Comparison and Combination of Textual and Visual Features for Interactive Image Retrieval

*ImageCLEF 2004: Volume I, pp. 621-630*

Pei-Cheng Cheng, Jen-Yuan Yeh, Hao-Ren Ke,
Been-Chian Chien and Wei-Pang Yang

Database Lab.,
Dept. of Computer & Information Science, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, TAIWAN 30050, R.O.C.
Outline

- The user-centered search task
- What helps find out relevant images?
- Interactive search process
  - Cross-language image retrieval
  - Relevance feedback
- Proposed systems
  - T_ICLEF: Textual based retrieval system
  - VCT_ICLEF: Textual & Visual based Retrieval system
- Experimental results
- Conclusion
The user-centered search task

- **Goal:** to investigate how native speakers of languages other than English interact with a CL image retrieval system.

- **Aims:** to investigate research issues, such as browsing support, automatic and interactive query expansion, relevance feedback, query formulation using both image and text, presentation of search results.
St. Andrews collection

- St. Andrews University Library photographic collection.
- Photos are primarily historic in nature from areas in and around Scotland.

Dataset Overview
- 28133 SGML documents consist of text and images.
- 946 categories
What helps find out relevant images?

- **Query by keyword**
  - A user must understand the background of a target image to describe it correctly
  - Different users use different keywords.

- **Query by visual features of image**
  - Hard to fully describe a visual query
  - Trivial: Most people have common visual perception

- **User feedback: Learning a user’s need**
  - Help to indicate the system “real-relevant” images, and find out an image in a fewer iterations
## Textual or visual?

<table>
<thead>
<tr>
<th></th>
<th>Textual query</th>
<th>Visual query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic query</td>
<td>yes</td>
<td>very little</td>
</tr>
<tr>
<td>Easily used for a user</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Easy to formula a query</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Background knowledge</td>
<td>needed</td>
<td>not require</td>
</tr>
<tr>
<td>Cross-language problems</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Visual description</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Matching metric</td>
<td>strict</td>
<td>vague</td>
</tr>
</tbody>
</table>

### Example

Query:

燈塔  
(A lighthouse)
Interactive search process

\[ Q = (Q_T, Q_I) \]

Query \( \rightarrow \) Retrieved Images \( \rightarrow \) Relevant Images \( \rightarrow \) Reformulated Query

Image Retrieval \( \rightarrow \) Relevance Feedback \( \rightarrow \) Query Reformulation

\[ Q' = (Q'_T, Q'_I) \]

An image is represented as \( P = (P_T, P_I) \)
Cross-language image retrieval
Textual vector representation

\[ P_T = \langle w_{t_1}(P_T), \ldots, w_{t_n}(P_T), w_{c_1}(P_T), \ldots, w_{c_m}(P_T), w_{y_1}(P_T), \ldots, w_{y_k}(P_T) \rangle \]

Term: \[ w_{t_i}(P_T) = \frac{tf_{t_i,P_T}}{\max tf} \times \log \frac{N}{n_{t_i}} \]

Category: \[ w_{c_i}(P_T) = \begin{cases} 1 & \text{if } P \text{ belongs to } c_i, \\ 0 & \text{otherwise} \end{cases} \]

Temporal: \[ w_{y_i}(P_T) = \begin{cases} 1 & \text{if } P \text{ was published in } y_i, \\ 0 & \text{otherwise} \end{cases} \]
Textual vector representation (cont.)

\[ Q_T = \langle w_{t_1}(Q_T), \ldots, w_{t_n}(Q_T), w_{c_1}(Q_T), \ldots, w_{c_m}(Q_T), w_{y_1}(Q_T), \ldots, w_{y_k}(Q_T) \rangle \]

\[ w_{t_i}(Q_T) = \begin{cases} \displaystyle \frac{tf_{t_i,Q_T}}{\max tf} \times \log \frac{N}{n_{t_i}} & \text{if } Q_T \text{ contains } "Y年以前," \text{ and } y_i \text{ is before } Y, \\ 1 & \text{if } Q_T \text{ contains } "Y年之中," \text{ and } y_i \text{ is in } Y, \\ 1 & \text{if } Q_T \text{ contains } "Y年以後," \text{ and } y_i \text{ is after } Y, \\ 0 & \text{otherwise} \end{cases} \]

\[ w_{c_i}(Q_T) = \begin{cases} 1 & \text{if } \exists j, e_j \in \text{AfterDisambiguity}(Q_T) \text{ and } e_j \text{ occurs in } c_i, \\ 0 & \text{otherwise} \end{cases} \]
Temporal operator

\[
w_{y_i}(Q_T) = \begin{cases} 
1 & \text{if } Q_T \text{ contains "年以前," and } y_i \text{ is BEFORE } Y, \\
1 & \text{if } Q_T \text{ contains "年之中," and } y_i \text{ is IN } Y, \\
1 & \text{if } Q_T \text{ contains "年以後," and } y_i \text{ is AFTER } Y, \\
0 & \text{otherwise}
\end{cases}
\]

- 1900年以前拍攝的愛丁堡城堡照片？
  
  (Pictures of Edinburgh Castle taken before 1900)
Image vector representation

- Color histogram
- HSV color space (18*2*4+4 gray values)
  - 18 hues, 2 saturations, and 4 values, with additional 4 levels of gray

\[ P_I = \langle h_{c_1}(P_I), ..., h_{c_m}(P_I) \rangle \]

\[ Q_I = \langle h_{c_1}(Q_I), ..., h_{c_m}(Q_I) \rangle \]

\[ h_{c_i}(P_I) = \frac{\sum_{p \in P_I} |\alpha_p - \beta_p|}{\delta} \]

Partial pixel contribution (PPC)

\[ \alpha_p \rightarrow c_{i-1} \rightarrow c_i \rightarrow c_{i+1} \rightarrow \beta_p \]
Similarity metric

- **Strategy 1 (T_ICLEF):** Based on the textual similarity

  \[ P_T = \langle w_{t_1}(P_T), \ldots, w_{t_n}(P_T), w_{c_1}(P_T), \ldots, w_{c_m}(P_T), w_{y_1}(P_T), \ldots, w_{y_k}(P_T) \rangle \]

  \[ Q_T = \langle w_{t_1}(Q_T), \ldots, w_{t_n}(Q_T), w_{c_1}(Q_T), \ldots, w_{c_m}(Q_T), w_{y_1}(Q_T), \ldots, w_{y_k}(Q_T) \rangle \]

  \[ Sim_1(P, Q) = \frac{P_T \cdot Q_T}{|P_T||Q_T|} \]

- **Strategy 2 (VCT_ICLEF):** Based on both the textual and the image similarity

  \[ Sim_2(P, Q) = \alpha \cdot Sim_1(P, Q) + \beta \cdot Sim_3(P, Q) \]

  where

  \[ P_I = \langle h_{c_1}(P_I), \ldots, h_{c_m}(P_I) \rangle, \]

  \[ Q_I = \langle h_{c_1}(Q_I), \ldots, h_{c_m}(Q_I) \rangle, \]

  \[ Sim_3(P, Q) = \frac{H(P_I) \cap H(Q_I)}{|H(Q_I)|} = \frac{\sum \min(h_{c_i}(P_I), h_{c_i}(Q_I))}{\sum h_{c_i}(Q_I)} \]
Query reformulation

original query \( Q = (Q_T, Q_I) \)

new query \( Q' = (Q'_T, Q'_I) \)

\[
Q'_T = \alpha \cdot Q_T + \frac{\beta}{|\text{REL}|} \sum_{P_T \in \text{REL}} P_T - \frac{\gamma}{|\text{NREL}|} \sum_{P_T \in \text{NREL}} P_T
\]

\[
Q'_I = \frac{1}{|\text{REL}|} \sum_{P_I \in \text{REL}} P_I
\]

- REL: relevant images
- NREL: irrelevant images
VCT_ICLEF
Design of the color table

- A full color table is hard for a user to describe a visual query.
- The retrieval system will expand the visual query automatically.
Experiment

- An experimental procedure similar to iCLEF 2003.
- 8 users are asked to test each system with 8 topics.
- Topics and systems will be presented to a user in a latin-square combination.
Searcher background

- 5 male and 3 female searchers.
- Average age is 23.5, with the youngest of 22 and the oldest of 26.
- Three of them major in computer science, two major in social science, and the others are librarians.
- All of them have an average of 3.75 years accessing online search services.
- Only a half of them have experiences in using image search services.
Experiment results

- Average number of iterations spent by a searcher for each topic
Experiment results (cont.)

- The average time spent by a searcher for each topic
Experiment results (cont.)

- Number of searchers who did not find the target image for each topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_ICLEF</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>VCT_ICLEF</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

- Average steps to find the target image, and the average spent time

<table>
<thead>
<tr>
<th>Topic</th>
<th>Avg. Iterations (Not including not found)</th>
<th>Avg. Spent Time for each topic</th>
<th>Avg. percent of searchers who found the target image (#/4×100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_ICLEF</td>
<td>2.24</td>
<td>208.67s</td>
<td>56.25%</td>
</tr>
<tr>
<td>VCT_ICLEF</td>
<td>1.84</td>
<td>132.20s</td>
<td>89.00%</td>
</tr>
</tbody>
</table>
Relevance feedback example

- Results after the initial search: “長鬍子的男人 (A man with a beard)”
Relevance feedback example

- Results after 1 feedback iteration
Relevance feedback example

- Results after 2 feedback iterations
User Survey

- 5 searchers thought that color information was helpful
- 4 searchers preferred to search with a text query first
  - The same cases happened while using VCT_ICLEF.
  - But will indicate color information when the system returned images all in black and white.
- 3 searchers preferred to use color information first.
- 2 searchers hoped to use a text query to describe color information
  - e.g., 藍色 (Blue):
- No one search with a temporal query since it is hard to decide in which year the image was published.
Conclusion

- The VCT_ICLEF has a better performance than T_ICLEF
  - The visual features will improve the performance.
  - The visual features can offer some clues for image retrieval system.

- Future work: a browsing interface
  - SOM (Self-Organizing Map) for image clustering.
Q&A

Thank you