WR20071.
geometrical shape is not necessarily simple but rather arbitrarily structured. This complexity has motivated many researchers to develop histogram modal analysis techniques. Nonparametric procedures are often preferred since these methods do not have embedded assumptions. The Expectation-Maximization (EM) algorithm, the mode tree of Minnotte and Scott [35] and the mean shift method are popular methods for mode seeking. These three procedures will be briefly reviewed below.

Mclachlan et. al provide a comprehensive introduction to both the EM algorithm [33] and finite mixture models [34]. The EM algorithm is a general method that maximizes successive approximations of the likelihood function. It alternates between an approximation (E) step, and one that maximizes it (M step). The EM algorithm works well in fitting mixture models and as a method to estimate unknown parameters, especially when they depend on latent or unobserved variables. It is appropriate for standard non-binned data, when the number of modes is known a-priori and the data distribution can be described well by a mixture of density functions. When the number of modes is unknown and are not sufficiently separated, then estimating the number of modes and the unknown model parameters is an important problem [34]. The EM methodology may exhibit difficulties in such situations. These difficulties include over fitting or over smoothing which results from incomplete information about the large number of parameters that must be estimated in an unsupervised way. Other complications include local convergence, or repeated random starting locations required to overcome poor initialization or the merging of two very close peaks into a single mode [34]. Moreover, the fitting of mixture densities to multivariate binned data intensifies the complexity of the EM iterations since analytical expressions must be replaced with numerical integration at each step [4].

Minotte and Scott [35] discuss building a mode tree based on kernel density estimates as a method of visualizing the modal structure in a one dimensional data sample of size $M$. Density kernel estimators are usually radially symmetric with a single bandwidth parameter. The monograph of Silverman [42] and the histogram focused text of Scott [39] discuss multivariable kernel estimators. The survey by Scott and Sain [40] provides some additional focus on multi-dimensional density estimators. The construction of the mode tree exploits a rare property of the univariate Gaussian kernel that insures modal continuity, a behavior described by Silverman [41]. The Gaussian kernel density estimates for a range of bandwidths reveals a splitting relationship between new modes and old modes. This behavior provides a graphical means to visualize the structure of an unknown data set. While the resulting mode tree is informative for 1-D histograms, visualizing the structure even for bivariate data is challenging.

The mean shift estimate of the gradient of a density function was developed by Fukunaga and Hostetler [24] with later convergence refinements by Cheng [6] and extensions to discrete data by Comaniciu and Meer [7]. It is a clever method to locate the modes without estimating the density. It locates modes by finding the stationary points of a kernel estimator. The analysis depends on the mean shift vector being proportional to the derivative of a kernel density estimator. When the directions of the mean shift and the gradient vector are aligned, then the mean shift will always point to a local maxima. It is a gradient ascent algorithm that avoids the step size pitfalls of traditional gradient based methods because under mild conditions the iterates will converge maintaining an adaptive step size. Artifacts induced by the termination thresholds must be eliminated by post-processing. In
order to accommodate the hidden multiple local maxima, Comaniciu and Meer recommend tessellating the entire feature space by the kernels. As with all kernel based methods, the choice of bandwidth is critical since an inappropriate choice can lead to either merged peaks or generate false peaks [43].

Unlike the methods described above, the Helmholtz principle can be applied to two dimensional binned histograms to develop a direct and fully automated modal detection strategy. The method described in this paper does not require repeated random starting or adjustable bandwidths to overcome incomplete information. Also, post processing is not needed to adjust for truncation effects. Furthermore, the level sets and the associated graphical tree structure is a complete representation of a two dimensional histogram. Direct connectivity of the FLST sets guarantee that no false peaks are introduced and that the integrity of the modal shape geometry in a neighborhood of a peak is preserved since no smoothing, over-fitting or under-fitting is introduced. In fact, the estimated modal shape is not constrained by the kernel size or support near a peak, but rather its shape will be a discrete set of bins that satisfy the Helmholtz selection criterion.

A form of the Helmholtz operational guidelines is summarized in the following quote: “the Helmholtz principle states that whenever some large deviation from randomness occurs, a structure is perceived” [20]. This principle has been developed and applied extensively to perceptual gestalt theory by Desolneux et al. They successfully found alignments, shapes and other important geometric structures in digital images [15, 18, 14, 16, 19]. Recently, using the Helmholtz principle, Delon et al. [11] constructed an automatic and parameter-free methodology of one dimensional histogram segmentation. This procedure consisted of an initial screening step and a refinement step. Candidate modes were identified by the merging of histogram bins into meaningful intervals ordered by the number of false alarms (NFA). A maximal principle was then applied to refine the family of candidate modes into a collection of disjoint sets. The Helmholtz principle combined with a unimodal hypothesis provided a further improvement. This technique segments a one dimensional histogram into intervals that contain only one mode. It provides a complete segmentation of the histogram, and it also finds modes that are small compared to the largest mode. While their process segments the histogram, the unimodal construction developed in their work is not easily generalizable to two dimensional histograms.

A unimodal hypothesis for two dimensional histograms was not pursued in this work, but several of the features common with the work of Delon et al. are maintained. The maximal meaningful mode algorithm, presented in Section 3, consists of an initial screening step followed by further refinement steps. The initial screening step locates groups of bins in which the data concentrates. Such groups of bins are in accordance with a preconceived notion of a mode. The further refinement steps motivated by a maximal principle will locate disjoint sets where the data concentrates. The two dimensional domain of the histogram compounds the complexity of the algorithm when compared with the one dimensional case. This increased complexity introduces a merging problem which can be understood in the context of a maximal principle.

The Helmholtz principle has the benefit of being weakly dependent on the parameters. The dependence on the parameters is reduced by the application of a maximal principle to find disjoint groups of independent and identically distributed random variables that minimize the $NFA$. Meaningful groups that contain or are
contained in other meaningful groups are common. For example, a large peak of the histogram may contain two or more separate peaks. It is then natural to consider whether the larger peak should be decomposed into the two smaller peaks, or whether the smaller peaks should be merged into the larger peak resulting in a single mode. A level set graphical structure that embeds the peaks in a containment tree structure is used in this work to define a maximal principle and address this merging problem.

The merging problem mentioned above in data grouping was tackled by Cao, Delon, Desolneux, Muse, and Sur [5]. Their method merges groups of data if the NFA for the merged group is less than the NFA for a pair of groups. This procedure requires the use of a trinomial since three outcomes are possible. In the analysis of two dimensional histograms, a countable number of outcomes is possible. The multinomial becomes an essential component of the analysis. The two dimensional nature of the problem and the need to use multinomial distributions introduces some difficulties.

The family of all possible merging candidates in a two dimensional histogram is combinatorially large. The number of merging candidates is reduced by imposing a data structure. Cao et al. uses bounding rectangles to reduce their search space. In this work, this problem is approached by applying the Fast Level Set Transform to the histogram, thus extracting a graphical structure. When combined with the Höeffding inequality, this methodology provides a merging and clustering framework for mode selection. Furthermore, the Fast Level Set Transform has the added benefit of detecting sets with arbitrary shapes. [1, 2, 36, 37].

The candidate modes are identified by using a maximal principle. The maximal principle merges meaningful histogram peaks into other meaningful peaks according to their NFA. Initially, the NFA is defined in terms of binomial and multinomial probability distributions resulting in numerous calculations of multinomial tail sums. The multiple calculations of binomial and multinomial distributions is computationally expensive, and this computational complexity motives a modification of the NFA definition. This modification uses the Höeffding bound to define and merge meaningful peaks. A major contribution of this paper is a derivation of the necessary mathematical theory of a new multivariate Höeffding inequality for the multinomial tail probabilities. This inequality generalizes Höeffding’s approximation for independent bounded univariate densities. With these new bounds, the Helmholtz principle, within the Cao et al. framework, can be extended to multi-dimensional histograms. The multinomial Helmholtz bound will be shown to have the following desirable properties: it is the minimization of an upper bound on the multinomial distribution, it inherits the conditional probability factoring of the multinomial distribution, and it has a natural relation to the mode of the multinomial distribution. Moreover, the new bound is always less than the standard binomial product upper bound of Mallows and Joag-Dev et al.. Furthermore, when some typical parameter values are examined, the new bound is many orders of magnitude tighter [32, 26].

The rest of the paper is organized as follows. The main line of reasoning is presented in Sections 2.1 through 2.2 and Section 3. Sections 2.1 through 2.2 reviews the Fast Level Set Transform and discusses the Helmholtz principle as applied to two dimensional histograms. The notion of a family of ϵ-meaningful modes in terms of the generalized Höeffding bound is also defined. Section 3 continues by defining maximal ϵ-meaningful modes, and the maximal meaningful mode algorithm
Figure 1. The tree of connected shapes obtained through the lower level set transform, upper level set transform and FLST is shown. The root of the tree in all three cases is the entire image. All three tree structures are ordered by containment, but the upper tree and lower tree both allow shapes that contain holes. The FLST tree closes all the holes.

is presented. This algorithm decomposes a histogram into its maximal $\epsilon$-meaningful modes.

Sections 2.3, 2.4, and 4.1 provide supporting material. A new relative entropy upper bound approximation for multinomials is derived in 2.3. This upper bound reduces to the Chernoff-Hoeffding bound for the binomial distributions. Section 4 provides some insight into the maximal $\epsilon$-meaningful mode definition, and it presents some theorems that decrease the computational burden. The first theorem in this section furnishes a new simple algebraic test that allows the direct comparison of two sets and predicts their comparative relative entropy ordering.

Experiments and applications are addressed in section 5. The effect of binning on mode location is discussed, and the number of $\epsilon$-meaningful modes found as a function of the parameter $\epsilon$ is explored. Two applications are presented, namely color segmentation and multiple target location in correlation plane pattern recognition. Additionally, the effectiveness of the approach using some of the examples in the works of Delon et al. [10, 11] is shown.

2. Preliminaries.

2.1. The fast level set transform. The image level set decomposition of Monasse [36] is a morphological image representation that readily adapts to two dimensional histogram analysis. Two essential features of this representation are exploited in the following. First, at a fixed gray level, this representation partitions the image into connected non-overlapping regions. Second, as the gray level changes, the connected components are ordered by a containment tree structure. This tree structure will be utilized in this work to guide the search for two dimensional histogram modes.
Consider the level sets of a continuous real valued function $h : \mathbb{R}^2 \rightarrow \mathbb{R}$. Let $\lambda \in \mathbb{R}$ and define the upper level set $X_{\lambda}$ and the lower level set $X^{\lambda}$ of $h$ as

$$X_{\lambda} = \{ x \in \mathbb{R}^2 | h(x) \geq \lambda \},$$
$$X^{\lambda} = \{ x \in \mathbb{R}^2 | h(x) < \lambda \}.$$

Since $h$ is a function it is clear that the level sets have a natural inclusion structure given by

$$\lambda_1 \geq \lambda_2 \iff X_{\lambda_1} \subseteq X_{\lambda_2},$$
$$\lambda_1 < \lambda_2 \iff X^{\lambda_1} \subseteq X^{\lambda_2}.$$  

Both the upper and lower level sets are contrast invariant morphological representations of an image. These representations decompose the image into distinct regions for fixed $\lambda$, but the level sets by themselves lack some useful topological information. Monasse’s level set tree structure, the Fast Level Set Transform (FLST) tree, organizes the connection between the upper and lower level sets as well as the level set containment properties. The set containment topological information is encoded through the parent-child relationship.

The graph theoretic definition of a tree is inadequate to describe the topology of upper and lower level sets since a level set parent may have an infinite number of children. Monasse overcomes this difficulty by defining a tree structure using a partial order relation. The following definition is from [36].

**Definition 2.1.** Let $E$ be a family of sets and $\preceq$ a partial order relation on $E$. We say that $\preceq$ induces a tree structure in $E$ if the following two conditions hold:

1. $\exists R \in E$, $\forall E \in E, E \preceq R$;
2. $\forall A, B, C \in E$, if $A \preceq B$ and $A \preceq C$ $\implies$ $B$ and $C$ are comparable.

A tree structure is guaranteed to have a root by the first condition, while the second condition prevents closed loops.

The tree structure in Definition 2.1 is sufficient to build upper level set and lower level set trees where the partial order relation is set containment. For example, the upper level set $X_{\lambda}$ can be represented as a union of connected component equivalence classes. The inclusion relationship in Equation 2 demonstrates that for $\lambda_1 \geq \lambda_2$ the connected components of $X_{\lambda_1}$ are contained in the connected components of $X_{\lambda_2}$. This inclusion relation on connected components is sufficient to satisfy condition 2 of Definition 2.1. Furthermore, if $h$ is bounded, then there exists a $\lambda < \infty$ such that $X_{\lambda}$ can serve as the root of an upper level set tree; thus both conditions for a tree structure are satisfied. Similar statements can be made about the creation of a lower level set tree.

Both the upper and lower level sets are complete representations of an image, but there is redundant information if both level sets are used. These two representations have been combined into one tree structure by Monnase. A complete discussion of his construction is too detailed and would distract from the main line of reasoning, but a few comments are essential. Ignore rigor for a moment and consider a bounded connected component, $A$, of an upper level set. The set $\mathbb{R}^2 \setminus A$ can be decomposed into its connected components, and the bounded connected components of $\mathbb{R}^2 \setminus A$ are surrounded by the set $A$. These sets will be called holes. The set obtained by taking the union of $A$ with all of its holes forms a new set. Following Monnasse’s dissertation, these sets will be called shapes. The shapes derived from the upper and lower level sets can be partially ordered using set inclusion, and the domain
of definition of $h$ can be used as the root of a tree structure. The tree structure obtained by this procedure is called the FLST tree structure. A rigorous approach to the building of this tree structure is given in Monasse’s dissertation [36].

Figure 1 demonstrates the tree of shapes obtained from the lower level set, the upper level set, and the FLST where lighter colors represent larger gray level values. All three constructions are ordered by set containment, but the FLST shapes do not have any holes.

The FLST tree structures for functions defined on a continuous domain can be constructed using the procedure outlined above, but images and histograms are defined on discrete domains. An image or histogram can be extended to the continuous domain using interpolation methods, and the interpolation method has consequences in the FLST tree construction. The interpolation techniques used by Monasse result in an upper semicontinuous function, and this interpolation scheme produces a simple relationship between connected sets in the discrete and continuous domains. In particular, upper level sets are eight connected while lower level sets are four connected.

There are many details not included in this discussion necessary for a rigorous approach to building the tree of shapes. The delicate aspect of the algorithm is deciding the connectedness of the shape and determining if there are holes in the shape. Rosenfeld showed that the number of holes can be counted using information derived from a pixel’s neighborhood [38]. This local knowledge is sufficient to determine the number of holes after the whole region is grown. The thesis by Monasse discusses this issue and provides a rigorous development of the topic and an effective algorithm [36].

2.2. Meaningful modes. The Helmholtz principle states that a group of features is “meaningful” whenever, given an a-contrario model, the group of features is unlikely to occur due to the a-contrario model. The works of Cao, Delon, Desolneux, Morel, Muse, and Sur [5, 9, 10, 11, 12, 13, 15, 18, 20, 14, 16, 19, 21] develop a mathematical procedure to find perceptual features based on the Helmholtz principle. Their implementation of the Helmholtz principle requires three coupled ingredients. First, an a-contrario background probability model must be defined. This background model can be given or learned from the data. Second, a procedure to group independent and identically distributed (IID) random variables must be given. The background model and the grouping procedure must work in conjunction to describe the absence of the feature of interest. Third, a function mapping the group of IID random variables to the real numbers, the NFA, must be defined. The NFA must be an upper bound on the expected number of feature groups that would occur due to the a-contrario model. The NFA must also scale with the sample size and the grouping procedure. Developing these three ingredients for to 2-D histogram analysis will be discussed in more detail.

A sample space, a set of events, and a probability function are the essential elements needed to describe the a-contrario probability model. The definition of such a space without any reference to the experiment it describes obscures the relationships between the probability model, the grouping procedure, and the NFA. For this reason, a constructive approach will be used to develop the probability model.

The classic probability model of randomly placing $M$ indistinguishable objects into $N$ bins is a natural model for two dimensional histogram creation. If each object is equally likely to be placed in any bin, then the probability of a ball ending
up in a given bin is $1/N$. Since the balls are indistinguishable, the simple events correspond to the $N$-tuples $(n_1, n_2, \ldots, n_N)$ where $n_i$ is the number of balls in the $i$th bin. The model will be completed by deriving the probabilities for the compound events corresponding to the grouping procedure.

The grouping procedure must work in conjunction with the $a$-contrario hypothesis model to describe a feature’s absence. Fundamentally, the grouping procedure identifies compound events that accurately characterize the feature. The features of interest in this work are modes of a histogram, and, as stated in the introduction, the modes of a histogram can be qualitatively understood as connected sets in the histogram’s domain where the data concentrates. This qualitative notion of modes suggests the following compound event. Consider a histogram $h : \mathcal{N} \to \mathbb{Z}$ where $\mathcal{N} = \{(1,1), (1,2), \ldots, (L_1, L_2)\}$ is the histogram domain. Let $I \subset \mathcal{N}$ be a connected set, and define

$$k(I) = \sum_{x \in I} h(x)$$

The set function $k(I)$ is the number of objects in $I$. The notation $|I|$ will represent the number of elements in the set $I$. The probability that at least $k(I)$ elements are in the connected region $I$ is given by the binomial tail

$$B(M, k(I), p(I)) = \sum_{j=k(I)}^{M} \binom{M}{j} p(I)^j (1-p(I))^{M-j},$$

where $p(I) = |I|/N$. This model is a Bernoulli scheme with $M$ trials and a probability of success $p(I)$.

Briefly, let $R$ denote the family of all connected sets in the histogram domain $\mathcal{N}$. This definition of $R$ will be modified in the sequel. Let $|R|$ equal the number of elements in $R$. The number of false alarms ($NFA$) of the set $I \in R$ is given by the following definition.

**Definition 2.2.** ($NFA$ definition 1) Let $I \in R$, $p(I) = |I|/N$, and let $k(I)$ be the observed number of objects in $I$. The number of false alarms of $I$ is

$$NFA(I) = |R| B(M, k(I), p(I)).$$

This definition leads directly to the notion of an $\epsilon$-meaningful mode.

**Definition 2.3.** A set $I \in R$ is an $\epsilon$-meaningful mode if and only if $k(I)/M > p(I)$ and

$$NFA(I) < \epsilon.$$ 

When $\epsilon = 1$ the set $I$ will be called a meaningful mode. The condition $k(I)/M > p(I)$ is present since a mode should contain more elements than is expected.

A driving force for mode detection using the Helmholtz principle is a natural interpretation of the $NFA$. The $NFA$ serves as an upper bound on the expected number of occurrences for the set $I \in R$ given the $a$-contrario model. Furthermore, the $|R|$ scale factor implies that the $NFA$ scales appropriately with the number of histogram bins. This statement is made precise in the following theorem proven in the appendix.

**Theorem 2.4.** The expected number of $\epsilon$-meaningful modes is less than or equal to $\epsilon$. 
The above definition of \( NFA \) suffers from two computational difficulties. First of all, the binomial tail is computationally expensive to calculate. This difficulty is being mitigated through the use of the tight Hoeffding upper bound on the binomial tail [22]. This bound is given by

\[
B(M, k(I), p(I)) \leq \exp \left[ -M \left( r(I) \log \frac{r(I)}{p(I)} + (1 - r(I)) \log \frac{1 - r(I)}{1 - p(I)} \right) \right],
\]

where \( r(I) = k(I)/M \). Define the function

\[
H(r, p) = r \log \frac{r}{p} + (1 - r) \log \frac{1 - r}{1 - p}.
\]

The Hoeffding inequality motivates the following definition.

**Definition 2.5.** (NFA definition 2) Let \( I \in R, p(I) = |I|/N \), and let \( k(I) \) be the observed number of objects in \( I \). Suppose \( r(I) = k(I)/M \), then the number of false alarms of \( I \) is

\[
NFA(I) = |R|e^{-MH(r(I), p(I))}.
\]

The second difficulty with definition 2.2 is subtler, and it will require the definition of \( R \) to be changed. A quick computation reveals that the number of connected sets in the histograms domain is large for large \( N \). Assume that the histogram has \( L_1 \) rows and \( L_2 \) columns, then the number of rectangular regions in the histogram is \( L_1L_2(L_1 + 1)(L_2 + 1)/4 \). The number of connected regions will be substantially greater. A search of all the rectangular regions in a histogram of \( L_1 = L_2 = 100 \) bins would therefore consist of at least \( L_1^2(L_1 + 1)^2/4 \approx 2.5 \times 10^7 \) regions to search. Performing all the computations would be a time prohibitive procedure. This computational dilemma will be mitigated through the use of the FLST tree.

As shown in section 2.1, the FLST constructs a topographic map that combines the interpolated histogram’s upper and lower level sets. The idea is to only search the histogram’s FLST shapes. In effect, the FLST will guide the search for meaningful modes. There are no known estimates of the number of shapes in a FLST tree, but in all experiments performed by the authors, a 100x100 histogram contains less than 500 shapes. This is several orders of magnitude less than the above estimate. The FLST tree, however, does not consist of connected sets, but rather connected sets without holes. The definition of \( R \) needs to be modified to take this change into account. Hereafter and in Definitions 2.2, 2.3, and 2.5, \( R \) will be the set of all connected sets without holes.

The preceding Helmholtz principle concepts can be extended to the case of Bernoulli trials with more than two outcomes, and this extension will play an essential role in Section 3. The following notational conventions will be used. The 1-norm of a non-negative \( q \)-tuple \( \mathbf{w} \) will be denoted by \( |\mathbf{w}| = w_1 + w_2 + \cdots + w_q \). Recall that the number of ways in which a population of \( M \) elements can be divided into \( q + 1 \) subpopulations is given by the multinomial coefficient

\[
C^M_k = \frac{M!}{k_1!k_2!\cdots k_q!(M - |k|)!},
\]

where \( k = (k_1, k_2, \ldots, k_q) \) is a \( q \)-tuple of non-negative integers subject to the condition \( |k| < M \).

Let \( \{G_i \in R | i = 1, 2, \ldots, q; G_i \cap G_j = \emptyset \text{ if } i \neq j\} \) be \( q \) mutually disjoint sets, then the set functions \( k_i = k(G_i) \) and \( p_i = p(G_i) \) are well defined. Furthermore,
the \( q \)-tuples \( \mathbf{k} = (k_1, k_2, \ldots, k_q) \) and \( \mathbf{p} = (p_1, p_2, \ldots, p_q) \) satisfy the relations
\[
|\mathbf{k}| < M, \\
p_i \geq 0 \quad \forall i \quad \text{and} \quad \sum_{i=1}^{N} p_i \leq 1.
\]
(12)

The probability that in \( M \) trials there are \( k_1 \) elements in \( G_1 \), \( k_2 \) elements in \( G_2 \), etc., is the multinomial distribution
\[
m_q(M, \mathbf{k}, \mathbf{p}) = C_M^{k_1} p_1^{k_1} p_2^{k_2} \cdots p_q^{k_q} (M - |\mathbf{p}|)^{M - |\mathbf{k}|};
\]
thus the probability of at least \( k_1 \) elements in \( G_1 \), \( k_2 \) elements in \( G_2 \) etc., is given by the multinomial tail
\[
\mathcal{M}(M, \mathbf{k}, \mathbf{p}) = \sum_{x_1=k_1}^{M} \sum_{x_2=k_2}^{M-x_1} \cdots \sum_{x_q=k_q}^{M-x_1-\cdots-x_{q-1}} m_q(M, \mathbf{x}, \mathbf{p}).
\]
(14)

Whenever \( r \) and \( p \) are 1-tuples, the multinomial tail \( \mathcal{M}(M, k, p) \) is equal to the binomial tail \( B(M, k, p) \).

The shorthand \( \{G_i \in R \mid i = 1, 2, \ldots, q; G_i \cap G_j = \emptyset \text{ if } i \neq j\} = \{G_i\}_{i=1}^{q} \) will be used to simplify notation. Also, to limit confusion, the family of sets \( \{G_i\}_{i=1}^{q} \) will be called a \( q \)-family of disjoint sets whenever \( q \) is a fixed number. In other words, all \( q \)-family of disjoint sets must consist of exactly \( q \) mutually disjoint non-empty sets. The multinomial tail motivates the following definition of the number of false alarms for a \( q \)-family of disjoint sets:

**Definition 2.6.** (NFA definition 3) Let \( \{G_i\}_{i=1}^{q} \) be a \( q \)-family of disjoint sets, \( p(G_i) = |G_i|/N \), and let \( k(G_i) \) be the observed number of objects in \( G_i \). Suppose
\[
\mathbf{k} = (k(G_1), k(G_2), \ldots, k(G_q)), \\
\mathbf{p} = (p(G_1), p(G_2), \ldots, p(G_q)),
\]
(15)
then the number of false alarms of \( \{G_i\}_{i=1}^{q} \) is
\[
\text{NFA}(\{G_i\}_{i=1}^{q}) = (M + 1)^{q-1} \binom{|R|}{q} \mathcal{M}(M, \mathbf{k}, \mathbf{p}).
\]
(16)

This definition is the same as 2.2 when \( q = 1 \). A couple of comments on the scale factor are required. The \( (M + 1)^{q-1} \) term arises because the multinomial is a multi-dimensional probability distribution. The proofs in the Appendix elucidate its presence. The multiplicative factor \( \binom{|R|}{q} \) is the number of \( q \)-families of sets, not necessarily disjoint, chosen from \( R \). This term over counts the total possible number of \( q \)-families of disjoint sets. The NFA, however, is still an upper bound on the expected number of false alarms, and the experiments in Section 5 also indicate that this over counting is not critical.

The definition of an \( \epsilon \)-meaningful mode follows *mutatis mutandis* for an \( \epsilon \)-meaningful \( q \)-family of disjoint sets.

**Definition 2.7.** Let \( \{G_i\}_{i=1}^{q} \) be a \( q \)-family of disjoint sets, \( p(G_i) = |G_i|/N \), and let \( k(G_i) \) be the observed number of objects in \( G_i \). Suppose
\[
\mathbf{r} = (k(G_1), k(G_2), \ldots, k(G_q)) \frac{1}{M}, \\
\mathbf{p} = (p(G_1), p(G_2), \ldots, p(G_q)),
\]
(17)
then the $q$-family $\{G_i\}_{i=1}^q$ is an $\epsilon$-meaningful family of modes if and only if $p < r$
and
\begin{equation}
NFA(\{G_i\}_{i=1}^q) < \epsilon.
\end{equation}

The following theorem, of which theorem 2.4 is a special case, will be proven in the appendix.

**Theorem 2.8.** The expected number of $\epsilon$-meaningful $q$-families of disjoint sets is less than $\epsilon$.

Multinomial tail calculations are more computationally expensive than their binomial equivalent. For computational ease, the $NFA$ of a $q$-family of sets is given below in terms of an apparently newly generalized Hoeffding bound for the Multinomial. This upper bound can be rewritten as the exponential of a relative entropy function, and using the definition $r = k/M$ it is given by
\begin{equation}
M(M, k, p) \leq \exp(-MH_q(r, p))
\end{equation}
where
\begin{equation}
H_q(r, p) = \sum_{i=1}^q r_i \log \frac{r_i}{p_i} + (1 - |r|) \log \frac{1 - |r|}{1 - |p|}.
\end{equation}

The subscript $q$ of a subscripted function (e.g. $H_1(r_1, p_1)$, $H_2(r_1, r_2, p_1, p_2)$) identifies a particular $q$-state relative entropy function. Two qualities of the function $H_q(r, p)$ are worth noting. By the log sum inequality [8], the functions $H_q(r, p)$ are jointly convex in both variables. The functions $H_q(r, p)$ for $q > 1$ are symmetric under the interchange action (e.g. $H_2(r_1, r_2, p_1, p_2) = H_2(r_2, r_1, p_1, p_2)$) while $H_1(r_1, p_1)$ is symmetric about $1/2$. (The proof of this upper bound will be delayed until Section 2.3.

The following definition is a generalization of Definition 2.5.

**Definition 2.9.** ($NFA$ definition 4) Let $\{G_i\}_{i=1}^q$ be a $q$-family of disjoint sets, $p(G_i) = |G_i|/N$, and let $k(G_i)$ be the observed number of objects in $G_i$. Suppose
\begin{equation}
\begin{array}{c}
r = (k(G_1), k(G_2), \ldots, k(G_q)) \frac{1}{M}, \\
p = (p(G_1), p(G_2), \ldots, p(G_q)),
\end{array}
\end{equation}
then the number of false alarms of $\{G_i\}_{i=1}^q$ is
\begin{equation}
NFA(\{G_i\}_{i=1}^q) = (M + 1)^{q-1} \binom{|R|}{q} e^{-MH_q(r, p)}.
\end{equation}

This definition is used in all subsequent analysis.

The Helmholtz principle has been formulated above for a single set and for a $q$-family of disjoint sets. Also, as mentioned in the text, the $q$-family formulation reduces to the single set formulation when $q = 1$. As a consequence, there is a redundancy between many of the definitions and theorems. This redundancy is intentional since the $NFA$ of a single set will play a different role than the $NFA$ of a $q$-family of disjoint sets. The different roles will be discussed more thoroughly in Section 3 when the notion of a maximal $\epsilon$-meaningful node is given. A brief discussion on potential algorithms to find modes using the Helmholtz principle will highlight the rational behind this disaffiliation.

A direct application of the Helmholtz principle would calculate the family of disjoint sets $\{G_i\}_{i=1}^q$ that minimizes the $NFA$. A simple estimate will highlight the
impracticality of such a procedure. The number of \( q \)-families of sets, not necessarily disjoint, is given by the binomial term \( \binom{|R|}{q} \). The sum of this term over \( 1 \leq q \leq M \) is \( 2^{|R|} \); thus a simple estimate on the total number of families of sets to search is given by substituting the number of rectangular regions \( L_1 L_2 (L_1 + 1) (L_2 + 1) / 4 \) for \( |R| \). A histogram consisting of \( N = 25 \) bins consists of approximately \( 1.0863 \times 10^{17} \) searches. Performing such a search is an impractical task. This estimate is a rough approximation since the term \( \binom{|R|}{q} \) is an upper bound on the number of sets while the number of rectangular regions in \( R \) is a lower bound. A more accurate estimate, however, would not make this procedure practical. Instead of searching all possible sets, the topological tree structure created using the FLST tree will be utilized to guide the search for meaningful modes. The details of this methodology will be presented in section 3.

In the next two sections, Sections 2.3 and 2.4, a proof of the generalized Hoeffding upper bound of equation 20 is given and a relation between the ordering of \( H_1(r, p) \) and \( H_q(r, p) \) is discussed. This relation will be used to simplify the algorithms, but both sections can be omitted upon a first reading.

2.3. Generalizing the Hoeffding upper bound. A common requirement throughout all applications by Desolneux et al. is the need for repeated computation of the tail probabilities for the binomial and trinomial distributions [15, 18, 14, 16, 19]. For the binomial distribution, the Chernoff-Hoeffding inequality provides a reasonably tight and mathematically useful relative entropy upper bound, but the authors are unaware of a Hoeffding non-asymptotic type bound for the multinomial distribution. Such a bound simplifies all computations.

The following theorems show that

\[
U(r, p) = \frac{p_1^{k_1} p_2^{k_2} \cdots p_q^{k_q} (1 - |p|)^{M - |k|}}{r_1^{k_1} r_2^{k_2} \cdots r_q^{k_q} (1 - |r|)^{M - |k|}}
= \frac{m_q(M, k, p)}{m_q(M, k, r)}
\]

(23)

are upper bounds to \( \mathcal{M}(M, k, p) \), and the choice \( r = k/M \) minimizes this upper bound.

**Theorem 2.10.** Let \( M \) denote the number of trials, and suppose \( r \) and \( p \) are \( q \)-tuples. If \( 0 < p \leq r \) and \( |r| < 1 \), then

\[
\mathcal{M}(M, k, p) \leq U(r, p)
\]

and \( r^* = k/M \) minimizes \( U(r, p) \).

**Proof.** Define \( G(M, k, p) = \mathcal{M}(M, k, p) / m_q(M, k, p) \). The function \( G(M, k, p) \) can be written as

\[
G(M, k, p) = \frac{1}{\binom{|k|}{k}} \sum_{x_1=k_1}^{M-x_1} \cdots \sum_{x_q=k_q}^{M-x_q} \binom{M}{x} p_1^{x_1-k_1} \cdots p_m^{x_m-k_m} \times (1 - |p|)^{|k|-|x|}
\]

(25)

A simple calculus exercise demonstrates that this equation is an increasing function in \( p_1, p_2, \ldots, p_m \) because the exponents \( x_1 - k_1, \ldots, x_m - k_m \) and \( |x| - |k| \) are all
nonnegative. This observation combined with the inequality \( p \leq r \) motivates the second relation in the following set of equations:

\[
\mathcal{M}(M, k, p) = \frac{m_q(M, k, p)}{m_q(M, k, p)}
\]

\[
\leq \frac{M(M, k, r)}{m_q(M, k, r)} m_q(M, k, p)
\]

\[
\leq \frac{M(M, 0, r)}{m_q(M, k, r)} m_q(M, k, p)
\]

\[
= \frac{m_q(M, k, p)}{m_q(M, k, r)}.
\]

(26)

The last two relations are true since \( \mathcal{M}(M, k, r) \) is the tail of the multinomial distribution.

The minimal value is derived by noticing that the multinomial ratio bound holds for any \( r \in [0,1]^q \) such that \( p \leq r \). It is a simple calculus exercise to show that the gradient of \( U(r, p) \) with respect to \( r \) is zero at \( r^* = k/M \). Moreover for fixed \( p \) the map \( r \to U(r, p) \) is convex in \( r \). Thus the global minimizer of \( U(r, p) \) occurs at \( r^* = k/M \).

The generalized Hoeffding bound is much tighter than the Joag-Dev/Proschain product bound typically used for multinomial random variables. Their product bound is appropriate since the random variables of a multinomial distribution are negatively associative; therefore, for the \( q \)-tuples \( p \) and \( r \) that satisfy \( p < r \), this bound can be written as [26]

\[
\mathcal{M}(M, k, p) \leq \prod_{i=1}^q B(M, k_i, p_i).
\]

(27)

Combining the Hoeffding inequality and the previous equation yields the result

\[
\mathcal{M}(M, k, p) \leq \frac{p_1^{k_1} (1-p_1)^{M-k_1} \cdots p_q^{k_q} (1-p_q)^{M-k_q}}{r_1^{k_1} (1-r_1)^{M-k_1} \cdots r_q^{k_q} (1-r_q)^{M-k_q}}.
\]

(28)

The next theorem shows that the ratio bound \( U(r, p) \) is less than or equal to this product bound for a multinomial random variable.

**Theorem 2.11.** Suppose \( M \) is the number of trials, \( p \) and \( k \) are \( q \)-tuples, and define \( r = k/M \). Assume \( 0 < p \leq r \) and \( |r| < 1 \) then

\[
\frac{p_1^{k_1} \cdots p_q^{k_q} (1-|p|)^{M-|k|}}{r_1^{k_1} \cdots r_q^{k_q} (1-|r|)^{M-|k|}} \leq \frac{p_1^{k_1} (1-p_1)^{M-k_1} \cdots p_q^{k_q} (1-p_q)^{M-k_q}}{r_1^{k_1} (1-r_1)^{M-k_1} \cdots r_q^{k_q} (1-r_q)^{M-k_q}},
\]

and the inequality is strict when \( p_i < r_i \) for some \( i \).

**Proof.** Define

\[
f(k, p) = \frac{p_1^{k_1} p_2^{k_2} \cdots p_q^{k_q} (1-|p|)^{M-|k|}}{p_1^{k_1} (1-p_1)^{M-k_1} \cdots p_q^{k_q} (1-p_q)^{M-k_q}}
\]

(29)

\[
= \frac{(1-|p|)^{M-|k|}}{(1-p_1)^{M-k_1} \cdots (1-p_q)^{M-k_q}}.
\]

(30)

See Inverse Problems and Imaging Volume 2, No. 4 (2008), 485–525
The function $f$ must be shown to be an increasing function of $p$. This is equivalent
to showing that $\log f(k, p)$ is increasing in $p$. Define $g(k, p) = \log f(k, p)$ then
$g(k, p)(\#\{M - |k|\} \log (1 - |p|)) - (M - k_1) \log (1 - p_1) - \ldots - (M - k_q) \log (1 - p_q)$.

Use $|k| = k_1 + \ldots + k_q$, $|p| = p_1 + \ldots + p_q$, $p_i \leq r_i = \frac{k_i}{M} = p_i + \epsilon_i$ to get
\[
\frac{\partial g}{\partial p_i} = \frac{M - k_1 - \ldots - k_q}{1 - p_1} - \frac{M - k_1 - \ldots - k_q}{1 - p_1 - \ldots - p_q} \\
= M \left( \frac{1 - r_1}{1 - p_1} - \frac{1 - r_1 - \ldots - r_q}{1 - p_1} \right) \\
= M \left( \frac{1 - p_1 - \epsilon_1}{1 - p_1} - \frac{1 - p_1 - \ldots - p_q - \epsilon_1 - \ldots - \epsilon_q}{1 - p_1 - \ldots - p_q} \right) \\
= M \left( \frac{(1 - p_1)(\epsilon_2 + \ldots + \epsilon_q) + \epsilon_1(p_2 + \ldots + p_q)}{(1 - p_1)(1 - p_1 - \ldots - p_q)} \right)
\]

(32)

The same argument shows that $\frac{\partial g}{\partial p_i} \geq 0$ for $i > 1$. The inequality is strict if any
$\epsilon_i$ is greater than zero; therefore the inequality is strict if $p_i < r_i$ for some $i$. $\square$

Table I shows that the standard product bounds can be several orders of magni-
tude less accurate than the true tail sum bound. In contrast, the ratio bounds are
approximately within a factor of ten.

| $M$ | $k$ | True Tail Sum | Product Bound | Ratio Bound |
|-----|-----|---------------|---------------|------------|
| 10  | 4   | $3.7 \times 10^{-3}$ | $1.2 \times 10^{-1}$ | $3.5 \times 10^{-2}$ |
| 20  | 9   | $9.9 \times 10^{-4}$  | $1.7 \times 10^{-3}$ | $1.7 \times 10^{-5}$ |
| 30  | 14  | $1.8 \times 10^{-10}$ | $2.5 \times 10^{-6}$ | $4.0 \times 10^{-9}$ |
| 40  | 19  | $2.9 \times 10^{-14}$ | $2.7 \times 10^{-9}$ | $7.6 \times 10^{-13}$ |

Table 1. Multinomial with 2 categories: $k_1 = k_2 = k$,
$p_1 = p_2 = 0.2$, $r_i = k_i/M$.

2.4. Family of sets. Just as the multinomial distribution can be factored into a
product of a binomial and conditional multinomial distribution, the ratio bound
can be factored resulting in a recursive sum of relative entropies [27, 28]; therefore
the ratio bound preserves the binomial factoring. This factoring helps demonstrate
that the entire family of subsets almost always has greater relative entropy than a
smaller collection of the sets. This observation will be made more precise below,
and it will greatly reduce the number of tests needed in order to find the maximal
$\epsilon$-meaningful modes defined in section 3.

Theorem 2.12. Suppose $M$ is the number of trials, $p$ and $k$ are $q$-tuples, and
define $r = k/M$. Let $\tilde{r} = (r_1, r_2, \ldots, r_{q-1})$ and $\tilde{p} = (p_1, p_2, \ldots, p_{q-1})$ denote the
$(q - 1)$-tuples obtained by deleting $q^{th}$ component. If $0 < p < r$ and $|r| < 1$, then
\[
H_q(r, p) = H_{q-1}(\tilde{r}, \tilde{p}) + (1 - |\tilde{r}|)H_1 \left( \frac{r_q}{1 - |\tilde{r}|}, \frac{p_q}{1 - |\tilde{p}|} \right).
\]

(33)

Inverse Problems and Imaging

Volume 2, No. 4 (2008), 485–525
Proof. The ratio bound can be factored as follows
\[
\frac{\tilde{p}_1^{k_1} \tilde{p}_2^{k_2} \cdots \tilde{p}_q^{k_q} (1 - |\tilde{p}|)^{M - |k|}}{r_1^{k_1} r_2^{k_2} \cdots r_q^{k_q} (1 - |r|)^{M - |k|}} = \left( \frac{\tilde{p}_1^{k_1} \tilde{p}_2^{k_2} \cdots \tilde{p}_q^{k_q}}{r_1^{k_1} r_2^{k_2} \cdots r_q^{k_q}} \right) \left( \frac{1 - \tilde{p}_q/(1 - |\tilde{p}|)}{1 - r_q/(1 - |r|)} \right)^{M - |k|}
\]
\[
\times \left( \frac{1 - |\tilde{p}|}{1 - |r|} \right)^{M - |k|}
\]
\[
= \left( \frac{p_1^{k_1} p_2^{k_2} \cdots p_{q-1}^{k_{q-1}} (1 - |\tilde{p}_{q-1}|)^{M - |k_{q-1}|}}{r_1^{k_1} r_2^{k_2} \cdots r_{q-1}^{k_{q-1}} (1 - |r_{q-1}|)^{M - |k_{q-1}|}} \right)
\]
\[
\times \left( \frac{p_q/(1 - |\tilde{p}_q|)}{r_q/(1 - |r_q|)} \right)^{k_q} \left( \frac{1 - p_q/(1 - |\tilde{p}_q|)}{1 - r_q/(1 - |r_q|)} \right)^{M - |k|}
\]
(34)
Using the rules of logarithms, these factors can be converted to yield the relative entropy sum.
\[
\text{Theorem 2.13. Let } \{G_i\}_{i=1}^q \text{ be a } q\text{-family of disjoint sets, } p(G_i) = |G_i|/N, \text{ and let } k(G_i) \text{ be the observed number of objects in } G_i. \text{ Suppose}
\]
\[
r = (k(G_1), k(G_2), \ldots, k(G_q)) \frac{1}{M},
\]
(35)
\[
p = (p(G_1), p(G_2), \ldots, p(G_q)).
\]
Let \(\tilde{r} = (r_1, r_2, \ldots, r_{q-1})\) and \(\tilde{p} = (p_1, p_2, \ldots, p_{q-1})\) denote the \((q-1)\)-tuples obtained by deleting the \(q\)th component. If
\[
\frac{\log(|r| + 1) + \log(|r| + 1 - q) - \log(q)}{M} < H_q(r, p) - H_{q-1}(\tilde{r}, \tilde{p})
\]
then
\[
NFA(\{G_i\}_{i=1}^q) < NFA(\{G_i\}_{i=1}^{q-1}).
\]
(37)
Proof. Definition 2.9 of NFA(\(\{G_i\}_{i=1}^q\)) implies that Equation (37) is true if
\[
\frac{NFA(\{G_i\}_{i=1}^q)}{NFA(\{G_i\}_{i=1}^{q-1})} = \frac{(|R| + 1)(|R| + 1 - q)}{q} e^{-M[H_q(r, p) - H_{q-1}(\tilde{r}, \tilde{p})]}
\]
is less than 1. Taking the natural logarithm and rearranging terms produces
\[
\frac{\log(M + 1) + \log(|R| + 1 - q) - \log(q)}{M} < H_q(r, p) - H_{q-1}(\tilde{r}, \tilde{p}).
\]
This is the same as Equation 36.
\[
\text{The last two theorems illustrate that removing one set from the } q\text{-family } \{G_i\}_{i=1}^q \text{ of disjoint sets almost always results in an increase in the NFA in practical histogram applications. Theorem 2.13 asserts that the NFA ordering in Equation 37 is almost always implied by the ordering between } H_q(r, p) \text{ and } H_{q-1}(\tilde{r}, \tilde{p}) \text{ for large } M. \text{ This claim can be further understood by noting that the left hand side of Equation 36 depends on } M \text{ and } |R| \text{ while the right hand side does not. Moreover, } |R| \text{ is a function of the histogram domain and does not depend on the number of samples } M. \text{ As shown in Section 3.1, an upper bound on } |R| \text{ in terms of the number of bins } N \text{ is } |R| < 2^N. \text{ As a consequence, the left hand side of Equation 36 decreases with}
\]
\text{Inverse Problems and Imaging}
\text{Volume 2, No. 4 (2008), 485-525}
increasing $M$ faster than $cN/M$ for large $M$ and some constant $c$. Furthermore, in a typical histogram application, there will be many more samples than there are bins; thus $N/M << 1$.

Theorem 2.12 demonstrates that the difference $H_q(r, p) - H_{q-1}(\bar{r}, \bar{p})$ is a positive quantity. Combining the last two observations indicates that removing one set almost always increases the NFA for $M >> N$. This comment will be employed to simplify the algorithm in Section 3.1.

3. **Maximal $\epsilon$-meaningful modes.** The collection of all $\epsilon$-meaningful modes typically contain many sets that overlap, but, to maintain coherence with the notion of a mode, a reduced set of disjoint $\epsilon$-meaningful modes needs to be identified. This issue of overlapping $\epsilon$-meaningful sets occurs in all previous applications of the Helmholtz principle. It is typically resolved using a maximal $\epsilon$-meaningful containment condition. This containment condition locates disjoint $\epsilon$-meaningful sets that are not contained in and do not contain sets with a lower NFA. The meaningful grouping procedure of Cao et.al [5], however, added another merging criteria to verify if two $\epsilon$-meaningful sets should be merged into one or kept separate. The discussion in [5] demonstrates the necessity of this added test, and the same conditions can exist in the algorithm presented below. For this reason, an $\epsilon$-meaningful mode will have to satisfy the containment conditions as well as a merging criteria. Before the maximal $\epsilon$-meaningful definition is given, however, a short discussion of the interplay between the FLST tree structure and the notion of $\epsilon$-meaningful will be presented.

Throughout all applications of the maximal principle to image analysis, events are ordered using the NFA. A clear interpretation exists for this ordering, namely, given that the image follows the a-contrario hypothesis model, the lower the NFA of an event the less the expected number of such events. As mentioned in Section 2.2, a search for a disjoint set of modes with a minimal NFA is computationally unreasonable. For this reason, the FLST tree structure is used to guide the search for maximal $\epsilon$-meaningful modes. The following two definitions will simplify the discussion.

**Definition 3.1.** Let $I \in R$ be a FLST tree structure node. The FLST tree structure node $G \in R$ is a descendant of $I$ if the shortest path from $G$ to the root crosses $I$.

**Definition 3.2.** Let $I \in R$ be a FLST tree structure node. The FLST tree structure node $G \in R$ is an ascendant of $I$ if the shortest path from $I$ to the root crosses $G$.

Section 2.1 notes that two aspects of the FLST tree structure will be utilized. To reiterate, at a fixed gray level, the FLST tree structure partitions the image into connected non-overlapping regions, and as the gray level changes, the connected components are ordered in a containment tree structure. These two observations indicate that the maximal $\epsilon$-meaningful containment conditions can be met by comparing the NFA of a set $I$ with the NFA of its ascendants and descendants. The merging criteria poses a more complex problem.

Many solutions to the merging problem that exploit the FLST tree structure exist. The solution presented below is a compromise between computational efficiency and completeness. The most obvious solution would be to consider all possible collections of descendants. A node would then be labeled as not maximal $\epsilon$-meaningful when the NFA of the node is less than the NFA of any family of descendant disjoint sets. Again, the computational burden of such a procedure would be too large.
Instead, a maximal meaningful set will be defined in terms of any \( q \)-family of disjoint sets such that all the sets are siblings.

In summary, maximal \( \epsilon \)-meaningful sets will be found using the FLST tree structure, and these sets will satisfy three conditions: they will be \( \epsilon \)-meaningful, they will satisfy a containment condition and they will satisfy a merging condition. The merging condition will be made more precise using the following two definitions.

**Definition 3.3.** Let \( G \in R \). The set \( G \) is indivisible with respect to a \( q \)-family of sets \( \{G_i\}_{i=1}^q \) if and only if \( G_i \subset G \) for all \( i \) and

\[
NFA(G) \leq NFA(\{G_i\}_{i=1}^q).
\]

The FLST tree structure and the notion of indivisibility motivates the definition of maximal \( \epsilon \)-meaningful modes. Each non-leaf node of the FLST has one or more children; therefore for each non-leaf node there exists a \( q \) such that the collection of all children form a \( q \)-family of disjoint sets. This topological structure of the FLST tree inspires the following definition of a family of sibling descendants of a tree node \( G \).

**Definition 3.4.** Let \( I \in R \) be a FLST tree structure node. The \( q \)-family of disjoint sets \( \{G_i\}_{i=1}^q \) is a family of sibling descendants of \( I \) if there exists a FLST tree structure node \( H \subset I \) such that \( G_i \) is a child of \( H \) for all \( 1 \leq i \leq q \).

The notions of indivisible, ascendent and descendent allow for the following definition of a maximal \( \epsilon \)-meaningful mode.

**Definition 3.5.** A FLST tree structure node \( G \in R \) is a maximal \( \epsilon \)-meaningful mode if and only if \( r(G) > p(G) \) and

1. \( NFA(G) < \epsilon \).
2. \( G \) is indivisible with respect to any family of sibling descendants of \( G \).
3. For all indivisible descendants \( I \), \( NFA(I) \geq NFA(G) \).
4. For all indivisible ascendants \( J \), \( NFA(J) > NFA(G) \) or there exists an indivisible descendants \( I \) of \( J \) such that \( NFA(I) < NFA(J) \).

The next lemma shows that two maximal \( \epsilon \)-meaningful modes are indeed disjoint.

**Lemma 3.6.** Let \( G_1 \) and \( G_2 \) be two maximal \( \epsilon \)-meaningful FLST tree structure nodes, then \( G_1 \) and \( G_2 \) are disjoint.

**Proof.** Suppose \( G_1 \) and \( G_2 \) are maximal \( \epsilon \)-meaningful modes and \( G_1 \cap G_2 = J \) with \( J \neq \emptyset \), then by the FLST tree structure \( G_1 \subset G_2 \) or \( G_2 \subset G_1 \). (This is proposition 2.4 in [36]). Without loss of generality assume that \( G_1 \subset G_2 \). Note that, by the definition of maximal \( \epsilon \)-meaningful, \( G_1 \) is indivisible with respect to any family of sibling descendants of \( G_1 \) and \( G_2 \) is indivisible with respect to any sibling descendants of \( G_2 \). Since \( G_1 \subset G_2 \), then \( G_1 \) is a descendant of \( G_2 \). There are three possible cases:

1. \( NFA(G_1) > NFA(G_2) \),
2. \( NFA(G_2) > NFA(G_1) \), and
3. \( NFA(G_1) = NFA(G_2) \).

Case 2 contradicts item three in definition 3.5 of maximal \( \epsilon \)-meaningful. Cases 1 and 3 contradict item four in definition 3.5 of maximal \( \epsilon \)-meaningful. Therefore \( G_1 \cap G_2 = \emptyset \). \( \square \)
3.1. The meaningful mode algorithm. The algorithm to detect maximal \( \epsilon \)-meaningful modes is a direct application of the definitions given above, but two computational issues need to be addressed before the algorithm is stated. First, the value used for \(|R|\) will be discussed. The FLST transform motived defining \( R \) as the family of sets without holes. A direct count of the number of elements in this family is difficult to obtain. A lower bound on \(|R|\) was used in this work, and the effect of using this lower bound is discussed. Second, the algorithm was simplified using the results in section 2.4. More information on this issue follows.

The quantity \(|R|\) is used in two different situations. It is used when a set is determined \( \epsilon \)-meaningful and when a set \( G \) is determined to be indivisible with respect to a \( q \)-family of sets. The consequences on the definition of \( \epsilon \)-meaningful modes when \(|R|\) is changed will be discussed first. Definition 2.3 for a \( \epsilon \)-meaningful mode clearly demonstrates that changing \(|R|\) is equivalent to changing \( \epsilon \). In this paper the lower bound

\[
|R| = \frac{L_1(L_1 + 1)L_2(L_2 + 1)}{4}
\]

was used for the size of \(|R|\). The relationship between \( \epsilon \) and \(|R|\) implies that by using a lower bound more modes will be determined meaningful. Furthermore, figure 5, discussed in more detail in section 5.1, indicates that this choice of \(|R|\) is consistent with the notion of \( \epsilon \)-meaningful given in Theorem 2.4.

The quantity \(|R|\) is also used to determine if a set is indivisible. Definition 3.3 for a indivisible set in conjunction with definition 2.7 imply that a set \( G \) is indivisible with respect to a \( q \)-family of disjoint sets \( \{G_i\}_{i=1}^{q} \) if

\[
(H_1(r, p) > H_q(r, p) - \frac{1}{M} \log\left(\frac{|R|}{q}\right) + (q - 1) \log(M + 1) - \log |R|)
\]

with \( r = r(G) \), \( p = p(G) \), \( r = (r(G_1), r(G_2), \ldots, r(G_q)) \), and \( p = (p(G_1), p(G_2), \ldots, p(G_q)) \). The binomial term \( \binom{|R|}{q} \) is bounded from below by \( (|R|/q)^q \); therefore \( G \) is indivisible with respect to \( \{G_i\}_{i=1}^{q} \) if

\[
(H_1(r, p) > H_q(r, p) - \frac{1}{M} \log(\frac{|R|(M + 1)) - q \log q).)
\]

An upper bound on \(|R|\) is \( 2^{L_1L_2} \) where \( L_1 \) and \( L_2 \) are the number of rows and columns of the histogram respectively. This upper bound is easily derived by noting that there are \( \binom{L_1L_2}{k} \) ways of choosing \( k \) bins in the histogram; thus the number of sets without holes of size \( k \) is less than \( \binom{L_1L_2}{k} \). Summing \( \binom{L_1L_2}{k} \) over \( k \) gives the upper bound. After substituting this upper bound into the equation 43 then for the second right hand term

\[
\frac{1}{M} [(q - 1) \log(\frac{|R|(M + 1)) - q \log q] < (q - 1) \frac{L_1L_2}{M} \log 2 + \frac{q - 1}{M} \log(M + 1) - \frac{q}{M} \log q.
\]

Note that \( q \) is typically much smaller than \( L_1L_2 \); therefore the \((q/M) \log q \) term is small compared with the other term. Two comments can be deduced from this equation. First, the factor \((L_1L_2)/M \) is the ratio of the number of bins to the number of objects; therefore by choosing the number of bins to be significantly smaller than the number of samples the second term in the right hand side of equation 42 is small. Second, the choice of \(|R|\) in equation 41 is smaller than the number of sets. The right hand side of equation 42 is, therefore, larger than the true...
value and the calculations will err on the side of not merging sets. This completes the discussion on $|R|$.

The result given in Theorem 2.13 demonstrates that if a set $G$ is indivisible with respect to the $q$-family of sets $\{G_i\}^q_{i=1}$, then $G$ is almost always indivisible with respect to sub-collection of the sets $\{G_i\}^l_{i=1}$ for $l < q$. These results motivate the following simplification in the algorithm. Definition 3.5 states that a node $G$ is maximal $\epsilon$-meaningful only if it is indivisible with respect to any family of sibling descendents. This definition implies that for a node with $q$ children, the algorithm should check if $G$ is indivisible with respect to all combinations of the $q$ children. Consider a sub-collection $\{G_i\}^l_{i=1}$ taken from the $q$-family of sets $\{G_i\}^q_{i=1}$. Theorem 2.13 implies that it is unlikely that $G$ is indivisible with respect to the $l$-family $\{G_i\}^l_{i=1}$ of disjoint sets and $G$ is not indivisible with respect to the complete $q$-family $\{G_i\}^q_{i=1}$ of disjoint sets. As a consequence, for each descendent node, indivisibility will be checked only for all of the nodes’ children $\{G_i\}^q_{i=1}$. The authors believe that this simplification does not substantially reduce the algorithm’s effectiveness.

The meaningful mode algorithm consists of two essential parts. The first part is to calculate the FLST transform. The FLST algorithm implemented in this work is due to Monasse, and the details of his methodology will be left to the reference [36]. The choice of quantization levels is the only aspect of his algorithm that requires some discussion. The FLST algorithm can be performed for an arbitrary number of quantization levels. The largest bin value was chosen as the number of quantization levels for all histogram applications presented in this paper, because with this choice of quantization, no information is lost.

The second part of the algorithm locates maximal $\epsilon$-meaningful modes, and this algorithm’s pseudo-code is presented in the pseudo-code below. The only required input to the algorithm is a FLST tree structure with the gray-level of each node recorded. The algorithm consists of four loops. The first loop, lines 2 to 10, calculates important quantities for each node. This loop also determines if a set is $\epsilon$-meaningful. A set cannot be maximal $\epsilon$-meaningful if it is not $\epsilon$-meaningful; therefore the other loops only consider the meaningful sets. The second loop tests the indivisibility condition for each $\epsilon$-meaningful mode. The third loop confirms if any of the $\epsilon$-meaningful modes contains a mode with a lower $NFA$. Any mode in which all the descendants have higher $NFA$ were called dominant ancestors. The last loop verifies that each indivisible $\epsilon$-meaningful mode is not contained in another $\epsilon$-meaningful mode that is both indivisible and a dominant ancestor. Any tree searching technique (depth first, etc.) that visits each node once for each loop will suffice to perform the algorithm.

The algorithm’s complexity will be discussed next. Define the height of the tree, $h$, as the maximum distance from a leaf node to the root of the tree, where the distance is measured by the minimum number of parents that must be traversed to go from the leaf node to the root. The complexity of the algorithm can be estimated in terms of four integers: the height of the tree $h$, the number of leaves $l$, the number of objects $M$, and the number of shapes $s$. Since there are no assumptions on the number of modes or the shape of the modes, the height of the tree, the number of leaves, and the number of shapes are random variables with unknown distributions.

Discussion on the complexity of the FLST algorithm is given in Monasse’s Ph.D. dissertation [36], and it will not be deliberated further in this work. The first loop makes calculation for each node and therefore has on the order of $s$ calculations. The number of operations in the remaining three loops is bounded above by $s^2$. 
Meaningful Mode Algorithm

1: Input ← FLST tree structure.
2: FOR each node $I$ of the tree
3: \hspace{1em} CALCULATE $k(I)$.
4: \hspace{1em} CALCULATE and SAVE $H_1(I)$.
5: \hspace{1em} CALCULATE and SAVE $NFA(I)$.
6: \hspace{1em} CALCULATE and SAVE $H_q(\{G_i\}_{i=1}^q)$ where $\{G_i\}_{i=1}^q$ are the children of $I$.
7: \hspace{1em} IF $NFA(I) < \epsilon$
8: \hspace{2em} LABEL $I$ as meaningful.
9: \hspace{1em} END IF
10: END FOR
11: FOR each meaningful node $I$ of the tree
12: \hspace{1em} FOR each family of sibling descendants $\{G_i\}_{i=1}^q$ of $I$
13: \hspace{2em} IF $NFA(\{G_i\}_{i=1}^q) \leq NFA(I)$
14: \hspace{3em} LABEL $I$ as not meaningful. // A book keeping shortcut
15: \hspace{2em} BREAK
16: \hspace{1em} END IF
17: END FOR
18: END FOR
19: FOR each meaningful node $I$ of the tree
20: \hspace{1em} FOR each descendant $G$ of $I$
21: \hspace{2em} IF $NFA(G) < NFA(I)$
22: \hspace{3em} LABEL $I$ as not a dominant ancestor.
23: \hspace{2em} BREAK // exits the FOR loop
24: \hspace{1em} END IF
25: \hspace{1em} LABEL $I$ as a dominant ancestor.
26: \hspace{1em} END FOR
27: END FOR
28: FOR each meaningful node $I$ of the tree
29: \hspace{1em} FOR each meaningful and dominant ancestor ascendent $G$ of $I$
30: \hspace{2em} IF $NFA(G) \leq NFA(I)$
31: \hspace{3em} LABEL $I$ as not meaningful. // A book keeping shortcut
32: \hspace{3em} BREAK
33: \hspace{2em} END IF
34: \hspace{1em} END FOR
35: END FOR
36: output → FLST tree structure with maximum meaningful nodes tagged as meaningful.

Combining these estimates, an upper bound on the total number of operations, $O$, is

\[(45) \quad O < s + 3s^2.\]

Estimating the number of shapes in a histogram is difficult; therefore an estimate of the complexity of the algorithm given the number of samples, or the number of bins in the histogram is more useful, but this estimate clearly overestimates the complexity. The difficulty in making a more precise estimate is due to the random
nature of $h, l$, and $s$. The number of shapes, $s$, must be bounded above by the number of leaves times the height of the tree, or $s < hl$; therefore $O < hl + 3(hl)^2$. The height of the tree is bounded above by the number of objects, $h < M$. For the case where the number of bins of the histogram, $N$, is less than the total number of objects, $M$, then the number of leaves is bounded above by the number of bins. Combining these terms, the total number of operations is bounded above by

\[(46) \quad O < M^2 + 3M^4.\]

The estimates for the complexity appears to be quite large at face value, but only sets that are $\epsilon$-meaningful need to be checked in the last three loops, and only $\epsilon$-meaningful and indivisible sets need to be checked in the last two loops. It is impossible to predict \textit{a priori} the number of $\epsilon$-meaningful, indivisible sets in a histogram, but, in all applications presented in this paper, the number of sets that were both $\epsilon$-meaningful and indivisible was much less than the number of sets found during the FLST. The meaningful shapes found in Section 5 were located at least ten times faster than the computational time of the FLST despite the reduced theoretical complexity of the FLST tree. Better estimates on the complexity of the algorithm could be obtained if there were better estimates on the height $h$ of the FLST tree given the total mass of the histogram and the number of histogram bins. The authors are not aware of any such estimates.

The next section collects results that increase the efficiency of the algorithm as well as discusses properties of $\epsilon$-meaningful modes. This section is outside the main line of reasoning, but is included to add an understanding of $\epsilon$-meaningful modes.

4. Höeffding bound relations. Calculations of the NFA based on Höeffding bounds serve as a basis for locating maximal $\epsilon$-meaningful modes. Computational gains can be achieved by finding simple algebraic relations that infer the ordering between Höeffding bounds since multiplications are computationally inexpensive compared with logarithmic calculations. Furthermore, properties of the generalized Höeffding bound in terms of the variables $r$ and $p$ and relationships between $H_q(r, p)$ and $H_l(r, p)$ with $q \neq l$ yield insights into the properties of $\epsilon$-meaningful modes. This section begins with a simple algebraic test for the ordering of $H_1(r, p)$, and continues with two results about $\epsilon$-meaningful modes. The relationship between the maximal modes of the multinomial and the variables $p$ and $r$ is also discussed.

4.1. Algebraic tests. The calculation of logarithms can be an order of magnitude more time consuming than simple multiplication; therefore algebraic tests that compare $q$-state relative entropy functions are important for fast calculations of $\epsilon$-meaningful modes. The following theorem was motivated by the Desolneux, Moisan, Morel mean value algebraic tests developed in their book \cite{20}. They asked the question: Is there a fast way to infer the monotone ordering $H_1(r_1, p_1) < H_1(r_2, p_2)$ and what properties does it reveal about meaningful intervals. They showed that the comparison of the 1-state relative entropy can be predicted by a simple equality test based on the mean value ratio $\mu(r, p) = r/p$ combined with the ordering $p_2 < p_1$. Their mean value test is insightful since it reveals the type of scaling ratio that is sufficient to infer ordering \cite{20}.

The next theorem provides a quick test for the ordering of $H_1(r_1, p_1)$ with $H_1(r_2, p_2)$ based on the ratios $\mu = r/p$ and $\omega = (1 - r)/(1 - p)$. To simplify notation, let $\mu_i = \mu(r_i, p_i)$ and $\omega_i = (1 - r_i)/(1 - p_i)$ for $i = 1, 2$. The method of
proof and the \( \omega(r,p) = (1-r)/(1-p) \) test are apparently new. The proof is in the appendix.

**Theorem 4.1.** Let \( r_1, r_2, p_1, p_2 \in [0,1] \) with \( r_1 > p_1, r_2 > p_2, 0 < r_1 + r_2 < 1, \) and \( 0 < p_1 + p_2 < 1 \).

1. If \( p_1 < p_2 \) and \( \mu_1 \geq \mu_2 \) then \( H_1(r_1, p_1) < H_1(r_2, p_2) \).
2. If \( p_1 = p_2 \) and \( \mu_1 \leq \mu_2 \) then \( H_1(r_1, p_1) \leq H_1(r_2, p_2) \).
3. If \( p_1 > p_2 \) and \( \omega_1 \geq \omega_2 \) then \( H_1(r_1, p_1) > H_1(r_2, p_2) \).
4. If \( p_1 = p_2 \) and \( \omega_1 \geq \omega_2 \) then \( H_1(r_1, p_1) \geq H_1(r_2, p_2) \).

Monte Carlo simulations show that the \( \mu(r,p) \) test is valid 81% of the time in the region \( 0 < r < p < 1 \), which can reduce the computation time by one third over computing \( H_1(r,p) \) for every value of \( r \) and \( p \).

4.2. \( \epsilon \)-meaningful mode properties. The next series of results provides information on the set properties of maximal meaningful modes. To simplify notation, let \( r(G_i) = r_i \) and \( p(G_i) = p_i \).

**Theorem 4.2.** Let \( G_1 \) and \( G_2 \) denote two \( \epsilon \)-meaningful disjoint sets and \( G_3 = G_1 \cup G_2 \). If \( r_1, r_2, p_1, p_2 \in [0,1] \) with \( r_1 > p_1, r_2 > p_2, 0 < r_1 + r_2 < 1 \) and \( 0 < p_1 + p_2 < 1 \) then

1. \( \min\{H_1(r_1, p_1), H_1(r_2, p_2)\} < H_1(r_3, p_3) \), or
2. when \( \mu(r_1, p_1) = \mu(r_2, p_2) \) then \( \max\{H_1(r_1, p_1), H_1(r_2, p_2)\} < H_1(r_3, p_3) \).

**Proof.** Since \( G_1 \) and \( G_2 \) are disjoint, both the area probability and frequency count are additive, which implies that \( \min\{\mu(r_1, p_1), \mu(r_2, p_2)\} < \mu(r_3, p_3) \). When \( \mu(r_1, p_1) = \mu(r_2, p_2) \) then \( \max\{\mu(r_1, p_1), \mu(r_2, p_2)\} \leq \mu(r_3, p_3) \)

The previous theorem illustrates the relationship between a parent shape and its decomposition into children sets. Statement 1 of this theorem indicates that when a set is decomposed into two other sets, then one of the other sets has a lower NFA than the parent set. Statement 2 of this theorem demonstrates that when the relative frequency \( \mu(r, p) \) is equal for the two children sets then the NFA of one of the children sets is greater than the parent.

Simple and insightful tests to predict the comparative ordering between \( H_1(r, p) \) evaluated at the union of two sets and \( H_2(r, p) \) evaluated on the respective separate sets are unknown. However, under the hypothesis of the following theorem, their interrelation can be compared for the important case when the two sets are disjoint. The \( q \) set counterpart of this theorem, which clearly exists, is left unstated.

**Theorem 4.3.** Let \( G_1 \) and \( G_2 \) denote two meaningful cells such that \( r_1 + r_2 < 1 \), \( p_1 + p_2 < 1 \), and \( G_3 = G_1 \cup G_2 \). If \( G_1 \cap G_2 = \emptyset \) then

\[
H_2(r_1, r_2, p_1, p_2) \geq H_1(r_3, p_3).
\]

**Proof.** Consider the difference

\[
H_2(r_1, r_2, p_1, p_2) - H_1(r_3, p_3) = r_1 \log \frac{r_1}{p_1} + r_2 \log \frac{r_2}{p_2} + (1 - r_1 - r_2) \log \frac{1 - r_1 - r_2}{1 - p_1 - p_2} - r_3 \frac{r_3}{p_3} -(1 - r_3) \frac{1 - r_3}{1 - p_3}.
\]

\[
= r_1 \log \frac{r_1}{p_1} + r_2 \log \frac{r_2}{p_2} - (r_1 + r_2) \log \frac{r_1 + r_2}{p_1 + p_2}
\]
implies that if the children of
therefore
This statement proves the theorem.

The previous theorem combined with Theorem 2.13 implies that if the children of a set \( G \) partitions the set into two disjoint subsets \( G_1 \) and \( G_2 \) such that \( G = G_1 \cup G_2 \), then the set \( G \) is almost always indivisible with respect to \( \{G_1, G_2\} \).

### 4.3. Relationship to the maximal mode

The boundary between the inequalities \( 0 < p < r < 1 \) and \( 0 < r < p \) can be simply and directly related to the maximal term of a binomial random variable. The multinomial characterization of the maximal term is more complex, and for completeness only the highlights are summarized in this section. For this discussion, both the probability \( q \)-tuple \( p \) and the outcome \( q \)-tuple \( k \) are augmented with the additional term \( p_{q+1} = 1 - |p| \) and \( k_{q+1} = M - |k| \) respectively. This is not unreasonable since any one of the terms in the multinomial distribution is determined by the remaining terms and either formulation is recognized. Among all the possible occurrence \( q + 1 \)-tuples \( k \) with \( \sum_{i=1}^{q+1} k_i = M \), the mode is the one for which \( m_q(M, k, p) \) is a maximum for fixed \( M \) and \( p \). In the case of non-uniqueness, the joint modes are equiprobable neighbors. The inequalities for the multinomial maximal term are derived by considering the ratios of successive local neighbors of the multinomial distribution \( m_q(M, k, p) \). It can be directly verified that a \( q + 1 \)-tuple \( k \) is maximal if and only if for all pairs \((i, j)\)

\[
(50) \quad k_j p_i \leq (k_i + 1) p_j
\]

for all \( i, j \in \{1, 2, \ldots, q, q + 1\} \). Since the inequality is trivially satisfied if \( i = j \), there are only \( q(q + 1) \) distinct inequalities.

According to an exercise in Feller, these inequalities can be combined to yield the following sharp set of inequalities attributed to P.A.P. Moran:

\[
(51) \quad M p_i - 1 < k_i \leq (M + q) p_i
\]

for all \( i \in \{1, 2, \ldots, q, q + 1\} \). In the region \( r > p \), if all of the components of the augmented \( q + 1 \) tuples \( p \) and \( k \) satisfy the Moran inequalities, then they are possible modal candidates; otherwise, the probabilities \( m_{q+1}(M, k, p) \) are always less than the mode(s). It is interesting to note that the minimum Moran inequality implies that by increasing each component of the modal \( q + 1 \)-tuple by one it falls in the region \( r > p \) and if \( M p_i < 1 \), then \( k_i = 0 \) can be a component of the mode(s). Finucan demonstrated that the modal region is unique if

\[
(52) \quad \max_{i \in \{1, 2, \ldots, q+1\}} \frac{k_i}{p_i} < \min_{i \in \{1, 2, \ldots, q+1\}} \frac{k_i + 1}{p_i}
\]

or includes joint neighboring modes, which occur all in one cluster [23], whenever

\[
(53) \quad \max_{i \in \{1, 2, \ldots, q+1\}} \frac{k_i}{p_i} = \min_{i \in \{1, 2, \ldots, q+1\}} \frac{k_i + 1}{p_i}
\]

It is instructive to rewrite the binomial example in the augmented form with \( p_2 = 1 - p_1, k_2 = M - k_1 \) and reinterpret the modal analysis. When the condition \( M p_1 = k_1 \) is satisfied then it is easy to verify that the 2-tuple \((k_1, k_2)\) is the maximal mode, because the identity \( k_2 = M - M p_1 = M p_2 \) yields the inequality \( (k_2+1)/p_2 = \)

---

**Inverse Problems and Imaging** Volume 2, No. 4 (2008), 485–525
Figure 2. The average number of meaningful modes found in a $L_1 = L_2 = 100$ random histogram is plotted. For each $\epsilon = 2^j$ with $j = 1, \ldots, 50$, a hundred 2D random histograms were created where the $x$ and $y$ values where uniform random variables ranging from 0 to 1. The number of meaningful modes for each $\epsilon$ were then averaged.

$(M_{p_2 + 1})/p_2 > k_1/p_1 + 1/p_2$, which is equivalent to Finucan’s. Thus, for any other occurrence 2-tuple $(k_1, k_2)$ satisfying the meaningful inequality $0 < p < r < 1$, its probability $B(M, k_1, p_1) = P[X = k_1]$ is less than the maximal probability. By similar reasoning the condition for the multinomial becomes $M_{p_i} = k_i$ for every $i \in \{1, 2, \ldots, q\}$ such that the $q + 1$-tuple $(k_1, k_2, \ldots, k_q, k_{q+1})$ is the modal $q + 1$-tuple. The condition $(M + 1)p_1 = k_1$ implies that the 2-tuples $(k_1 - 1, k_2)$ and $(k_1, k_2 - 1)$ constitute a joint mode, because the identity $k_2 + 1 = M - (M + 1)p_1 + 1 = (M + 1)p_2$ yields the equality pairing needed in equation 53. This example motivates the following observation by Le Gall for the multinomial [30]. If for every $i \in \{1, 2, \ldots, q\}$, $(M + 1)p_i = k_i$, then there are $q + 1$ joint modal $q + 1$-tuples $(k_1, k_2, \ldots, k_q, k_{q+1}) - \gamma_j$ with $j \in \{1, 2, \ldots, q + 1\}$ where $\gamma_j$ is a $q + 1$-tuple whose only nonzero component is the $j$th one. Since these arguments are reversible it follows that the equality $(M + 1)p_1 = k_1$ is equivalent to the existence of a joint mode for the binomial. The meaningful inequality $M_{p_1} < (M + 1)p_1 = k_1$ is also satisfied; therefore an equiprobable maximal mode can also be meaningful. Thus the interrelation between simple algebraic constraints associated with the meaningful inequality and their interconnection with the modal behavior of the multinomial distribution are similar to those of the binomial distribution. Moreover, the role of the meaningful inequality test for the q-state entropy is consistent.

5. Experiments. The experimental section is divided into three subsections. The first subsection investigates the effects of binning, sampling, and the support of the histogram. Two test problems are presented in this section. The first test problem
consists of finding meaningful modes in a uniform random histogram. When the \(a\)-\textit{contrario} model is correct then the number of \(\epsilon\)-meaningful sets should be less than or equal to \(\epsilon\). The second test model is the geyser data analyzed by Venables and Ripley [45]. This data set illustrates the binning and sampling aspects of the algorithm.

The second and third subsections demonstrate two applications of the meaningful mode detection algorithm. The first application automatically finds dominant color features of an image. These features are established by transforming a color image to the CIELAB colorspace and analyzing the projection of this colorspace onto the \((a, b)\) plane. The dominant color features located in this manner are shown to be robust to illumination changes. The third subsection extends the meaningful mode detection to correlation planes. A correlation plane is not a histogram, but this difficulty is overcome by using the Höeffding and ratio bounds instead of multinomial tail sums. This demonstrates that the meaningful mode algorithm extends to arbitrary distributions because the \(q\)-tuples \(r\) and \(p\) are the only variables of the relative entropy functions. The number of gray scale values for the FLST is reduced in the correlation plane problem from all the possible values to a fixed number. The algorithm is shown to be robust through a large range of quantization levels.

5.1. Test models. Theorem 2.4 asserts that for a random histogram the expected value of the number of \(\epsilon\)-meaningful modes is less than or equal to \(\epsilon\), but an approximation to \(|R|\) was used and the FLST locates a reduced family of sets in \(R\). The dependance of the number of false alarms on \(\epsilon\) was investigated by creating 100 two dimensional histograms with \(M = 25,000\) points, and \(L_1 = L_2 = 100\). The \(x\) and \(y\) values of these histograms were distributed uniformly between 0 and 1. The average number of meaningful modes found as a function of \(\epsilon\) ranging from 1 to 2\(^{50}\) on a logarithmic scale is shown in Figure 5. In the figure the average number of meaningful modes is always less than the parameter \(\epsilon\), and, in particular, for \(\epsilon = 1\) the average number of meaningful modes is less than one. This result is consistent with the interpretation of \(\epsilon\) as an upper bound of the expected number of false alarms given in Theorem 2.4.

This paper does not attempt to address the problem of histogram binning. The result of bin size on the number of maximal meaningful modes found is investigated in the second test example, the well known Old Faithful geyser data analyzed by Venables and Ripley [45]. This data is freely obtainable from the R software package. The Old Faithful data, shown in Figure 3, consists of the waiting time in minutes between eruptions and the duration in minutes of the first eruption. A scatter plot of the data shows many possible groupings. The scatter plot is a little misleading since there are several data points with the same value.

Three histograms are shown indicating the result of binning on the number of modes found. When the bin size shrinks the number of modes increase as more peaks are resolved. The two limiting cases of too coarse of a bin size and too fine of a bin size can be easily understood. If the histogram bin size is too coarse, then all the data will sit in one, or a few connected bins. In this case, one maximal meaningful shape will be found. If each of the data points consists of a unique value, then it is possible to choose a fine enough bin size such that each bin contains only one point. In this case the there will be no maximal meaningful modes. If there are repeated values in the data then some bins will contain more than one point, and enough repeated values will make that bin meaningful. These results are consistent with Venables and Ripley [45].
Maximal Meaningful Boundaries

Figure 3. The effect of binning the histogram is shown. The upper left is a scatter plot of the data. The upper right figure contains a coarse grained histogram with $L_1 = 15$ and $L_2 = 14$, and there were two maximal meaningful modes found. The lower left figure shows a finer grained histogram with $L_1 = 29$ and $L_2 = 28$, and three maximal meaningful modes were found. The lower right figure shows a finer grained histogram with $L_1 = 57$ and $L_2 = 56$, and five maximal meaningful modes were found.

5.2. Color feature extraction. Two dimensional histogram analysis applied to color image segmentation was investigated. Image color information is often stored in the RGB color space as a three dimensional vector per pixel, but for two dimensional histogram segmentation one of the color channels must be omitted. For this reason other color spaces were considered, and the CIELAB color space was chosen. The CIELAB color space attempts to be more perceptually linear and can be obtained from the sRGB color space through a nonlinear transformation. After transforming to the CIELAB color space, an image is still a set of pixels with a three dimensional vector representing the color value. Denote this vector as $(l, a, b)$ with $l \in [0, 1]$ indicating the luminous, $a \in [-1, 1]$ representing the value between magenta and green, and $b \in [-1, 1]$ representing the position between yellow and blue. In this color space an attempt is made to represent the color of the object.
Figure 4. The histogram of the beans image in Figure 5 is shown. The predominant central peak is the background color, and the four smaller surrounding peaks are the beans of different colors.

in the two dimensional projection onto the \((a, b)\) coordinates. The \(l\) vector determines the brightness of the color, thereby reducing, but not eliminating, the effect of shadows on the color feature extraction.

The two dimensional analysis presented here can be contrasted with the Delon et al. one dimensional histogram segmentation [11]. They were able to correctly segment the image of beans, given in Figure 5, using a one dimensional histogram of the hue. The color features extracted from the beans image using the two dimensional histogram segmentation is also shown in Figure 5. The background is the predominant color feature with a \(\log(\text{NFA}) = -1000645\). The different color beans are then extracted. The yellow, green, purple, and red beans had values for the \(\log(\text{NFA})\) equal to -8741.11, -4438.61, -3413.64, -3374.42 respectively. In the analysis of Delon et al., all the pixels were assigned a color depending on the mode the pixel belonged to. In the analysis presented in this work, a pixel that does not belong to any maximal meaningful group was not assigned a color.

Figure 6 shows the two dimensional \((a, b)\) histogram that was obtained by converting the RGB data of the peppers image shown in Figure 7 to the \((l, a, b)\) color space and then suppressing the \(l\) value. Through inspection of the histogram it is difficult to determine the number of modes or the functional form of the modes. The dominant mode is not symmetric; thus an isotropic Gaussian approximation is not appropriate for this mode.

There were six maximal meaningful modes found in the segmentation, and the first five ordered by increasing \(\text{NFA}\) are shown in Figure 7. The most meaningful
Color Segmentation of Beans Image

Figure 5. An example of a two dimensional histogram color image segmentation is shown. The predominant colors were extracted, and the different color beans belonged to separate maximal meaningful groups. The top panel shows the original image. For easier comparison with Delon et al., all the pixels that belonged to a maximal meaningful group were combined into the bottom panel.

mode, with $\log(NFA) = -432180$, corresponded to the purple background. The other meaningful modes in order of increasing false alarm rates were then the red peppers, $\log(NFA) = -33094.8$, the green peppers, $\log(NFA) = -31225.9$, the onion and garlic, $\log(NFA) = -16830.1$, and then the left most orange pepper, $\log(NFA) = -12867.7$. The last mode not shown extracted the shadows and did not correspond to an interesting feature. All of the peppers’ pixels were not included in the maximal meaningful sets, but a large majority were.

The robustness of this method to contrast changes in extracting dominant color features is shown in Figure 8. The top panel shows an image of a crowd, and the
Figure 6. The log of the histogram corresponding to the \((a, b)\) projection of the CIELAB LAB color space of the peppers image shown in Figure 7 is shown. It is difficult from this image to determine the correct number of modes of the histogram. The logarithm was used for visualization only, and the analysis was performed on the original histogram. This histogram had \(L_1 = 500\) and \(L_2 = 500\).

dominant color feature was the skin tones. The image in the top panel underwent two different transformations. The middle panel was obtained by scaling the luminous value \(l\) down by a factor of four before the dominant color features were found. The luminous value is independent of the \((a, b)\) plane; therefore the algorithm is invariant to this change. Images, however, are typically stored in RGB format. To simulate the luminous change in real images, the \(l\) value was scaled, and then converted back to 8 bit sRGB values, before being re-analyzed. This transformation introduces some quantization error due to the 8 bit quantization. This error explains the slight differences in the top and middle right hand images. The bottom panel of Figure 8 was obtained by scaling the sRGB values by a factor of four before the dominant color features were extracted. The most dominant color feature was still the skin tones, but not all of the same pixels were extracted. The difference in the pixels being extracted can be explained by the nonlinear transformation from the sRGB to the CIELAB color space.

The theory presented for two dimensional histogram segmentation is readily adaptable to histograms of other dimensions. The only difficulty is in deriving an algorithm that will quickly construct a containment tree structure analogous to
Color Segmentation of Peppers Image

Figure 7. An example of two dimensional histogram color image segmentation is shown. The original histogram is shown in Figure 6. Six maximal meaningful modes were found, and the first five are shown in increasing order of false alarm rate. The sixth mode corresponded to dark regions, or shadows.

the FLST transform. For one dimensional histograms the algorithm is obvious and the application to a one dimensional histogram color segmentation is shown in Figure 9. In this example the RGB color space was first transformed to the HSV color space and then only the hue (H) histogram was analyzed to extract the lady bug from the image. With a two dimensional histogram, the lady bug is not extracted since the red pixels of the lady bug are spread over many bins, while in the one dimensional histogram the red pixels were concentrated in few bins.

Delon et al. analyzed the same lady bug image to find an automatic color palette using K-means clustering and the Helmholtz principle applied to one dimensional histograms [10]. Their work assigns a color to each pixel, and it is able to automatically determine the correct color to pixel mapping to obtain an accurate representation of the image. The three colors extracted by the method presented in this paper is analogous to the three modes in the Hue histogram found in the work of Delon et al.

5.3. Correlation pattern recognition. Adapting the point of view that $r$ are the observed probabilities and $p$ are the a-priori probabilities, then the Höeffding
Robust Under Contrast Change

Figure 8. An example of the robustness under contrast change is shown in the above figure. The top left panel is the original image and the mode with the least number of false alarms extracts the skin tones as shown in the top right panel. The luminous value $l$ is scaled down by a factor of four in the middle left panel, and the skin tones are still extracted. The RGB values were all scaled down by a factor of four in the bottom left panel. The image was then transformed to the CIELAB color space, and slightly fewer skin tones were extracted as shown in the bottom right panel. The dominant skin tones were still found.

and ratio tail bounds can be extended to analyze correlation plane for pattern recognition. The $NFA$ definition is stated in terms of binomial or multinomial tail sums, and it depends on the variables $r$, $p$, $|R|$, $M$; therefore the calculation of the binomial or multinomial tail sums require knowledge of the total number of objects in the histogram, the count of the elements in a given set, and the probability of an element to be in the set. The Hoeffding upper bound and the ratio tail upper bounds, however, were used in the operational definition of the $NFA$, and these
Color Segmentation of Lady Bug Image

Figure 9. Using a one dimensional histogram, the lady bug is extracted from this image. Analysis of the one dimensional histogram gives three maximal meaningful modes. The first two modes correspond to the background leaves, and the third mode is the lady bug. A two dimensional histogram will miss the lady bug.

Correlation pattern recognition consists of using a reference signal as a template, and calculating the cross-correlation between the template and a input scene. The output is a correlation plane. Local maxima of the correlation plane correspond to potential targets, but a measure of the peak height and quality must be used to distinguish clutter from targets. There are many measures of peak quality, but one of the most popular is the peak to side lobe ratio (PSR) [29]. The PSR is measured by considering a square annulus around the local maxima consisting of two square windows of size $R_1$ and $R_2$ with $R_1 < R_2$. If $p$ is the location of the peak, then the annular region, $A$, is defined by

$$A = \{ x \mid |p - x| > R_1, |p - x| < R_2 \}.$$ (54)

Let $c(x)$ be the correlation plane, and define the mean of the annular region, $m$, by

$$m = \frac{1}{|A|} \sum_{x \in A} c(x)$$ (55)

and the standard deviation, $s$, as

$$s = \sqrt{\frac{1}{|A|} \sum_{x \in A} (c(x) - m)^2}.$$ (56)

The PSR is then given by

$$PSR = \frac{c(p) - m}{s}.$$ (57)

If the PSR is above a certain threshold, then the peak is identified as a targets. More details on the MACH filter and the PSR can be found in Kumar et al. [29].
Figure 10. The application of finding maximal meaningful modes in a correlation plane is shown. A template was used to identify multiple targets in a single image. The targets are identified as peaks in the correlation plane that have a high peak to side lobe ratio (PSR). The top panel shows the original Lidar data with the potential targets in blue boxes. The middle panel illustrates the results using the PSR values of local maxima. The bottom panel shows the targets found in the correlation plane using the meaningful mode algorithm. The meaningful modes procedure reduced the false alarms by three, and none of the targets were repeated. The local maxima technique found 117 locations for targets, and moreover most of the targets were found multiple times using the local maxima technique.

There are many different types of correlation filters, and the one that will be investigated in this work is the maximum average correlation height (MACH) filter. This filter has been used to locate multiple target of the same type in synthetic aperture radar (SAR) and laser radar (ladar) imagery [29, 31, 44]. The MACH filter was chosen since it produces improved performance and reduced computational load over many other filters [44].

A problem exists when local maxima are used as a screening step in correlation planes generated with MACH filters because the correspondence between local
Figure 11. The robustness of the correlation plane algorithm under a change in quantization levels is explored. The threshold for the peak to side lobe ratio (PSR) was fixed at 5.3. There are slight differences in the number of false alarms and the number of missed targets until 12 quantization levels. The meaningful mode algorithm does not have adequate resolution to distinguish between separate peaks at a low quantization level since many of the peaks are merged into one.

Maxima with high PSR and targets is not one to one. This lack of pairing between local maxima and targets is partially due to mismatch between the template and the output signal, which is created due to a rotation and scaling with respect to the reference template. The mismatch between the reference template and the output signal results in smaller modes of arbitrary shape. Just as the meaningful mode algorithm can find modes of arbitrary shape in histograms, the algorithm will extend to find modes of arbitrary shape in correlation planes.

Since the meaningful mode algorithm uses the FLST and it requires a probability distribution as input, the correlation plane underwent two transformations before it was analyzed. First it was shifted by a constant such that the minimum was zero and then it was quantized. The shift does not change the location of local maxima; therefore the local maxima still correspond to potential targets. Because the maximum number of values in a discrete grid is the number of grid points, the number of quantization levels for the FLST can be as high as the number of grid points in the digital correlation plane. The number of quantization levels was reduced to 250 to quicken the analysis. Figure 11 demonstrate that there was little difference in the results when the quantization was changed from 250 levels to 25. Errors in finding targets arose when 12 quantization levels were chosen and separate peaks were merged into one. After a quantization level is chosen, to preserve the probability interpretation of $r$, $M$ must be the number of quantization levels times the area under the shifted correlation plane.

An example of the meaningful mode algorithm applied to a correlation plane is given in Figure 10. The top panel is a rendered image of ladar data with targets and clutter present in the image. The imaged area has 19 potential targets, and the targets are displayed in blue boxes. The boxes are for display purposes only, and were not used in the analysis. The local maxima method identified 117 potential
target locations, many of them duplicates, missed two of the targets, and had five false alarms as shown in the middle panel. The bottom panel shows the improved ability of the meaningful mode algorithm to correctly identify the targets. A total of 17 of the 19 targets were found with the meaningful modes algorithm, and two false alarms were found.

6. Conclusions. A novel method of finding two dimensional histogram modes based on the Helmholtz principle and the FLST tree structure has been presented. This method was then applied to the problems of image color segmentation and correlation plane pattern recognition. The Helmholtz principle was used to create a parameter free technique to locate the modes while the FLST tree structure was exploited to coordinate the search for modes. Its inclusion structure was utilized to find disjoint modes of the histogram that are unlikely to occur in a uniform histogram.

The development of maximal $\epsilon$-meaningful modes using the FLST tree structure required advancements in multinomial tail bounds. These advancements include a generalization of the Hoeffding relative entropy bound to the multinomial distribution. The containment genealogy ordering in the tree of shapes incorporates parents with a variable number of descendants, something the new $q$-state relative entropy bound can accommodate. Moreover, the $q$-state bound is always a tighter bound than the Joag-Dev/Proschan product bound. Finally, the maximal $\epsilon$-meaningful mode criterion depends on the $q$-state entropy, which fits within the Cao et al. Helmholtz framework.

Generalized algebraic mean value ratio tests to predict the monotone ordering of the 1-state relative entropy function $H_1(r, p)$ are derived, which provide algorithmic shortcuts for the determination of maximal $\epsilon$-meaningful modes. Algebraic factoring of the $q$-state relative entropy function $H_q(r, p)$ and a simple test that predicts the monotone ordering between the 1-state entropy and 2-state entropy functions also facilitate the maximal $\epsilon$-meaningful mode determination.

The meaningful mode algorithm was applied to color image segmentation and to correlation pattern recognition. A useful two dimensional histogram of color pixels was created by transforming the RGB color space into the CIELAB color space and suppressing the luminous value. This created a color histogram that is less susceptible to shadows. The resulting color image segmentation was demonstrated in two pictures. The extracted colors in both images were very uniform and captured most of the pixels that visually represented that color. The color feature extraction also benefited from the morphological nature of the FLST tree structure. The FLST tree structure combined with the invariance of the variables $r$ and $p$ upon multiplication of the histogram by a constant integer generated a morphologically invariant approach. This morphological invariance translated to robustness under contrast changes for color image segmentation.

An alternative multiple target identification procedure based on MACH filters was also described and demonstrated. This technique was able to reliably locate multiple unique targets, and, without the addition of new parameters, it lowered the false detection rate compared with previous methods.

Applications of the meaningful mode algorithm is not restricted to the applications presented here. The method has also been used by Hewer et al. to analyze two dimensional wavelet histograms to generate an image cartoon-texture decomposition. See reference [25] for a complete discussion on their procedure. The algorithm
also naturally extends to all families of correlation filters. The work of Kumar, Mahalanobis, and Juday discuss other correlation filters and correlation pattern recognition \[29\].

Meaningful gaps, introduced in \[20\], are defined as intervals that contain fewer points than the expected average. While meaningful gaps are not included in the discussion, they could be accommodated within the \(q\)-state entropy framework. Moreover, the \(\epsilon\)-maximal meaningful mode analysis is readily extensible to three or more dimensional histograms. Unfortunately, a three dimensional FLST algorithm does not exist at this time.

Acknowledgments. We would like to thank Professor Julie Delon at CNRS, Télécom, Paris, France for the beans and lady bug image used in the color feature extraction examples. We would also like to thank Dr. Alan Van Nevel at China Lake, CA for many useful discussions about the correlation plane application used in this work. The authors would also like to thank the reviewer for many useful comments.

Appendix.

6.1. Expected number of false alarms. The interpretation of the \(NFA\) as an upper bound on the expected number of \(\epsilon\)-meaningful modes serves as an argument for its use. The proof that the expected number of false alarms is less than \(\epsilon\) is given below. Only the multinomial theorem will be proven since the binomial is a special case. First a necessary lemma will be stated and proven. The proof of this lemma is modeled after a lemma in \[5\].

\[\text{Lemma 6.1.}\] Let \(F(k) = \mathcal{M}(M, k, p)\) be the multinomial tail probability sum, then for all \(\epsilon \in (0, 1)\)
\[
P(F(K) < \epsilon) \leq (M + 1)^{(q-1)} \epsilon. \tag{58}
\]

**Proof.** Let \(B = \{k \in \mathbb{Z}^q|F(k) < \epsilon, |k| \leq M\}\). By the definition of the multinomial probability tail then
\[
P(F(K) < \epsilon) = \sum_{k \in B} m_q(M, k, p). \tag{59}
\]

The binomial tail is well defined on the set \(C = \{k \in \mathbb{Z}^q|k_i > 0, k_1 \leq M, k_2 \leq M - k_1, \ldots, k_q \leq M - k_1 - k_2 - \cdots - k_{q-1}\}\). Extend \(m_q(M, k, p)\) to \(\mathbb{Z}^q\) by defining \(m_q(M, k, p) = 0\) for \(k \notin C\).

Define the function \(k(k) = (k_2, k_3, \ldots, k_q)\) as the projection of \(k\) onto the last \(q - 1\) elements and let
\[
A(\epsilon, \tilde{k}) = \{0 \leq k_1 \leq M|\mathcal{M}(M, (k_1, \tilde{k}), p) < \epsilon, k \in C\}. \tag{60}
\]

Observe that for fixed \((k_2, k_3, \ldots, k_q)\), the set \(A(\epsilon, \tilde{k})\) is the set of \(k_1\) such that \(\mathcal{M}(M, (k_1, \tilde{k}), p)\) is less than \(\epsilon\). The function
\[
\kappa(\epsilon, \tilde{k}) = \begin{cases} 
\inf A(\epsilon, \tilde{k}) & \text{if } A \neq \emptyset \\
M + 1 & \text{if } A = \emptyset
\end{cases} \tag{61}
\]
will be used to correctly determine the summation indexes. Notice that $F(k)$ is decreasing as $k_1$ increases; thus $k_1 \in A(\epsilon, k)$ if and only if $k_1 \geq \kappa(\epsilon, k)$ and $|k| + k_1 \leq M$.

The extension of $m_q(M, k, p)$ and the definition of $\kappa(\epsilon, k)$ yields

$$\sum_{k \in B} m_q(M, k, p) = \sum_{k_2=0}^{M} \sum_{k_3=0}^{M} \cdots \sum_{k_{q-1}=0}^{M+1} m_q(M, k, p).$$

The $M + 1$ limit on the last sum is necessary since $\kappa(\epsilon, k)$ is $M + 1$ when $A = \emptyset$. The extension of $m_q(M, k, p)$ guarantees that this addition does not contribute to the sum. Adding positive terms to the sum only increases the sum, then

$$\sum_{k_2=0}^{M} \sum_{k_3=0}^{M} \cdots \sum_{k_{q-1}=0}^{M+1} m_q(M, k, p) \leq \sum_{k_2=0}^{M} \sum_{k_3=0}^{M} \cdots \sum_{k_{q-1}=0}^{M+1} m_q(M, k, p).$$

There are no problems with the limits of the summations in the previous equation since $m_q(M, k, p) = 0$ when $k \notin C$.

To finish the proof, notice that the last $q$ sums in equation (63) is just the multinomial tail sums $F(k)$; thus

$$\sum_{k \in B} m_q(M, k, p) \leq \sum_{k_2=0}^{M} \sum_{k_3=0}^{M} \cdots \sum_{k_{q-1}=0}^{M+1} F(k) \leq (M + 1)^{q-1}/\epsilon.$$

(64)

**Theorem 6.2.** The expected number of $\epsilon$-meaningful $q$-families $\{G_i\}_{i=1}^{q}$ of disjoint sets is less than $\epsilon$.

**Proof.** Define $D$ as the collection of all the $q$-families of disjoint sets of the form $\{G_i\}_{i=1}^{q}$, then the number of $q$-families in $D$ consisting of disjoint sets is less than $|\mathcal{R}|$. The random variable $K = (K(G_1), K(G_2), \ldots, K(G_q))$ represents the random number of elements in each set and define $p = (p(G_1), p(G_2), \ldots, p(G_q))$ with $p(G_i) = |G_i|/N$ where $N$ is the number of bins.

Denote by $Y(\{G_i\}_{i=1}^{q})$ the binary random variable equal to 1 if $\{G_i\}_{i=1}^{q}$ is $\epsilon$-meaningful and to 0 otherwise. The random variable $Z = \sum_{\{G_i\}_{i=1}^{q} \in D} Y(\{G_i\}_{i=1}^{q})$ counts the number of $\epsilon$-meaningful modes. By linearity of the expectation

$$E(Z) = E \left( \sum_{\{G_i\}_{i=1}^{q} \in D} Y(\{G_i\}_{i=1}^{q}) \right) = \sum_{\{G_i\}_{i=1}^{q} \in D} E(Y(\{G_i\}_{i=1}^{q})).$$

(65)
Since \( Y(\{G_i\}_{i=1}^q) \) is a Bernoulli variable,
\[
\mathbb{E}(Y(\{G_i\}_{i=1}^q)) = P(Y(\{G_i\}_{i=1}^q) = 1) = \sum_{\{G_i\}_{i=1}^q \in \mathcal{D}} P(Y(\{G_i\}_{i=1}^q) = 1 \mid \{G_i\}_{i=1}^q) P(\{G_i\}_{i=1}^q) = \sum_{\{G_i\}_{i=1}^q \in \mathcal{D}} P(M(M, K, p, \epsilon) < \frac{\epsilon}{M^q-1} q) P(\{G_i\}_{i=1}^q) \leq \frac{\epsilon}{|R|} \sum_{\{G_i\}_{i=1}^q \in \mathcal{D}} P(\{G_i\}_{i=1}^q),
\]
where the last inequality is true by the previous lemma. Using this result, then for equation \( 65 \)
\[
\sum_{\{G_i\}_{i=1}^q \in \mathcal{D}} \mathbb{E}(Y(\{G_i\}_{i=1}^q)) < \epsilon.
\]

Let \( r = \frac{\mathcal{H}}{M^q} \), then the inequality \( M(M, K, p, \epsilon) < \exp(-MH_q(r, p)) \) implies that
\( (M + 1)^{q-1} \frac{\epsilon}{|R|} M(M, K, p) < \epsilon \) whenever \( (M + 1)^{q-1} \frac{\epsilon}{|R|} \exp(-MH_q(r, p)) < \epsilon \),
and the expected number of \( \epsilon \)-meaningful modes is less than \( \epsilon \). \( \square \)

6.2. **Algebraic tests proof.** The following is a proof of theorem 4.1. The algebraic tests in this theorem can reduce the computational time to calculate if a set is \( \epsilon \)-meaningful. Recall the definitions \( \mu(r, p) = r/p \) and \( \omega(r, p) = (1 - r)/(1 - p) \). The subscripts accompanying \( \mu_i \) and \( \omega_i \) is a shorthand to distinguish between the variables \( p_i \) and \( r_i \).

**Theorem 6.3.** Let \( r_1, r_2, p_1, p_2 \in [0, 1] \) with \( r_1 > p_1, r_2 > p_2, 0 < r_1 + r_2 < 1, \) and
\( 0 < p_1 + p_2 < 1 \).

1. If \( p_1 < p_2 \) and \( \mu_1 \leq \mu_2 \) then \( H_1(r_1, p_1) < H_1(r_2, p_2) \).
2. If \( p_1 = p_2 \) and \( \mu_1 \leq \mu_2 \) then \( H_1(r_1, p_1) \leq H_1(r_2, p_2) \).
3. If \( p_1 > p_2 \) and \( \omega_1 \geq \omega_2 \) then \( H_1(r_1, p_1) > H_1(r_2, p_2) \).
4. If \( p_1 = p_2 \) and \( \omega_1 \geq \omega_2 \) then \( H_1(r_1, p_1) \geq H_1(r_2, p_2) \).

**Proof.** (1) The curves \( r_{\mu_1}(p) = \mu_1 p \) and \( r_{\omega_1}(p) = \omega_1(p - 1) + 1 \) preserve the ratios \( \mu_1 = r_1/p_1 \) and \( \omega_1 = (1 - r_1)/(1 - p_1) \) respectively, and moreover in the region \( 0 < p < r < 1, \mu/\omega > 1 \). Note that for any curve \( r(p) \), the total derivative of \( H_1(r(p), p) \) is
\[
\frac{dH_1}{dp} = \frac{\partial H_1}{\partial r} \frac{dr}{dp} + \frac{\partial H_1}{\partial p} = \log \frac{\mu}{\omega} \frac{dr}{dp} + \omega - \mu.
\]
Let \( r_{H_1}(p) \) be the curve that preserves the relative entropy \( H_1(r_{H_1}(p), p) \) given by
\[
\frac{d}{dp} H_1(r_{H_1}(p), p) = 0,
\]
with the initial condition \( r_{H_1}(p_1) = r_1 \). From equation 68 this curve solves the equation
\[
\frac{d}{dp} r_{H_1}(p) = \frac{\mu(r, p) - \omega(r, p)}{\log \mu(r, p) - \log \omega(r, p)}.
\]
The logarithmic mean satisfies the inequalities
\[
\omega = \min \{ \mu, \omega \} \leq \frac{\mu - \omega}{\log \mu - \log \omega} \leq \max \{ \mu, \omega \} = \mu.
\]
The total derivative of the difference, \( D(p) = H_1(r_{H_1}(p), p) - H_1(r_{H_1}(p), p) \), between the relative entropies satisfies
\[
\frac{dD}{dp} = \log \frac{\mu_1}{\omega_1} \frac{dr_{H_1}}{dp} + \omega_1 - \mu_1
\]
\[
= \mu_1 \log \frac{\mu_1}{\omega_1} + \omega_1 - \mu_1
\]
\[
= \log \left( \frac{\mu_1}{\omega_1} \right) \left( \mu_1 - \frac{\mu_1 - \omega_1}{\log \mu_1/\omega_1} \right).
\]

Since \( \mu_1/\omega_1 > 1 \) in the interval \( 0 < p < r < 1 \), the outer inequalities in equation 71 are strict so that
\[
\frac{dD}{dp} > 0,
\]
and the inequality \( H_1(r_{H_1}(p_2), p_2) > H_1(r_{H_1}(p_1), p_1) \) holds for \( p_2 > p_1 \).

The inequalities \( p_1 < p_2 \) and \( \mu_1 < \mu_2 \) imply that \( r_2 \geq r_{H_1}(p_2) \) because
\[
H_1(r_{H_1}(p_2)) - H_2(p_1) = \int_{p_1}^{p_2} \frac{dr_{H_1}(p)}{dp} dp
\]
\[
\leq \int_{p_1}^{p_2} \mu_1 dp
\]
\[
= \mu_1 p_2 - \mu_1 p_1.
\]

In addition \( \frac{dH_1}{dp}(r, p) > 0 \), thus the asserted inequality
\( H_1(r_{H_1}(p_2), p_2) > H_1(r_{H_1}(p_1), p_2) = H_1(r_1, p_1) \) is valid.

The proof for the \( \omega \) inequality follows by similar reasoning. Note the \( D_{\omega}(p) \) is the difference between the relative entropies evaluated along the two curves \( r_{H_1}(p) \) and \( r_{H_1}(p) \). After applying the differential inequality for the lower bound \( \omega(r, p) \), it follows that the total derivative is now negative. The final inequality \( H_1(r_1, p_1) > H_1(r_2, p_2) \) is satisfied because the needed inequality \( r_2 \leq r_{\omega}(p_2) = \omega_1(p_2-1)+1 < r_1 \) follows from the inequalities \( p_1 > p_2 \) and \( \omega_2 \leq \omega_1 \).

The proof of (2) and (4) follow from obvious continuity considerations of \( H_1(r, p) \).
[5] F. Cao, J. Delon, A. Desolneux, P. Musé and F. Sur, “A unified framework for detecting groups and application to shape recognition,” J. Math. Imaging Vision, 27 (2006), 91–119.
[6] Y. Cheng, Mean shift, mode seeking and clustering, IEEE Trans. on Pattern Analysis and Machine Int., 17 (1995), 790–799.
[7] D. Comaniciu and P. Meer, Mean Shift: A Robust Approach toward Feature Space Analysis, IEEE Trans. Pattern Analysis Machine Intell., 43 (2002), 603–619.
[8] Thomas M. Cover and Joy A. Thomas, “Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing),” Wiley-Interscience, 2006.
[9] J. Delon, A. Desolneux, J. L. Lisani and A. B. Petro, “Pattern Recognition and Image Analysis,” Springer, Berlin, 2005.
[10] J. Delon, A. Desolneux, J. L. Lisani and A. B. Petro, “Non-parametric approach for histogram segmentation,” IEEE Trans. on Image Process, 15 (2007), 253–261.
[11] J. Delon, A. Desolneux, J. L. Lisani and A. B. Petro, Histogram analysis and its applications to fast camera stabilization., IWSSIP’04, (2004), 431–434.
[12] A. Desolneux, Evenements significatifs et applications a l’analyse d’images., Ph.D. thesis, Ecole Normale Supérieure de Cachan, 2000.
[13] A. Desolneux, L. Moisan and J. M. Morel, Edge detection by helmholtz principle, J. of Math. Imaging Vision, 14 (2001), 271–284.
[14] J. Delon, “Computational gestalts and perception thresholds,” J. Physiology-Paris, 97 (2003), 311–324.
[15] A. Desolneux, Maximal meaningful events and applications to image analysis, Ann. of Statist., 31 (2003), 1822–1851.
[16] J. Delon, Maximal meaningful events and applications to image analysis, Ann. of Statist., 31 (2003), 1822–1851.
[17] J. Delon, A grouping principle and four applications, IEEE Trans. on Pattern Analysis and Machine Int., 26 (2004), 1105–1111.
[18] J. Delon, Meaningful alignments, J. Computer Vision, 40 (2004), 7–23.
[19] W. Feller, “An Introduction to Probability Theory and its Applications,” vol. 1, John Wiley and Sons, New York, NY, 1968.
[20] H. M. Finucan, The mode of a multinomial distribution, Biometrika, 51 (1964), 513–517.
[21] K. Fukunaga and L. D. Hostetler, The estimation of the gradient of a density function with applications in pattern recognition, IEEE Trans. on Inform. Theory, 21 (1975), 32–40.
[22] G. Hewer, W. Kuo and G. Hanson, Double density complex wavelet based image cartoon-texture decomposition, Proceedings of ACSSC, (2006), 861–868.
[23] N. Lloyd Johnson, Samuel Kotz, Adrienne W Kemp and Norman Lloyd Johnson, “Univariate Discrete Distributions,” Wiley, New York, Chichester, 1992.
[24] N. S. Kambo and S. Kotz, On exponential bounds for binomial probabilities, Ann. Inst. Stat. Math., 18 (1966), 277–287.
[25] B. V. K. Vijaya Kumar, Abhijit Mahalanobis and Richard D. Juday, “Correlation Pattern Recognition,” Cambridge University Press, 2006.
[26] F Le Gall, Determination of the modes of a multinomial distribution, Statistics and Probability Letters, 62 (2003), 325–333.
[27] A. Mahalanobis, D. Carlson and BVK Vijaya Kumar, Evaluation of mach and docf filters for sar atr using the mstar public database, Proc. SPIE, 3370 (2001), 424–431.
[28] C. L. Mallows, An inequality involving multinomial probabilities, Biometrika, 55 (1968), 422–424.
[29] Geoffrey J. McLachlan and T. Krishnan, “The em Algorithm and Extensions,” vol. Wiley series in probability and statistics, Applied probability and statistics, Wiley, New York, 1997.
[30] Geoffrey J. McLachlan and David Peel, “Finite Mixture Models,” Wiley, New York, 2000.
[31] M. C. Minnotte and D. W. Scott, The mode tree: a tool for visualization of nonparametric density features, J. of Comput and Graph. Statist., 2 (1993), 51–68.
[32] P. Monasse, “Représentation Morphologique D’images Numériques et application au recalage,” Ph.D. thesis, Université de Paris-Dauphine, 2000.
[37] P. Monasse and F. Guichard, *Fast computation of contrast-invariant image representation*, IEEE Trans. Image Process, 9 (2000), 860–872.

[38] A. Rosenfeld, *On connectivity properties of grayscale pictures*, Pattern Recognition, 16 (1983), 47–50.

[39] D. W. Scott, “Multivariate Density Estimation,” Wiley, New York, 1992.

[40] D. W. Scott and S. R. Sain, “Handbook of Statistics,” vol. 23, Springer, Amsterdam, 2004.

[41] B. W. Silverman, *Using kernel density estimates to investigate multimodality*, J. R. Stat. Soc. Ser. B Stat. Methodol., 43 (1981), 97–99.

[42] _____, “Density Estimation for Statistics and Data Analysis,” vol. Monographs on statistics and applied probability; 26, Chapman & Hall/CRC, Boca Raton, 1998.

[43] C. Sun, H. Talbot, S. Ourselin and T. Adriaansen (eds.), “False-peak-avoiding Mean Shift Method for Unsupervised Peak-Valley Sliding Image Segmentation,” Sydney, Australia, 2003.

[44] A. Van Nevel and A. Mahalonobis, *Comparative study of maximum average correlation height filter variants using ladar imagery*, Optical Engineering, 42 (2003), 541–550.

[45] W. N. Venables, “Modern Applied Statistics with s,” Springer-Verlag, 2002.

Received May 2007; revised September 2008.

E-mail address: arjuna.flenner@navy.mil
E-mail address: gary.hewer@navy.mil