The Spatial Vision Tree: A Generic Pattern Recognition Engine—Scientific Foundations, Design Principles, and Preliminary Tree Design

Zia-ur Rahman[†], Daniel J. Jobson[‡], Glenn A. Woodell[‡]
[†]Old Dominion University, Norfolk, VA 23529
[‡]NASA Langley Research Center, Hampton, Virginia 23681

ABSTRACT

New foundational ideas are used to define a novel approach to generic visual pattern recognition. These ideas proceed from the starting point of the intrinsic equivalence of noise reduction and pattern recognition when noise reduction is taken to its theoretical limit of explicit matched filtering. This led us to think of the logical extension of sparse coding using basis function transforms for both de-noising and pattern recognition to the full pattern specificity of a lexicon of matched filter pattern templates. A key hypothesis is that such a lexicon can be constructed and is, in fact, a generic visual alphabet of spatial vision. Hence it provides a tractable solution for the design of a generic pattern recognition engine. Here we present the key scientific ideas, the basic design principles which emerge from these ideas, and a preliminary design of the Spatial Vision Tree (SVT). The latter is based upon a cryptographic approach whereby we measure a large aggregate estimate of the frequency of occurrence (FOO) for each pattern. These distributions are employed together with Hamming distance criteria to design a two-tier tree. Then using information theory, these same FOO distributions are used to define a precise method for pattern representation. Finally the experimental performance of the preliminary SVT on computer generated test images and complex natural images is assessed.

1. INTRODUCTION

The rise in the use of personal computers and digital cameras has brought almost everyone into daily contact with multitudes of digital images. Even television has now crossed into the domain of digital acquisition, storage, and transmission. Hence, computer processing of images has naturally assumed an increasingly important role when compared to its use in the past. Foremost among the challenges associated with this role is for computers to autonomously make "visual sense" of the large archives of still imagery and of live video data. Even more exciting is the challenge of embedding this technology in real-time "smart" imaging sensors to enable previously unimaginable vision-based automation that can interact with the changing visual world environment. The state-of-the-art for generic pattern recognition is fairly primitive, so we feel that it is time to explore and exploit fundamental new ideas to begin to establish this kind of major technology advance.

Our approach to visual pattern recognition is fundamentally new, but has its roots in important ideas and results from information theory and natural vision, noise reduction, and natural image statistics. We seek to define *generic* pattern recognition methods rather than ones targeted to specialized applications. This generic processing is intended, however, to form a core platform from which a wide array of specialized processing functions can be derived for specific uses.

At first glance, it seems impossible to reconcile a purely statistical view of spatial vision with generic pattern recognition due to the immense number of pattern permutations that are possible in the two-dimensional "random" image space. Indeed in order to have a viable solution to the daunting problem of designing generic computer vision algorithms, one is required to avoid this combinatorial explosion at every step of the design process. Fortunately, results from natural image statistics research and from information theoretic studies of

Contact: Zia-ur Rahman (zrahman@odu.edu) is an Associate Professor at ODU. He is affiliated with the Electrical and Computer Engineering Department and the Virginia Modeling, Analysis and Simulation Center (VMASC). Daniel Jobson (daniel.j.jobson@nasa.gov) and Glenn Woodell (glenn.a.woodell@nasa.gov) are with the Electromagnetics and Sensors Branch at the NASA Langley Research Center.

natural vision, show that the pattern space for natural images is highly redundant, and not at all random. As Ruderman[1] states: "Natural images are distinctive because they contain particular types of structure. They are far from random: images constructed randomly on a computer practically never contain a naturalistic scene—or even a tree. Natural images are thus very rare among the huge space of all possible images." Our natural vision system takes advantage of this redundancy to achieve its impressive performance in object recognition. We will use this insight into developing our general pattern space which is sparse, yet *complete* in the sense that we can represent (almost) any naturally occurring pattern with patterns from this space.

In order to do this, we find with respect to natural images, that we must go beyond the notion of developing basis functions that cover the full pattern space to the idea of a sparse, but complete, pattern lexicon that provides the same functionality. This notion arises from the idea of using matched filters[2–4] to achieve *maximum* noise reduction (see later discussion). A pattern in the pattern lexicon is similar to a matched filter because it provides a maximum response to a like-pattern in the image space. The pattern lexicon consists of 5×5 patterns that were found to recur frequently in a large set of natural images. Application of the lexicon to natural images (almost) fully captures the structural elements in the images. The matching pattern is the one that provides the maximum response within a 5×5 image window.

Our approach to generic pattern recognition evolved obliquely from our previous work on the development of generic image enhancement methods. This development culminated in the creation of the non-linear image enhancement method known as the multi-scale retinex with color restoration (MSRCR).[5–8] While the MSRCR performs exceedingly well for most image classes, we felt that its passive nature was a restriction on the performance it could achieve for extremes of poor visibility conditions. Hence, we added an active control system—visual servo (VS)—to augment the passive MSRCR.[8–10] While this addition succeeded in improving the overall performance of the MSRCR in a completely automatic and generic sense, we saw that the visibility improvement in the enhanced images was limited by the intrinsic sensor noise. This is intuitively obvious, but was made graphically dramatic by numerous image enhancement examples of high turbidity and low light level scenes, i.e., the poorest visibility imaging conditions. An example of turbid imagery and the associated enhancement is shown in Figure 1. As can be seen, while the enhanced image brings out a lot of the details not readily observable in the original data, the "enhanced" noise masks several of the line features, blending them into their environment. The best way of breaking through this visibility performance barrier is by applying one of the most powerful noise reduction methods: matched filtering.[2–4] However, since the relevant scenes are highly complex natural images, a single matched filter is insufficient. Hence, we arrived at the idea of *generic* matched filtering where a

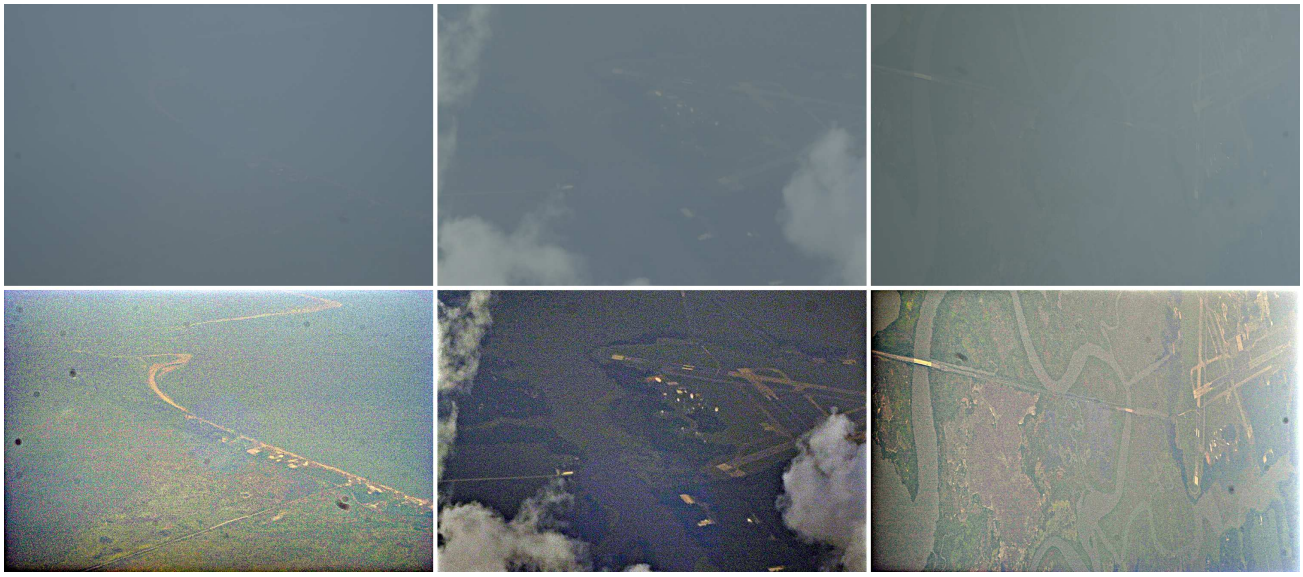

Figure 1. Turbidity: (left) original; (right) enhanced image. The sensor noise is evident in the enhancement.

lexicon of matched filters is used for pattern recognition, and hence noise reduction. This gives rise to the basic, yet powerful, realization that

MAXIMUM NOISE REDUCTION ≡ GENERIC PATTERN RECOGNITION

This intrinsic connection seems obvious in retrospect, but we have not found any explicit emphasis of this very important point in previous research. Noise reduction and pattern recognition seem to have been treated in isolation as completely distinct technical problems. Yet it is quite clear that any attempt to maximally reduce noise inexorably leads to generic pattern recognition via matched filtering.

Taking Ruderman's logic from global down to gross regional scales in images one step further, we arrive at local regions in the image and the notion that naturally occurring visual pattern information at this scale is likewise going to be a small (and hopefully tractable) subset of of all possible patterns in a purely statistical or random sense. At this point we immediately conceived of the possibility that some scale exists for which there is a visual *alphabet* of patterns that are explicit matched filter templates: These would provide the powerful noise reduction of matched filters and the substrate for completely generic pattern recognition. So our challenge is to define the scale, and the set of patterns at this scale, i.e., the visual alphabet, which form a complete set of all visually significant structures. In keeping with the idea of pattern significance, *a la* natural image statistics, we also need to characterize the pattern lexicon by experimentally estimating the frequency of occurrence (FOO) of each pattern over a large set of complex natural imagery. In order to fully capture small variations in the significant patterns, the lexicon is laid out in the form of a tree where the roots are comprised of the significant set of patterns, and the branches under the roots are comprised of the minor variations in the patterns at that root. We call this layout the spatial vision tree (SVT).

We envision using the SVT (Figure 2) and the VS as key components of future smart avionics imaging sensors. In addition to visibility improvements, the VS also provides pattern constancy: a report and pattern constancy related image galleries are available on `http://dragon.larc.nasa.gov/VIP/pattern_constancy.html`. Pattern constancy refers to the ability of the VS to produce enhanced imagery across dramatic lighting and turbidity variations as well as serious exposure errors. Scene patterns derived from VS enhanced versions of these wide ranging classes of images, and wide ranging extraneous image acquisition conditions, prove to be dramatically stable. These properties encourage us to think of the VS processing as a "universal front-end" both for immediate pilot enhanced vision display, and for pattern recognition processing such as the SVT described here. Specialized processing for scene analysis for various aviation safety purposes follows the SVT. The results of this specialized processing then feed into a flight monitor/pilot advisor. Since the SVT also performs noise reduction, it can also be employed for direct noise improved enhanced vision display to the pilot. In addition, a real-time noise-reduced pattern vision is provided to the pilot during extreme poor visibility imaging (and flight) conditions. In Section 2 we will describe the general design of the matched filters, the specifics of the lexicon itself. In Section 3 we describe the tree design of the SVT based on the FOO data. And, finally in Section 4 we describe the information theoretic system of pattern representation.

2. BASIC ELEMENTS OF MATCHED FILTER DESIGN AND PATTERN SELECTION

Our first task was to define the largest regional scale possible to achieve as much noise reduction as possible while still avoiding the combinatorial explosion of too many visually significant pattern forms. An initial exercise of examining real images at the micro-pattern scale led us very quickly to conclude that a regional scale of 5×5 pixels was as large as we could use without rapidly increasing the pattern space. Smaller scales did not provide sufficient pattern specificity over simple edge detection nor did they achieve significant noise reduction.

With this scale of the 5×5 pixel region established, we examined the visually significant structural patterns in images, with particular emphasis on fine printed text—font sizes between 10–12. This latter arose from the logic that reading fine print is the most demanding high spatial acuity task of visual perception, so we should encompass the structures that occur in this process, as well as those that occur in natural images. Further, fine printed text appears to be the densest and most complex pattern information that is visually significant. We

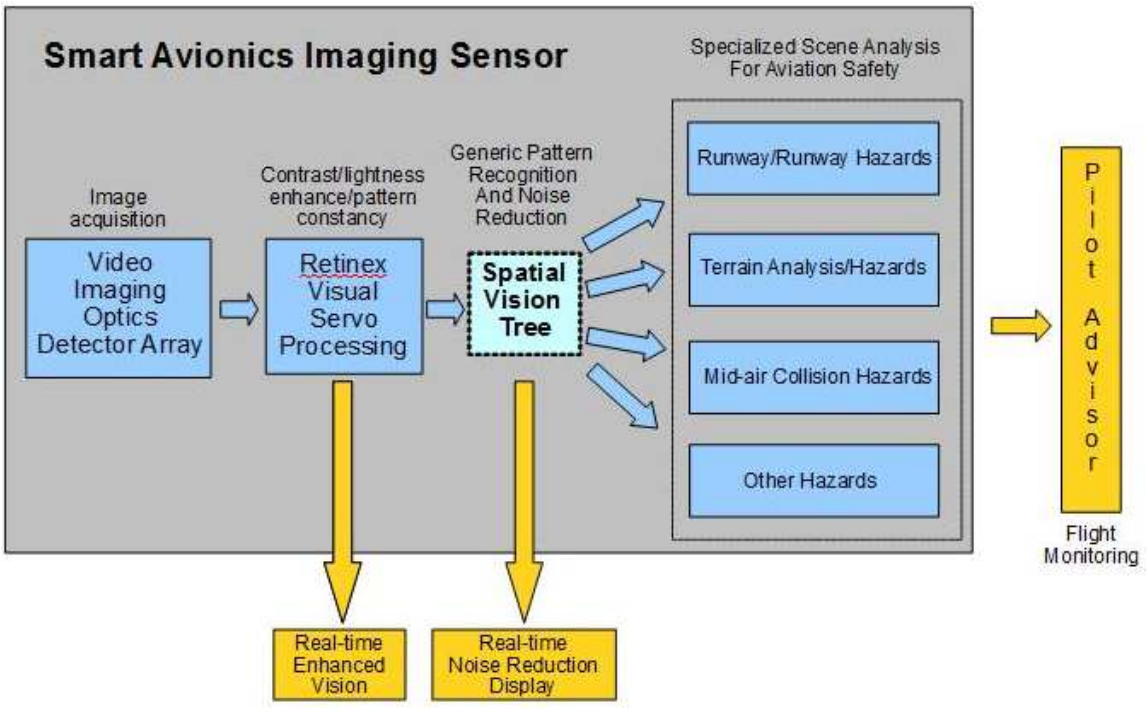

Figure 2. A Smart Avionics Imaging Sensor Concept (inside box)

simply used our own visual perception to judge whether micro-patterns were visually significant. This exercise led to an initial lexicon of 690 binary filters shown in Figure 3.

Peromaa and Laurinen,[11] and Perna and Morrone[12] provide psycho-physical evidence that brightness perception and edge structure perception split between high and low spatial frequency "channels." With this in mind, we designed the matched filters to be high-pass filters that have a response of zero in regions of uniform intensity. This was done by letting the filter coefficients take values ± 1, i.e., $f_{ij} \in \{+1, -1\}$, and adjusting the weights w_{ij} associated with the coefficients such that

$$\sum_{i=0}^{5} \sum_{j=0}^{5} w_{ij} f_{ij} = 0 \tag{1}$$

The lexicon constructed this way is a set of explicit pattern templates intended to be a *complete* set of all visually significant forms. The implication of this is that all patterns not in the lexicon, whether random noise patterns or deterministic structures, are not treated as signal. This pattern lexicon was run on a battery of complex natural images. An example of the processing results is shown in Figure 4. While the pattern richness and accuracy that was captured was encouraging, a closer examination revealed a large number of missing patterns. Many of these were due to pattern line width variations due to image sampling errors, but others were more finely graduated orientation (Figure 5) and shape variants of simple line features and fine structure (one pixel wide) linear "tightly looped" patterns we did not pick up in our initial foray into the study of diverse natural images. As we compiled and added all of the cases of missing patterns, we increased the pattern lexicon to $\sim$1250 (not illustrated because the filters are too small to discern in a reasonable sized figure). This set seems to be close to a complete set of visually significant patterns. We will however continue to examine pattern processing results for complex natural images as well as computer-generated text and other artificial test targets to determine whether there are any other missing patterns. The lexicon construction is clearly an evolving iterative process, but we do not anticipate having to add more than a handful of additional patterns. Therefore we feel this represents a solid first cut at the lexicon.

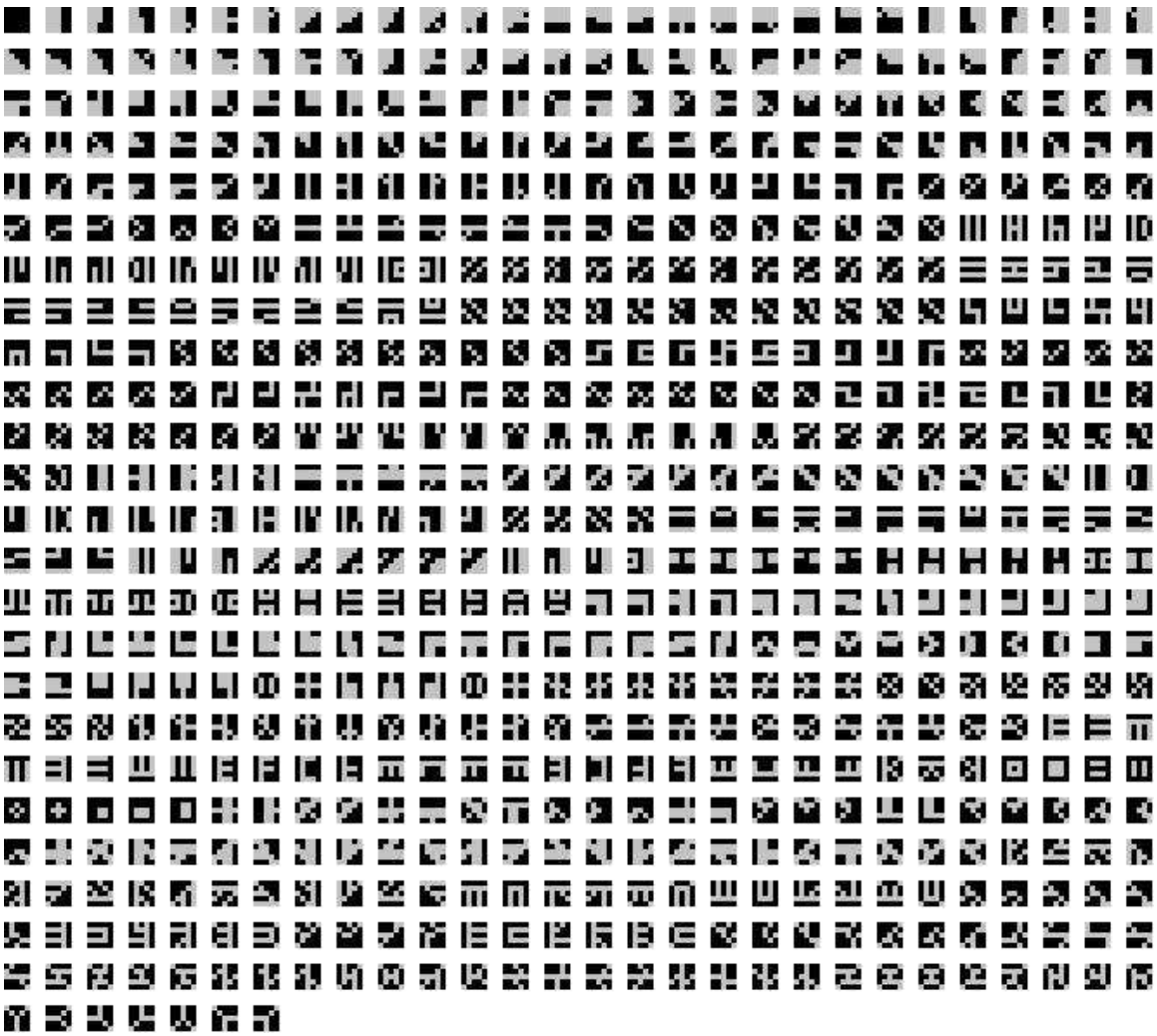

Figure 3. The original pattern lexicon of 690 patterns

2.1 A Cryptographic Manual of Spatial Vision

As we strove to achieve a practical computation for eventual real-time implementation, it was obvious the number of matched filtering operations based on the 1250 filter lexicon was too large. Hence, as was hinted at in Section 2, we developed a tree structure for the lexicon, with the main structural filters occupying the roots, and their variants occupying the branches under the respective roots. While manual implementation of this idea is viable for a small set of filters, it becomes tedious, and susceptible to error as the number of filters increases. For this reason, we decided to construct the tree using a cryptographic approach: namely that the layout of the tree should be based upon the FOO of patterns in a large statistical aggregate of highly diverse natural images. The FOO data were aggregated by using the 1250 pattern lexicon on a set of 1450 diverse natural images. The size of the set was large enough to provide a reasonable degree of statistical convergence.

Figure 6 shows a key trend in the FOO data. As is clear from Figure 6, each addition of a structural element to the matched filter, i.e., an increase in structural complexity, leads to about an order of magnitude drop in the FOO of that pattern. Note, that the logarithmic scale used in Figure 6 visually diminishes this dramatic fall-off, but also shows an almost linear decrease in the FOO as a function of pattern complexity. This has important ramifications as we will show later, but for now this is the strongest factor in guiding the design process of a two-tier tree. Some other trends are that FOO falls off linearly with increasing edge or line shape, and the FOO

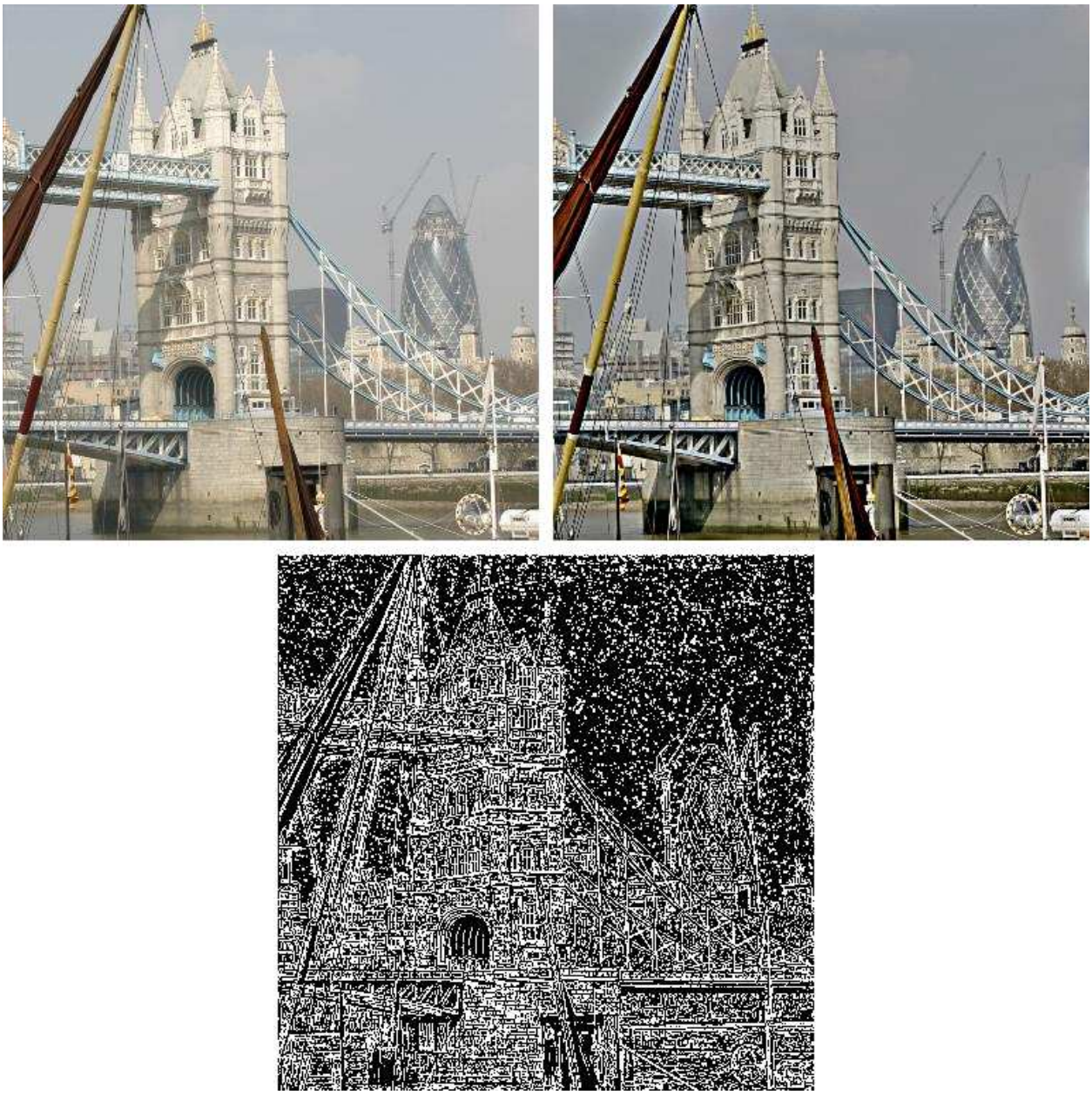

Figure 4. Example of SVT pattern processing: (top-left) Original image (top-right) RVS enhanced; (bottom) SVT pattern result.

data exhibited very strong orientation preferences for vertical and horizontal pattern forms. The lattermost result was unexpected and has been traced to two sources:

1. For normal terrestrial images, people automatically take images aligning the frame with dominant horizontal or vertical scene elements.
2. The square lattice detector array sampling grid pattern is imprinted on all image data.

As we compiled FOO data for nadir viewing aerial and orbital imagery, the orientation preference diminished by about half but was still a strong feature of the data. This remaining orientation preference is due to the

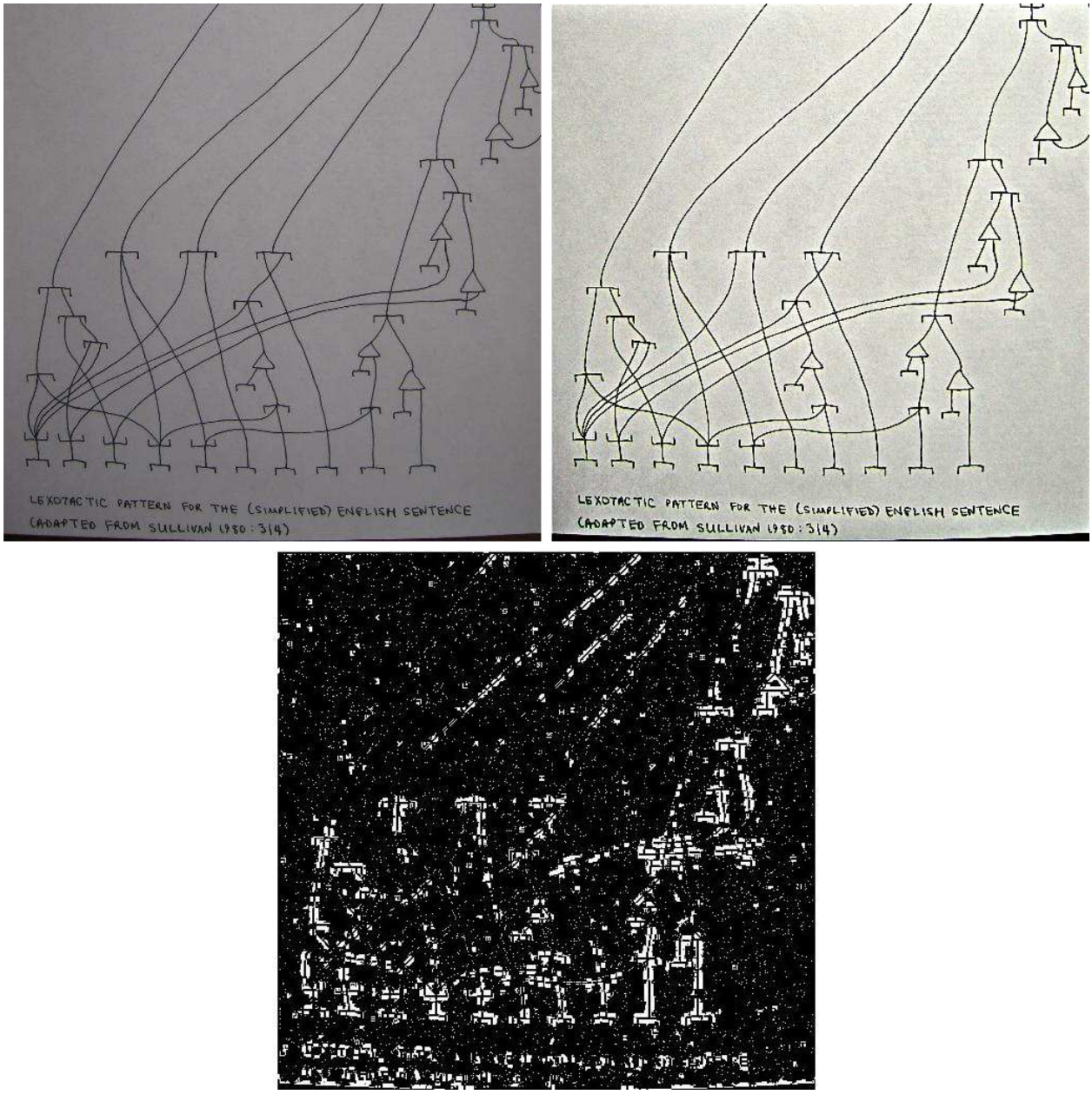

Figure 5. Example of missing finely graduated orientation patterns: (top-left) Original image (top-right) RVS enhanced; (bottom) SVT pattern result.

second of the two sources outlined above, i.e., the sampling lattice. For forward looking imaging, of relevance to aviation sensors, the vertical orientation preference diminished, but not the horizontal. This is entirely expected and consistent with most flying being done with aircraft in a horizontal orientation. Note that the 1450 image data set used to compute the FOO data was comprised of high quality JPEG images, so JPEG block artifacts due to lossy JPEG should not be a concern. A visual examination of the image pattern results also did not exhibit systematic block artifact patterns.

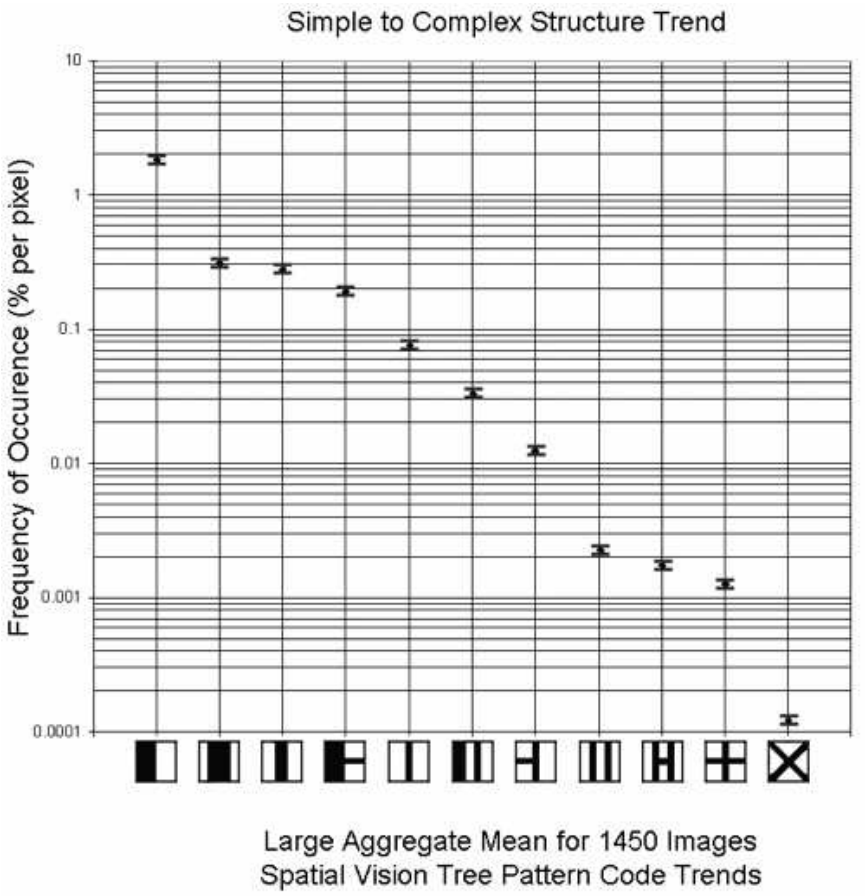

Figure 6. Key trends in FOO data: the FOO drops of exponentially with increasing structural complexity. The displayed patterns include the full range of pattern complexity, from some of the most frequently to some of the most rarely observed.

3. THE DESIGN OF THE SPATIAL VISION TREE

The goals of the tree design are two-fold: (1) achieve accurate regional pattern detection with fewest small kernel spatial convolution operations, and (2) generate the SVT design as automatically as possible—ideally fully-automated. The first goal can be assessed by comprehensive experimental testing, so we turn to the second goal, namely the tree design. The following approach to the two-tier tree design was employed:

1. The patterns are rank-ordered by decreasing FOO: the pattern with the highest FOO ranks as #1, and the one with the lowest FOO as #1250.
2. A list of roots and children is maintained at each presentation of a pattern, with each root possessing a list of children. The process commences with a single root corresponding to the highest FOO.
3. As each new pattern is presented, it is compared with all the roots in the root list. The comparison is performed using the Hamming distance (HD), which measures bit-wise similarity between two binary patterns.
   (a) If the pattern has an HD > 4 with all the roots in the roots list, then it is classified as a root, and added to the root list.
   (b) Otherwise, it is classified as a child of the root with which it has the smallest HD.
   (c) In the case where the HD is the same with several roots, the child is assigned to root with the highest FOO.

This process automatically assigns each pattern either to a root or to a child while preserving the FOO order. It also guarantees that children have a smaller FOO than their parent roots, and roots are arranged from left to right in decreasing order of FOO. The process led to a tree design with 418 first tier members and a maximum population in the second tier of 9 members. Thus, the maximum number of convolutions during tree operation is $418 + 9 = 427$ compared to the 1250 convolutions for treeless pattern detection. This fully-automated tree design did result in a significant amount of "blurring" of structural classes, because HD only measures structural differences and not structural similarities. This made it possible for children to be assigned to a non-optimal parent root.

In an attempt to retain a clearer scientific understanding of visual structural groupings, the fully-automated design process was replaced by laying out the first tier manually. The process still uses the FOO to arrange the first tier patterns, but makes sure that patterns representing each structural class are represented in the first tier. This implies that instead of starting with an initial tree that has a single root, the process starts with a tree that has several roots in the roots list. The second tier is then populated by using step 3 of the algorithm described above. For example, if we manually set a one pixel wide vertical line pattern as a root, the automated search should assign all other patterns that have HD ≤ 4 to that root. This provides a significantly different tree design. For this approach the tree design results in 363 roots, and a maximum population in any second tier of 27 members. So the maximum number of convolutions for this alternate tree design is $363 + 27 = 390$, slightly fewer than the fully automated design. Since the structural relationships within this tree design are much easier to study, it is experimentally easier to interpret. However, we will go forward with both design options until it becomes clear which has superior performance in experimental testing on a large aggregate of quite diverse images. Another design iteration of both of these designs may be necessary if we uncover any missing patterns that alter the basic layout of the first tier roots and that reshuffles second tier memberships.

4. AN INFORMATION THEORETIC APPROACH TO PATTERN REPRESENTATION

The previous exposition has centered upon what we would call the first stage of the SVT pattern detection. But a second stage of pattern representation is also required. In the first stage, each *pixel* in the image is examined, and the pattern matching the 5×5 neighborhood about that pixel is used to classify that pixel to a pattern. So a 5×5 region actually has 25 possible pattern matches after the first stage. However, in any 5×5 region, there should only be a single pattern that produces a maximum response. This implies that of the 25 patterns that are obtained from the first stage, we need to find the one that has the maximum response in any given 5×5 window. Part of the problem in determining this "winner" has to do with the basic sampling constraints on the 5×5 regional pattern detection process. The problem arises from the fact that the maximum response within a 5×5 window does not necessarily correspond with the center pixel within that window, and may correspond to a pixel in a neighboring 5×5 space. Hence, the sampling space of winners is not necessarily on a uniformly spaced rectangular grid.

In addition, while we made every effort to eliminate spatial translational variants by 2 pixels or less from the pattern lexicon, this is not possible for such highly symmetric patterns as say a vertical straight line pattern which is identical under vertical translation. But more importantly, it is obvious from Figure 7 that once we move over three pixels in image space we would encounter a possible new pattern overlapped with the pattern three pixels back. These two patterns should, in an ideal world, match exactly in the overlap zone. But given that there will be errors due to noise and pattern approximation errors inevitable in the pattern detection stage, we must have some scientific foundation for giving a pattern priority to one pattern over others in these overlap zones. Information theory and the FOO data supplies just this tool.

In information theory, the information conveyed by the occurrence of an event is inversely proportional to its probability of occurrence. We can use this idea as a tie-breaker to determine a winner within a window. If we find more than one pattern has a maximum response within a 5×5 window, then the one with the largest information content, i.e., the smallest FOO is declared the winner. Interestingly, this inverts the rule used for tree design which places patterns with higher FOOs higher within the tree structure. We have not explored this idea, aside from some preliminary graphical studies. We will, however, test this hypothesis fully as a second stage of SVT processing after the pattern detection stage. For the SVT results shown in this paper, a simple expedient method

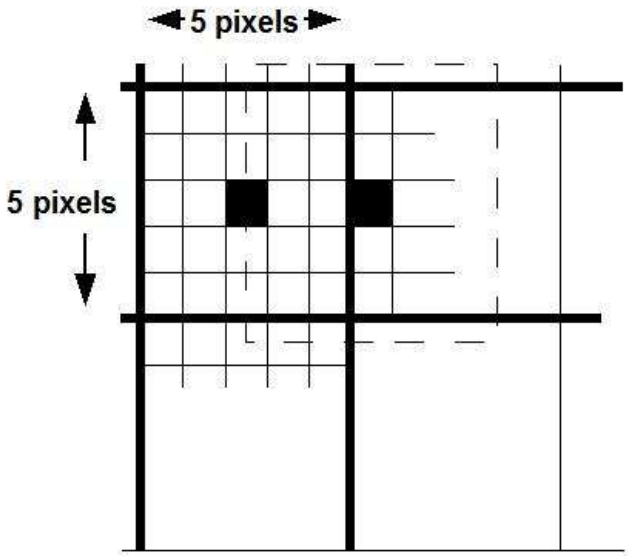

Figure 7. Basic sampling in the first stage of pattern detection.

was used which divided the image space into a grid of 5×5 pixels with abutting (not overlapping) regions, and the pattern detected closest to the center pixel of each region was selected as the detected pattern unless that pattern did not meet basic threshold criteria for sufficient pattern contrast. The information theoretic approach provides a solid foundation for a pattern sorting order which we expect to have far reaching consequences as we move in the future from micro-pattern recognition to the macro-recognition object and scene recognition based upon the SVT lexicon.

5. CONCLUSIONS

While the results we have described here are in an early stage of development and testing, we are encouraged that our original hypothesis is correct—that a generic spatial vision pattern lexicon exists and is tractable in terms of a sufficiently small number of patterns being able to capture all of the visually significant structures that can occur in complex and arbitrary natural images at the regional scale of 5×5 pixels. On-going research addresses (1) further experimental testing of alternate tree designs to assess their performance over wide-ranging input imagery in terms of pattern representation accuracy, (2) discovery of any missing patterns that would have to be added to lexicon, and (3) the implementation and testing of the information-theoretic approach to pattern representation. Further, we plan to explore the noise reduction aspect of the SVT pattern's matched filtering but this latter topic is secondary at this point to testing the pattern recognition aspect of the SVT, and defining how this micro-pattern recognition can be fleshed out into a full, generic object and scene recognition system.

ACKNOWLEDGMENTS

The authors wish to thank the NASA Aviation Safety Program for the funding which made this work possible. In particular, Dr. Rahman's work was partially supported under NASA cooperative agreement NNL07AA02A.

REFERENCES

1. Ruderman, D., "The statistics of natural images," *Network: Computation in Neural Systems* **5**(4), 517–548 (1994).
2. North, D. O., "Analysis of the factors which determine signal/noise discrimination in radar," Tech. Rep. Rept. PTR-6C, RCA Laboratories, Princeton, N. J. (June 1943).

3. North, D. O., "An analysis of the factors which determine signal/noise discriminations in pulsed-carrier systems," *Proc. IEEE* **51**, 1016–1027 (1963).
4. Turin, G., "An introduction to matched filters," *IRE Transactions on Information Theory* **6**(3), 311–329 (1960).
5. Rahman, Z., Jobson, D. J., and Woodell, G. A., "Retinex processing for automatic image enhancement," *Journal of Electronic Imaging* **13**(1), 100–110 (2004).
6. Jobson, D. J., Rahman, Z., and Woodell, G. A., "A multi-scale Retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing: Special Issue on Color Processing* **6**(7), 965–976 (1997).
7. Woodell, G. A., Jobson, D. J., Rahman, Z., and Hines, G. D., "Advanced image processing of aerial imagery," in [*Visual Information Processing XV*], Rahman, Z., Reichenbach, S. E., and Neifeld, M. A., eds., Proc. SPIE 6246 (2006).
8. Jobson, D. J., Rahman, Z., and Woodell, G. A., "Feature visibility limit in the nonlinear enhancement of turbid images," in [*Visual Information Processing XII*], Rahman, Z., Schowengerdt, R. A., and Reichenbach, S. E., eds., Proc. SPIE 5108 (2003).
9. Jobson, D. J., Rahman, Z., Woodell, G. A., and Hines, G. D., "The automatic assessment and reduction of noise using edge pattern analysis in nonlinear image enhancement," in [*Visual Information Processing XIII*], Rahman, Z., Schowengerdt, R. A., and Reichenbach, S. E., eds., Proc. SPIE 5438 (2004).
10. Woodell, G. A., Jobson, D. J., Rahman, Z., and Hines, G. D., "Enhancement of imagery in poor visibility conditions," in [*Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense IV*], Carapezza, E., ed., Proc. SPIE 5778 (2005).
11. Peromaa, T. and Laurinen, P., "Separation of edge detection and brightness perception," *Vision Research* **44**(16), 1919–1925 (2004).
12. Perna, A. and Morrone, M., "The lowest spatial frequency channel determines brightness perception," *Vision Research* **47**(10), 1282–1291 (2007).