Research Methods in computer science

Fall 2013

Lecture 4

Omprakash Gnawali

September 5, 2013

Agenda

Research Conference Updates

Experiment Design

Deployment Experiments

Feedback from HW2

HW3

Experiments

Hypothesis

Scenarios

Measurements

Conclusions

Types of Experiments

Model / Analysis

Simulations

Testbed (Real word ^{lite})

“Real world”

Which one to use when?

Scenarios

Types of inputs

Types of configurations

Try to keep the number of scenarios small while covering normal and meaningful corner cases.

A new image recognition system…

What inputs should we use?

Random library from Flickr

Algorithm specific

Standard datasets

Hypothesis

Experiments: hypothesis testing

Bias in hypothesis

Examples

Metrics

Systems	Classification	HCI
Throughput	Precision	Accuracy
Latency	Recall	Latency
Overhead	Running time	“Discomfort”
Reliability		

Conclusions from Experiments

Strict interpretation

Extrapolate

Touché: Enhancing Touch Interaction on Humans, Screens, Liquids, and Everyday Objects

[Sato '12]

What hypothesis, scenarios, and metrics should we expect to see in this paper?

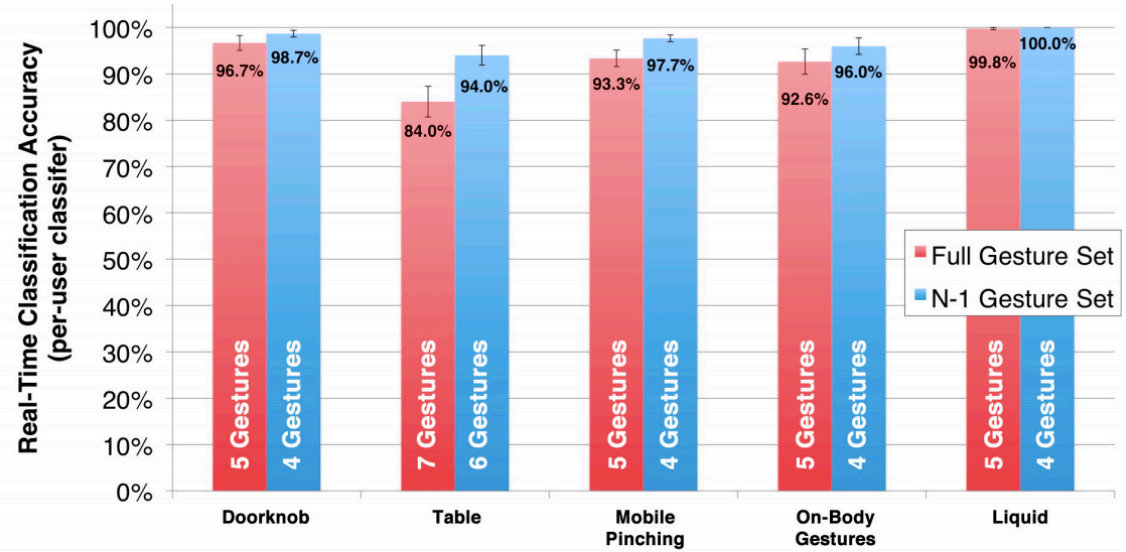


Figure 11. Real-time, per-user classification accuracy for five example applications.

[Sato '12]

What are the (missing) scenarios and (missing) metrics? What can we conclude?

Fast, Accurate Detection of 100,000 Object Classes on a Single Machine

[Dean ’13]

What hypothesis, scenarios, and metrics should we expect to see in this paper?

	arp	bike	bird	boat	bttl	bus	car	cat	chr	cow	tbl	dog	hrs	mbke	prsn	plnt	shp	sofa	trn	tv	Mean
Ours	0.19	0.48	**0.03**	0.10	0.16	**0.41**	0.44	0.09	0.15	**0.19**	0.23	**0.10**	**0.52**	0.34	0.20	**0.10**	0.16	**0.28**	**0.34**	0.34	0.24
[6] (base)	**0.29**	**0.55**	0.01	**0.13**	**0.26**	0.39	**0.46**	**0.16**	**0.16**	0.17	**0.25**	0.05	0.44	**0.38**	**0.35**	0.09	**0.17**	0.22	0.34	**0.39**	**0.26**

Table 1. Comparison of the hashing-based and baseline algorithms on the PASAL VOC 2007 dataset

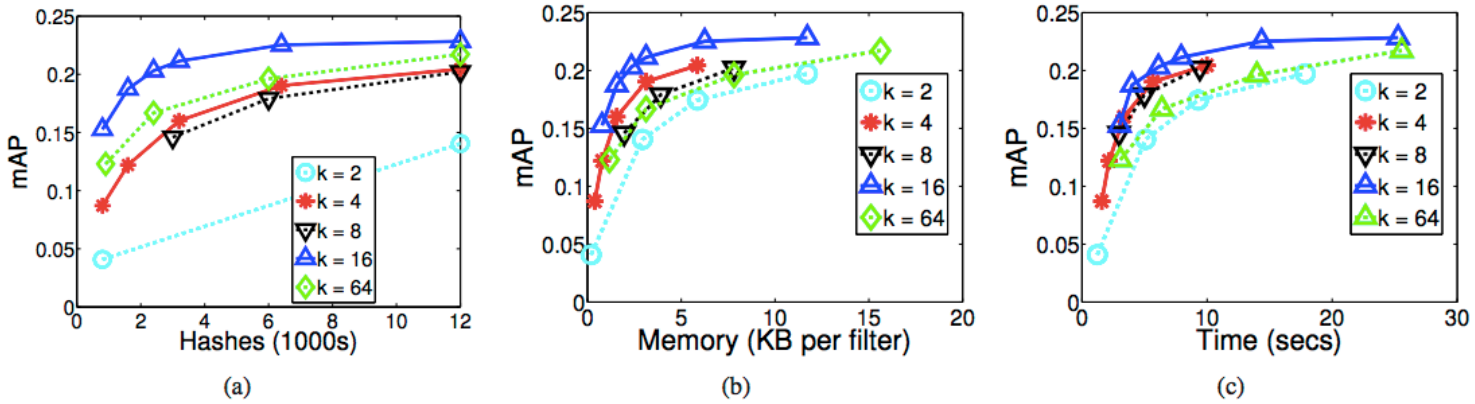


Figure 3. Effect of hashing parameters on the accuracy, speed and memory required by the system.

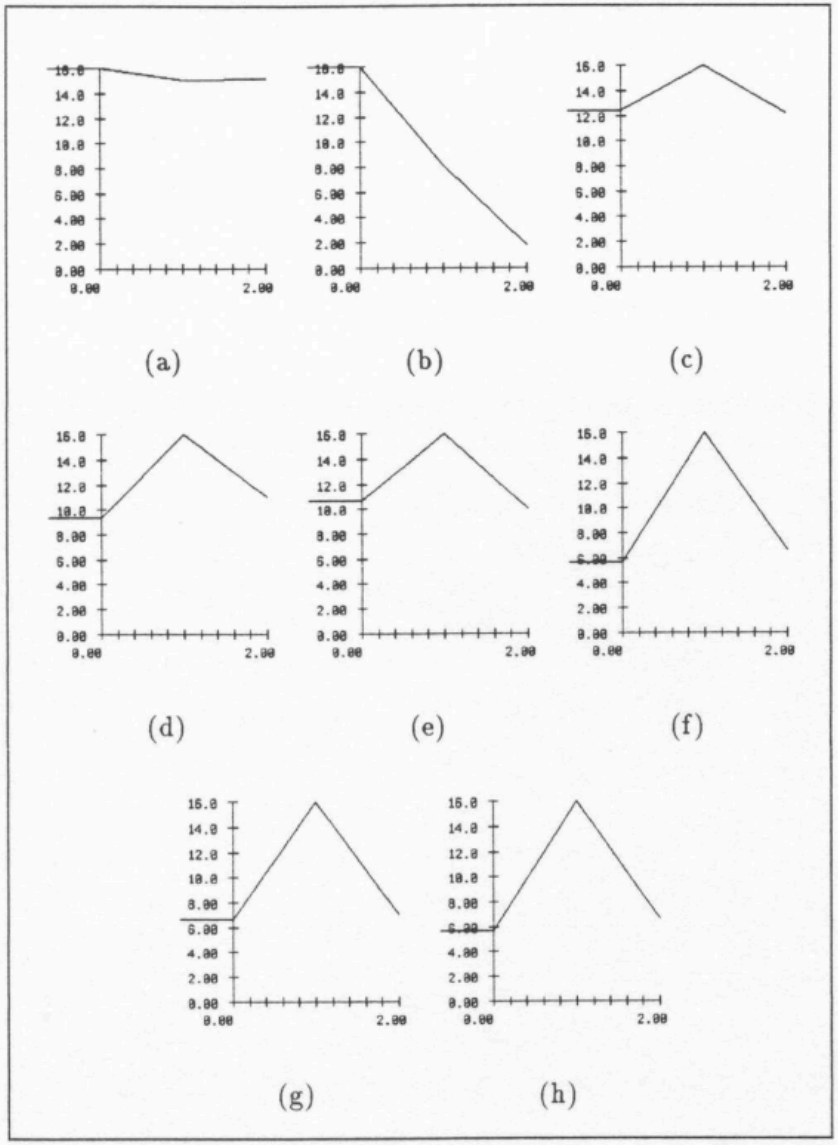
[Dean '13]

What are the (missing) scenarios and (missing) metrics? What can we conclude?

Eigenfaces for Recognition

[Turk ’91]

“We have developed a near-real-time computer system that can locate and track a subject’s head, and then recognize the person by comparing the characteristics of the face to those of known individuals.”



Scenarios and metrics from [Turk '91]

Figure 9. Results of experiments measuring recognition performance using eigenfaces. Each graph shows averaged performance as the lighting conditions, head size, and head orientation vary—the *y*-axis depicts number of correct classifications (out of 16). The peak (16/16 correct) in each graph results from recognizing the particular training set perfectly. The other two graph points reveal the decline in performance as the following parameters are varied: **(a)** lighting, **(b)** head size (scale), **(c)** orientation, **(d)** orientation and lighting, **(e)** orientation and size (#1), **(f)** orientation and size (#2), **(g)** size and lighting, **(h)** size and lighting (#2).

The Anatomy of a Large-Scale Hypertextual Web Search Engine

[Brin and Page ’98]

What hypothesis, scenarios, and metrics should we expect to see in this paper?

5 Results and Performance

The most important measure of a search engine is the quality of its search results. While a complete user evaluation is beyond the scope of this paper, our own experience with Google has shown it to produce better results than the major commercial search engines for most searches. As an example which illustrates the use of PageRank, anchor text, and proximity, Figure 4 shows Google's results for a search on "bill clinton". These results demonstrates some of Google's features. The results are clustered by server. This helps considerably when sifting through result sets. A number of results are from the whitehouse.gov domain which is what one may reasonably expect from such a search. Currently, most major commercial search engines do not return any results from whitehouse.gov, much less the right ones. Notice that there is no title for the first result. This is because it was not crawled. Instead, Google relied on anchor text to determine this was a good answer to the query. Similarly, the fifth result is an email address which, of course, is not crawlable. It is also a result of anchor text.

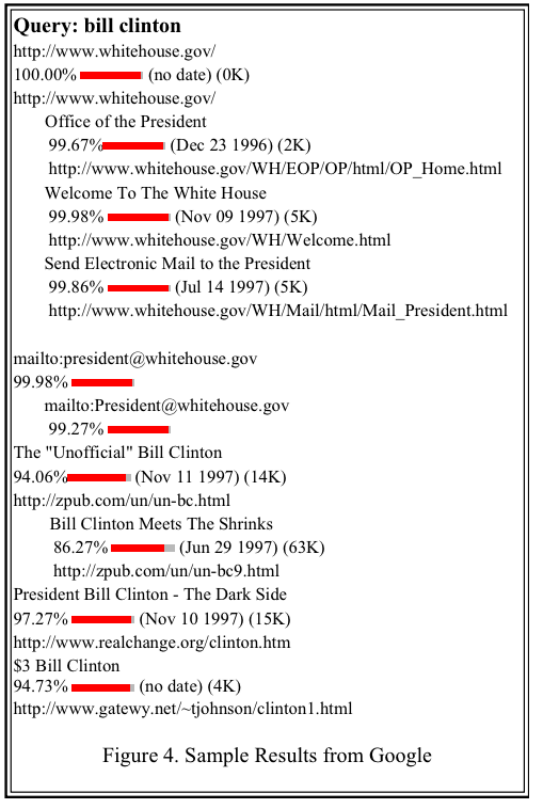


Figure 4. Sample Results from Google

All of the results are reasonably high quality pages and, at last check, none were broken links. This is largely because they all have high PageRank. The PageRanks are the percentages in red along with bar graphs. Finally, there are no results about a Bill other than Clinton or about a Clinton other than Bill. This is because we place heavy importance on the proximity of word occurrences. Of course a true test of the quality of a search engine would involve an extensive user study or results analysis which we do not have room for here. Instead, we invite the reader to try Google for themselves at http://google.stanford.edu.

[Brin and Page '98]

Storage Statistics	
Total Size of Fetched Pages	147.8 GB
Compressed Repository	53.5 GB
Short Inverted Index	4.1 GB
Full Inverted Index	37.2 GB
Lexicon	293 MB
Temporary Anchor Data (not in total)	6.6 GB
Document Index Incl. Variable Width Data	9.7 GB
Links Database	3.9 GB
Total Without Repository	**55.2 GB**
Total With Repository	**108.7 GB**

Web Page Statistics	
Number of Web Pages Fetched	24 million
Number of Urls Seen	76.5 million
Number of Email Addresses	1.7 million
Number of 404's	1.6 million

Table 1. Statistics

[Brin and Page '98]

Why did the authors decide to report these measurements?

HW3

Write one-paragraph summary of the proposed project.

Write a paragraph addressing each “Research Formulation” questions. The complete writeup should not be longer than two pages.