

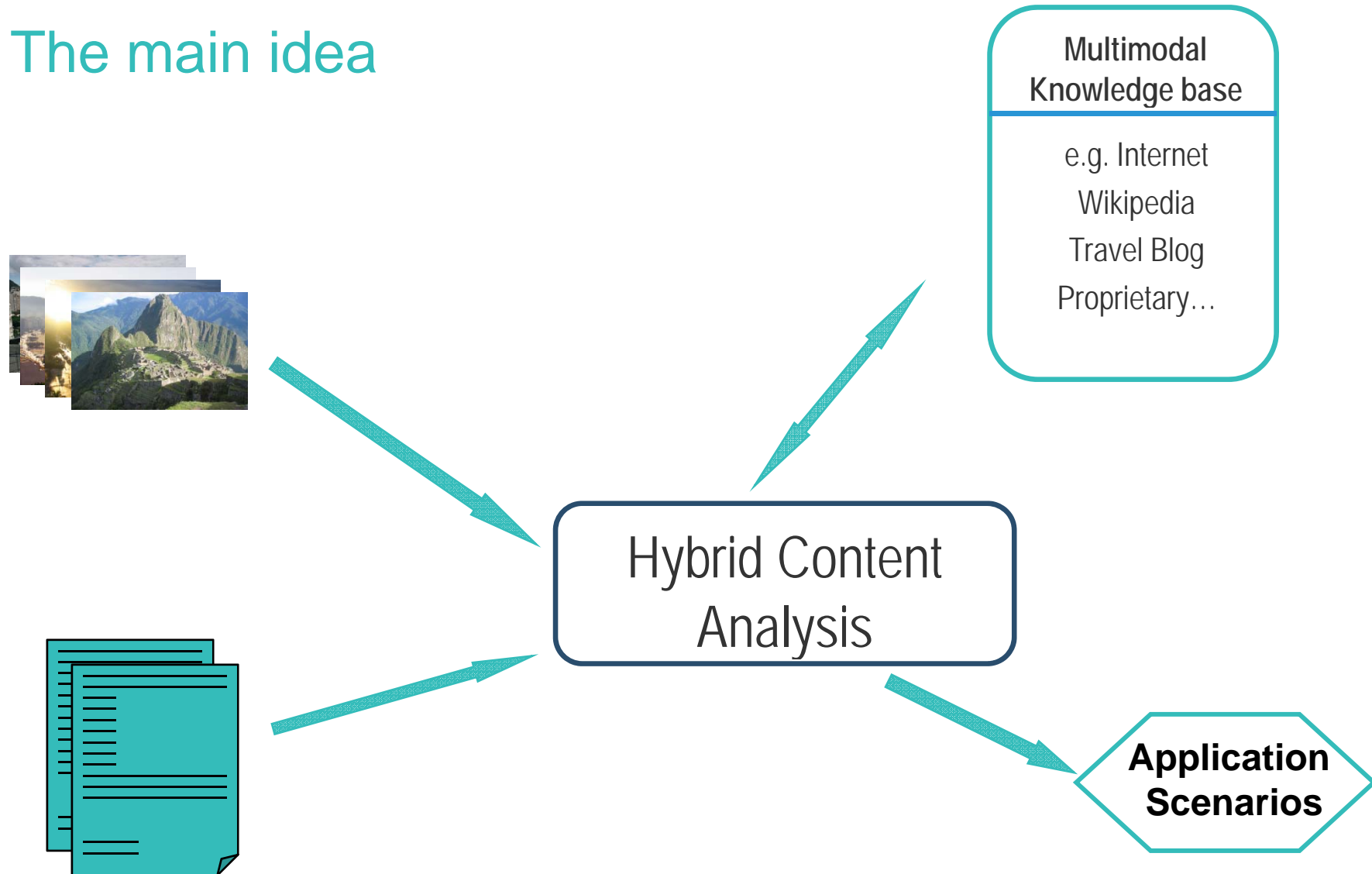
Crossing textual and visual content in different application scenarios

Marco Bressan, Stephane Clinchant, Gabriela Csurka, Yves Hoppenot and Jean-Michel Renders

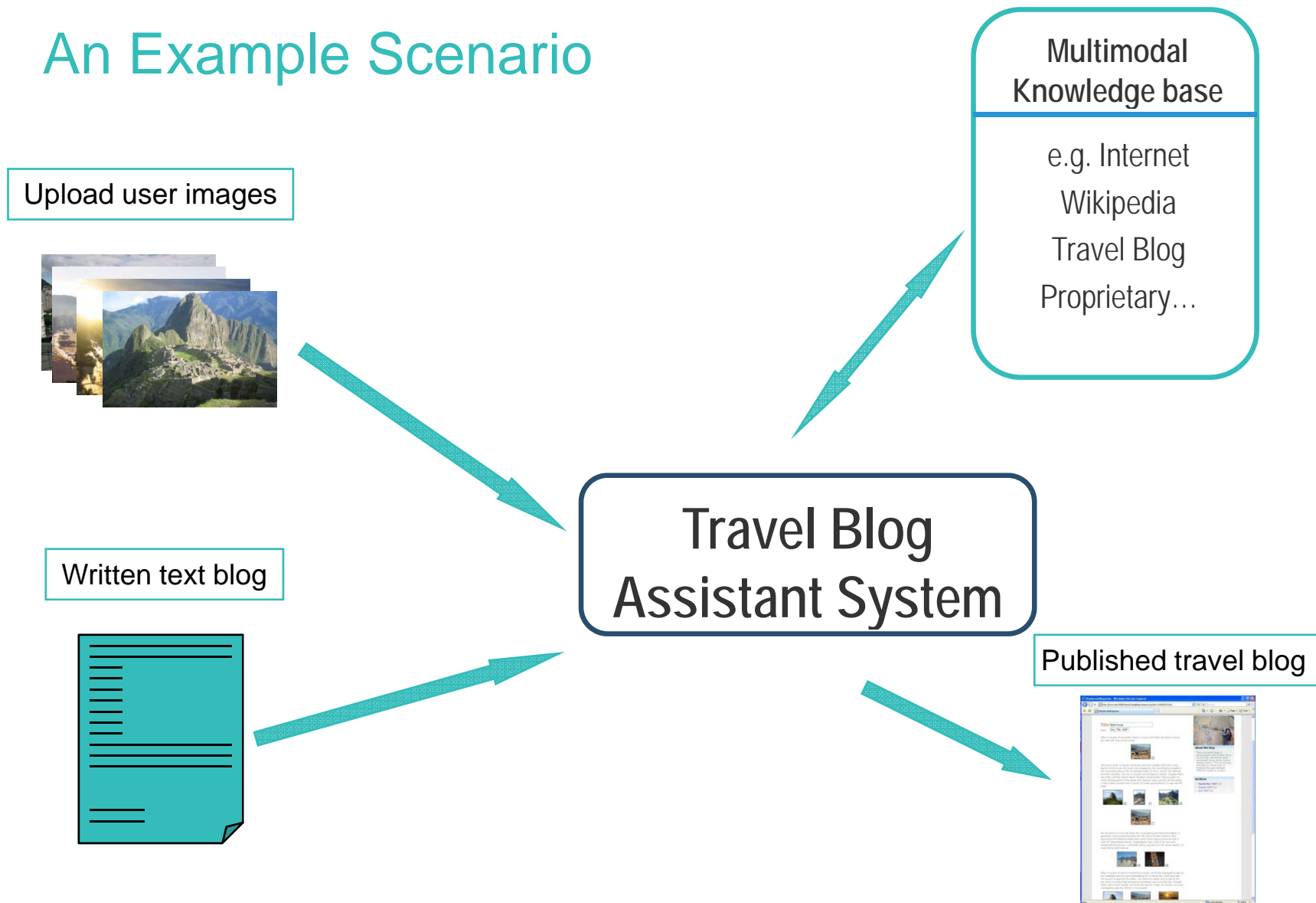
Xerox Research Center Europe
6 chemin de Maupertuis
38240 Meylan, France



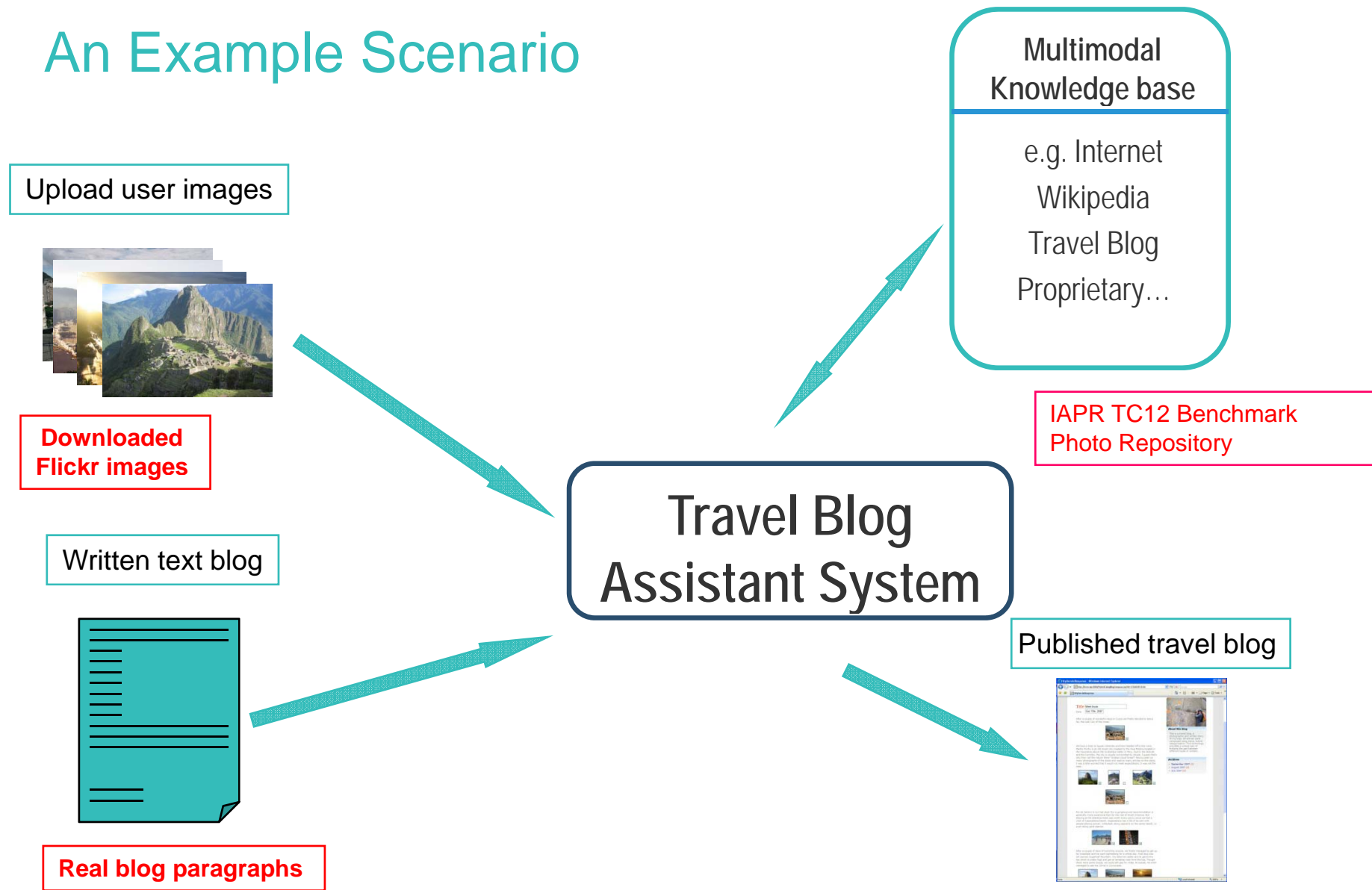
The main idea



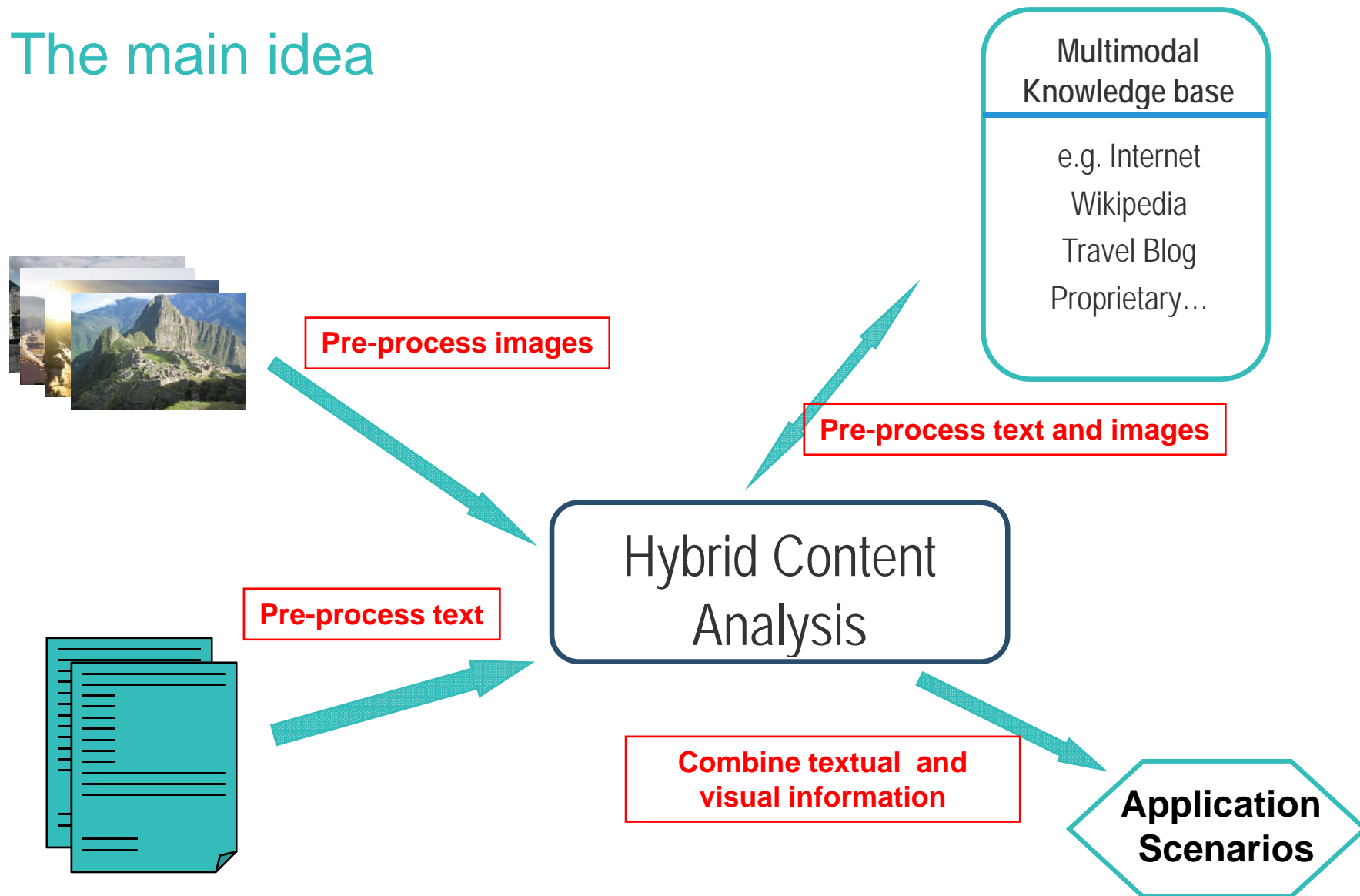
An Example Scenario



An Example Scenario



The main idea



Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

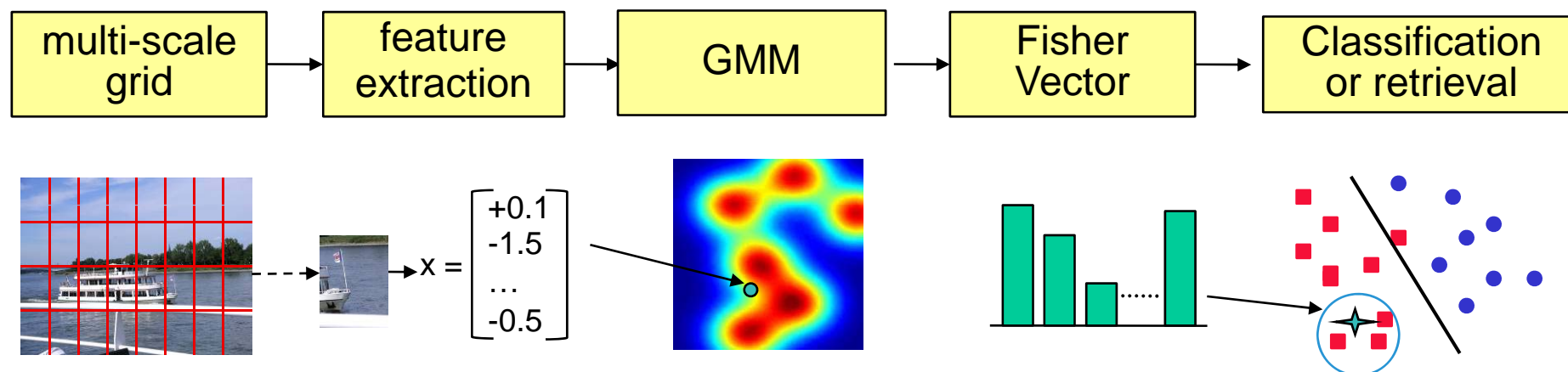
Image Similarity

- The goal is to define an image similarity measure that is able to “best” reflect a “semantic” similarity of the images.

– E.g.

$$\text{sim}\left(\begin{array}{c} \text{Pyramid} \\ \text{Pyramid} \end{array} \right) > \text{sim}\left(\begin{array}{c} \text{Sunset} \\ \text{Pyramid} \end{array} \right)$$

- Our proposed solution (detailed in next slides) is:



Low-level features

- They are extracted on regular grids at different scales
- We used two types of features:
 - Color features (local RGB statistics)
 - Texture features (local histograms of gradient orientations)
- They are handled independently and fused at late stages

Visual Vocabulary with a GMM

- Modeling the visual vocabulary in the feature space with a GMM:

$$p(x_t|\lambda) = \sum_{i=1}^N w_i p_i(x_t|\lambda)$$
$$p_i(x|\lambda) = \frac{\exp \left\{ -\frac{1}{2} (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) \right\}}{(2\pi)^{D/2} |\Sigma_i|^{1/2}}$$

- Occupancy probability:

$$\gamma_t(i) = p(i|x_t, \lambda) = \frac{w_i p_i(x_t|\lambda)}{\sum_{j=1}^N w_j p_j(x_t|\lambda)}$$

- The parameters λ of the GMM are estimated by EM algorithm maximizing the log-likelihood on the training data*:

$$\log p(X|\lambda^u)$$

* *Adapted Vocabularies for Generic Visual Categorization*, F. Perronnin, C. Dance, G. Csurka and M. Bressan, ECCV 2006.

The Fisher Vector

- Given a generative model with parameters λ (GMM)
 - the gradient vector

$$\nabla_{\lambda} \log p(I | \lambda)$$

- normalized by the Fisher information matrix

$$F_{\lambda} = E \left[\nabla_{\lambda} \log p(I | \lambda) \cdot (\nabla_{\lambda} \log p(I | \lambda))^T \right]$$

- leads to a unique “model-dependent” representation of the image, called **Fisher Vector***

$$\mathbf{f} = F_{\lambda}^{-1/2} \nabla_{\lambda} \log p(I | \lambda)$$

* *Fisher Kernels on Visual Vocabularies for Image Categorization*, F. Perronnin and C. Dance, CVPR 2007.

Similarity between images

- As similarity between images we used the L1-norm between the normalized Fisher vectors :

$$sim_{IMG}(I, J) = sim(\mathbf{f}_I, \mathbf{f}_J) = norm_{\max} - \|\hat{\mathbf{f}}_I - \hat{\mathbf{f}}_J\| = norm_{\max} - \sum_k \|\hat{\mathbf{f}}_I^k - \hat{\mathbf{f}}_J^k\|$$

- Where $\hat{\mathbf{f}}$ is obtain from \mathbf{f} by normalizing it to L1-norm 1.

Note: for color images the Fisher vectors obtained for color and texture features are first concatenated to obtain \mathbf{f} .

* Fisher Kernels on Visual Vocabularies for Image Categorization, F. Perronnin and C. Dance, CVPR 2007.

Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Example of retrieved images in our TBAS



Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Image Metadata examples in TBAS using GVC



**Clouds and Sky,
Mountain**



Individual



Aerial



Aerial



**Clouds and Sky,
Mountain**



**Buildings,
Night, Urban**



**Automobile,
Train**



**Clouds and Sky,
Sunrise-sunset**



**Beach,
Clouds and Sky**



**Clouds and Sky,
Mountain, Ocean**

The Classifier was trained for 44 classes such as: Aerial, Baseball, Beach, Boat, Desert, House, Forest, Flower, Individuals, Motorcycle, Waterfall, etc

Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Text representation

- We used the Language Model (LM) obtained as follows:
 - Consider the frequency of words in d :

$$P_{ML}(w|\theta_d) = \#(w,d)/|d|$$

- The probabilities are smoothed by Jelinek-Mercer interpolation:

$$\theta_{dw} \equiv P(w|\theta_d) = \lambda P_{LM}(w|\theta_d) + (1-\lambda)P_{LM}(w|\theta_C)$$

- using the corpus language model:

$$P_{ML}(w|\theta_C) = \sum_d \#(w,d)/|C|$$

- The similarity between texts is given by the cross-entropy:

$$sim_{TXT}(q,d) = CE(\theta_q|\theta_d) = \sum_w P_{LM}(w|\theta_q) \log(P(w|\theta_d))$$

Outline

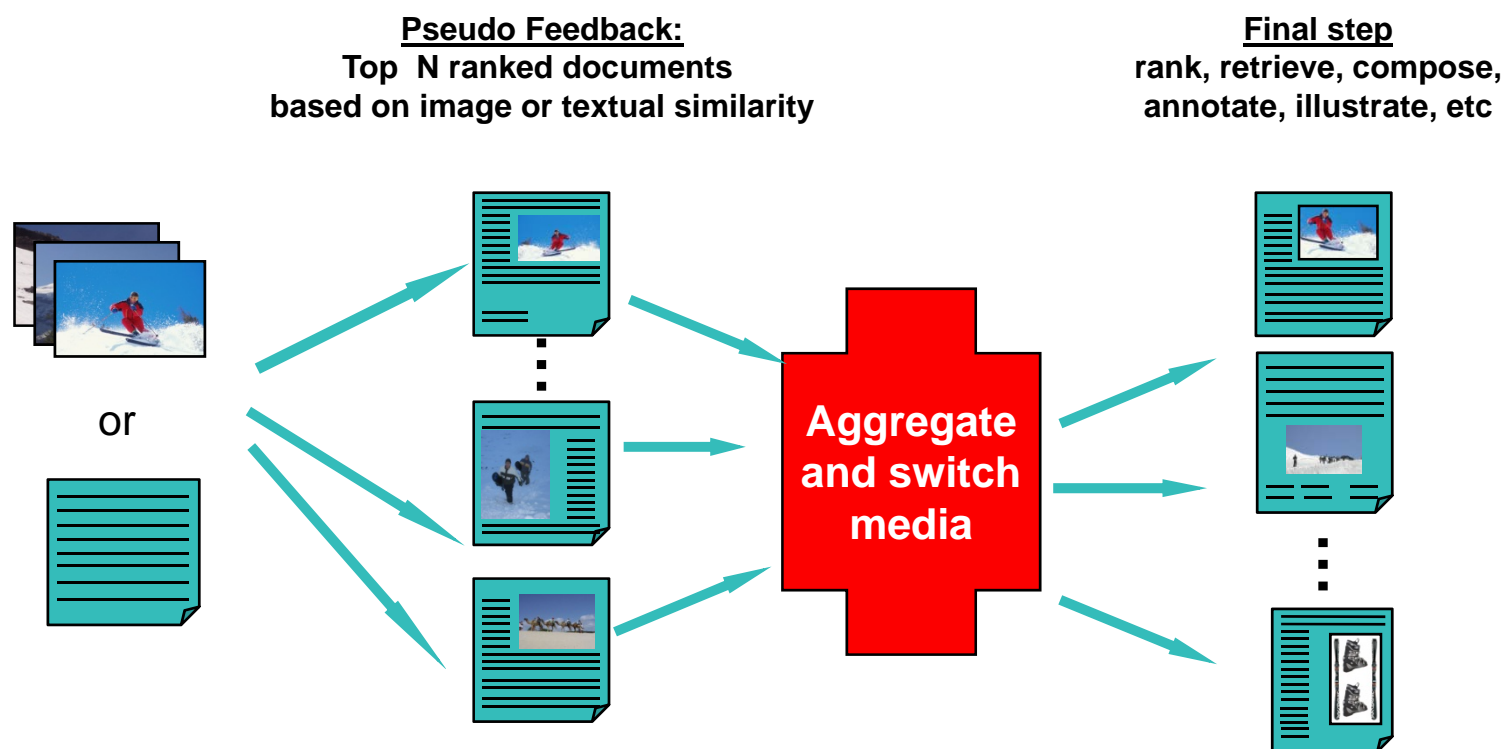
- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Fusion between image and text

- Early fusion:
 - Simple concatenation of image and text features (e.g. bag-of-words with bag-of-visual-words)
 - Estimating the co-occurrences or joint probabilities between textual and visual features (Mori et al, Vinokourov et al, Duygulu et al, Blei et al, Jeon et al, etc)
- Late fusion
 - Late score combination of mono-media results (Maillot et al, Clinchant et al)
- Intermediate level fusion
 - Relevance models (Jeon et al)
 - Trans-media (or intermedia) feedback (Maillot et al, Chang et al)

Intermediate level fusion

- The main idea is to switch media during using pseudo feedback process:
 - use one media type to gather relevant multimedia objects from a repository
 - use the dual type to step further (retrieve, annotate, etc)



Pseudo Feedback (PF)

- Let $\{d_k\}$, $k=1..M$ be the multi-modal documents in the repository
 - Denote by $T(d_k)$ and $I(d_k)$ the textual and visual part of d_k
- Using image I_q as query:
 - Retrieve the N most similar documents $(d_1, d_2 \dots d_N)$ from the repository based on *image similarity* between I_q and $I(d_k)$
 - Consider their textual part and aggregate them
 - $N_{\text{TXT}}(I_q) = \{T(d_1), T(d_2) \dots T(d_N)\}$
- Using text T_q as query:
 - Retrieve the N most similar documents $(d_1, d_2 \dots d_N)$ from the repository based on *text similarity* between T_q and $T(d_k)$
 - Consider their visual part and aggregate them
 - $N_{\text{IMG}}(T_q) = \{I(d_1), I(d_2) \dots I(d_N)\}$

Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Text illustration

- Given the set of images $N_{\text{IMG}}(T)$ obtained by PF from repository with PF for T we can use the
 - the most similar image(s) to illustrate T
 - cluster them (using Fisher Vectors) and choose the most representative image (e.g. closest to the cluster center)

After dumping our bags at our pousada (two blocks from the beach) and flinging on our swim suits, we headed down to the world's most famous beach... Copacabana. Along with its neighbour Ipanema, it's been immortalised in a song and is synonymous with glamour and beautiful bodies.

Blog text



Images from the repository

Image annotation

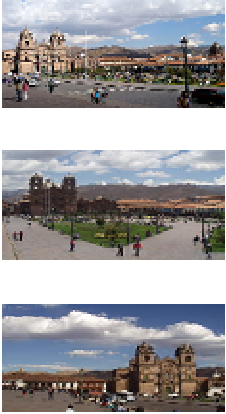

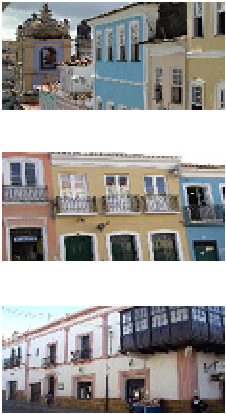

- Given the aggregated text $N_{\text{TXT}}(I)$ obtained by PF from repository with PF for I we can use the:
 - the most similar text as image title/caption
 - the most frequent words in the aggregated text $N_{\text{TXT}}(I)$ (*weighted by the idf*)
 - compute a Language Model* θ_F for $N_{\text{TXT}}(I)$ and use its peaks (relevant concepts) to annotate the image

$$P(\mathbf{F} | \theta_F) = \prod_{d \in N_{\text{TXT}}} \prod_w (\lambda p(w | \theta_F) + (1 - \lambda) P(w | \theta_R))^{\#(w,d)}$$

where $P(w|\theta_C)$ is word probability built upon the repository R

* Xrce's participation to ImageClefPhoto 2007, S. Clinchant, J.M. Renders and G. Csurka, CLEF 2007.

Examples of auto-annotation from the repository

 <p>Labels: Armas Plaza Lima</p>	 <p>Labels: Huayna Machu Picchu</p>
 <p>Labels: Pelourinho</p>	 <p>Labels: Catalina Monastery Santa</p>

Annotations obtained for test (flickr) images from the aggregated text (titles) of the 4 top ranked images retrieved by PF

Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Information Retrieval

1. Complementary Feedback*

- We can estimate the Language Model θ_F of the aggregated text $N_{\text{TXT}}(I_q)$ and
 - use the cross-entropy between θ_F and the LM θ_u of a documents u in retrieval
 - or first to interpolate θ_F with the LM of the query text (*if any*) before retrieval

$$\theta_{\text{new_query}} = \alpha \theta_q + (1 - \alpha) \theta_F$$

2. Trans-media document re-ranking*

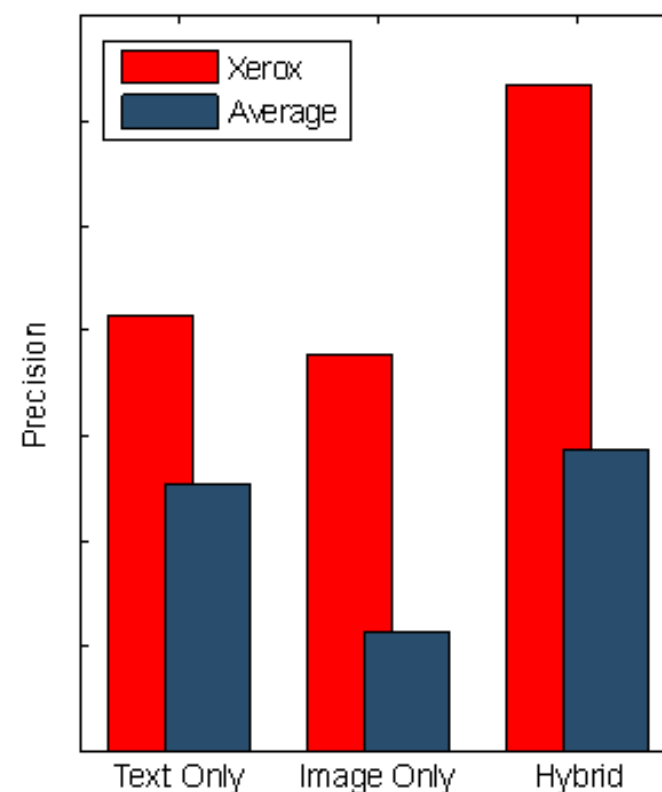
- We define the similarity between the aggregate of objects $N_{\text{TXT}}(I_q)$ and the textual part of a document u to re-rank the documents:

$$\text{sim}(N_{\text{TXT}}(I_q), u) = \sum_{T(d_k) \in N_{\text{TXT}}(I_q)} \text{sim}_{\text{TXT}}(T(d_k), T(u))$$

* Xrce's participation to ImageClefPhoto 2007, S. Clinchant, J.M. Renders and G. Csurka, CLEF 2007.

Retrieval Results of ImageClefPhoto

- All our systems performed significantly better than the average and we win the pure image and mixed text + image retrieval task
- In contrast to other systems:
 - both combining methods we proposed allowed for a significant improvement (about 50% relative) over mono-media (pure text or pure image) systems .



Outline

- Image Representation
 - Image Similarity
 - Image Retrieval
 - Image Classification
- Text Representation
 - Textual Similarity
- Crossing textual and visual content
 - Text illustration and image auto-annotation
 - Ranking and retrieval
 - Relate text with images through a repository
- Conclusion

Relating text and image through a repository

- Based on the PF, we can define the following similarity measures between an image I and a given text T (none of them being in the repository):
 - Using I as query in the PF:

$$sim(I, T) = sim(N_{\text{TXT}}(I), T) = \sum_{N_{\text{TXT}}(I)} sim_{\text{TXT}}(T(d_k), T)$$

- Using T as query in the PF:

$$sim(I, T) = sim(I, N_{\text{IMG}}(T)) = \sum_{N_{\text{IMG}}(T)} sim_{\text{IMG}}(I, I(d_k))$$

- Using both as queries and combine the results :

$$sim(I, T) = \alpha \cdot sim(N_{\text{TXT}}(I), T) + (1 - \alpha) \cdot sim(I, N_{\text{IMG}}(T))$$

Examples of text and images linked by the TBAS

Our plans to hit Copacabana beach the next day and check out hot Brazilian girls in skimpy bikinis were ruined by the weather. It rained all day! Can you believe that. I think we'll be heading to another place mid-week for some beach time.

There is a lot of tourists there from around ten until three, but it didn't feel as crowded as we'd feared. We started there for 12 hours- saw the sunrise and sunset, and walked the citadel twice. It is an awesome site in the proper sense of the word (Yanks take note). Bloody magic. Some archeologists reckon that Machu Picchu could have predated the Inca but that they did a lot of improvements.

Blog texts



Flickr images

Conclusion

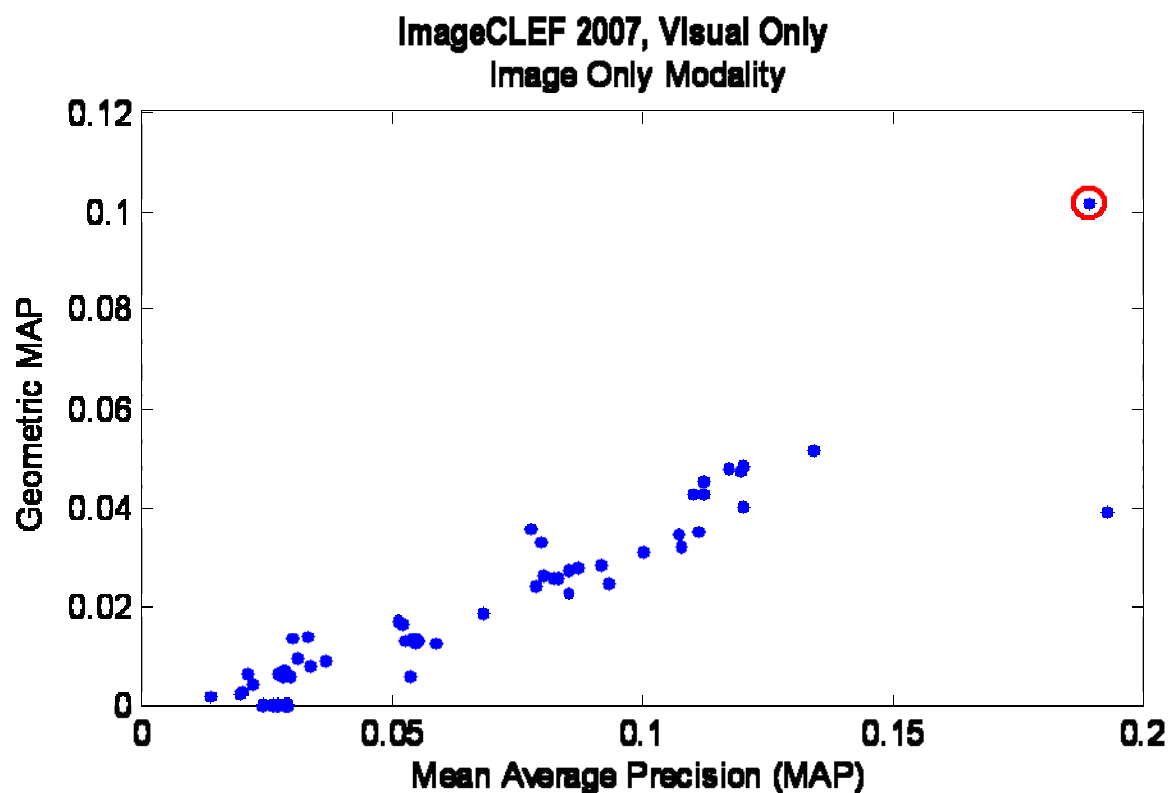
- We designed a system that:
 - uses rich and generic text and image representations and related metrics
 - Good retrieval and categorization performances obtained at different evaluation forums (Pascal, ImageClefPhoto)
 - handles very efficiently cross-modal relations
 - Combining text and images allowed for about 50% (relative) improvement over mono-media (pure text or pure image) results
- The technology developed has been shown to have potential in :
 - Multi-modal information retrieval
 - Enriching images with text (image annotation)
 - Enriching text with images (illustration)
 - Relating text and images based on a multi-modal knowledge base



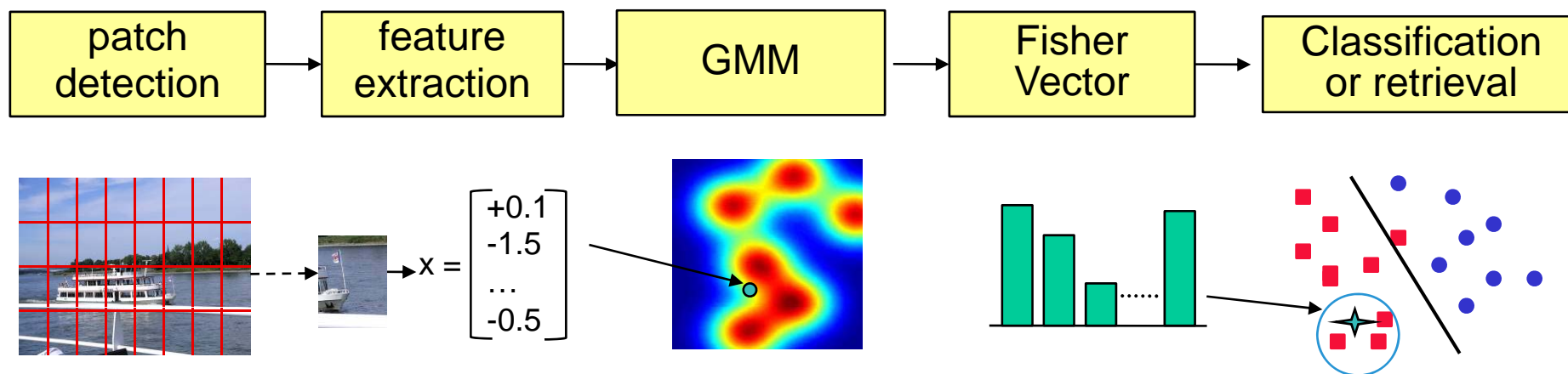
Back-up slides

Image Retrieval

- Our system was the best performing “Visual Only” system at the ImageClefPhoto 2007 Evaluation Forum



Generic Visual Categorizer (GVC)



Visual Categorization

- Our image categorizer (GVC) is composed by
 - one-against-all binary classifiers trained on labeled Fisher Vectors
 - one classifier is trained per feature type and the classification scores are combined (late fusion)
- Main advantages*
 - very efficient
 - low computational cost (fast)
 - universal

* Fisher Kernels on Visual Vocabularies for Image Categorization, F. Perronnin and C. Dance, CVPR 2007.

Categorization experiments with TBAS

- GVC can be used by the TBAS to add image metadata (class names) to the users uploaded images:
- To show it, we trained our GVC system on:
 - an independent in-house set of 38800 images
 - multi-labeled with 44 different labels such as:
 - Aerial, Beach, Baseball, Desert, House, Forest, Flower, Individuals, Motorcycle, Waterfall, etc
- Then:
 - the test images (flickr) were categorized by the GVC
 - all classes above a probability score (0.65) were automatically added to the image metadata

Performance of our GVC

- Third system, second institution in the VOC Pascal Challenge 2007
 - categorization of 20 object classes

