

INTERACTIVE SEARCH FOR IMAGE CATEGORIES BY MENTAL MATCHING

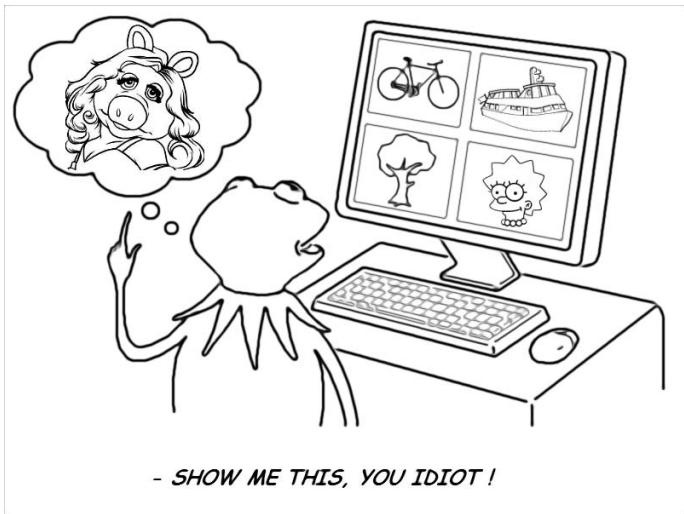
Donald Geman
Johns Hopkins University

Frontiers in Computer Vision
M.I.T., August 2011

A Statistical Framework for Image Category Search from a Mental Picture

Marin Ferecatu and Donald Geman, *Senior Member, IEEE*

SCENARIO



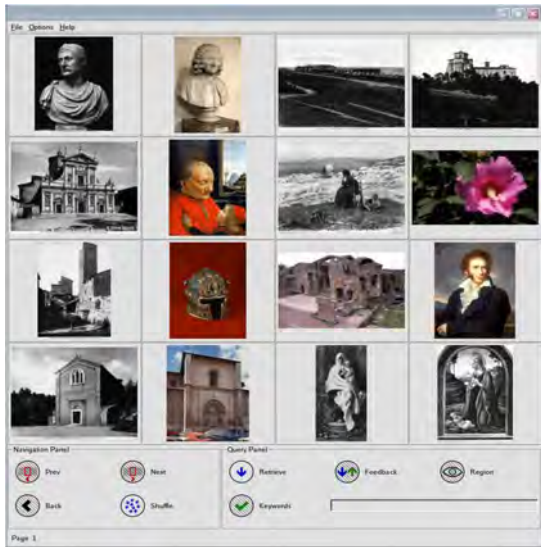
OUTLINE

- ▶ Standard Image Retrieval
- ▶ Mental Matching
- ▶ Experiments
- ▶ Statistical Framework (*maybe*)
- ▶ Modeling Human Behavior (*maybe*)

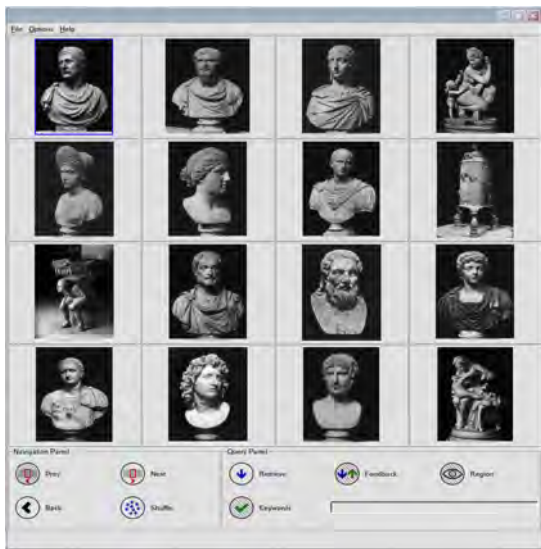
CONVENTIONAL QUERY-BY-EXAMPLE (QBE)

- ▶ Start from a query image **in a database**. Find other images which are “close” or “closest”
 - ▶ in overall color, texture or shape, or
 - ▶ *in a semantic sense*, or . . .
- ▶ Matching is performed **by the system**.
- ▶ Good results in limited domains, e.g., comparing paintings, plants and landscapes.

EXAMPLE: IKONA SEARCH ENGINE (INRIA)



EXAMPLE (CONT)



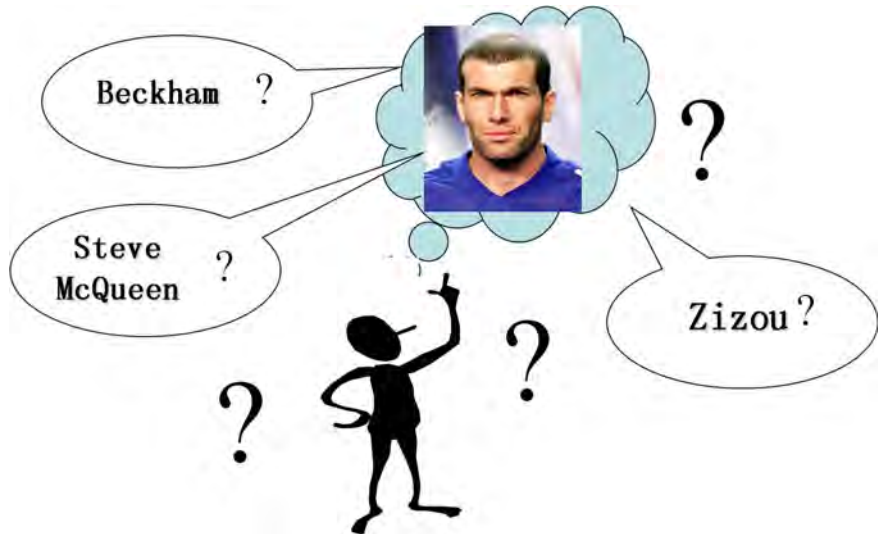
“PAGE ZERO” PROBLEM

- ▶ QBE requires a starting point - a query image.
- ▶ *Dilemma*: Without a starting point, random sampling a large database is too slow in practice.

EXTERNAL IMAGES

- ▶ **Mental Picture:** The user has a picture “in mind”, e.g., a face or painting or house.
- ▶ **Viewed Image:** The user is looking at a picture, e.g., in a magazine or on the web.
- ▶ **Physical Object:** The user is holding an object.

WHO IS THAT PERSON ?



MENTAL CATEGORY SEARCH

- ▶ Assume this “external query” is represented in our database, either by
 - ▶ a version of the same image (e.g., same person), or
 - ▶ variations on a theme, i.e., a category of images (e.g., similar houses).
- ▶ **Objective:** Find an efficient way to display this version or representatives of this category.
- ▶ **Applications:** Image retrieval (“page zero”); web browsing; security; art management; plant science; e-commerces; blah blah blah.

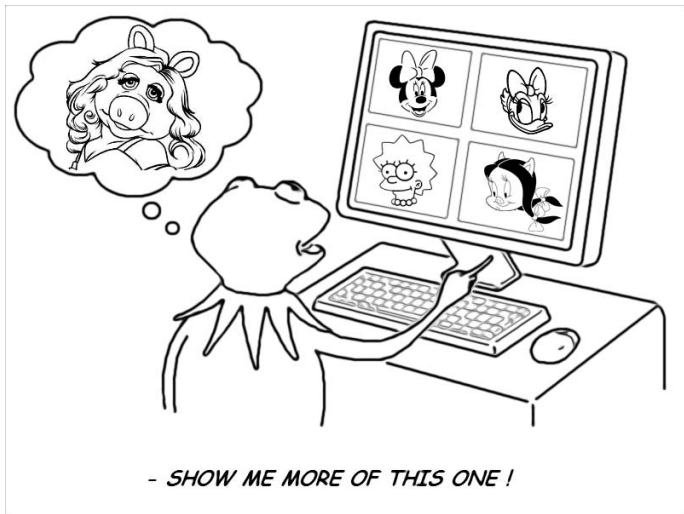
INTERACTIVE SEARCH

- ▶ The object of the search is a class S (variations on an image or theme).
- ▶ Single target search is the special $|S| = 1$.
- ▶ Assume the user always recognizes an instance of his target.
- ▶ At each iteration, some images are displayed, typically two to sixteen.
- ▶ The user responds by either
 - ▶ signaling a target if present; or
 - ▶ *choosing the one deemed “closest”*.

INTERACTIVE SEARCH (CONT)

- ▶ Based on this *feedback*, the system chooses another set of images to display.
- ▶ *Goal: Minimize the number of iterations until an exemplar of the target is displayed.*
- ▶ Then display other examples (“page zero”) for specialization and refinement.

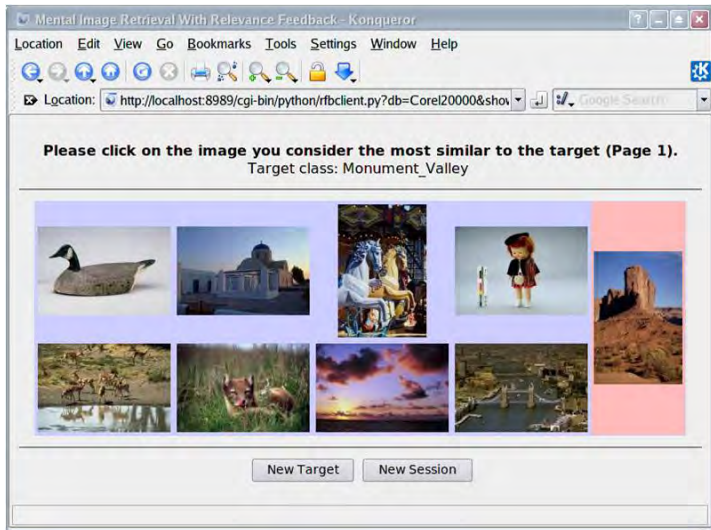
BACK TO KERMIT



COMPLICATIONS

- ▶ Mental matching involves human memory, perception and opinions.
- ▶ People are semantically oriented. However, images are indexed by low-level features (“semantic gap”).
- ▶ Interest in large databases, order 10,000 to 1,000,000.

THE USER INTERFACE



MEASURES OF PERFORMANCE

- ▶ T : number of iterations until S is displayed.
- ▶ $P(T < t)$: The probability distribution over some population of users.
- ▶ $E(T)$: The mean of this population.
- ▶ For a random search,

$$E(T) \cong N/(L(|S| + 1)),$$

where N is the size of the database and L is the number displayed per iteration.

- ▶ *Coherence*: The probability that the user selects the i 'th closest image to S .

EXPERIMENTAL DATABASES

- ▶ Corel: $N=60,000$ images
- ▶ Alinari: $N=20,000$ images

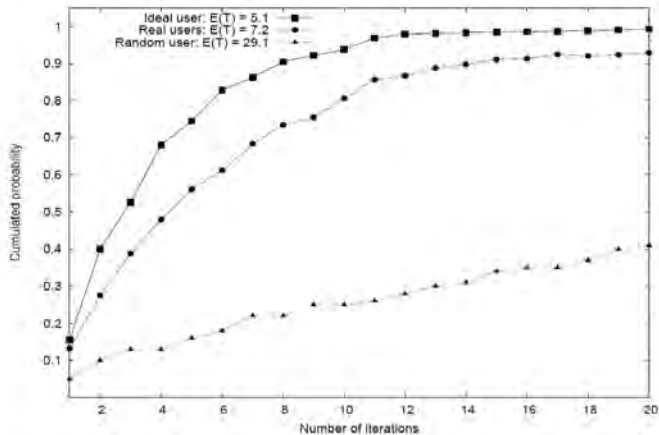
Ground truth: 10 semantic classes of ≈ 100 hand-chosen images

ALINARI DATABASE



“Madonna and Child” (top rows) and “Horse and Rider” (bottom rows).

PERFORMANCE: ALINARI



Search time distribution

CONCLUSIONS

- ▶ Rich possibilities for mathematical modeling in building efficient man-machine interfaces.
- ▶ Mixes geometry, probability, optimization and information theory.
- ▶ Solving the “vision problem” is probably not around the corner.
- ▶ Hence extending to databases of order 1,000,000 remains a challenge.

DATABASE AND IMAGE METRIC

- ▶ I ... an image
- ▶ $\Omega = \{1, 2, \dots, N\}$... a database of images
- ▶ We do **not** assume Ω is “structured” (partitioned into categories)
- ▶ $\{f(I_1), f(I_2), \dots, f(I_N)\}$... “features” in R^M .
- ▶ $d_f : R^M \times R^M \rightarrow [0, 1]$... a metric on features.
- ▶ $S \subset \Omega$... the category (semantic class) in the mind of the user, **a random set**.
- ▶ For each $k = 1, \dots, N$, define a binary random variable

$$Y_k = 1 \text{ if } k \in S$$

$$Y_k = 0 \text{ if } k \notin S$$

DISPLAY

- ▶ $D \subset \{1, 2, \dots, N\}$... a set of L distinct images.
- ▶ D_t ... the images displayed at time $t = 1, 2, \dots$
- ▶ X_D ... the response of the user to D .

For $D \cap S = \emptyset$, $X_D = i$ means i is “closest” to S ,
in the opinion of the user

SEARCH HISTORY

- ▶ History (“evidence”) after t steps:

$$\begin{aligned} B_t &= \{D_1 = d_1, X_{D_1} = i_1, \dots, D_t = d_t, X_{D_t} = i_t\} \\ &= \{D_1 = d_1, X_{D_1} = i_1, X_{D_2} = i_2, \dots, D_t = d_t, X_{D_t} = i_t\} \end{aligned}$$

because D_1 is chosen at random and D_{s+1} will depend only on D_1 and the previous answers (actually on the posterior).

- ▶ Given S and D_t , the answer X_{D_t} is independent of the search history:

$$P(X_{D_t} = i | S, B_t) = P(X_d = i | S, D_t = d)$$

DISPLAY CRITERION

$$D_{t+1} = \arg \max_D I(X_D; S|B_t)$$

SEPARATE BAYESIAN SYSTEMS FOR EACH $k \in \Omega$

- ▶ Prior model:

$$p_0(k) = P(Y_k = 1) = P(k \in S)$$

- ▶ Answer model: For $k \notin D, i \in D$,

$$q_+(i|k, D) = P(X_D = i | Y_k = 1)$$

$$q_-(i|k, D) = P(X_D = i | Y_k = 0)$$

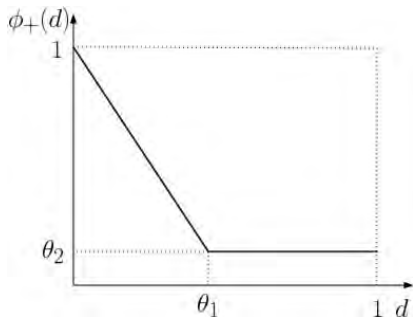
- ▶ Posterior distribution at step t :

$$p_t(k) = P(Y_k = 1 | B_t)$$

ANSWER MODELS

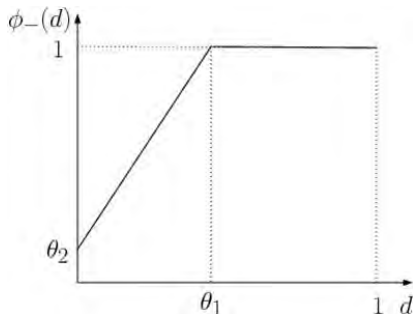
Positive Model

$$P(X_d = i | Y_k = 1) = \frac{\phi_+(d(i, k))}{\sum_{j \in D} \phi_+(d(j, k))}$$



Negative Model

$$P(X_d = i | Y_k = 0) = \frac{\phi_-(d(i, k))}{\sum_{j \in D} \phi_-(d(j, k))}$$



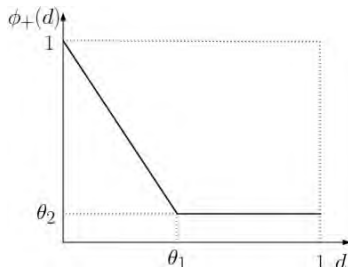
PARAMETER ESTIMATION (θ_1)

The positive model

Θ_1 : “no preference” threshold

Repeat M times:

1. Fix θ and $k \in \mathcal{S}$.
2. Choose two images i, j such that:
 - (a) $d(i, k) \approx \theta$
 - (b) $d(j, k)$ is chosen uniformly in $[\theta, 1]$
3. Display i, j and record the user's choice.



PARAMETER ESTIMATION (θ_1)

Consider two hypotheses:

- ▶ H_0 : “no preference”
- ▶ H_1 : “preference for i (closest)”

Let N^θ be the number of times the user chooses i . Under H_0 ,

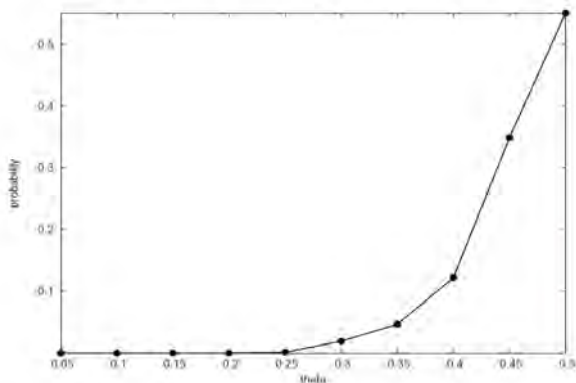
$$N^\theta \sim \text{Bin}(M, \frac{1}{2})$$

Let $p(\theta) = P(\text{Bin}(M, \frac{1}{2}) > N^\theta)$.

Choose the largest value of θ such that H_0 is rejected at $p = 0.05$.

PARAMETER ESTIMATION (θ_1)

θ	0.20	0.25	0.30	0.35	0.40	0.45
N_c^θ	181	143	136	133	129	123
p	≈ 0	0.0015	0.0194	0.0466	0.1226	0.3493



Corel Database

$M \sim 200$

$\theta_1 = 0.35$

PARAMETER ESTIMATION (θ_2)

The positive model

Θ_2 : degree of coherence with system metric

Repeat M times:

1. Fix θ and $k \in S$.
2. Choose a display D such that:
 - (a) One image i in D is very close to some $k \in S$;
 - (b) All the other images in D are more than θ_1 units away from k .
3. Display D and record the user's choice.

PARAMETER ESTIMATION (θ_2)

$$P(X_D = x_i | Y_k = 1) \cong \frac{1}{1 + (n-1)\theta_2}$$

$$\theta_2^+ \cong \frac{1}{n-1} \frac{P(X_D \neq x_i | Y_k = 1)}{P(X_D = x_i | Y_k = 1)}$$

Corel database (M=600):

$$\theta_2 = 0.065$$

UPDATE MODEL

The new posterior distribution is

$$p_{t+1}(k) = P(Y_k = 1 | B_{t+1})$$

which reduces to

$$\frac{P(X_{D_{t+1}} = i | Y_k = 1, D_{t+1})p_t(k)}{P(X_{D_{t+1}} = i | Y_k = 1, D_{t+1})p_t(k) + P(X_{D_{t+1}} = i | Y_k = 0, D_{t+1})(1 - p_t(k))}$$

which is finally

$$\frac{q_+(i|k, D_{t+1})p_t(k)}{q_+(i|k, D_{t+1})p_t(k) + q_-(i|k, D_{t+1})(1 - p_t(k))}$$

TAKING STOCK

- ▶ So mental category search reduces to two difficult tasks:
 - ▶ *An optimization problem:* Discover approximations to the optimal display.
 - ▶ *A modeling problem:* Discover answer models which match human behavior.

IDEAL USER

Suppose $d(i, S) < d(j, S)$ for each $j \in D, i \in D$. Ideal user:

$$P(X_D = i | S) = 1$$

Since S determines X_D :

$$\begin{aligned} D_{t+1} &\doteq \arg \max_D I(X_D; S | B_t) \\ &= \arg \max_D (H(X_D | B_t) - H(X_D | S, B_t)) \\ &= \arg \max_D H(X_D | B_t), \end{aligned}$$

which motivates the following choice of display:

OPTIMAL DISPLAY: THE VORONOI CELLS HAVE EQUAL MASS

