Comparisons of Gist Models in Rapid Scene Categorization Tasks

Manu Viswanathan, Christian Siagian, and Laurent Itti

Computer Science Department, University of Southern California, Los Angeles, California USA

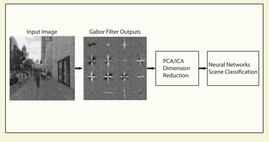
Introduction

The capacity of humans to perform a number of complex visual tasks such as scene categorization in as little as 100ms has been attributed to their ability to rapidly extract the "gist" of a scene [1 -3]. Following a brief presentation of a photograph, an observer is able to summarize the quintessential characteristics of an image, a process previously expected to require much analysis [4-5].

Various gist models, which utilize different types of low level features have recently been presented.

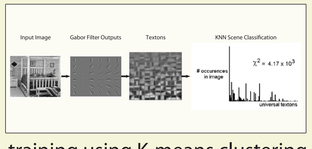
In this work we report a comparison study with a scene categorization task in 3 large scale and visually challenging outdoor environments, with multiple lighting conditions.

Models



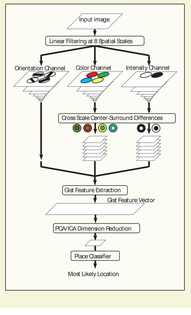
Oliva & Torralba 2002

The model computes magnitude spectrum of Windowed Fourier Transform, here we use 4 by 4 equally-sized, non-overlapping regions. We then reduce the feature dimension with Principal Component Analysis (PCA) before classifying the scenes.



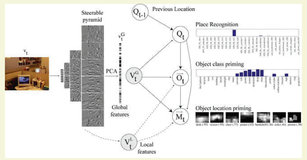
Renninger & Malik 2004

The model applies Gabor filters to an input image and then extracts 100 universal textons, selected from training using K-means clustering. The gist vector is a histogram of universal textons. Classification is using KNN based on Chi-square distance.



Siagian & Itti 2007

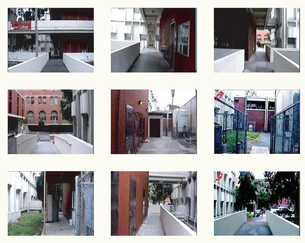
The model uses biologically plausible center surround features from orientation, color, and intensity channels. For each of the 34 subchannels, it then computes average values from a predefined 4 by 4 grid (16 values) for a total of 544 raw gist values. These values are then reduced using PCA/ICA to 80 features. These features are used to classify scenes using neural networks.



Torralba,et. al. 2003

The model uses wavelet image decomposition (we use equivalent Gabor filters) tuned to 6 orientations and 4 scales. The gist vector is computed by averaging each filter output over a 4x4 grid. These 384-dimensional vectors are then reduced to 80 dimensions using PCA. Classification is done by finding the minimum Euclidean distance between the gist vectors of the input images and those of the training set.

Testing Design



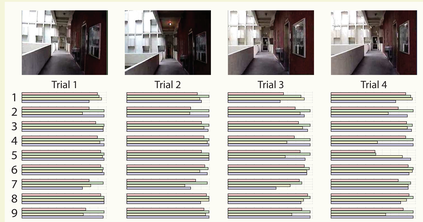
- 3 visually contrasting outdoor sites under various lighting conditions.
- 9 - 11 training runs for each site.
- 4 testing runs taken on separate days.
- Each site is divided into 9 segments to be classified.
- Site 1: ACB
  Building complex.
- Site 2: AnF
  Park full of trees.
- Site 3: FDF
  Open space.

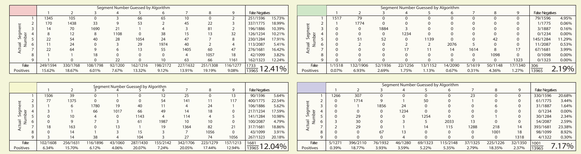
Results

ACB

- Oliva-Torralba 2001
- Renninger-Malik 2004
- Siagian-Itti 2007
- Torralba et.al. 2003

Bars indicate percentage of classification correctness between 0 - 100%.

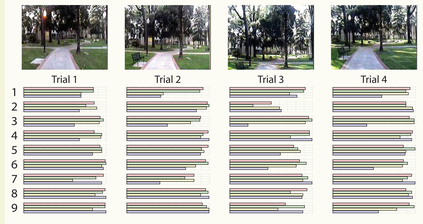


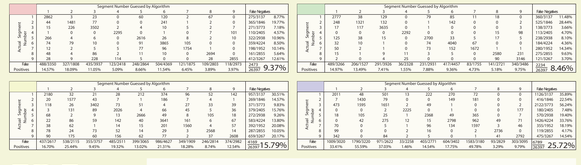


AnF

- Oliva-Torralba 2001
- Renninger-Malik 2004
- Siagian-Itti 2007
- Torralba et.al. 2003

Bars indicate percentage of classification correctness between 0 - 100%.

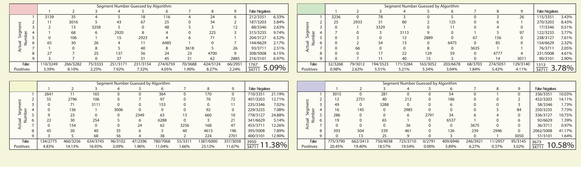




FDF

- Oliva-Torralba 2001
- Renninger-Malik 2004
- Siagian-Itti 2007
- Torralba et.al. 2003

Bars indicate percentage of classification correctness between 0 - 100%.



Combined

	ACB		AnF		FDF	
	False Positives	False Negatives	False Positives	False Negatives	False Positives	False Negatives
1	11.02%	16.54%	8.31%	8.58%	13.81%	3.49%
2	16.53%	10.08%	14.49%	15.28%	7.33%	4.81%
3	5.04%	12.14%	12.58%	10.31%	2.75%	1.79%
4	11.10%	5.92%	6.37%	2.87%	9.56%	6.37%
5	10.06%	11.53%	13.16%	4.32%	5.13%	5.47%
6	2.52%	5.46%	10.30%	6.58%	4.08%	5.61%
7	5.63%	19.27%	7.23%	13.93%	2.36%	4.10%
8	16.53%	7.19%	3.16%	11.94%	9.84%	8.85%
9	4.02%	11.49%	5.86%	12.98%	5.13%	9.96%
Overall	8.98%	11.19%	9.25%	9.29%	6.65%	5.72%
Error Rate	7.99%					

	ACB		AnF		FDF	
	False Positives	False Negatives	False Positives	False Negatives	False Positives	False Negatives
1	0.00%	39.47%	29.94%	19.73%	3.11%	13.58%
2	30.77%	1.13%	14.68%	31.69%	9.10%	10.15%
3	0.55%	3.71%	18.70%	5.17%	9.40%	5.80%
4	4.43%	0.16%	2.78%	5.41%	7.72%	9.43%
5	11.93%	22.98%	16.27%	17.67%	13.54%	5.85%
6	0.42%	9.30%	12.06%	6.23%	7.43%	9.05%
7	3.12%	29.80%	6.46%	18.34%	4.72%	1.08%
8	13.14%	0.73%	20.07%	16.29%	8.09%	14.44%
9	4.70%	1.81%	14.52%	5.94%	6.89%	14.16%
Overall	9.32%	12.49%	16.10%	12.61%	7.78%	9.40%
Error Rate	11.10%					

	ACB		AnF		FDF	
	False Positives	False Negatives	False Positives	False Negatives	False Positives	False Negatives
1	17.62%	14.47%	18.11%	18.55%	10.02%	38.08%
2	16.20%	19.27%	29.46%	17.50%	14.33%	15.08%
3	13.54%	6.89%	13.22%	3.45%	23.01%	10.70%
4	5.04%	6.81%	9.70%	9.85%	1.81%	5.84%
5	10.68%	15.97%	17.93%	9.19%	3.43%	3.81%
6	7.94%	10.01%	20.75%	16.29%	12.20%	14.89%
7	21.42%	32.78%	17.74%	29.20%	8.34%	8.52%
8	8.64%	20.11%	22.27%	10.75%	6.21%	9.48%
9	10.18%	22.68%	8.25%	12.52%	15.87%	5.80%
Overall	12.71%	16.28%	17.29%	13.32%	10.84%	12.63%
Error Rate	13.55%					

	ACB		AnF		FDF	
	False Positives	False Negatives	False Positives	False Negatives	False Positives	False Negatives
1	28.18%	22.06%	41.00%	36.34%	22.57%	23.63%
2	28.34%	4.56%	56.06%	22.97%	19.51%	28.75%
3	3.88%	1.64%	36.87%	58.15%	14.60%	14.85%
4	53.63%	0.16%	0.54%	7.69%	17.81%	9.71%
5	6.30%	2.65%	36.38%	21.78%	0.00%	15.96%
6	5.62%	4.26%	33.87%	39.49%	3.48%	13.71%
7	24.77%	38.01%	53.35%	18.39%	6.62%	9.89%
8	20.17%	12.10%	3.11%	4.94%	4.30%	43.05%
9	10.78%	4.91%	17.93%	27.85%	2.23%	9.48%
Overall	22.30%	10.21%	32.89%	29.02%	9.85%	19.44%
Error Rate	21.09%					

Discussions and Conclusions

- In terms of accuracy, all the models perform roughly the same (> 80%) on the one-shot scene classification task.
- All the models train about the same with maximum training durations in the range of two hours.
- Gist computation speed is less than 1ms/frame for all models, except for the Renninger-Malik model, which takes about 7 sec/frame because it needs to compute textons for each pixel.
- The specific features (wavelets, FFT, Gabor) or training algorithms (neural network, KNN) are not critical in implementing a successful gist classifier.
- The coarseness and low computation costs of extracting gist features facilitates effective usage in applications such as content-based image retrieval and robot vision localization.

References

[1] M. C. Potter, "Meaning in visual search," Science, vol. 187, no. 4180, pp. 965–966, 1975.
[2] I. Biederman, "Do background depth gradients facilitate object identification?" Perception, vol. 10, pp. 573 – 578, 1982.
[3] B. Tversky and K. Hemenway, "Categories of the environmental scenes," Cognitive Psychology, vol. 15, pp. 121 – 149, 1983.
[4] A. Oliva and P. Schyns, "Coarse blobs or ne edges? evidence that information diagnosticity changes the perception of complex visual stimuli," Cognitive Psychology, vol. 34, pp. 72 – 107, 1997.
[5] T. Sanocki and W. Epstein, "Priming spatial layout of scenes," Psychol. Sci., vol. 8, pp. 374 – 378, 1997.
[6] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.
[7] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in IEEE Intl. Conference on Computer Vision (ICCV), Nice, France, October 2003, pp. 1023 – 1029.
[8] C. Siagian and L. Itti, "Rapid biologically-inspired scene classi cation using features shared with visual attention," IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 29, no. 2, pp. 300–312, Feb 2007.
[9] L. Renniger and J. Malik, "When is scene identication just texture recognition?" Vision Research, vol. 44, pp. 2301–2311, 2004.

Acknowledgement

This research is funded by NSF, HFSP, NGA, and DARPA.