

Combining Textual- and Content-based Image Retrieval

Tutorial Image Retrieval

Thomas Deselaers, Henning Müller

Outline

- Why combination of text and images
- Late fusion
- Image Annotation and Textual Retrieval
- Direct fusion in the continuous approach
- Direct fusion in the discrete approach
- Query-based fusion (domain specific)
- Fusion by refining






Why Combinations?

- Textual and visual information are often orthogonal
- Content vs. context

Google [Advanced Image Search](#) [Preferences](#)
[Moderate SafeSearch is on](#)

Images Showing: Results 1 - 20 of about 119,000,000 for tree [definition]. (0.29 seconds)

Related searches: [cartoon tree](#) [oak tree](#) [tree clipart](#) [family tree](#)

				
More images from the category Tree. 600 x 400 - 129k www.freefoto.com [More from www.freefoto.com]	Take a tour of the beautiful trees ... 383 x 300 - 31k - jpg pws.byu.edu	Picture of Tree, Northumberland ... 600 x 400 - 97k - jpg www.freefoto.com	A Self-Adjusting Search Tree 782 x 961 - 213k - jpg www.link.cs.cmu.edu	Create your own Family Tree. 350 x 343 - 45k - jpg www.pbs.org

Combination of Textual and Visual Image Retrieval

- Google Image Search, Flickr, Youtube, ...
 - Access visual content using textual searches
 - Using textual information associated with the images
 - Meta data
 - Tags
 - Figure captions
 - ...
- Textual and visual information are often orthogonal
 - Visual features can be used to find similar images
 - Textual information can be used to find “semantically” annotated images
 - Textual information will often fail, due to
 - Unannotated images (creating a hidden web)
 - Images annotated in a different language

Combination of Textual and Visual Image Retrieval

- Try to benefit from textual and visual information
- E.g. find an initial set of images using a textual search and then find similar images using visual search
- Find an initial set of images and regroup using the respective other technique (e.g. to obtain diverse results)

Example: Textual Query visual Refinement Microsoft LiveSearch Images

Announced on Dec 1st, 2008



Advantages of Text-based Retrieval

- Find images with semantic concepts
 - Object recognition is an unsolved problem in particular in large-scale (number of concepts)
- Find images belonging to a certain event
 - E.g. show me all images from the US elections 2008
- Find images taken at a particular location
 - Although some researchers are trying to determine image location vision-based
- Many images are created with captions, description, tags, ... and are thus open to textual searches
 - And if someone took the effort to annotate images, why not use it ?

Advantages of Visual Retrieval

- Search images without any information
 - e.g. possible to find images showing a particular person, given a suitable face detection and recognition
- Find similar images
 - E.g. find images with certain colours (sunset, ...)
 - E.g. find illicit uses of “my” images (for copyright holders)
 - find duplicates to clean up a set of images

Late Fusion of Images and Text

- For a database $\mathcal{B} = \{x_1, \dots, x_n, \dots, x_N\}$
- Given the result of a textual retriever, i.e. the scores for each of the images

$$(s_1^t, \dots, s_n^t, \dots, s_N^t)$$

- And the scores of a visual retriever:

$$(s_1^v, \dots, s_n^v, \dots, s_N^v)$$

- Fuse these into a joint result:

$$(s_1, \dots, s_n, \dots, s_N)$$

Late fusion

- Weighted sum:

- Allows for arbitrary weighted combinations

$$s_n \leftarrow w_t s_n^t + w_v s_n^v$$
$$w_v = 1 - w_t, w_t \in [0, 1]$$

- Minimum:

- An image is similar if it is similar by text **or** content

$$s_n = \min\{s_n^t, s_n^v\}$$

- Maximum:

- An image is only similar if it is similar by text **and** content

$$s_n = \max\{s_n^t, s_n^v\}$$

- ...

Image Annotation and Textual Retrieval

- Google image search
 - Restrict to photos/images showing faces

The screenshot shows a Google Image Search interface. The search term is 'smith'. The search results are filtered to show only images containing faces. The interface includes search buttons, filters for image sizes and content, and a results count of approximately 144,000,000. Below the search bar, there are related search terms: 'anna nicole smith' and 'agent smith'. The main results area displays a grid of image thumbnails, each with a caption and source information. The captions include names like 'Anna Nicole Smith', 'Will Smith', 'Kim Smith', and 'Prophet Joseph Smith', along with their respective image dimensions and file sizes. A 'Training faces (show all images)' link is also visible.

Google Search Images Search the Web Advanced Image Search Preferences
Moderate SafeSearch is on

Showing: All image sizes Any content Results 1 - 20 of about 144,000,000 for smith

Related searches: [anna nicole smith](#) [agent smith](#)

Results 1 - 20 of about 7,690,000 for smith [definition]. (0.31 seconds)

Training faces (show all images)

Anna Nicole Smith
378 x 490 - 52k - jpg
www.askmen.com

Will Smith
300 x 400 - 27k - jpg
www.people.com

Will Smith Pictures
378 x 490 - 55k - jpg
www.askmen.com

Q&A with EP President Brian Smith, ...
300 x 418 - 69k - jpg
www.editorialphoto.com

niths I am not:
160 - 944k - jpg
le.tamu.edu

Wil Smith MySpace Pictures
1024 x 768 - 146k - jpg
www.coolfreepix.com

Kim Smith
1600 x 1200 - 232k - jpg
www.wallpaperbase.com

Smith's Parish
828 x 1122 - 449k - gif
www.bermuda-online.org

Will Smith learned how quickly the ...
300 x 400 - 28k - jpg
dellassouthblog.com

Anna Nicole Smith
531 x 411 - 38k - jpg
abcnews.go.com

Since the 1970's, Alexis Smith has ...
2931 x 4000 - 2616k - jpg
stuartcollection.ucsd.edu

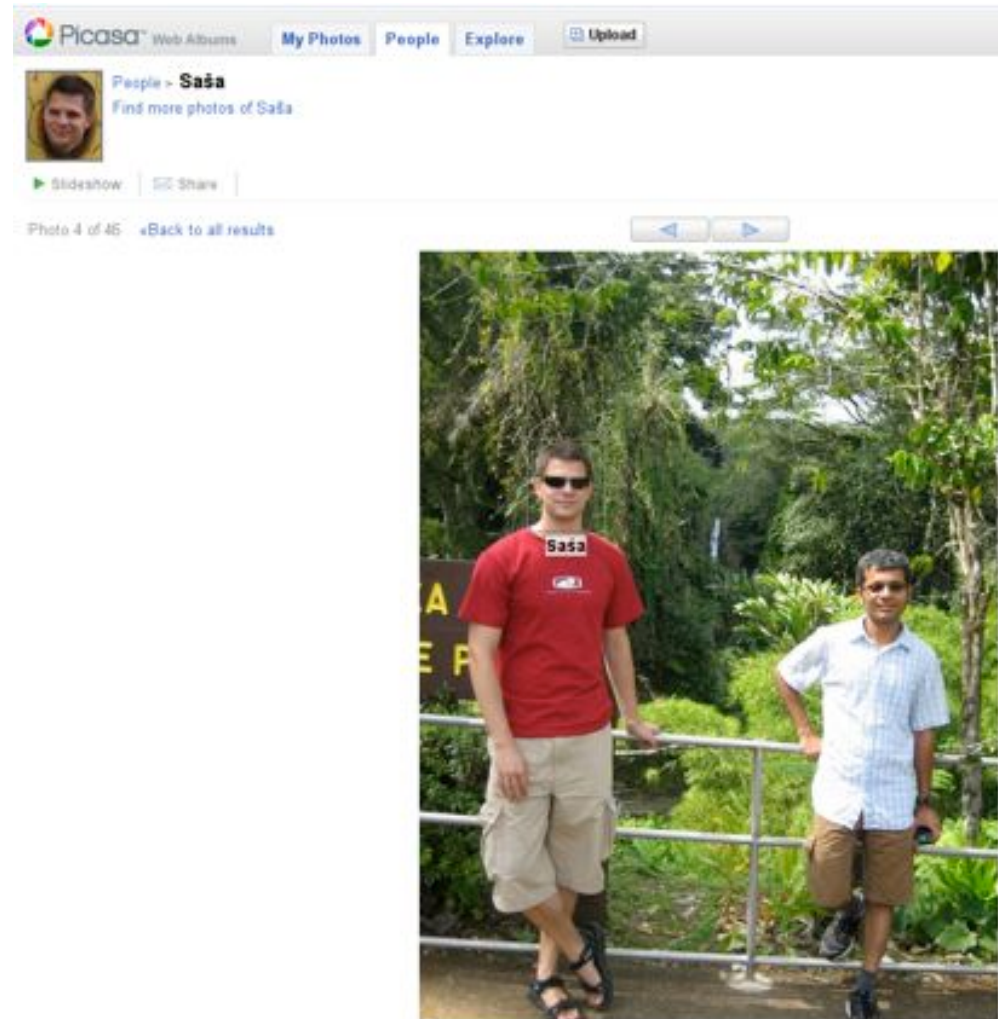
ig Smith
08 - 160k - jpg
anderbit.edu

Laurajane Smith, awaiting image
587 x 751 - 77k - jpg
www.york.ac.uk

Prophet Joseph Smith
385 x 550 - 41k
www.prophetjosephsmith.org

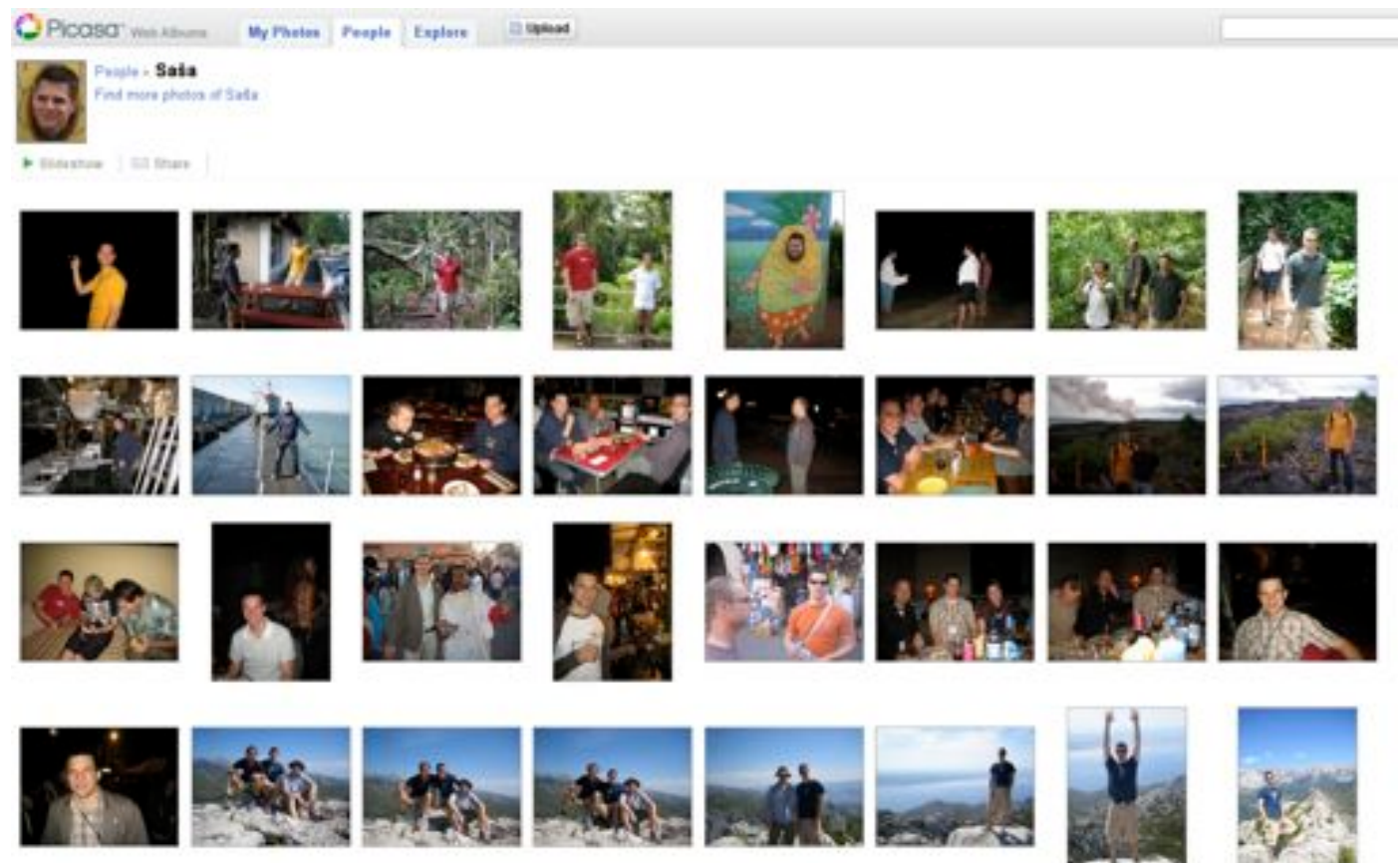
Picasa Name Tags

- Label Persons in **your** images



Picasa Name Tags

- Find images showing certain persons



Picasa Name Tags

- Refine the model for a certain person



Direct Fusion in the Continuous Approach

- Given the result of a textual retriever, i.e. the scores for each of the images

$$(s_1^t, \dots, s_n^t, \dots, s_N^t)$$

- Use the ranks or the scores as a distance function in the continuous approach and treat it like a normal visual feature
 - Textual data is a feature like any other visual feature
 - Very flexible

Direct Fusion in the Cont. Approach: Examples

Query with text and contextual information



Query by description

Direct Fusion in the Discrete Approach

- Given discrete image features and words
 - Create a joint index (i.e. inverted files) for both
 - Create individual indices for each and fuse later
 - GIFT uses a “*late*” fusion for 4 individual visual cues
 - Allows for a more flexible weighting

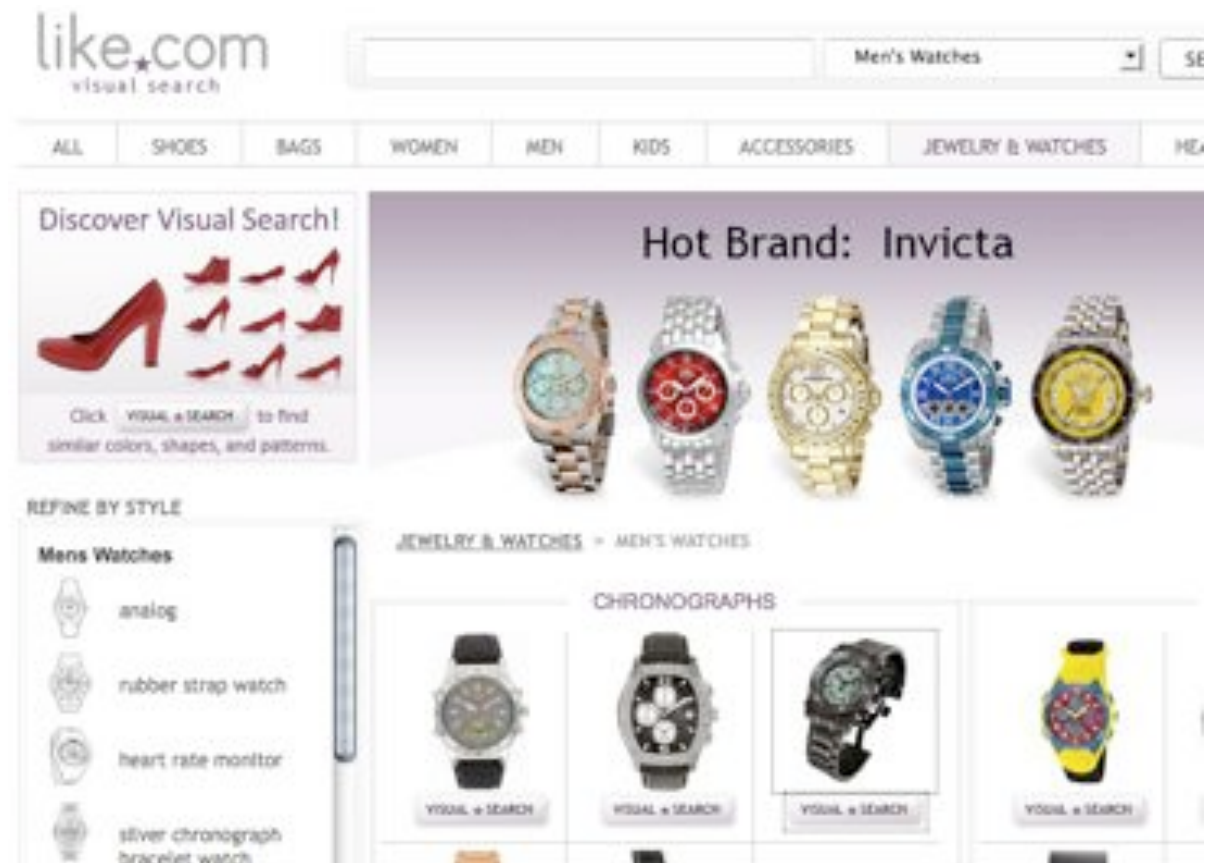
Query-based Fusion

- In certain domains, different queries might occur
 - E.g. in the medical domain, it can be distinguished between queries, that can be solved using
 - Visual techniques
 - Textual techniques
 - Mixed techniques

Fusion by Refining

Like.com

- Search in images from a certain category
- E.g. clothes, watches, ...



Fusion by Refining

- Query a database using a textual query with a clear concept in mind, e.g.

Children playing



Cell phone



- Textual search will lead to a large amount of images, most not matching “your” idea
- Use visual techniques to refine the results

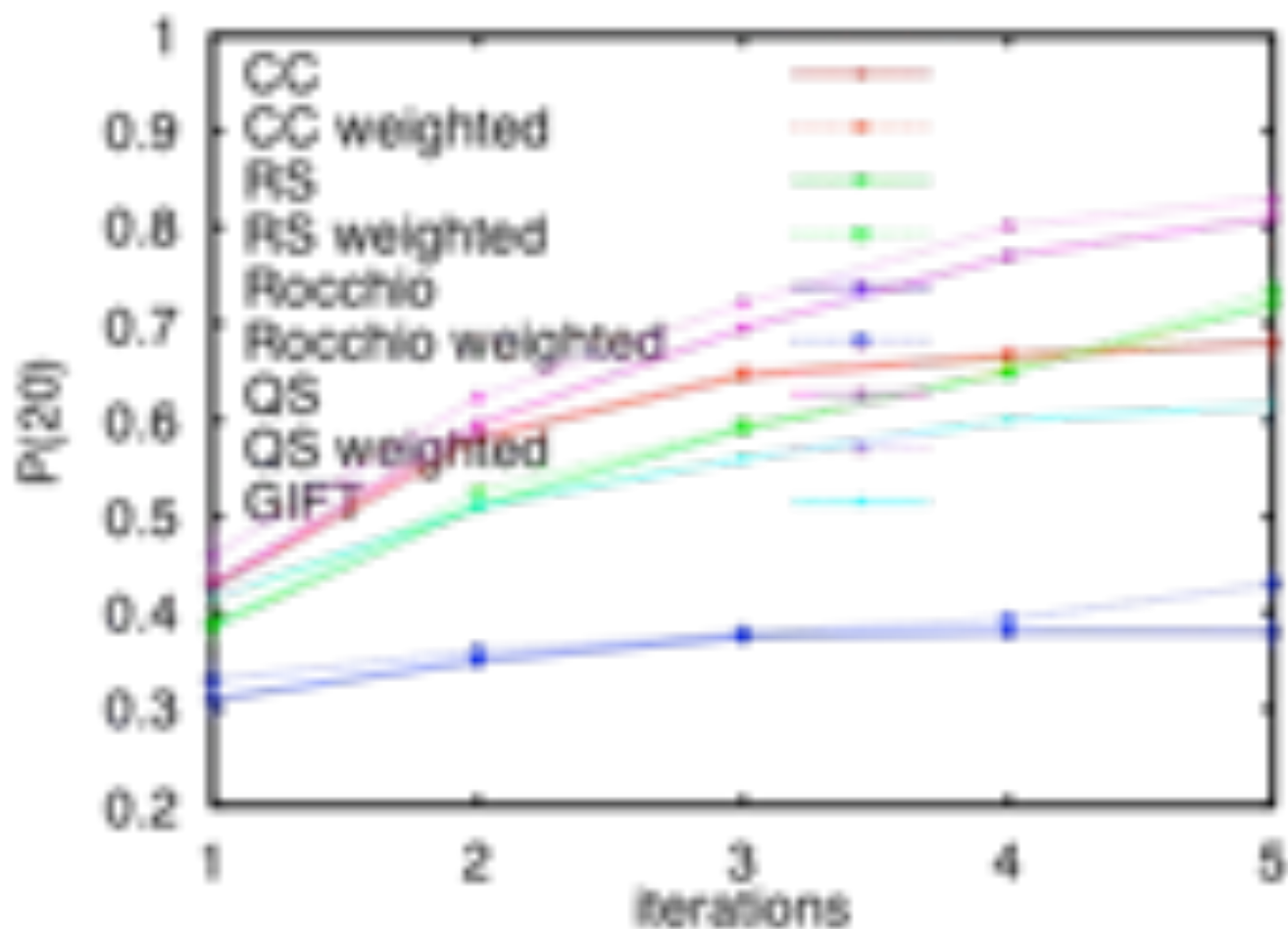
Some Results: Fusion by Refining

- Flickr Database
 - Created for evaluating fusion by refining
 - 10 queries

query	relevant	P_0	description
beach	36	0.2	beach scene, e.g. for illustration of a vacation catalogue
bike	38	0.15	single bikes
construction site	32	0.1	construction site where construction workers are working
dancing	90	0.25	energetic dancer
desert	68	0.2	lonely desert scenes
dog	62	0.15	portrait images of dogs/images where the dog is the central them
mountains	122	0.15	mountain scenes w/o people
people eating ice	158	0.65	images where it can clearly be recognised that people are eating
people running	120	0.45	people running e.g. in a sports event
teacher in classroom	66	0.2	classroom scenes where the teacher can clearly be recognised

Results on the Flickr Task

- Use Flickr's result of the query as starting point, then use relevance feedback with visual methods to re-order the images



Literature

- Proceedings of the ImageCLEF track in the CLEF Workshops 2003-2008
 - Photo retrieval
 - Medical Retrieval
 - www.imageclef.org
- iCLEF in CLEF 2007/2008
 - Multilingual retrieval in Flickr databases
 - Large-Scale Interactive Evaluation of Multilingual Information Access Systems - the iCLEF Flickr Challenge.
 - P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Artile, V. Peinado, Workshop on Novel Methodologies for Evaluation in Information Retrieval. 30th European Conference on Information Retrieval (ECIR 2008). 2008
- Clustering
 - Deselaers T., Keysers D., Ney H., "Clustering Visually Similar Images to Improve Image Search Engines", Informatiktage der Gesellschaft für Informatik, Bad Schussenried, Germany, Springer, pp. 302, 01/11/2003
- Direct fusion in the continuous approach
 - Deselaers T., Weyand T., Keysers D., Macherey W., Ney H., "FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval", CLEF Workshop 2005, vol. 4022, Vienna, Austria, Springer, pp. 652-661, 21/09/2005, 2006
- Fusion by refining
 - Paredes R., Deselaers T., Vidal E., "A probabilistic model for user relevance feedback on image retrieval", Workshop on Machine Learning and Multimodal Interaction, Utrecht, The Netherlands, pp. 260-271, 08/09/2008, 2008