

# Context-Inclusive Approach to Speed-up Function Evaluation for Statistical Queries : An Extended Abstract

Vijay Gandhi, James M. Kang, Shashi Shekhar  
University of Minnesota [gandhi, jkang, shekhar]@cs.umn.edu

Junchang Ju, Eric D. Kolaczyk, Sucharita Gopal  
Boston University [junchang, kolaczyk, suchi]@bu.edu

## Abstract

Many statistical queries such as maximum likelihood estimation involve finding the best candidate model given a set of candidate models and a quality estimation function. This problem is common in important applications like land-use classification at multiple spatial resolutions from remote sensing raster data. Such a problem is computationally challenging due to the significant computation cost to evaluate the quality estimation function for each candidate model. A recently proposed method of multiscale, multigranular classification has high computational overhead of function evaluation for various candidate models independently before comparison. In contrast, we propose a context-inclusive approach that controls the computational overhead based on the context, i.e. the value of the quality estimation function for the best candidate model so far. Experimental results using land-use classification at multiple spatial resolutions from satellite imagery show that the proposed approach reduces the computational cost significantly while providing comparable classification accuracy.

## 1 Introduction

We are interested in a probabilistic statistical query to find the preeminent candidate model from a set of candidate models using a quality estimation function. We refer to such a problem as the *best candidate model problem*. Formally, it can be stated as follows: Given a set of candidate models and a function to evaluate the quality of each candidate model, the goal is to find the best candidate model probabilistically. The evaluation of this measure is generally very expensive and thus minimizing the computation time is a key objective. One important example of the *best candidate*

*model problem* is in classification of a spectral image, obtained from a satellite, with domain-specific labels to produce a *thematic map*. *Thematic maps* are widely used in applications including agricultural monitoring, land cover change analysis, environmental assessment. Image classification at multiple spatial resolutions is an important application of spatial data mining. For example, NASA's Earth observation systems obtain a spectral image of land-use, which is then classified at multiple resolutions. The *best candidate model problem* to find the best classification label can be considered as a parameter estimation problem. Since estimating parameters at a spatial region is an important function in spatial data mining, the *best candidate model problem* is a sub-class of spatial data mining.

Figure 1 gives an example of classification based on land usage. Figure 1a is a synthetic remotely-sensed satellite image. Figure 1b is a set of domain-specific labels (also called classes) logically grouped as a hierarchy. Figure 1c gives the satellite image after classification. Each label in the hierarchy represents a candidate model. The goal is to classify each pixel in the satellite image to one of the labels based on a quality measure called *likelihood*. The *likelihood* measure is calculated using a function called Expectation Maximization (EM) which is expensive because of the large number of iterations till convergence.

Calculating the *likelihood* of each candidate model makes the problem computationally expensive. For instance, the work proposed by [3], takes about 7 hours of computation time to classify an image of size 512 x 512 pixels with 12 labels at varying *spatial resolutions*. About 80% of the total computation time is consumed to find the quality measure for each candidate model. Thus, as the image size grows the computation time increases, which makes this problem challenging.

Numerous studies in remote sensing has been done for multi-resolution land-use classification (e.g., [2, 4,

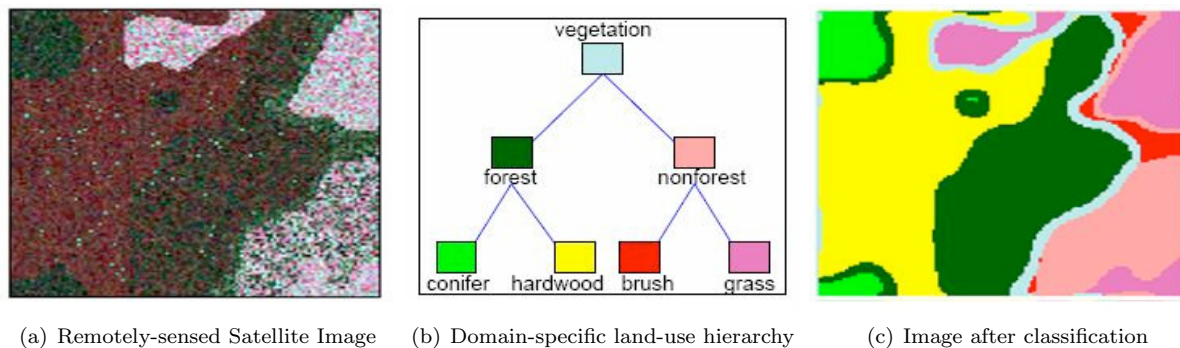


Figure 1: Example: Land-use classification (Courtesy: Boston University [3]. Best viewed in color)

5]). See [9] for a detailed discussion on various methods for multi-resolution classification. A statistical method to classify an image at varying spatial and categorical resolutions was proposed in [3]. This approach is context-exclusive based using a query tree to identify each candidate model independently. The maximum likelihood is used as a set operator among quality measures for each candidate model, thus limiting the fact to analyze candidate models together causing very high computation costs to identify the preeminent candidate model. We propose a context-inclusive approach to speed-up the evaluation of user-defined functions. This approach considers information from all models while evaluation of the user-defined function on a specific model to obtain the preeminent candidate model. This information utilizes each tuple of its respective candidate model to iteratively monitor the quality measure and exit once the appropriate candidate model is found. We also provide an insight to further reduce the computation time by introducing a limiting factor to calculate the best candidate model at a faster rate. Real and synthetic raster data sets from remote sensing are used for our experimental studies in the context-inclusive approach. Our contributions can be summarized as follows:

1. A context-inclusive function evaluation approach that exploits the natural relationship among all candidate models to distinguish the preeminent candidate model.
2. An insight to further reduce computation costs using a limiting factor to iteratively monitor the quality measures of all candidate models and exit when the appropriate candidate model is discovered.
3. Experimental evaluation to compare our approach with a previous approach [3].

The rest of the paper is organized as follows: Section 2

gives a detailed overview of our approach along with the major differences with previous work. Experimental results to compare the previous and the proposed approach are given in Section 3. Finally, Section 4 concludes this paper with a summary and future work.

## 2 Approaches

In this section, we present our proposed approach to address the *best candidate model problem* which may be represented by two tree based methods. The first tree based method uses information from the ancestors and siblings of a query tree to identify the appropriate strategy. Figure 2a gives an example of a query tree where the root node is a set/relational operator, the interior node is a table transformation, and the leaf is a single table. An example of a set operator at the root of this query tree is MAX. A variant to the query tree is a instance-level syntax tree where the main distinction is that the former has the whole table or relation as the leaf and the latter has multiple leaves consisting of individual tuples in a table (Figure 2b). Similar to the query tree, the root can also be represented as a MAX set operation. However, the instance-level syntax tree has many more children than the query tree where each node is represented by a quality measure (i.e., likelihood function) and each respective child as a distinct candidate model.

A comparison to a previous approach using a query tree is given in Section 2.1. Our proposed approach to reduce the computational complexity of the *best candidate model problem* using an instance-level syntax tree as it is applied to a land-use classification problem in the remote sensing domain is presented in Section 2.2. Also an insight to further reduce the computation time of identifying the best candidate model in Section 2.3.

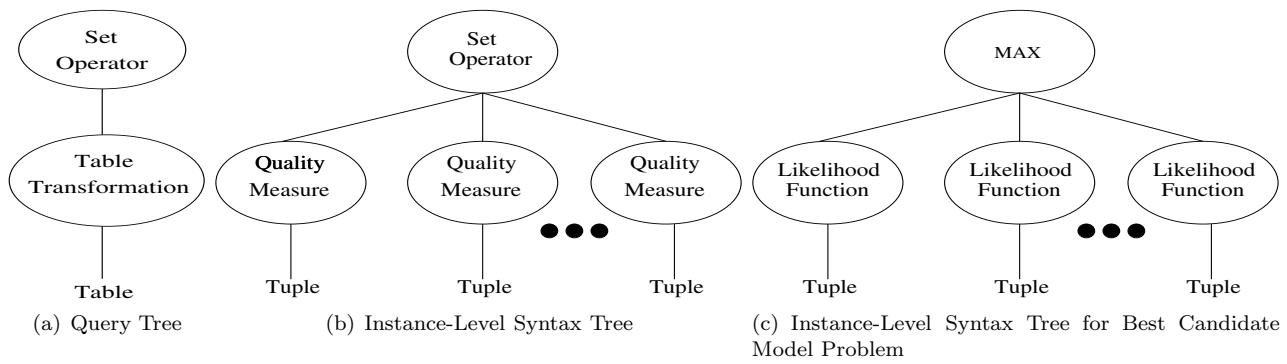


Figure 2: Query Tree and Instance-level Syntax Tree

---

### Algorithm 1 Context-Exclusive Approach

---

```

1: Function CONTEXTEXCLUSIVE(set Cand)
2: for each candidate model  $c \in Cand$  do
3:   repeat
4:     Refine quality measure for each candidate model
        $c \in Cand$ 
5:   until EM exit criteria is met
6: end for
7: Find candidate model  $c$  in  $Cand$  with the Maximum Likelihood
8: return  $c$ 

```

---

## 2.1 Context-Exclusive Approach

A query tree may evaluate an entire table independent from other tables and this relationship is referred as context-exclusive. Algorithm 1 gives the pseudo code of the context-exclusive approach from [3]. The input to Algorithm 1 is the set of candidate models including both *general* and *specific* classes, and the output is the maximum likelihood classification candidate model in  $Cand$ . The main objective in Algorithm 1 is to obtain the maximum likelihood classification for a spatial region. Each candidate model in  $Cand$  is analyzed independently to refine its quality measure (Line 4 in Algorithm 1). For *specific* class candidate models, the quality measures are found at a one time cost while the *general* classes are found iteratively until convergence (Line 5 in Algorithm 1). Finally, the candidate model with maximum likelihood is declared as the best candidate model for a spatial region (Line 7 in Algorithm 1).

## 2.2 Context-Inclusive Approach

Our proposed approach utilizes an instance-level syntax tree where each tuple is evaluated together to obtain the optimal candidate model which we refer this relationship as context-inclusive. Figure 2c gives an example of the instance-level syntax tree as it is applied

---

### Algorithm 2 Context-Inclusive Approach

---

```

1: Function CONTEXTINCLUSIVE(set Cand)
2: repeat
3:   Refine the quality measure for each candidate
       model  $c \in Cand$ 
4:   Prune inferior  $c \in Cand$  based on their quality measures
5: until one candidate model  $c \in Cand$  remains or EM exit
       criteria is met
6: return  $c$ 

```

---

within the remote sensing domain. The root node of the instance-level syntax tree represents the maximum likelihood set operator and the interior nodes are the likelihood function quality measures for each respective child consisting of candidate model tuples. Using remote sensing terminology, each candidate model tuple represents a classification consisting of either a *specific* (i.e., *Conifer* or *Hardwood*) or a *general* (i.e., *Forest*) class (see Figure 1b). The quality measure for a *specific* class is a single value whereas a *general* class consists of several proportions of *specific* classes. For example, a *general* class of type *Forest* may have several proportions of *Conifer* and *Hardwood* trees. A computationally very expensive function is used (i.e., EM) to identify the likelihood value for each candidate model. The main objective is to find the maximum likelihood value or best candidate model to represent a land-use classification.

Algorithm 2 gives the pseudo code for the context-inclusive approach to find the optimal candidate model as it is applied to the domain of remote sensing. The input to Algorithm 2 is the set of candidate models  $Cand$  for the *general* and *specific* classes for a spatial region, and the output is the maximum likelihood candidate models in  $Cand$ . Initially, the likelihood or quality measure is calculated for each tuple in the candidate model set (Line 3 in Algorithm 2). This measure is calculated once for each *specific* class, but calculated iteratively for the *general* class. For every iteration,

Algorithm 2 prunes *specific* classes that have a quality measure less than any other class, since these candidate models will never become the maximum or best candidate model (Line 4 in Algorithm 2). The iterations of calculating the quality measure and pruning steps will continue until a *general* class is the maximum of all candidate models or the likelihood value for *general* classes in EM converges. The former exit case utilizes all candidate model information at each iteration and the latter can be improved to reduce the number of iterations of EM, which is discussed in Section 2.3. Notice that the main distinction between Algorithm 2 and 1 is that the quality measures for each candidate model in Algorithm 1 are found independently (Line 4 in Algorithm 1), whereas the candidate models in Algorithm 2 are pruned based on the current state of each class (Line 4 in Algorithm 2). When the quality measure for a *general* class is the maximum, the number of iterations will be substantially less in our context-inclusive approach (Algorithm 2) than the previous context-exclusive (Algorithm 1) method since Algorithm 2 does not wait until all candidate models converge. Instead, our approach will end when the level of accuracy is appropriate.

To illustrate with an example, consider the hierarchy defined in Figure 1b. Computing the likelihood value for a *specific* class (*conifer*, *hardwood*, *brush*, *grass*) is not expensive while computing the likelihood value for a *general* class (*forest*, *non-forest*, *vegetation*) is. For a spatial region, initially the likelihood of *specific* classes are determined to find the one with the highest value. Assuming the class with the highest value was *conifer*, the candidate models to be compared are: *conifer*, *forest*, *non-forest*, *vegetation*. Assume the maximum likelihood values for *conifer*, *forest*, *non-forest*, *vegetation* were supposed to be 100, 120, 140, and 160 respectively. In a context-exclusive approach, the maximum likelihood are calculated first and then compared to find the best (*vegetation*). In the context-inclusive approach, the likelihood value for a class is calculated until it exceeds the current best likelihood value. To start with, the likelihood of *conifer* is the best and the likelihood of others are yet to be determined. The likelihood for *forest* is evaluated next using the EM. At each iterative step in EM, the current likelihood value of *forest* is compared to the likelihood value of the best so far i.e., *conifer*. As soon as the likelihood of *forest* exceeds the value of *conifer*, EM stops iterating further. For example, when the likelihood value of *forest* reaches 110, EM stops and *conifer* is pruned. This saves additional iterations by avoiding the need to calculate the actual maximum likelihood (120 for *forest*). As the

next step, the current best likelihood (110 for *forest*) will be used as comparison with the next class. The procedure repeats until all classes are considered.

### 2.3 Limiting Factor

We discovered an insight to reduce the computation time for the exit criteria (limiting factor) in the EM used in the previous approach [3]. The objective in this EM is to determine an accurate representation of the proportion sizes of each *specific* component in a *general* class. This EM computes the quality measure at each iteration to be used as input for subsequent iterations. For the quality measure at each iteration, a limiting factor is introduced to determine if the current measure represents the best proportions for the *general* class. In [3], the limiting factor was used at a very fine level of detail. Since the proportions are based on the underlying distributions of *specific* classes, we varied this limiting factor at lower levels, which reduced the number of iterations significantly while maintaining a consistent level of accuracy. Our experiments support our claim that using a lower limiting factor will reduce the computations without sacrificing a high level of accuracy.

## 3 Preliminary Experiments

In our experiments, we evaluate the context-inclusive and context-exclusive approaches in terms of computation and accuracy. All experiments were performed on two different datasets: (1) A synthetic input image of size 128 x 128 pixels, 7 total classes, and 3 *general* classes (Figure 1); and (2) A real dataset of Plymouth County, Massachusetts, consisting of an input image of size 128 x 128 pixels, 12 total classes, and 4 *general* classes. Outputs were obtained for varying spatial scales (resolutions). Spatial scale of 1 corresponds to spatial regions of size 2 x 2 pixels each and increases by a power of 2 for each subsequent spatial resolution i.e., a spatial scale of 6 indicates that the regions are of size 64 x 64 pixels each. All experiments were performed on an UltraSparc III 1.1 GHz processor with 1GB of RAM.

Figure 3 and 5 provides the number of iterations in EM for Dataset 1 and 2 respectively. As the spatial scale increases, the number of iterations also increases because of the size of regions. Note that Dataset 2 has more *general* classes than Dataset 1, and thus is more computationally expensive. The number of iterations are less for context-inclusive approach because the EM algorithm exits as soon as a candidate model with the

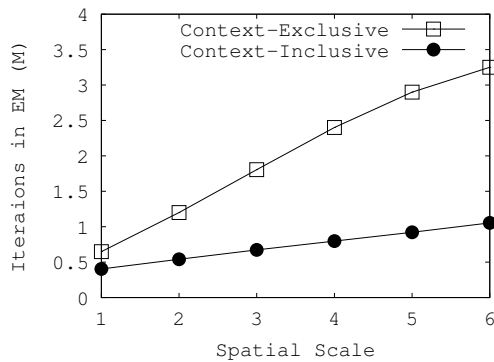


Figure 3: Iterations in EM for Dataset 1

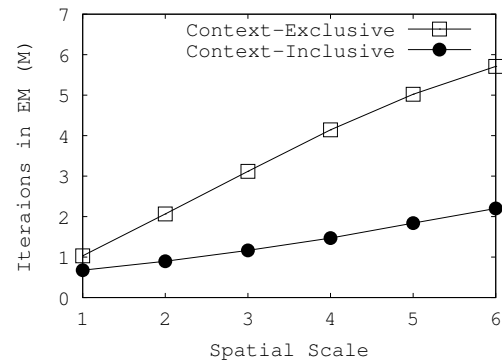


Figure 5: Iterations in EM for Dataset 2

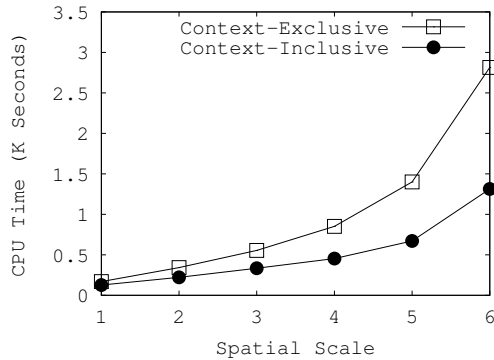


Figure 4: CPU Time for Dataset 1

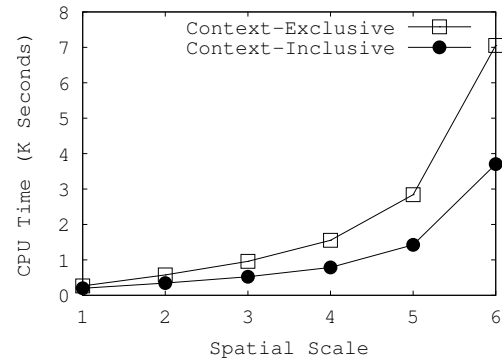


Figure 6: CPU Time for Dataset 2

best likelihood is found. As compared to the context-exclusive approach, at the spatial scale of 6, the number of iterations with the context-inclusive approach reduce by 67.6% for Dataset 1 and by 61.45% for Dataset 2.

Figure 4 and 6 provides the execution time taken for Dataset 1 and 2 respectively. Since the number of iterations taken by context-inclusive approach is less than that for context-exclusive approach, the execution time for context-inclusive approach reduces. At the spatial scale of 6, execution time for context-inclusive approach, as compared to the context-exclusive approach, is reduced by 53.34% and 47.48% for Dataset 1 and Dataset 2 respectively. The speedup is obtained without sacrificing accuracy. Table 1 provides the relative accuracy of the context-inclusive approach as compared to the results from the context-exclusive ap-

Dataset	Spatial Scale					
	1	2	3	4	5	6
Dataset 1	99.69	99.27	99.62	99.58	99.57	99.57
Dataset 2	99.59	98.43	98.74	98.74	98.74	98.74

Table 1: Relative Accuracy for Context-Inclusive

proach. For both the datasets, relative accuracy is always more than 98%.

EM computes the mixture proportions (likelihood) iteratively. The limiting factor for EM can be varied based on the desired accuracy for the final mixture proportion. Figure 7 and 8 compare the number of iterations and execution time, respectively, for different limiting factors. The total number of iterations in EM and execution time decreases linearly with the decrease in limiting factor. As shown in Figure 8 for Dataset 2, the execution time decreases from 133 minutes to 55 minutes for a change in limiting factor from 0.00001 to 0.01. Table 2 provides the accuracy results for a limiting factor of 0.01 as compared to the limiting factor of 0.00001 at different scales. These results show that our cost-effective approach does not sacrifice accuracy.

Dataset	Spatial Scale					
	1	2	3	4	5	6
Dataset 1	99.99	99.85	99.87	99.88	99.87	99.87
Dataset 2	99.78	99.76	99.79	99.82	99.82	99.82

Table 2: Relative Accuracy for Limiting Factor of 0.01

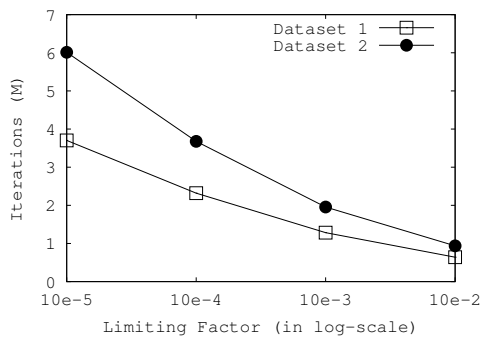


Figure 7: Iterations with change in Limiting Factor

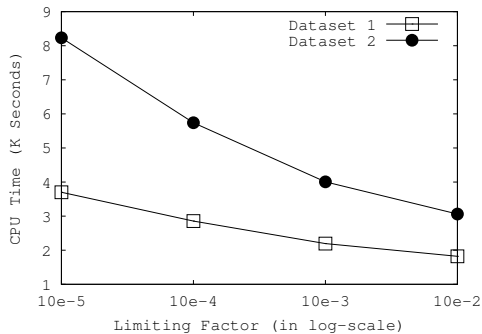


Figure 8: CPU Time with change in Limiting Factor

## 4 Conclusion and Future Work

In this paper, we presented a context-inclusive approach in an instance-level syntax tree where each tuple is evaluated together to obtain the optimal candidate model. This approach is different from previous work which uses a context-exclusive method in a query tree [3]. Although a context-inclusive approach has been applied to query trees before [6], but was never applied to an instance-level syntax trees. A limiting factor is also introduced that reduces the number of iterations toward convergence in EM. Our experiments show that our approach has reduced the computational complexity over previous approach while maintaining comparable classification accuracy.

Other types of context may be explored. For example, spatial context i.e., the correlation of a variable with space [7], may be used. Classification of spatial data based on an extended regression model called the Spatial Auto-regression model (SAR) is provided in [8].

As discussed, most of the execution time is spent in calculating the quality measure using EM. EM is used to find the best candidate model; more specifically, EM is used to find the best Gaussian mixture model in the

case of land-use classification. The execution time of finding the best Gaussian mixture model can be reduced by using *KD-Trees* [1]. A *KD-Tree* is a space partitioning data structure used for organizing points in a *K*-dimensional space. In our case, a *KD-Tree* can be used to organize the mean and the variance of the mixture at each node.

Our proposed approach uses a bottom-up strategy where information from finer spatial scales is used for coarser spatial scales. Enhancements can be made to consider a top-down approach where a pruning procedure may be used to reduce computation.

## Acknowledgements

This work was a result of research supported in part by National Science Foundation grants BCS 0079077 and 0318209, Office of Naval Research award N00014-99-1-0219, IGERT and SEI.

## References

- [1] A. Ihler. *Inference in Sensor Networks: Graphical Models and Particle Methods*, chapter Ph.D. Thesis, pages 26–32. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2005.
- [2] J. R. Irons, B. L. Markham, R. Nelson, D. L. Toll, and D. Williams. The effects of spatial resolution on the classification of Thematic Mapper data. In *International Journal of Remote Sensing*, volume 6, pages 1385–1403, 1985.
- [3] E. D. Kolaczyk, J. Ju, and S. Gopal. Multiscale, Multi-granular Statistical Image Segmentation. In *Journal of the American Statistical Association*, volume 100, pages 1358–1369, 2005.
- [4] B. L. Markham and J. R. G. Townshend. Land cover classification accuracy as a function of sensor spatial resolution. In *Proceedings of the Fifteenth International Symposium on Remote Sensing of Environment*, pages 1075–1090, 1981.
- [5] V. S. Raptis, R. A. Vaughan, and G. G. Wright. The effects of scaling on land cover classification from satellite data. In *Computers & Geosciences*, volume 29, pages 705–714, 2003.
- [6] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access Path Selection in a Relational Database Management System. In *Proceedings of 1979 ACM-SIGMOD International Conference on Management of Data*, June 1979.
- [7] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.
- [8] S. Shekhar, P. Schrater, R. Vatsavai, W. Wu, and S. Chawla. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. In *IEEE Transaction on Multimedia*, 2002.
- [9] A. Willsky. Multiresolution Markov models for signal and image processing. In *Proceedings of the IEEE 90*, volume 8, pages 1396–1458, 2002.