Appendix: ShapeToVec: Encoding Polygonal Shapes with Extreme Area Variability for Effective Approximate Jaccard Similarity Queries

1 Feature Vector Max Recall Rate

The recall rate metric estimates the number of true positives expected while searching over the index. Since our approach preprocesses the input polygonal data into feature vectors, errors can arise during the polygon encoding and indexing phases. We designed the following experiment to estimate an upper bound on recall after the encoding phase independent of indexing methods. We used bit-encoded feature vectors in this experiment.

The brute force all-to-all comparison approach is the most suitable method for this evaluation. However, we did not use it since it is impractical on large datasets. First, we retrieved the 500 most similar polygons for each test polygon from the ground truth data. Subsequently, we computed the Jaccard similarity between each test polygon and the retrieved polygons using their corresponding feature vectors. Next, we arranged them in descending order based on their Jaccard similarity, which was calculated using corresponding feature vectors, to identify the 50 most similar polygons. Finally, we compared the top 50 similar polygons from the ground truth dataset with the subset of 50 polygons computed based on vectors. These recall rates are recorded in Table 1 and indicate that the feature vectors produced using the quad tree-based approach are much more accurate than those encoded using the uniform grid-based approach.

Table 1: Max recall rates independent of the index method.

Encoding method	**Grid size**	**Recall for K=50**
Uniform grid-based	6,084	11%
	18,225	11%
Quad tree-based	6,004	77%
	18,220	79%

2 Number of Nonzero elements in a feature vector

The number of nonzero (NNZ) elements in a feature vector is significant when comparing two vectors as these elements contain shape information about the polygons. Low NNZ elements in the vectors may lead to insufficient information during the comparison operation, potentially compromising high recall.

Figure 1 depicts a plot of NNZ elements in the feature vectors over the Parks dataset. We maintain a grid size of approximately 35k in each polygon encoding technique. The feature vectors encoded using the uniform grid approach tend to have lower NNZ elements, whereas those generated using the quad tree-based approach tend to have higher NNZ elements. This illustrates that the quad tree-based method can incorporate more information in the feature vectors to represent a polygon, suggesting its appropriateness for handling real-world datasets.

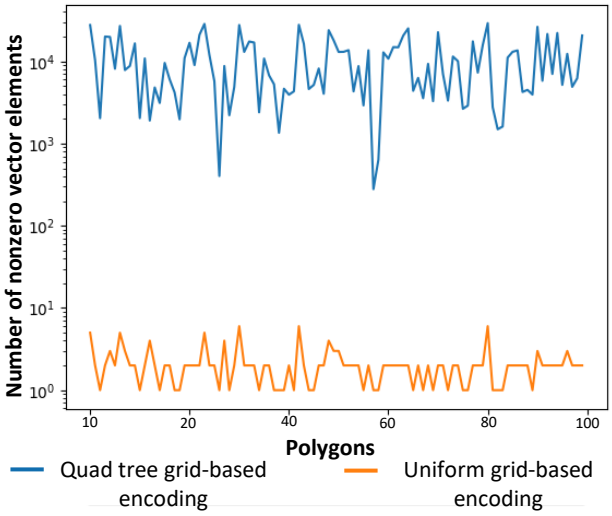

Figure 1: The number of nonzero elements in the feature vectors over 100 polygons from the Parks dataset. Grid resolution is 35K.

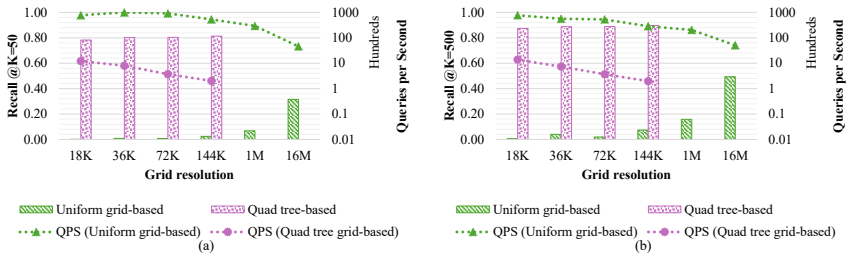

Figure 2: Recall rate and query throughput comparison for different grid sizes over uniform grid-based vectors and quad tree-based bit vectors using 64 threads. A subset of 50k records from the Parks dataset was used for indexing (80%) and testing (20%). (a) Recall at K=50. (b) Recall at K=500.

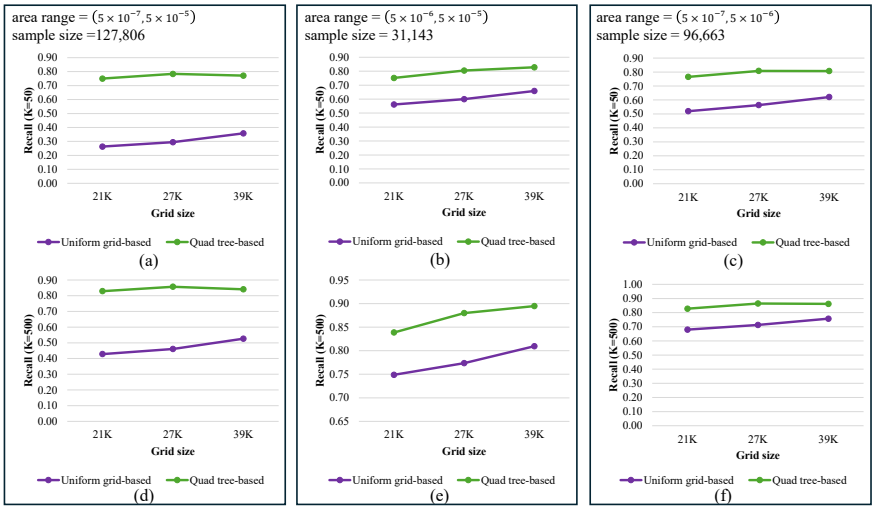

Figure 3: Recall rate comparison over different area ranges.

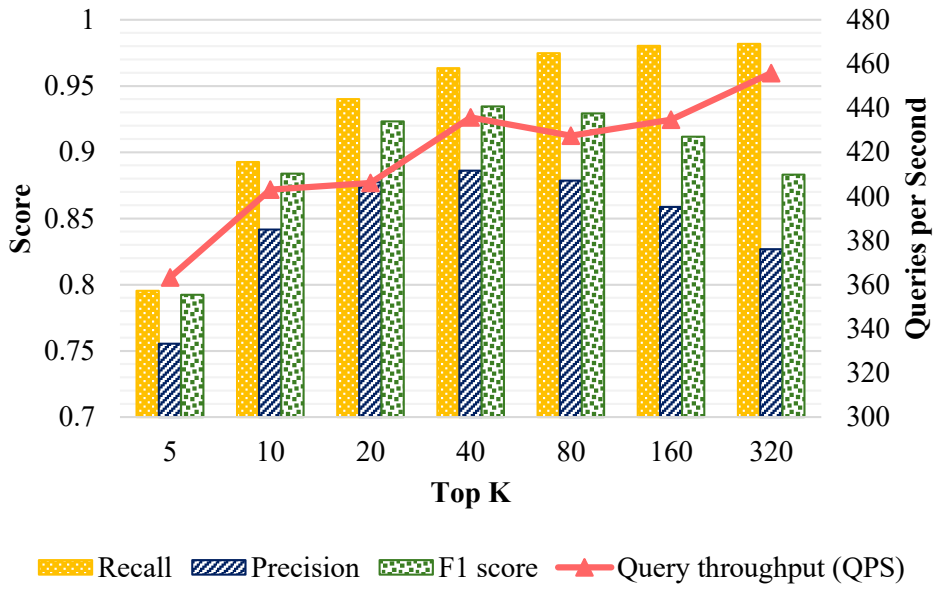

Figure 4: Performance metrics versus the number of nearest neighbors (K) on the Parks dataset. The plot illustrates the trade-offs between accuracy (Recall, Precision, F1-Score) and query throughput.

Table 2: Evaluation of Quad tree-based encoding (K=500).

		Parks			Water bodies			Sports		
		3k	**6k**	**12k**	**3k**	**6k**	**12k**	**3k**	**6k**	**12k**
Bit encoding	**Recall**	80%	81%	82%	76%	74%	78%	63%	66%	79%
	Precision	63%	65%	65%	60%	58%	63%	62%	65%	76%
	F1 score	68%	70%	70%	67%	65%	70%	63%	66%	77%
	Query throughput (queries/s)	1,642	1,692	1,572	1,876	1,642	1,192	3,937	2,865	1,199
Floating-point encoding	**Recall**	97%	97%	97%	97%	98%	98%	95%	96%	96%
	Precision	79%	79%	79%	78%	78%	79%	91%	92%	92%
	F1 score	85%	85%	85%	86%	87%	87%	93%	93%	93%
	Queries per second	1,115	696	357	246	512	286	974	657	410