Representing and Detecting Geons by Joint Statistics of Steerable Pyramid Decomposition*

Xiangyu Tang, Kazunori Okada, Christoph von der Malsburg†
Laboratory of Computational & Biological Vision
Computer Science Department, University of Southern California
HNB228 Los Angeles, CA 90089 USA
{tangx, kazunori, malsburg}@organic.usc.edu

Abstract

We present a parametric method to represent and detect geons. The parameters are extracted from joint statistical constraints defined on complex wavelet transform. We first review how steerable pyramid may be used in multi-scale multi-orientation image decomposition. Then, four stages of object recognition theory is adopted to support the choice of joint statistical constraints, which characterize geons in a high dimensional parameter space. This parametric representation is examined in detail under circumstances when geons change in orientation, location, and size. Constraint-wise similarity is introduced to describe the corresponding statistics variations. Finally, we present details of a geon detection system, and demonstrate successful experimental results as well as system limitations.

1 Introduction

The term "geon" refers to geometric shape primitive that serves as basic building element for more complex structures. Biederman [1] differentiates geons by shape of cross-section, curvature of axis, and size of cross-section, etc.; the full family of geons thus has 24 members. For simplicity and easy illustration, in this paper we focus on three typical geons: cylinder, cone and bended cuboid. But the method and results is also applicable to other geons.

Implied by three-geon sufficiency theory [1], objects can be quickly and accurately recognized if an arrangement of three geons (or less) composing the object is recovered from the image. The essential issue is then how to detect these geons in the objects. A successful detection must address both identification and localization problems, namely "what is it" and "where is it". A good geon representation is the key to answer the questions. Suppose we have pre-computed such representation instances for each single geon as model data, then we can evaluate each geon component of the object by scanning the image at local positions. Local image region that is the most similar to a model provides the answer to the "what is it" question, while its location answers the "where is it" question.

There has been extensive efforts focusing on recovery of geon-like structures for object recognition. Hummel and Biederman [2] proposed a neural network model of viewpoint invariant visual recognition. The model's structural representation specifies both an object's visual attributes (e.g., edges, vertices) and the relations among them, however it only concentrates on line drawing objects. In [3], Dickinson and Metaxas decouple the processes of object recognition and localization by selectively integrating the qualitative and quantitative shape recovery components. Although flexible, the system only uses perfect geon-like objects and avoids fine structural details. Wu and Levine [4] introduced a superquadric model called parametric geons representation, which can also deal with imperfect geon-like object inputs, while having comparable performance as in [3]. In this paper, we present a general parametric model to characterize and detect geons based on statistical constraints that are originated from popular object recognition theories.

Portilla and Simoncelli proposed a parametric model based on joint statistics of complex wavelet coefficients [5], and successfully demonstrated their algorithms' capability of analyzing and synthesizing visual textures. A brief review of this method would be given in section 2. Then, in section 3, we apply this parametric model as a representation for geons,

*The authors wish to thank Irving Biederman for all helpful hints and discussions.

†Also at Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum Germany.

and reason the appropriateness. Following these arguments, we perform geon detection task in section 4. Experimental results and system limitations will be discussed in section 5.

2 Joint Statistics Approach

Statistical models have been used widely to characterize images. Particularly, visual textures are of primary concern because they are spatially homogeneous and contain repeated elements, which subject to statistical description. Based on the use of linear kernels at multiple scales and orientations, recent year's development of wavelet representations enabled design of more practical and powerful statistical models.

2.1 Steerable Pyramid

In [5], Portilla and Simoncelli designed a universal statistical model that enforces four statistical constraints on a multi-scale multi-orientation wavelet decomposition of images. The set of wavelet filters they adopted is known as "steerable pyramid" [6, 7], named after their properties of steerability (multi-orientation) and scalability (multi-scale). Steerable pyramids recursively split an image into a set of oriented subbands and a lowpass residual, thus resulting in independent representation of scale and orientation. In the frequency domain, the filters implemented for this transformation are polar-separable. Analytically they can be written as:

$$L(r,\theta) = \begin{cases} 2\cos\left(\frac{\pi}{2}\log_2\left(\frac{4r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 2, & r \le \frac{\pi}{4} \\ 0, & r \ge \frac{\pi}{2} \end{cases}$$

$$B_k(r,\theta) = H(r)\cdot G_k(\theta), k \in [0, K-1]$$

where r and θ are polar frequency coordinates, K stands for the total number of orientations. $L(r,\theta)$ is the lowpass filter, that downsamples an image by 2 along both axes. The bandpass filter at each orientation k is composed of radial part $H(r)$ and angular part $G_k(\theta)$:

$$H(r) = \begin{cases} \cos\left(\frac{\pi}{2}\log_2\left(\frac{2r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 1, & r \ge \frac{\pi}{2} \\ 0, & r \le \frac{\pi}{4} \end{cases}$$

$$G_k(\theta) = \begin{cases} \alpha_K[\cos(\theta - \frac{\pi k}{K})]^{K-1}, & |\theta - \frac{\pi k}{K}| < \frac{\pi}{2} \\ 0, & \text{otherwise.} \end{cases}$$

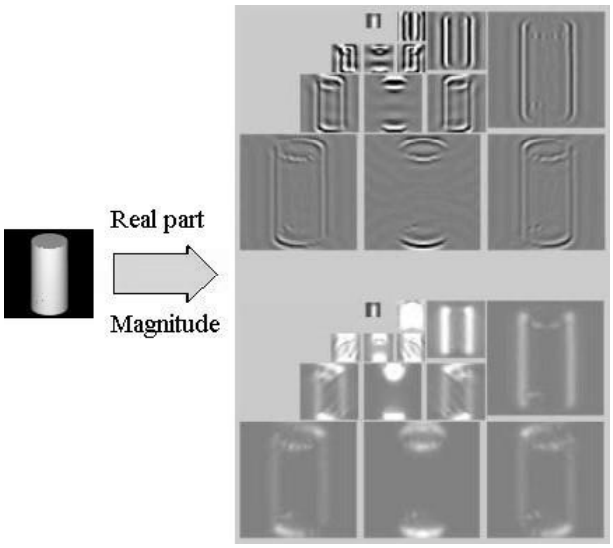

Figure 1: Complex steerable pyramid representation of a cylinder-like geon. The upper half is the real parts of the filtered images, the lower half is the corresponding magnitude representation. Both are pyramids of 3 levels and 4 orientations.

Here, α_K is a constant for a fixed K, it's defined as

$$\alpha_K = 2^{K-1}\frac{(K-1)!}{\sqrt{K[2(K-1)]!}}.$$

Before we recursively build the steerable pyramid, the process is initialized by splitting the input image into lowpass and highpass bands without downsampling, using the filters:

$$L_0(r,\theta) = L(\frac{r}{2},\theta)/2$$
$$H_0(r,\theta) = H(\frac{r}{2},\theta).$$

Figure 1 gives an example of steerable pyramid decomposition of an image containing a cylinder. Since the wavelet coefficients are complex, we show the pyramids of both the real part and the corresponding magnitude. Each pyramid has 3 levels and 4 orientations.

2.2 Statistical Constraints

In order to extract the key features from images' steerable pyramid representations, Portilla and Simoncelli [5] defined four statistical constraints on the complex coefficients of this decomposition. Besides traditional marginal statistics and correlation measures, these statistical constraints also include

joint statistics [8] motivated by nonparametric models as in [9]. Such inclusion gives visually impressive synthesis results. The statistical constraints are described as follows

• Marginal Statistics: Mean, variance, skewness, kurtosis, minimum and maximum values of the gray level image pixels, skewness and kurtosis of the lowpass images at each scale, and variance of the highpass band. These capture pixel intensity distribution. There are totally $6 + 2(N + 1) + 1$ parameters, where N is the number of pyramid levels.

• Raw Coefficient Correlation: Autocorrelation of the partially reconstructed lowpass images at each scale. Since the steerable pyramid decomposition is highly overcomplete, such autocorrelation measure is quite redundant, a more efficient method only considers central samples of a $M \times M$ area. These capture periodic or globally oriented structures. There are totally $(N + 1)\frac{M^2+1}{2}$ parameters.

• Coefficient Magnitude statistics: Central sample of magnitude autocorrelation in each subband, cross-correlation of magnitudes in each subband with those in the same pyramid level of different orientations, and cross-correlation of subband magnitudes with those of all oriented subbands in a coarser scale. This statistical constraint captures important structural information such as edges, corners, etc. It includes totally $N \cdot K \cdot \frac{M^2+1}{2} + N \cdot \frac{K(K+1)}{2} + K^2(N - 1)$ parameters.

• Cross-Scale Phase Statistics: Cross-correlation of coefficients' real part with both the real and imaginary parts of phase doubled coefficients of all oriented subbands in the next coarser scale. The constraint captures effects of illumination gradients due to objects' 3D appearance. It contains totally $2K^2(N - 1)$ parameters.

The parameter sets derived from the above four constraints form a *universal* image representation. It generates 710 parameters when we choose $N = 4$, $K = 4$, $M = 7$. For a 64×64 image, this is already an economic model.

3 Representing Geons

The joint statistics model reviewed in the last section is originally proposed to characterize texture images. However, the entire parameter set not only captures periodicity or repeated structures, but also pays attention to global orientations, edges, corners, and 3D lighting gradients. These properties make the parameter representation also effective to distinguish images containing geon-like structures.

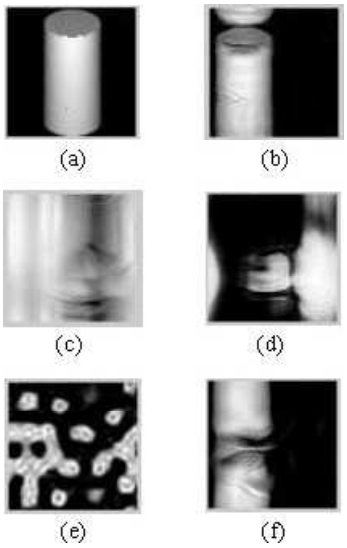

Figure 2: Synthesis results. (a) the original geon image containing a cylinder; (b) synthesis by using all of the four statistical constraints; (c) synthesis by removing the marginal statistics constraint; (d) synthesis without the raw coefficient correlation constraint; (e) synthesis without the coefficient magnitude statistics constraint; (f) synthesis without the cross-scale phase statistics constraint.

In comparison with other object recognition theories, the model can be viewed as an original development and implementation of David Marr's four stages of object representation [10] described as follows

• Pixel based: light intensity;

• Primal sketch: local geometrical structures, detection of illumination effects;

• $2\frac{1}{2}D$ sketch: orientation and depth information of surfaces;

• $3D$ model: shape representation, spatial relationship.

The *marginal statistics* constraint characterizes exactly pixel based intensity distribution, which serves the first stage of Marr's object recognition process. At the stage of primal sketch, recognition involves detecting and analyzing local geometric structures as well as illumination effects. These are what the constraints of *coefficient magnitude statistics* and $cross - scale\ phase\ statistics$ try to extract from an image's steerable pyramid decomposition. The correlation measure enforced by the *raw coefficient correlation* constraint captures oriented structures, that is similar to the $2\frac{1}{2}D$ sketch. But since it is not specific for surface orientation, the analogy is not strict. At the final stage, the $3D$ shape representation is formed by utilizing the information ob-

tained in previous processing stages. In other words, the parameters resulted from the analysis of the four statistical constraints define a multidimensional representation of our $3D$ geons. This can be realized in the image synthesis process.

Although the recognition task particularly depends on the analysis process to capture the key features from the input images, successful synthesis results help to confirm that the choice of the parameter set is appropriate. The synthesis algorithm introduced in [5] recursively imposes the statistical constraints parameter set on the lowpass and oriented bandpass subbands, which are initially constructed by decomposing an image containing Gaussian white noise. Different from Marr's four stages of object recognition theory, this synthesis process is performed in parallel fashion, such that in each iteration one can choose to simultaneously impose a few constraints while neglecting the others. Figure 2 illustrates synthesis results of a 64×64 geon image containing a single cylinder. Part (a) is the original image, (b) is synthesized by using all of the four statistical constraints. Visually, the result is very close to (a), it captures both the top and the front surfaces, and preserves the lighting effects due to the geon's $3D$ appearance. The dislocated position of the synthesized cylinder is caused by the shift-invariant property of the pyramid decomposition. Actually, the same geon located in different positions should share the same parameter values, and this can benefit geon detection task, which will be discussed in the next section. Part (c) is synthesized with all but the *marginal statistics* constraints. Clearly, the pixel intensity distribution doesn't agree with (a). The result of synthesis without *raw coefficient correlation* is shown in (d), it fails to recover the global continuous structure of the geon. The synthesis result in (e) has no continuous edge or distinguishable corner, this is caused by omitting *coefficient magnitude statistics*. Although (f) roughly improves reconstruction quality, it can't tell $3D$ details or correct lighting effects without *cross − scale phase statistics* constraint. The successful synthesis in (b) and failures in (c), (d), (e) and (f) manifest the necessity of including all of the four statistical constraints and the appropriateness of using the computed parameter set as a geon representation.

An immediate question would then be how good this representation is. In other words, when the geons rotate in direction, translate in location, or scale in size, can the model tolerate the changes? And to what degree? In order to investigate this issue, we compute a series of parameter sets extracted from geon images, in which geons gradually change in

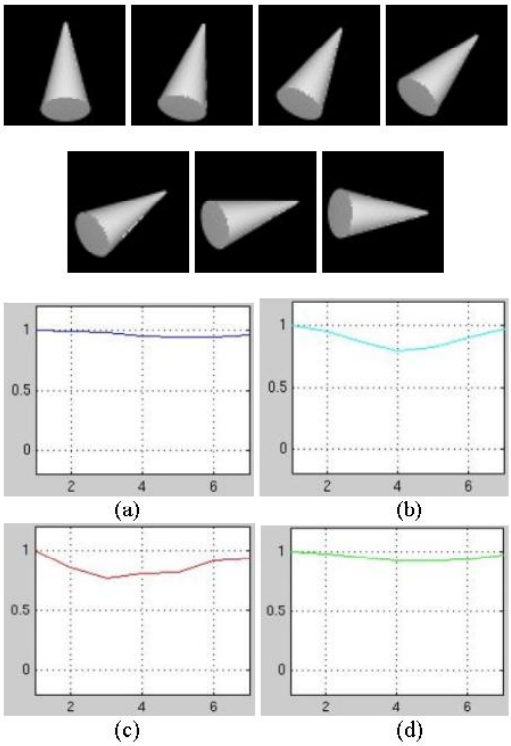

Figure 3: Constraint-wise similarity plots when a cone rotates from vertical to horizontal. Totally 7 inputs are used, each is compared with the first one. (a) is the similarity plot for the marginal statistics subset, (b) is for raw coefficient correlation, (c) is for coefficient magnitude statistics, and (d) is for cross-scale phase statistics.

orientation, location, or size, and then compare the parameter sets by calculating their similarity values. The comparison is constraint-wise. Namely, we'll have four similarity values between two images, and each of them is the comparison result of the parameter subset for one of the four statistical constraints. Thus, we can study and dynamically assign each constraint's contribution to the overall similarity. Suppose we have $\mathcal{A}_{ms}$ and $\mathcal{B}_{ms}$ as the parameter subsets defined by *marginal statistics* constraint of image $\mathcal{A}$ and image $\mathcal{B}$, then the similarity between them can be given by:

$$\mathcal{S}_{ms} = \frac{\|\vec{\mathcal{A}_{ms}}\| + \|\vec{\mathcal{B}_{ms}}\|}{\|\vec{\mathcal{A}_{ms}}\| + \|\vec{\mathcal{B}_{ms}}\| + \|\vec{\mathcal{A}_{ms}} - \vec{\mathcal{B}_{ms}}\|},$$

here the parameter subsets are treated as high dimensional vectors. This formula takes into account magnitude of the vectors as well as their angle relation. The value goes 1 only when $\vec{\mathcal{A}_{ms}}$ is identical to $\vec{\mathcal{B}_{ms}}$. The similarity values corresponding to other three constraints will also be similarly computed.

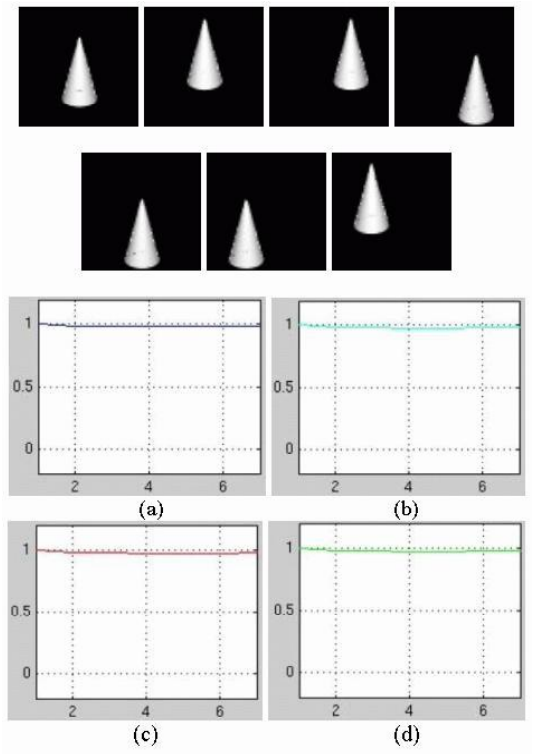

Figure 4: Constraint-wise similarity plots when a cone translates in location. The arrangement of the plots is the same as in Figure 3.

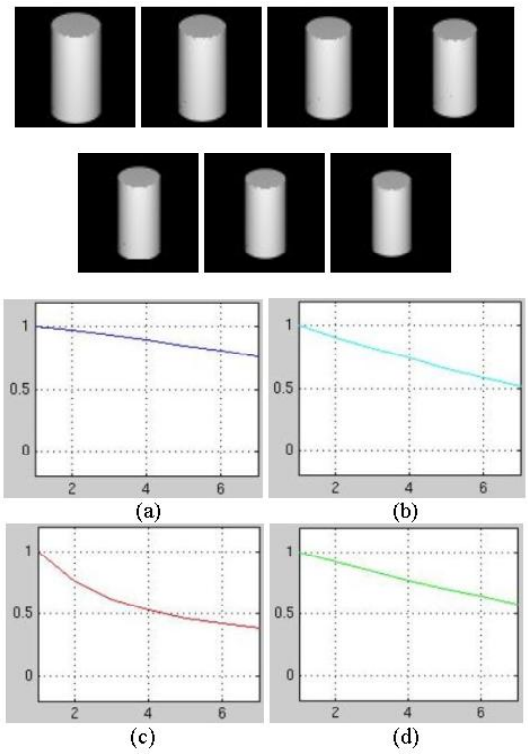

Figure 5: Constraint-wise similarity plots when a cylinder scales down in size. The arrangement of the plots is as the same as in Figure 3.

Figure 3 – 5 are constraint-wise similarity plots when the input geon varies in orientation, location, and size. In each case, 7 input images are used, each is compared with the first one. The four similarity values all stay high in Figure 3, and the average always exceeds 0.85 (the similarity threshold we choose). The rotation-invariant property of steerable pyramid decomposition plays a pivotal role here. The valleys in (b) and (c) are affected by the choice of K, the number of subband orientation, which is 4 in our case. In Figure 4, all of the four similarity values are almost always 1.0. This suggests geons of only different locations actually share the same parametric representation. As mentioned before, this is because the pyramid decomposition we use is also shift-invariant. However, for the scaling case, the parametric representation doesn't tolerate large variations. As shown in Figure 5, if a cylinder linearly scale down 25%, the average similarity drops to 0.5, which is below the similarity threshold. Only in the linear range of $\pm 10\%$, the similarity is above 0.85, and the geon is close to the original input in the parametric space. So, in order to detect geons with various sizes, we need to have model geons of every 10% difference in size. An alternative method will be discussed in section 5.

4 Detecting Geons

Having the parametric geon representation, we can then detect geons in a scene that contains several geon-like structures. The idea is that 1) pre-analyze the parameter sets of single model geons, and put them in memory; 2) slide a window, which is of the same size of the model geons, over the input image; 3) compute the parameter set at each position every a few pixels away; 4) constitute similarity saliency maps by comparing the parameter set with the ones in memory; 5) if the maximum similarity value(s) exceeds some threshold, then a corresponding geon is detected at the position given by the saliency map. Figure 6 shows the systematic view of how it works.

The pre-analyzed memory sets we use in this experiment are constraint-wise parameters extracted from three typical geons: cylinder, cone, and bended cuboid, which are shown in Figure 7. The image size of these model geons is 64×64. The inputs to be searched on are synthetic 256×256 images that contain several such geons with various orientations and locations as well as a little size changes. The scanning step is set to 8 or 16 pixels, such that the detection task can be completed in 1–5 minutes on a Pentium III PC with a 800MHz processor.

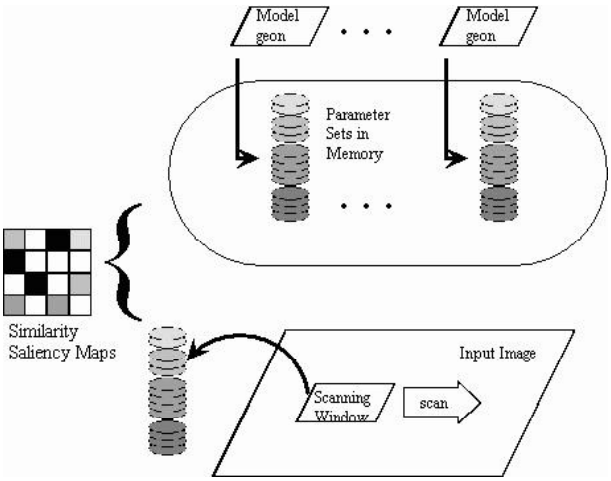

Figure 6: Detect geons with the parametric representation.

Figure 7: Three model geons used.

Figure 8 illustrates a detection result. The first row is the input image we searched. It has three separate geons comparable to the model geons such that the cylinder is rotated near 90° degrees, and the cone is scaled down 5%. The second row is the similarity saliency maps built by comparing the constraint-wise parameter set, which is extracted from the scanning window at every position, with the sets in memory. Each comparison gives four values due to the four statistical constraints. In this study, we use the average value of these four similarities to construct the saliency maps. In a saliency map, the brighter the region is, the closer the average similarity is to 1. The maximum value in a highlighted region corresponds to the location in the input image that may contain a known geon as in memory. If this value exceeds a threshold (0.85 here), then a successful detection occurs. Each saliency map is dedicated to one of the model geons, for instance, the second saliency map helps to detect the cone, but inhibits detection of other geons by darkening their corresponding regions. The third row shows the detected geons captured from the original input with the references of the similarity saliency maps.

Since we don't search the input image pixel by pixel, the captured geons may translate in location when the window slide over them. But the centers

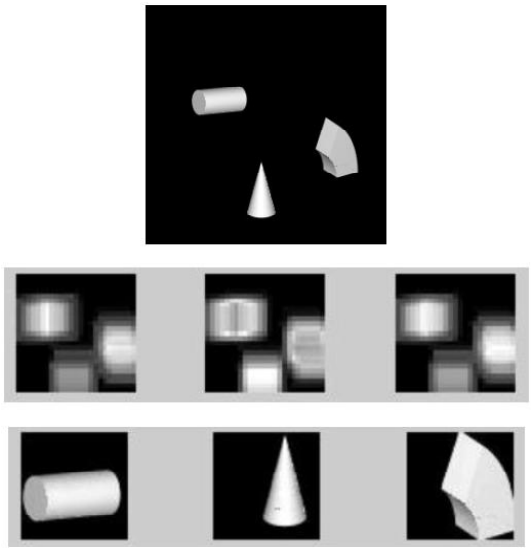

Figure 8: A geon detection result. The first row is the input image; the second row shows the similarity saliency maps dedicated to each model geon; the third row captures the detected geons from the input image.

of the images in the third row still serve as good approximations of their actual locations. Thus, a successful detection can tell both what and where the geon is. More precise localization can be expected by using finer resolution maps with trade-off for longer running time.

More examples are given in Figure 9 - 10. In Figure 9, geons are closely placed to each other, look more like parts of an object. The cone and the bended cuboid are partially occluded by the rotated cylinder, also the bended cuboid is slightly rotated in depth. However, the system still can capture each geon quite well. The input of Figure 10 has a texture background and 50% random noise as surface marking. By lowering the threshold to 0.8, successful detection can still be maintained. Since the bended cuboid is not included in Figure 10, no value in the corresponding similarity saliency map exceeds the threshold, thus no geon is captured in the last subgraph of Figure 10.

5 Discussion

We tested 40 trials, 34 of them (85%) give satisfactory detection results. The system performance can be affected by several aspects:

• The resolution of the similarity saliency map decides how precisely the geons are located. A coarse resolution may fail to contain a full single geon, thus

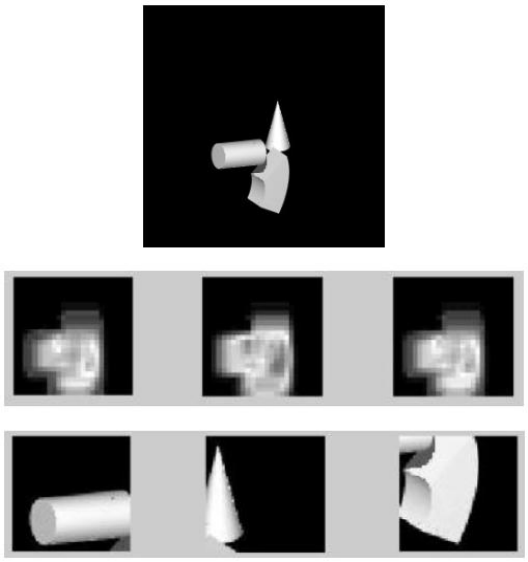

Figure 9: A geon detection result. The subgraph arrangement references Figure 8.

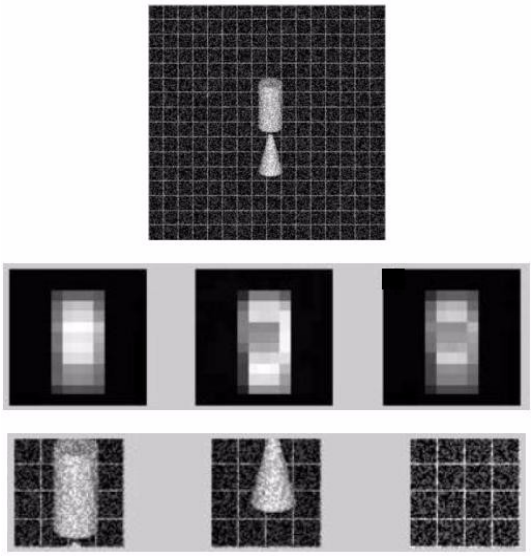

Figure 10: A geon detection result. The subgraph arrangement references Figure 8.

also lead to false recognition. Our choices of the scanning step generate 25×25 or 13×13 maps, as shown in Figure 8-10.

• A cylinder's corners (in 2D images) or edges from different viewpoint may give people false impression that it is a cone or cuboid. Such "corner effect" could generate brighter lines around a dark area in the saliency maps. Most times they do not exceed the threshold to affect detection results. But, in order to reduce the extreme cases, we can smooth the maps by convolving them with a small Gaussian window.

• Since the steerable pyramid decomposition is both shift-invariant and rotation-invariant, geons arranged at different orientations and locations can be detected easily. However, we have to limit size changes in $\pm 10\%$ as mentioned in section 3. To deal with larger scaling cases, we experimented multiple window search. Namely, a set of scanning windows of various sizes are used, before analyzing the statistical parameters, each window is adjusted to model geons' size by truncating or expanding its Fourier representation to 64×64. Although valid, this method is quite time consuming.

• In our experiment, we set the average of the four constraint-wise similarity values as the overall similarity, illustrated in the saliency maps. However, for a specific detection task, we can change the contribution of each constraint-wise similarity to the overall value. For instance, $cross - scale\ phase\ statistics$ varies sharply in a noisy background, we can then weigh down the corresponding similarity in order to make the system more noise tolerable.

In our future work, we plan to incorporate more model geons, study their statistical behavior when one morphs to another, and find their distances in the parametric space. On the other hand, although the system can survive slight rotation in depth, it's still desirable to find more effective statistical constraints to address such viewpoint invariant properties for geon detection.

References

[1] I. Biederman, Recognition-by-components: A theory of human image understanding, Psychological Review, 1987.

[2] J. E. Hummel, I. Biederman, Dynamic Binding in a Neural Network for Shape Recognition, Psychological Review, Vol. 99, No. 3, 480-517, 1992.

[3] S. J. Dickinson, D. Metaxas, Integrating Qualitative and Quantitative Shape Recovery, International Journal of Computer Vision, Vol. 13, No. 3, 1-20, 1994.

[4] Kenong Wu, M. D. Levine, 3-D Object Representation Using Parametric Geons, Centre for Intelligent Machines, McGill University, 1993.

[5] J. Portilla, E. P. Simoncelli, A Parametric Texture Model based on Joint Statistics of Complex Wavelet Coefficients, International Journal of Computer Vision, Vol. 40, issue 1, pages 49-71, 2000.

[6] E. P. Simoncelli, W. T. Freeman, The Steerable Pyramid: A Flexible Architecture for Multi-scale Derivative Computation, Second Annual IEEE Inter-

national Conference on Image Processing, Vol. III, pages 444-447, 1995.

[7] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, D. J. Heeger, Shiftable Multi-scale Transforms, IEEE Trans. Information Theory, 38(2): 587-607, Special Issue on Wavelets, 1992.

[8] R. W. Buccigrossi, E. P. Simoncelli, Image Compression via Joint Statistical Characterization in the Wavelet Domain, IEEE Trans Image Proc, 8(12): 1688-1701, Dec 1999.

[9] J. De Bonet, P. Viola, A Non-Parametric Multi-Scale Statistical Model for Natural Images, Neural Information Processing, Vol. 9, Dec 1997.

[10] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W. H. Freeman and Company, 1982.