Medical Task
ImageCLEF 2011

Henning Müller¹, Theodora Tsikrika¹, Steven Bedrick², Jayashree Kalpathy-Cramer², Ivan Eggel¹, Alba Garcia Seco de Herrera¹

¹ University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
² Oregon Health and Science University (OHSU), Portland, OR, USA
Support

CHORUS+
AUDIO-VISUAL SEARCH

khresmoi
MEDICAL INFORMATION ANALYSIS & RETRIEVAL

MANY

promise

FNS SNF

Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation
1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
The Medical track consists of three tasks:

1. Modality classification
2. Image-based retrieval
3. Case-based retrieval
Modality classification consists of identifying an image’s modality using some combination of textual or visual features:
Modality classification consists of identifying an image’s modality using some combination of textual or visual features:
Modality classification consists of identifying an image’s modality using some combination of textual or visual features:
Modality classification consists of identifying an image’s modality using some combination of textual or visual features:

- CT Scan
- X-Ray
- Photograph (Endoscopic?)
Modality Classification

Why do we care?
Modality Classification

Why do we care?

An image’s modality is one of its most important features!
Why do we care?

An image’s modality is one of its most important features!

Previous studies have taught us that end users consider modality to be a very useful and important search facet.
Modality Classification

The task:
The task:

Given an image and its metadata (caption, etc.)...
The task:

Given an image and its metadata (caption, etc.)...

... map the image to one of a set of possible modalities.
Standard TREC-style retrieval task:
Standard TREC-style retrieval task:

Data: Images along with annotations and metadata
Standard TREC-style retrieval task:

Data: Images along with annotations and metadata

Topics (text, along with sample images) representing a real-world clinical information needs
Image-based Retrieval

Standard TREC-style retrieval task:

Data: Images along with annotations and metadata

Topics (text, along with sample images) representing a real-world clinical information needs

Unit of retrieval: individual images
Standard TREC-style retrieval task:

Data: Images along with annotations and metadata

Topics (text, along with sample images) representing a real-world clinical information needs

Unit of retrieval: individual images

Expert relevance judgments as “ground truth”
Case-based Retrieval

Key differences from image-based retrieval:
Case-based Retrieval

Key differences from image-based retrieval:

Textual topics representing detailed descriptions of clinical cases from a teaching collection
Case-based Retrieval

Key differences from image-based retrieval:

Textual topics representing detailed descriptions of clinical cases from a teaching collection

Unit of retrieval: whole articles
1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
What’s new in 2011?

1. New, larger test collection
2. New set of image modalities
What’s new in 2011?

1. New, larger test collection
2. New set of image modalities
The 2011 test collection consisted of ≈231,000 images from PubMedCentral.
The 2011 test collection consisted of ≈231,000 images from PubMedCentral.

The collection includes images from a wider variety of journals than in previous years...
The 2011 test collection consisted of \( \approx 231,000 \) images from PubMedCentral.

The collection includes images from a wider variety of journals than in previous years...

... which added new challenges to this year’s tasks.
This year’s images:
This year’s images:

1. Have “noisier” captions...
This year’s images:

1. Have “noisier” captions...

2. Include many more “irrelevant” images (tables, flowcharts, etc.)...
New Test Collection

This year’s images:

1. Have “noisier” captions...

2. Include many more “irrelevant” images (tables, flowcharts, etc.)...

... than those used in previous years’ campaigns.
What’s new in 2011?

1. New, larger test collection
2. New modality hierarchy
The 2010 modality classification task used a set of eight classes.
The 2010 modality classification task used a set of eight classes.

CT, MR, XR, US, PX, GX, NM, PET
The 2010 modality classification task used a set of eight classes.

CT, MR, XR, US, PX, GX, NM, PET

This year’s larger and more diverse image collection needed more categories.
**New Modality Set**

This year’s larger and more diverse image collection needed more categories, 18 in all:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>3d reconstruction</td>
</tr>
<tr>
<td>AN</td>
<td>angiography</td>
</tr>
<tr>
<td>CM</td>
<td>compound figure (more than one type of image)</td>
</tr>
<tr>
<td>CT</td>
<td>computed tomography</td>
</tr>
<tr>
<td>DM</td>
<td>dermatology</td>
</tr>
<tr>
<td>DR</td>
<td>drawing</td>
</tr>
<tr>
<td>EM</td>
<td>electronMicroscopy</td>
</tr>
<tr>
<td>EN</td>
<td>endoscopic imaging</td>
</tr>
<tr>
<td>FL</td>
<td>fluorescence</td>
</tr>
<tr>
<td>GL</td>
<td>gel</td>
</tr>
<tr>
<td>GX</td>
<td>graphs</td>
</tr>
<tr>
<td>GR</td>
<td>gross pathology</td>
</tr>
<tr>
<td>HX</td>
<td>histopathology</td>
</tr>
<tr>
<td>MR</td>
<td>magnetic resonance imaging</td>
</tr>
<tr>
<td>PX</td>
<td>general photo</td>
</tr>
<tr>
<td>RN</td>
<td>retinograph</td>
</tr>
</tbody>
</table>
We provided participants with 1,000 training and 1,000 test images, and participants annotated the remaining 229,000 images in the collection.
We provided participants with 1,000 training and 1,000 test images, and participants annotated the remaining 229,000 images in the collection.

This larger collection will be used as a resource in future years’ tasks— stay tuned!
17 teams submitted a total of 207 runs!
(+Δ33% from 2010)
Participation

17 teams submitted a total of 207 runs! (+Δ33% from 2010)

34 Modality classification
17 teams submitted a total of 207 runs!
(+Δ33% from 2010)

34 Modality classification

130 Image-based retrieval
Participation

17 teams submitted a total of 207 runs! (+Δ33% from 2010)

34 Modality classification
130 Image-based retrieval
43 Case-based retrieval
• 55 registrations, 17 groups submitting results (*=new groups)
  ○ BUAA AUDR (China)*
  ○ CEB, NLM (USA)
  ○ DAEDALUS UPM (Spain)
  ○ DEMIR (Turkey)
  ○ HITEC (Belgium)*
  ○ IPL (Greece)
  ○ IRIT (France)
  ○ LABERINTO (Spain)*
  ○ SFSU (USA)*
  ○ medGIFT (Switzerland)
  ○ MRIM (France)
  ○ Recod (Brazil)
  ○ SINAI (Spain)
  ○ UESTC (China)*
  ○ UNED (Spain)
  ○ UNT (USA)
  ○ XRCE (France)
1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
Six teams submitted a total of 34 runs.
Six teams submitted a total of 34 runs.

Two purely-textual
Six teams submitted a total of 34 runs.

Two purely-textual

Seventeen purely-visual
Six teams submitted a total of 34 runs.

Two purely-textual

Seventeen purely-visual

Fifteen “mixed”
This year’s scores were lower than last year’s:
This year’s scores were lower than last year’s:

In 2010, several runs had classification accuracies > 0.9...
This year’s scores were lower than last year’s:

In 2010, several runs had classification accuracies > 0.9...

... this year, however, the best run had an accuracy score of 0.8691.
This year’s scores were lower than last year’s:

In 2010, several runs had classification accuracies > 0.9...

... this year, however, the best run had an accuracy score of 0.8691.

Why?
This year’s modality classification task was more difficult than last year’s.


## 2011 Modality Classification

<table>
<thead>
<tr>
<th>Run</th>
<th>Group</th>
<th>Run Type</th>
<th>Classification</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE_all_MIX_semiLM.txt</td>
<td>XRCE</td>
<td>Mixed</td>
<td></td>
<td>0.8691</td>
</tr>
<tr>
<td>XRCE_Testset_MIX_semiL50.txt</td>
<td>XRCE</td>
<td>Mixed</td>
<td></td>
<td>0.8642</td>
</tr>
<tr>
<td>2011.06.10-02.38.40.test.prediction.trec</td>
<td>HITEC</td>
<td>Mixed</td>
<td></td>
<td>0.8603</td>
</tr>
<tr>
<td>2011.06.09-18.36.25.test.prediction.trec</td>
<td>HITEC</td>
<td>Mixed</td>
<td></td>
<td>0.8564</td>
</tr>
<tr>
<td>2011.06.08-19.58.41.test.prediction.trec</td>
<td>HITEC</td>
<td>Mixed</td>
<td></td>
<td>0.8515</td>
</tr>
<tr>
<td>2011.06.10-00.01.26.test.prediction.trec</td>
<td>HITEC</td>
<td>Mixed</td>
<td></td>
<td>0.7685</td>
</tr>
<tr>
<td>image_text_test_result_multilevel.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.7412</td>
</tr>
<tr>
<td>2011.06.10-03.25.40.test.prediction.trec</td>
<td>HITEC</td>
<td>Mixed</td>
<td></td>
<td>0.7412</td>
</tr>
<tr>
<td>image_text_test_result_sum_ext.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.6025</td>
</tr>
<tr>
<td>image_text_test_result_sum.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.5966</td>
</tr>
<tr>
<td>image_text_test_result_multilevel_ext.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.5917</td>
</tr>
<tr>
<td>image_text_test_result_sum.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.5917</td>
</tr>
<tr>
<td>image_text_test_result_sum.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.5820</td>
</tr>
<tr>
<td>image_text_test_result_original.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.5820</td>
</tr>
<tr>
<td>image_text_test_result_ext.dat</td>
<td>CEB</td>
<td>Mixed</td>
<td></td>
<td>0.5439</td>
</tr>
<tr>
<td>ICLEF2011_MED_MODALITY_09062011_1500.txt</td>
<td>IPL</td>
<td>Textual</td>
<td></td>
<td>0.7041</td>
</tr>
<tr>
<td>ICLEF2011_MED_MODALITY_09062011_1600.txt</td>
<td>IPL</td>
<td>Textual</td>
<td></td>
<td>0.4765</td>
</tr>
<tr>
<td>XRCE_Testset_MIX_semiL25.txt</td>
<td>XRCE</td>
<td>Visual</td>
<td></td>
<td>0.8593</td>
</tr>
<tr>
<td>XRCE_all_VIS_semiL25.txt</td>
<td>XRCE</td>
<td>Visual</td>
<td></td>
<td>0.8359</td>
</tr>
<tr>
<td>XRCE_Testset_VIS_semi20_CBIR.txt</td>
<td>XRCE</td>
<td>Visual</td>
<td></td>
<td>0.8349</td>
</tr>
<tr>
<td>XRCE_all_VIS_semi20_CBIR.txt</td>
<td>XRCE</td>
<td>Visual</td>
<td></td>
<td>0.8339</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_357l</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6972</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_V1</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6943</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_V1NoR</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6904</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_VsNoR</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6835</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_Vs</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6806</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_343s</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6787</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_370l</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6787</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_370s</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6767</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_343l</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6748</td>
</tr>
<tr>
<td>recod_imageclefmed_ModCla_357s</td>
<td>Recod</td>
<td>Visual</td>
<td></td>
<td>0.6669</td>
</tr>
<tr>
<td>classificationResults_GIFT.txt</td>
<td>medGIFT</td>
<td>Visual</td>
<td></td>
<td>0.6220</td>
</tr>
<tr>
<td>image_text_test_result_original.dat</td>
<td>CEB</td>
<td>Visual</td>
<td></td>
<td>0.5712</td>
</tr>
<tr>
<td>image_text_test_result_ext.dat</td>
<td>CEB</td>
<td>Visual</td>
<td></td>
<td>0.4853</td>
</tr>
</tbody>
</table>
2011 Modality Classification

Visual Techniques:
Visual Techniques:

Many runs (including Xerox’s top-performers) used the Scale Invariant Feature Transform...
2011 Modality Classification

Visual Techniques:

Many runs (including Xerox’s top-performers) used the Scale Invariant Feature Transform...

... a wide variety of other textural, spatial, and color features made appearances, as well.
2011 Modality Classification

Visual Techniques:

Many runs (including Xerox’s top-performers) used the Scale Invariant Feature Transform...

... a wide variety of other textural, spatial, and color features made appearances, as well.

Often, the best results came from combining several feature sets.
Visual Techniques:
Visual Techniques:

Several teams used off-the-rack tools such as the GNU Image Finding Tool or the Lucene Image Retrieval library...
Visual Techniques:

Several teams used off-the-rack tools such as the GNU Image Finding Tool or the Lucene Image Retrieval library...

... whereas other teams “rolled their own” toolchains:
Visual Techniques:

Several teams used off-the-rack tools such as the GNU Image Finding Tool or the Lucene Image Retrieval library...

... whereas other teams “rolled their own” toolchains:

kNN, k-means, SVM, genetic programming, etc.
Textual Techniques:
Textual Techniques:

Only one team (IPL) submitted text-only runs for modality classification.
Textual Techniques:

Only one team (IPL) submitted text-only runs for modality classification.

Note that their Lucene-based approach outperformed many visual runs!
Mixed Techniques:
Mixed Techniques:

As in past years, the best performance came from mixing visual and textual techniques.
Mixed Techniques:

As in past years, the best performance came from mixing visual and textual techniques.

In addition to the grab-bag of visual features, some teams made use of binary features derived from captions and other metadata, including MeSH mapping.
1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
14 teams submitted a total of 130 runs.
14 teams submitted a total of 130 runs.

64 textual runs
14 teams submitted a total of 130 runs.

64 textual runs

26 visual runs
2011 Image-based Retrieval

14 teams submitted a total of 130 runs.

64 textual runs

26 visual runs

40 “mixed” or “multimodal” runs
14 teams submitted a total of 130 runs.

64 textual runs

26 visual runs

40 “mixed” or “multimodal” runs

Almost all runs were “automatic” (≈11 “feedback”, 1 “manual”, several “Not Applicable”)
A set of 30 textual topics that arose from a user study designed to solicit realistic information needs from clinicians.
A set of 30 textual topics that arose from a user study designed to solicit realistic information needs from clinicians.

“photographs of benign or malignant skin lesions”
A set of 30 textual topics that arose from a user study designed to solicit realistic information needs from clinicians.

“photographs of benign or malignant skin lesions”

“medial meniscus MRI”
A set of 30 textual topics that arose from a user study designed to solicit realistic information needs from clinicians.

“photographs of benign or malignant skin lesions”

“medial meniscus MRI”

“microscopic images of tissue from the cerebellum”
A set of 30 textual topics that arose from a user study designed to solicit realistic information needs from clinicians.

“photographs of benign or malignant skin lesions”

“medial meniscus MRI”

“microscopic images of tissue from the cerebellum”

“images of findings with Alzheimer's Disease”
As usual, the best results came from multimodal approaches:

Best MAP runs:

<table>
<thead>
<tr>
<th>Team</th>
<th>Type</th>
<th>map</th>
<th>bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMIR</td>
<td>Mixed</td>
<td>0.2372</td>
<td>0.2738</td>
</tr>
<tr>
<td>LABERINTO</td>
<td>Textual</td>
<td>0.2172</td>
<td>0.2402</td>
</tr>
<tr>
<td>IPL</td>
<td>Visual</td>
<td>0.0338</td>
<td>0.0717</td>
</tr>
</tbody>
</table>
Visual Techniques:
2011 Image-based Retrieval

Visual Techniques:

Participants used similar visual features as seen in the modality classification task (SIFT, etc.), as well as Edge Histogram Descriptors and Brightness Texture Histograms.
Visual Techniques:

Participants used similar visual features as seen in the modality classification task (SIFT, etc.), as well as Edge Histogram Descriptors and Brightness Texture Histograms.

Lucene (via LIRE) and GIFT performed best among the visual systems.
Textual Techniques:
Textual Techniques:

As in past years, Lucene was the most commonly-used search engine.
Textual Techniques:

As in past years, Lucene was the most commonly-used search engine.

Several groups used external language resources (MeSH/UMLS, MetaMap, etc.)
Textual Techniques:

As in past years, Lucene was the most commonly-used search engine.

Several groups used external language resources (MeSH/UMLS, MetaMap, etc.)

The best-performing groups all used some form of query expansion, although this approach caused problems for some teams.
Multimodal Techniques:
Multimodal Techniques:

On average, textual techniques outperformed multimodal techniques- fusion is hard!
Multimodal Techniques:

On average, textual techniques outperformed multimodal techniques- fusion is hard!

The best multimodal runs may not have been purely automatic, and so may have incorporated manual feedback.
Multimodal Techniques:

On average, textual techniques outperformed multimodal techniques—fusion is hard!

The best multimodal runs may not have been purely automatic, and so may have incorporated manual feedback.

Often, rank-based fusion outperformed score-based fusion.
There was considerable between-run variation in MAP:
Furthermore, some topics had higher MAP than others:

Why?
There was considerable between-topic variation in the number of relevant results...
... although this did not seem to strongly affect MAP:
For example, topic 20 ("CT liver abscess") had very few relevant results (n=6), but had a relatively high average MAP (≈0.29)...
For example, topic 20 ("CT liver abscess") had very few relevant results (n=6), but had a relatively high average MAP (≈0.29)...

... whereas topic 27 ("gastrointestinal neoplasm") had many potentially relevant results (n=204), but had a low average MAP (≈0.05).
The topics were quite variable in terms of complexity:
The topics were quite variable in terms of complexity:

Topic 16: “images with multinucleated giant cells”

*No modality limit, simple pathology, etc.*
The topics were quite variable in terms of complexity:

Topic 16: “images with multinucleated giant cells”
*No modality limit, simple pathology, etc.*

Topic 17: “microscopic pathology images of the kidney”
*Modality and anatomy limits*
The topics were quite variable in terms of complexity:

Topic 16: “images with multinucleated giant cells”
   *No modality limit, simple pathology, etc.*

Topic 17: “microscopic pathology images of the kidney”
   *Modality and anatomy limits*

Topic 12: “microscopic pathologies of cases with chronic myelogenous leukemia”
   *Modality, pathology, and disease state (“chronic”)*
What about other metrics?

map by bpref, All topics
Very few runs have bprefs out of line with their MAPs.

What about on a per-topic basis?
Differences between a run’s MAP and its bpref suggest that the run contained a significant number of un-judged images.

This sort of analysis may prove useful for identifying pooling problems in future campaigns.
General Overview

1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
6 teams submitted a total of 43 case-based runs.
6 teams submitted a total of 43 case-based runs.

35 textual runs
6 teams submitted a total of 43 case-based runs.

35 textual runs

4 visual runs
6 teams submitted a total of 43 case-based runs.

35 textual runs

4 visual runs

4 “mixed” or “multimodal” runs
6 teams submitted a total of 43 case-based runs.

35 textual runs

4 visual runs

4 “mixed” or “multimodal” runs

Virtually all runs were “automatic” (42 auto, 1 “manual”)

This year’s case-based topics were also re-used.
This year’s case-based topics were also re-used. The ten topics are derived from the Casimage radiology teaching file, and range widely in complexity:
This year’s case-based topics were also re-used.

The ten topics are derived from the Casimage radiology teaching file, and range widely in complexity:

“Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity...”
This year’s case-based topics were also re-used.

The ten topics are derived from the Casimage radiology teaching file, and range widely in complexity:

“Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity…”

“Young female gymnast presents with leg pain.”
This year’s case-based topics were also re-used.

The ten topics are derived from the Casimage radiology teaching file, and range widely in complexity:

“Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity...”

“Young female gymnast presents with leg pain.”

Unlike the image-based retrieval tasks, the unit of retrieval for case-based topics was an entire article.
Virtually all of the teams used text-only techniques; as with the image-based retrieval, these techniques outperformed visual or multimodal techniques.

Best MAP runs:

<table>
<thead>
<tr>
<th>Team</th>
<th>Type</th>
<th>map</th>
<th>bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>medGIFT</td>
<td>Mixed</td>
<td>0.0754</td>
<td>0.0958</td>
</tr>
<tr>
<td>UESTC</td>
<td>Textual</td>
<td>0.1297</td>
<td>0.1212</td>
</tr>
<tr>
<td>medGIFT</td>
<td>Visual</td>
<td>0.0204</td>
<td>0.0292</td>
</tr>
</tbody>
</table>
Virtually all of the teams used text-only techniques; as with the image-based retrieval, these techniques outperformed visual or multimodal techniques.
Virtually all of the teams used text-only techniques; as with the image-based retrieval, these techniques outperformed visual or multimodal techniques.

The most effective textual techniques indexed the full text of the articles represented in the collection (not just image captions!)
Virtually all of the teams used text-only techniques; as with the image-based retrieval, these techniques outperformed visual or multimodal techniques.

The most effective textual techniques indexed the full text of the articles represented in the collection (not just image captions!)

Only two groups attempted any kind of visual-textual fusion.
As with the image-based task, there was a great deal of between-topic variation in performance...
As with the image-based task, there was a great deal of between-topic variation in performance...

... but less variation in the number of potentially relevant results.
1. Task summaries
2. What’s new in 2011?
3. Modality classification
4. Image-based retrieval
5. Case-based retrieval
6. Judging
As in years past, our judges have been clinicians, and use our custom judging interface:
As in years past, our judges have been clinicians, and use our custom judging interface:
As in years past, our judges have been clinicians, and use our custom judging interface:
Nine of the image-based topics were judged by at least two judges.
Judging

Nine of the image-based topics were judged by at least two judges.

- Kappa scores were generally good, and sometimes very good…
Nine of the image-based topics were judged by at least two judges.

- Kappa scores were generally good, and sometimes very good…
  - Worst was topic #14 ("angiograms containing the aorta") with ≈0.43
Nine of the image-based topics were judged by at least two judges.

• Kappa scores were generally good, and sometimes very good...
  • Worst was topic #14 (“angiograms containing the aorta”) with ≈0.43
  • Best was topic #3 (“Doppler ultrasound images (colored)”) with ≈0.92
Nine of the image-based topics were judged by at least two judges.

- Kappa scores were generally good, and sometimes very good…
  - Worst was topic #14 (“angiograms containing the aorta”) with ≈0.43
  - Best was topic #3 (“Doppler ultrasound images (colored)”) with ≈0.92
- Kappas varied from topic to topic and judge-pair to judge-pair.
Nine of the image-based topics were judged by at least two judges.

- Kappa scores were generally good, and sometimes very good…
  - Worst was topic #14 (“angiograms containing the aorta”) with \( \approx 0.43 \)
  - Best was topic #3 (“Doppler ultrasound images (colored)”) with \( \approx 0.92 \)
- Kappas varied from topic to topic and judge-pair to judge-pair.
  - For example, on topic #2 (“images containing one or several full-body scintigraphies”):
Nine of the image-based topics were judged by at least two judges.

- Kappa scores were generally good, and sometimes very good…
  - Worst was topic #14 (“angiograms containing the aorta”) with ≈0.43
  - Best was topic #3 (“Doppler ultrasound images (colored)”) with ≈0.92
- Kappas varied from topic to topic and judge-pair to judge-pair.
  - For example, on topic #2 (“images containing one or several full-body scintigraphies”):
    - judges 6 and 5 had a kappa of ≈0.79…
Nine of the image-based topics were judged by at least two judges.

- Kappa scores were generally good, and sometimes very good...
  - Worst was topic #14 (“angiograms containing the aorta”) with ≈0.43
  - Best was topic #3 (“Doppler ultrasound images (colored)”) with ≈0.92
- Kappas varied from topic to topic and judge-pair to judge-pair.
  - For example, on topic #2 (“images containing one or several full-body scintigraphies”):
    - judges 6 and 5 had a kappa of ≈0.79...
    - … while judges 6 and 8 had a kappa of ≈0.56
The new, larger database presented new, larger retrieval challenges...
Conclusions

The new, larger database presented new, larger retrieval challenges...

... Modality classification is a useful technique for filtering results and limiting search space...
Conclusions

The new, larger database presented new, larger retrieval challenges...

... Modality classification is a useful technique for filtering results and limiting search space...

... Fusing visual and textual retrieval is a powerful tool for improving search performance... but it can be a double-edged sword...
The new, larger database presented new, larger retrieval challenges...

... Modality classification is a useful technique for filtering results and limiting search space...

... Fusing visual and textual retrieval is a powerful tool for improving search performance... but it can be a double-edged sword...

... blindly aggregating results across topics can be misleading, as much information remains hidden.
Thanks! Questions?

Stay tuned for news and discussion about ImageCLEF 2012!

Thanks to all of our funders, sponsors, and participants.