Travel Blog Assistant System (TBAS) - An Example Scenario of How to Enrich Text with Images and Images with Text using Online Multimedia Repositories

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

Xerox Research Centre Europe, 6, ch. de Maupertuis, 38240 Meylan, France

Abstract. In this paper we present a Travel Blog Assistant System that facilitates the travel blog writing by automatically selecting for each blog paragraph written by the user the most relevant images from an uploaded image set. In order to do this, the system first automatically adds metadata to the traveler's photos based both on a Generic Visual Categorizer (visual keywords) and by exploiting cross-content web repositories (textual keywords). For a given paragraph, the system ranks the uploaded images according to the similarity between the extracted metadata and the paragraph. The technology developed and presented here has potential beyond travel blogs, which served just as an illustrative example. Clearly, the same methodology can be used by professional users in the fields of multimedia document generation and automatic illustration and captioning.

1 Introduction

In only a few years, *the blogging* phenomenon on the web is becoming one of the most popular publishing media for a wide-range of audiences. Blogs allow users to interact and to reflect individual opinions, philosophies, experiences and emotions. As of September 2007, blog search engine [1] has tracked more than 106 million blogs.

Many of those blogs focus on particular genres such as politics, travel, fashion, projects, niche markets, legal topics, etc. In this paper, we are particularly interested in travel blogs. Often, these blogs can be understood as the digital version of the classical travel/road journals carried by a traveler for the purpose of documenting a journey. Due to the nature of online digital content, travel blogs also enable travelers to share any type of digital content, e.g. photos, and to update their readers on their location as the trip is taking place. Travel blogs are often composed and published while their authors are still traveling, using public Internet stations and laptops. This means that the traveler does not necessarily want to spend a long time editing it during his travel. One of the most time consuming parts of the *blog editing* is to select the right images from a non-organized set of recently uploaded images. Further, the traveler might want to add some specific information about the place or the monument he has just visited, the height of the mountain he had just climbed, etc. He could browse the web, but again, this would take time and he would either abandon or let this for later.

Consequently, in this paper, we propose a Travel Blog Assistant System (TBAS) that facilitates these tasks for the user (detailed in section 2). Its main purpose is to

Bressan M., Csurka G., Hoppenot Y. and Renders J. (2008).
Travel Blog Assistant System (TBAS) - An Example Scenario of How to Enrich Text with Images and Images with Text using Online Multimedia Repositories.
In *Metadata Mining for Image Understanding*, pages 90-104
DOI: 10.5220/0002338200900104
Copyright © SciTePress

automatically select the most relevant images from the uploaded set for each of the written paragraphs of the blog. Technically, this is done in two steps. First the system automatically adds metadata to the traveler's photos exploiting web repositories (shared and tagged photo repositories, other travel blogs, wikipedia, etc). Metadata have some textual form or, at least, features that make it comparable to textual data. Then, it uses these metadata to measure the similarity between the images and the given text.

In the first step, the system adds predefined keywords – related to a given set of visual categories (people, car, animal, ...) – to the metadata using a Generic Visual Categorizer (section 3.1). Secondly, based on the image features obtained by the categorizer, it uses a CBIR (Content Based Image Retrieval) technique to retrieve images similar (section 3.1) to the image to which it wants to add further metadata. Processing the aggregate of textual parts corresponding to the retrieved images allows the extraction of relevant concepts and topics (section 3.3). These emerging concepts and topics will be the "textual" keywords enriching the image metadata.

In the second step, we measure the similarity between the enriched metadata and a particular piece of text (section 3.2), which in our case is the written paragraph of the blog. According to this similarity measure, the images are ranked and the system is able to propose the most relevant images for a given paragraph, so that the traveler can choose what he thinks is the more appropriate for the final version. Furthermore, as it is the most costly part, the metadata extraction and indexation can be done offline; the computation of the image relevance scores with respect to the paragraphs (which has almost no cost with pre-prepared, indexed data) will be done on-line. So the system can propose the set of images for selection as soon as a paragraph is finished, which is very user friendly. Finally, another advantage of the system is obviously the metadata information added to each image. Indeed, the traveler can keep them and use them for further organization and retrieval in his private repository of pictures.

In order to illustrate the different steps and the performance of the TBAS system, we designed a prototype using a relatively small database (compared to the data we can get on the Web). We used as multi-media data repository the [2] as it contains travel images with additional textual information such as title, location and description. To obtain realistic traveler data, we downloaded a large set of images from the online photo sharing site Flickr [3]. For the travel text we collected blog paragraphs from two travel blog sites [4, 5]. In order to ensure the semantic correlation between images and blog texts, we used city names of two different travel destination (Peru and Brasil) as search tags to gather the images and blog texts.

The technology developed and presented here has potential beyond travel blogs. We used the Travel Blog Assistant System just as an illustrative example of a more general problem, which is complementing text with images, or vice-versa. Clearly, the same methodology can be used by professional users in the fields of multimedia document generation and automatic illustration and captioning, such as e.g. graphical designers, illustrators, journalists, pedagogical writers, etc.

2 The System Overview

The Travel Blog Assistant System (TBAS) is a plug-in system that can be integrated with any travelogue websites such as [6, 5, 7–9, 4, 10]. Its role is not to replace them, but to complement them and it can be integrated with other services these sites already propose, such as:

- publishing articles and guides focusing on travel related issues, comparing prices;
- providing the user with advices to plan a trip (trip planner);
- mapping functionalities that outline a trip and offer a graphic visualization of how the trip was undertaken [4, 9, 5].
- plotting automatically the trip steps across the globe, e.g using [11] which allows family and friends to see exactly where the traveler is [8].

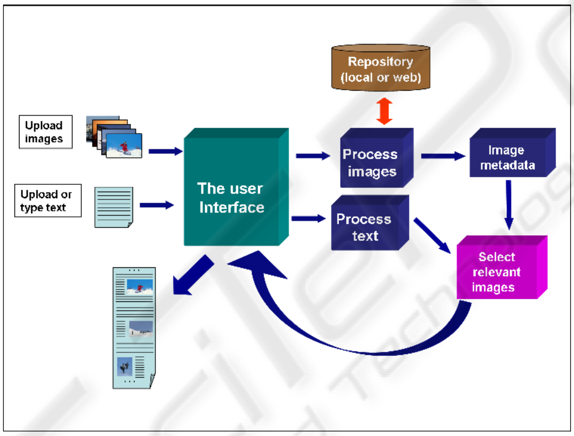


Fig. 1. The schema of a TBAS system.

Figure 1 shows the overall schema of the TBAS system. Its main steps are:

1. The user uploads a set of images to be considered. If he edits his blog during his travel he simply plugs his camera and uploads the images captured that day and the previous days. The images are then pre-processed and image metadata are added to each image. The metadata file can contain different type of information and is typically a structured xml file. First, information such as date, time, location (GPS) can be provided by the camera (exif file). Secondly, a pre-trained Visual Categorizer can provide annotations (short textual keywords) related to some generic visual aspects and objects of the image (see section 3.1). Finally, textual information (tags) are obtained using a Cross-content Information Retrieval System (CCIRS described in section 3.3) using a repository of multi-media objects (e.g. the database of the travelogue site itself).

SciTePress Science and Technology Publications

2. The typed or uploaded blog text is pre-processed at a paragraph level and “text” similarities (section 3.2) are computed between the paragraphs and image metadata. These similarities allow the system to rank the images according to a given paragraph. The top N ranked images (thumbnails) are shown to the user as illustrative examples of the paragraph. The user has naturally the possibility to select them, to reject them, to ask for the next N images or to re-initiate a new search based on this relevance feedback.
3. When all paragraphs are processed and the illustrative images selected, the system shows a preview of the composed “blog”. The page layout can be computed automatically or selected from a set of template layouts. When the user is satisfied with the result, he can simply publish the blog.

The components described above are the main components of the TBAS. However optional services/tools can be combined with them such as:

- *Photo filtering, by uploading the user’s planning/diary*. Having an electronic diary containing dates and main places (the planning of the trip) allows improving both the speed and the accuracy of the system. Indeed, combining this information with the dates of the image capture can limit the set of the photos to be ranked for a given paragraph. This can be useful for users who have the habit to take lots of photos at the same place.
- *Include links such as Wikipedia pages to images or words.* First, named entities are extracted and identified as being person names, events, geographic locations, organizations or monuments, etc. In our context, we are mainly focusing on geographic locations, organization or monument names, etc. In order to disambiguate the named entities we can use the context (group of named entities present in the text or metadata), ”image” similarities between images of the traveler and images of the page we want to link, ”textual” similarities between the paragraph (or metadata) and the page we want to link, or ”trans-media” similarities based on the web data repository .
- *Include further content*. This goes a step forward with respect to the first option (item 1), where we are not only seeking for a link, but we want to extract further information such as the height of a mountain, a short history of a monument, etc. from the relevant page. To do this, the candidate pages are decomposed in paragraphs and the similarities are computed for each paragraph. The selected paragraphs are proposed for further consideration. For example, if the user wants to include the height of Huayna Picchu, he types the text ”the height of the Huayna Picchu is” in the paragraph.
- *Include maps and draw trajectories*. Tools exist that allow the Automatic Assistant to include maps and drawings in the blog, assuming the GPS information is available (GeoMicro’s [12] is such a tool). If the camera has no GPS available, the system can still provide such information based on disambiguated named entities (see section 3.4).

3 Enabling Technologies

3.1 Image Similarity and GVC

As image signature (image representation), we use the Fisher kernel as proposed in [13]. This is an extension to the bag-of-visual-words (BOV) and the main idea is to characterize the image with the gradient vector derived from the generative probability model (visual vocabulary). This representation can then be subsequently fed to a discriminative classifier for categorization, or used to compute the similarities between images for retrieval.

The generative probability model in our case is the Gaussian mixture model (GMM) which approximates the distribution of the low-level features in images. It can be seen as a *Visual Vocabulary* where each Gaussian component $\mathcal{N}(\mu_i, \varSigma_i)$ models a visual word.

If we denote the set of parameters of the GMM by $\varPhi = \{w_i, \mu_i, \varSigma_i,\, i = 1...N\}$ (w_i, being the mixture's weight), we can compute the gradient vector of the likelihood that the image was generated by the model $\varPhi$:

$$\nabla_{\varPhi} \log p(I|\varPhi)\ . \tag{1}$$

This gradient of the log-likelihood describes the direction in which parameters should be modified to best fit the data (image features). One of its advantages is that it transforms a variable length sample (number of local patches in the image) into a fixed length representation (which we will call Fisher Vector) whose size is only dependent on the number of parameters in the model ($|\varPhi|$). Before feeding these vectors in a classifier or before computing similarities, each vector is first normalized using the Fisher Information matrix $F_{\varPhi}$, as suggested in [13] (see the paper for the computational details):

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

with

$$F_{\varPhi} = E_X \left[\nabla_{\varPhi} \log p(I|\varPhi) \nabla_{\varPhi} \log p(I|\varPhi)^T \right]\ .$$

This Fisher Vector is the basis of both our image categorizer and our CBIR system. In the first case, the classifier is learnt offline using pre-labeled training images. Due to the high dimensionality of the Fisher Vector, a set of linear one-against-all classifiers provides already good categorization results (see examples in [13]).

In the case of image retrieval, we define the similarity measure between two images as the the L1-norm of the difference between the normalized Fisher Vectors (each vector is itself re-normalized to have an L1-norm equal to 1):

$$\begin{aligned} &sim_{IMG}(I, J) = sim_{IMG}(\mathbf{f}_I, \mathbf{f}_J) \\ &= norm_{MAX} - ||\tilde{\mathbf{f}}_I - \tilde{\mathbf{f}}_J||_1 \\ &= norm_{MAX} - \textstyle\sum_i |\tilde{f}_I^i - \tilde{f}_J^i| \end{aligned} \tag{3}$$

where $\tilde{f}^i$ are the elements of the re-normalized Fisher Vector $\tilde{\mathbf{f}}$. The corresponding dissimilarity (distance) can be interpreted as the angle between the two Fisher Vectors.

We have to mention here that actually in both cases (categorization and retrieval), we build not one but two visual vocabularies: one for grey-level texture features (gradient histograms) and one for color features (local mean and variances in RGB). Both types of features are computed for image patches extracted on regular grids at 5 different scales. Hence, we obtain two Fisher Vectors for each image. In the case of classification the two vectors are fed into separate classifiers to estimate the probability that the image contains the object of the given class and a late fusion (simple mean) is used to estimate the final probability. In the case of retrieval, the two Fisher Vectors (color and texture) are simply concatenated before computing the similarity between two images.

The strength of such techniques was already demonstrated in image categorization and [13] and content based image retrieval [14].

3.2 Text Similarity

First the text is pre-processed including tokenization, lemmatization, word decompounding and standard stop-word removal. Then starting from a traditional bag-of-word representation (assuming independence between words), we adopt the language modeling approach to information retrieval. The core idea is to model a document d by a multinomial distribution over the words denoted by the parameter vector θ_d. A simple language model (LM) could be obtained by considering the frequency of words in d (corresponding to the Maximum Likelihood estimator):

$$P_{ML}(w|d) = \frac{\#(w,d)}{|d|} \ .$$

The probabilities could be further smoothed by the corpus language model:

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

using the Jelinek-Mercer interpolation :

$$\theta_{d,w} = \lambda \, P_{ML}(w|d) + (1-\lambda) \, P_{ML}(w|C) \ . \qquad (4)$$

Using this language model, we can define the similarity between two documents using the cross-entropy function:

$$sim_{TXT}(d_1, d_2) = \textstyle\sum_w P_{ML}(w|d_1) \log(\theta_{d_2,w})) \qquad (5)$$

Using this model we can rank the image metadata $q(I)$ according to the similarity between their language model $\theta_{q(I)}$ and the language model of a given paragraph in the blog.

3.3 Cross-Media Similarity

We want to use a cross-media retrieval approach in order to extract the image metadata (or enrich them if some are already present). In this context, we choose the so-called

”inter-media” fusion of image and text content through blind feedback approach. The main idea is to use the images (each image being a query) to rank multi-media documents (text+image) in a given repository. Denoting the textual and visual components of d by $T(d)$ and $V(d)$ respectively, the ranking of the documents is based on the visual similarity between the query image I and the image part $V(d)$ of d, using as features the corresponding Fisher Vectors (2).

If d_1, d_2 ,$\cdots$, d_N are the top N relevant documents (according to the mentioned ranking) corresponding to image I, we denote by $\mathcal{N}_{TXT}(I) = \{T(d_1), T(d_2), \cdots T(d_N)\}$ the set of their textual parts. Assuming that there is some underlying ”relevance concept F” in this set, we derive for it a corresponding language model θ_F as proposed in [14]:

$$P(\mathbf{F}|\theta_F) = \prod_{d_i \in \mathcal{N}_{TXT}(I)} \prod_{w} \psi(d_i, w) \tag{6}$$

with

$$\psi(d_i, w) = (\lambda P(w|\theta_F)) + (1-\lambda)P(w|\theta_{\mathcal{C}}))^{\#(w,d_i)}$$

Here $P(w|\theta_{\mathcal{C}}))$ is the word probability built upon the corpus (repository), λ (=0.5) is a fixed parameter, which can be understood as a noise parameter for the distribution of terms and $\#(w, d)$ is the number of occurrence of term w in document d. The model θ_F can be learnt by maximum likelihood with an Expectation Maximization algorithm

Finally, to choose the set of keywords to be added in the image metadata, it is sufficient to search for the components with the highest values in θ_F (or considering all values which are above a threshold).

The strength of such inter-media fusion techniques was already demonstrated in cross-content information retrieval, especially in the ImageCLEF Competition (see [14–16]).

3.4 Extraction of Named Geo-entities

First, named entities are identified and extracted in both the textual part of the documents of the multi-media reference repository and the user’s paragraphs. Here, we are mainly interested in named entities related to places. They can be names of natural points of interest, like rivers, beaches, mountains, monuments (map types) or businesses, like hotels, restaurants, hospitals (yellow page types). In this step all the other words are simply ignored and each document will be represented as the set of extracted named entities (locations). Then, we can compute the textual and cross-media similarities as described in section 3.3 and 3.2, but limiting the vocabulary (support of the distribution for the Language Models of each textual object) to this set of named entities. In particular, for extracting the metadata to be associated with an uploaded image, we can deduce a set of location names as the “peaks” of the “relevance language model” deduced from the top relevant documents with respect to this image.

Note that the extraction of named entities is also the key for linking paragraphs / image metadata to specific Knowkedge Bases. Still, this requires an additional step to disambiguate the named entities (in general, named entities could be associate to multiple entries of the Knowledge Base if their are only matched using the surface form). This disambiguation step will exploit the context information given by the whole paragraph (on the source side), by the whole text associated to the Knowledge Base entry (on the target size), or by the associated images. The knowledge bases could be external geographic knowledge databases such as [17], [18], [19] and/or [20]. Some of these databases ([19] for instance) give further information about the latitude and longitude of the place. This information can further be used to add indication on a map (e.g. using tools such as [12]) or adding a pointer to the geographic location on [11], etc.

4 Experiments

In order to illustrate the different steps and the performance of the TBAS system, we designed a prototype and tested it on a relatively small database (compared to the data we can get on the web). Actually, our multi-media data repository was the [2] database which contains travel images with a corresponding semi-structured caption consisting of a title, a free-text description of the semantic and visual contents of the image (e.g. *"a man is playing a pan pipe and another one a flute in front of a market sand with clothes; a woman on the left doesn't seem to enjoy the music"*) and a location (city or region and country). In our experiments we restricted ourselves to the title and location fields.

To simulate the traveler's image data, we downloaded a large set of images from the online photo sharing site [3]. For the travel blog text we collected real blog paragraphs from two travel blog sites [4] and [5]. In all cases, in order to ensure the semantic correlation between images and blog texts, we focused on two main destinations, Peru and Brasil, and used a few touristic names (cities, etc.) associated with them as keywords or tags for focusing our search in gathering the images and the blog paragraphs.

Firstly, we applied a pre-trained Visual Categorizer (section 3.1). The categorizer was trained on 44 classes and Figure 4 shows some of the example images with the obtained labels. The strength of this Visual Categorizer was already demonstrated at Pascal Visual Object Classes Challenge 2007 [21] (see Figure 2) and on several other databases (see [13]).

Secondly, we used our image similarity measure (see section 3.1) to retrieve the top N ranked images in our repository [2]. The strength of this image retrieval system was clearly showed at the ImageCLEFphoto 2007 Evaluation Forum [14], where it got far better retrieval results than all the other competing systems (see Figure3, left part). In Figure 5 we show retrieval results in our particular case. The query is the Flickr image (top left) and the most similar four images retrieved in the IAPR repository.

We further extracted and aggregated the text corresponding to the top N images and kept the most frequent words to enrich the metadata (see textual labels in Figure 5). In this example we didn't used the relevant topic search with the Language Model described in section 3.3, because of the poorness of our textual data. However, in a real situation when the textual data is much richer (e.g travel blogs, wikipedia, etc), this would be a more reasonable option.

The label examples in Figure 5 show that in some cases, the obtained annotations are general keywords, while in others they are more specific (to the location, not to the repository site). Indeed, when in the database several images of the same place (e.g. Machu Picchu) or the same building (Cuzco's cathedral) were found, the system is able to get the specific concept. Otherwise, it will average the information in the images and obtain more general concepts (such as sunset, cathedral). The bigger the repository is, the higher the probability is to get specific labels; it is also clear that the performance of the system increases when domain-specific knowledge bases are used.

Furthermore, additional information such as traveler's diary and GPS data can be further used to pre-filter the multi-media documents in the repository. For example, we can first restrict our database to the documents which contain either Peru or Brasil as location information. However, since we did not use this option here, incorrect labels can also be obtained (see last line in Figure 5).

Finally, the last step was to rank the images according to their similarities with respect to each blog paragraphs. To compute the textual similarity, we have combined the visual indexes (more precisely, the classes of our generic categorizer, used as a set of textual words) and the extracted keywords.

However, to associate images with paragraphs, it is not at all mandatory to explicitly extract textual metada. Indeed, as explained in the section 3.3, we can use a cross-content similarity measure to directly rank the uploaded images with respect to a given paragraph T_p. One effective way of computing this trans-media similarity measure is (see [14]):

$$sim_{IMGTXT}(I_i, T_p) = \sum_{d \in \mathcal{N}_{TXT}(I_i)} sim_{TXT}(T(d_i), T_p) \tag{7}$$

where T_p is the blog paragraph, sim_{TXT} is defined by (5) and $\mathcal{N}_{TXT}(I_i) = \{T(d_1), T(d_2), \cdots T(d_N)\}$ is the the set of texts corresponding to the N retrieved images in the repository. Note that this definition of trans-media similarity is an alternative to the one explained in section 3.3, that offers the advantage to be simpler to implement and faster, while providing equivalent performance.

This was shown at the ImageCLEFphoto 2007 Evaluation Forum [14], where our best Transmedia Reranking system given by (7) leads to slightly lower performance ($MAP = 0.302$) than the best variant of the method presented in 3.3 ($MAP = 0.3168$) and both were better than other hybrid competing systems at the Evaluation Forum (see also [15] and Figure 3 right).

In Figure 6 we show how this system was used in our TBAS example. The figure shows blog paragraph examples and the 4 top ranked Flickr images, where the similarity between an unlabeled Flickr image and a given blog was obtained using the similarity given by (7).

5 Discussion

In this paper we have shown an example scenario, the travel blogs assistant system, as an illustrative example of a more general problem, which is complementing text with images, or vice-versa.

The aim is to enrich an object (an image, a text) by auxiliary data of other types, thanks to a repository of multi-facet objects. An image could be enriched by text, named entities, GPS position (even if not given initially by the photo device), ratings. The idea is quite similar to unsupervised auto annotation e.g. [22–25] and more specially to the AnnoSearch system proposed by [26] which is an automatic annotation based on image search and the News image Annotation proposed by Jeon and Manmantha [27].

However, unlike our method, the system in [26] assumes that an accurate keyword of the image to annotate is available, the keyword being used to retrieve a set of semantically relevant images. [27] uses Normalized Continuous Relevance Models to compute joint probabilities between words and images, which is more an early fusion of the modalities. Our approach is an intermediate level of trans-media (or trans-type) enrichment which is based on a blind relevance feedback to the image query using a repository of multi-facet (multi-type, multi-media) objects similarly to [28, 29].

In our example, we enriched the text (paragraph) with related images using simple cross-media similarities between the text and a set of images through the medium of an external repository. However, the same methodology can be applied in a dual way too, where the repository is interrogated directly by the text (blog paragraph) and seeked for relevant texts or images without image query. In this case the text can be directly enriched by images from the repository (automatic illustration) or geographical information (height of a mountain, etc).

Acknowledgements

The authors want to thank Florent Perronin for his greatly appreciated help in applying some of the Generic Visual Categorizer (GVC) components and to thank Stephane Clinchant for his help in preprocessing the IAPR repository.

We would like also to acknowledge the following Flickr users whose photographs we reproduced here under Creative Common licences:

Tatiana Sapateiro	*/tatianasapateiro
Pedro Paulo Silva de Souza	*/pedrop
Leonardo Pallotta	*/groundzero
Laszlo Ilyes	*/laszlo-photo
Jorge Wagner	*/jorgewagner
UminDaGuma	*/umindaguma
Scott Robinson	*/clearlyambiguous
Gabriel Flores Romero	*/gabofr
Jenny Mealing	*/jennifrog
Roney	*/roney
David Katarina	*/davidkatarina
T. Chu	*/spyderball
Bill Wilcox	*/billwilcox
S2RD2	*/stuardo
Fred Hsu	*/fhsu
Abel Pardo López	*/sancho_panza

Cat	*/clspeace
Thowra_uk	*/thowra
Elena Heredero	*/elenaheredero
Rick McCharles	*/rickmccharles
Marília Almeida	*/68306118@N00
Gustavo Madico	*/desdegus
Douglas Fernandes	*/thejourney1972
James Preston	*/jamespreston
Rodrigo Della Fávera	*/rodrigofavera
Dinesh Rao	*/dinrao
Marina Campos Vinhal	*/marinacvinhal
Jorge Gobbi	*/morrissey
Steve Taylor	*/*st*

With * meaning http://www.flickr.com/photos and *st*=theboywiththethorninhisside.

Finally would like also to acknowledge the users who wrote the blog pharagraphs were used and reproduced here. These texts can be found at the folloing adresses:

*/cuzco-journals-j1879736.html
*/machu_picchu-journals-j5181463.html
*/rio-journals-j4669810.html
**/sarah_s_america/south_america/1140114720/tpod.html
**/rachel_john/roundtheworld/1146006300/tpod.html
**/eatdessertfirst/world_tour_05/1160411340/tpod.html
**/idarich/rtw_2005/1140476400/tpod.html
**/twittg/rtw/1132765860/tpod.html
**/emanddave/worldtrip2006/1155492420/tpod.html

With * meaning http://realtravel.com and ** http://www.travelpod.com/travel-blog-entries.

References

1. Technorati: http://technorati.com/pop/blogs/ (2007)
2. IAPR-TC12-Benchmark: http://eureka.vu.edu.au/ grubinger/iapr/tc12_benchmark.html (2006)
3. Flickr: http://www.flickr.com (2007)
4. Realtravel: http://realtravel.com/ (2007)
5. Travelpod: http://www.travelpod.com/ (2007)
6. Travelblog: http://www.travelblog.org/ (2007)
7. Trippert: http://trippert.com/ (2007)
8. Footstops: http://footstops.com/ (2007)
9. Travbuddy: http://www.travbuddy.com/ (2007)
10. Travellerspoint: http://www.travellerspoint.com (2007)
11. Google-Earth: http://earth.google.com (2007)
12. AltaMap: http://www.geomicro.com/gis/ (2007)
13. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
14. Clinchant, S., Renders, J.M., Csurka, G.: Xrce's participation to imageclefphoto 2007. In: Working Notes of the 2007 CLEF Workshop. (2007)

SciTePress
Science and Technology Publications

15. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Working Notes of the 2006 CLEF Workshop. (2006) http://www.clef-campaign.org/2006/working_notes/.
16. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEFphoto 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop. (2007) http://www.clef-campaign.org/2007/working_notes/.
17. World-Gazetteer: http://www.world-gazetteer.com (2007)
18. GNS: http://eart-hinfo.nga.mil/gns/html (2007) GEOnet Names Server.
19. Getty: http://www.getty.edu/research/conducting_research/ vocabularies/tgn (2007) Thesaurus of Geographical Names.
20. Wikipedia: http://www.wikipedia.org (2007)
21. Everingham, M., Zisserman, A., Williams, C., Gool, L.V.: The pascal visual object classes challenges (2007) http://www.pascal-network.org/challenges/VOC/.
22. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research **3** (2003)
23. Monay, F., Gatica-Perez, D.: Plsa-based image auto-annotation: Constraining the latent space. In: ACM MM. (2004)
24. Pan, J., Yang, H., Faloutsos, C., Duygulu, P.: Gcap: Graph-based automatic image captioning. In: CVPR Workshop on Multimedia Data and Document Engineering. (2004)
25. Feng, S., Lavrenko, V., Manmatha, R.: Multiple bernoulli relevance models for image and video annotation. In: CVPR. (2004)
26. Wang, X., Zhang, L., F. Jing, W.Y.M.: Annosearch: Image auto-annotation by search. In: CVPR. (2006)
27. Jeon, J., Manmatha, R.: Automatic image annotation of news images with large vocabularies and low quality training data. In: ACM MM. (2004)
28. Maillot, N., Chevallet, J.P., Valea, V., Lim, J.H.: Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In: CLEF 2006 Working Notes. (2006)
29. Chang, Y.C., Chen, H.H.: Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. In: CLEF 2006 Working Notes. (2006)

Figures

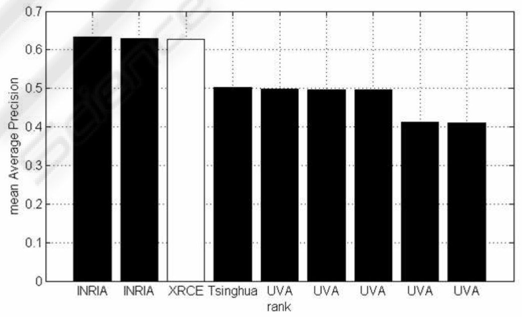


Fig. 2. Results of the Pascal Visual Object Classes Challenge 2007.

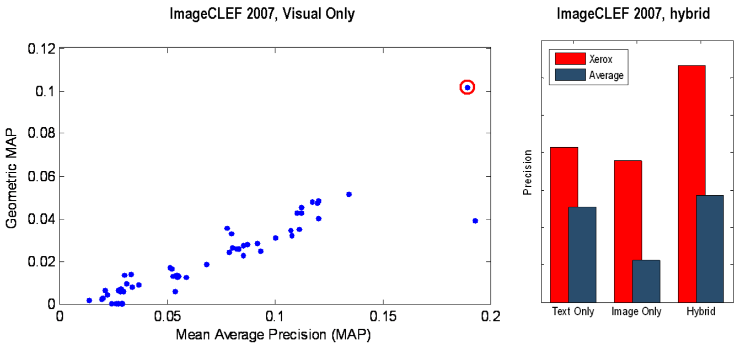


Fig. 3. Results of our system at ImageCLEFphoto 2007 Evaluation Forum.

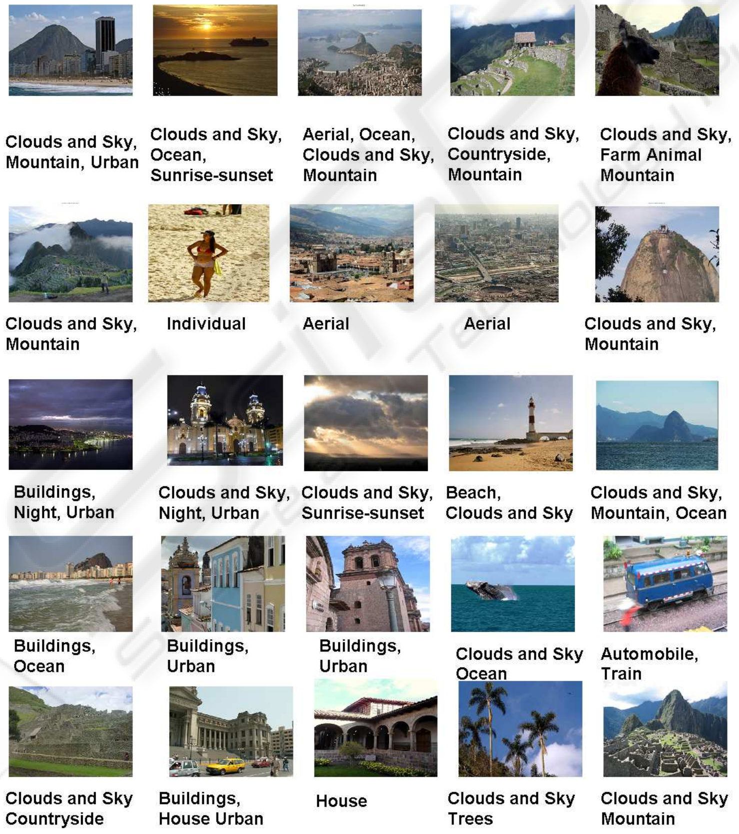

Clouds and Sky, Mountain, Urban
Clouds and Sky, Ocean, Sunrise-sunset
Aerial, Ocean, Clouds and Sky, Mountain
Clouds and Sky, Countryside, Mountain
Clouds and Sky, Farm Animal Mountain
Clouds and Sky, Mountain
Individual
Aerial
Aerial
Clouds and Sky, Mountain
Buildings, Night, Urban
Clouds and Sky, Night, Urban
Clouds and Sky, Sunrise-sunset
Beach, Clouds and Sky
Clouds and Sky, Mountain, Ocean
Buildings, Ocean
Buildings, Urban
Buildings, Urban
Clouds and Sky Ocean
Automobile, Train
Clouds and Sky Countryside
Buildings, House Urban
House
Clouds and Sky Trees
Clouds and Sky Mountain

Fig. 4. Example images with visual labels provided by the visual categorizer.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

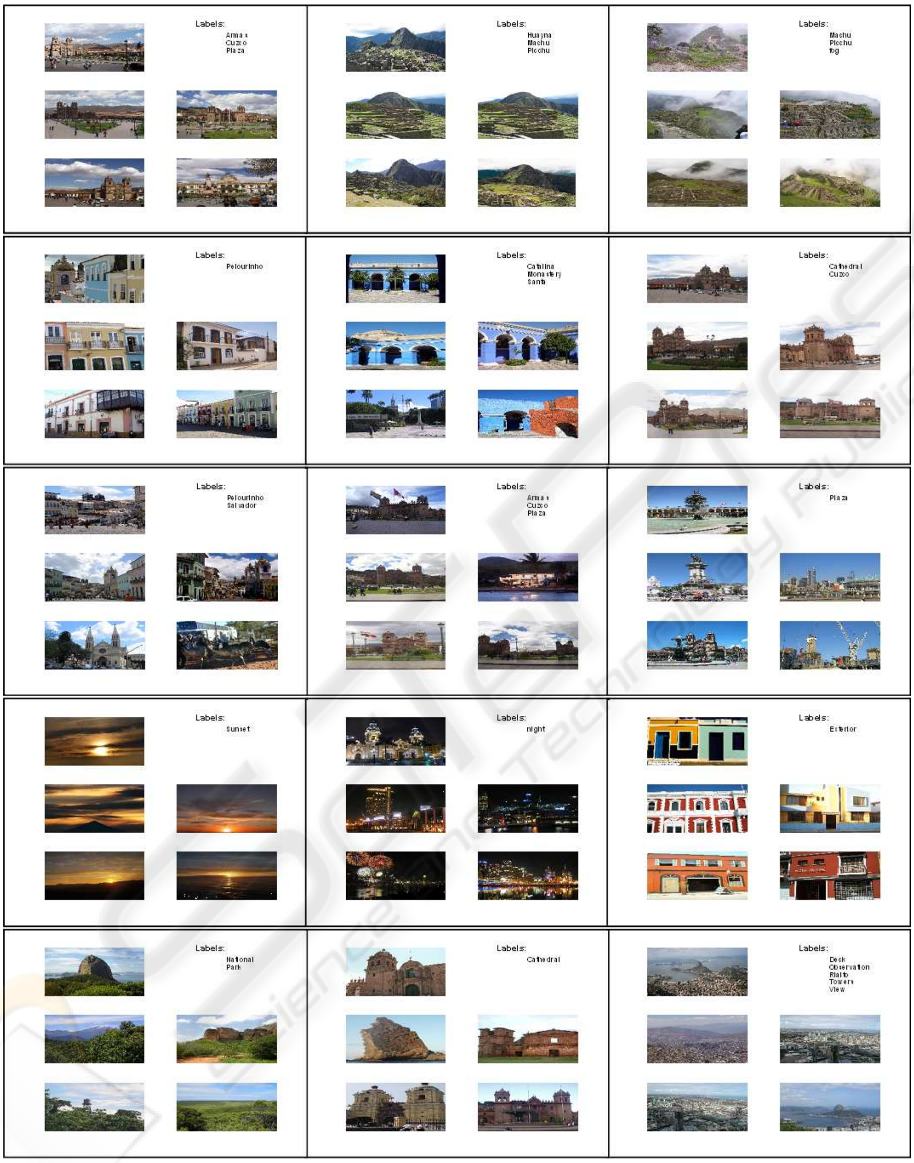
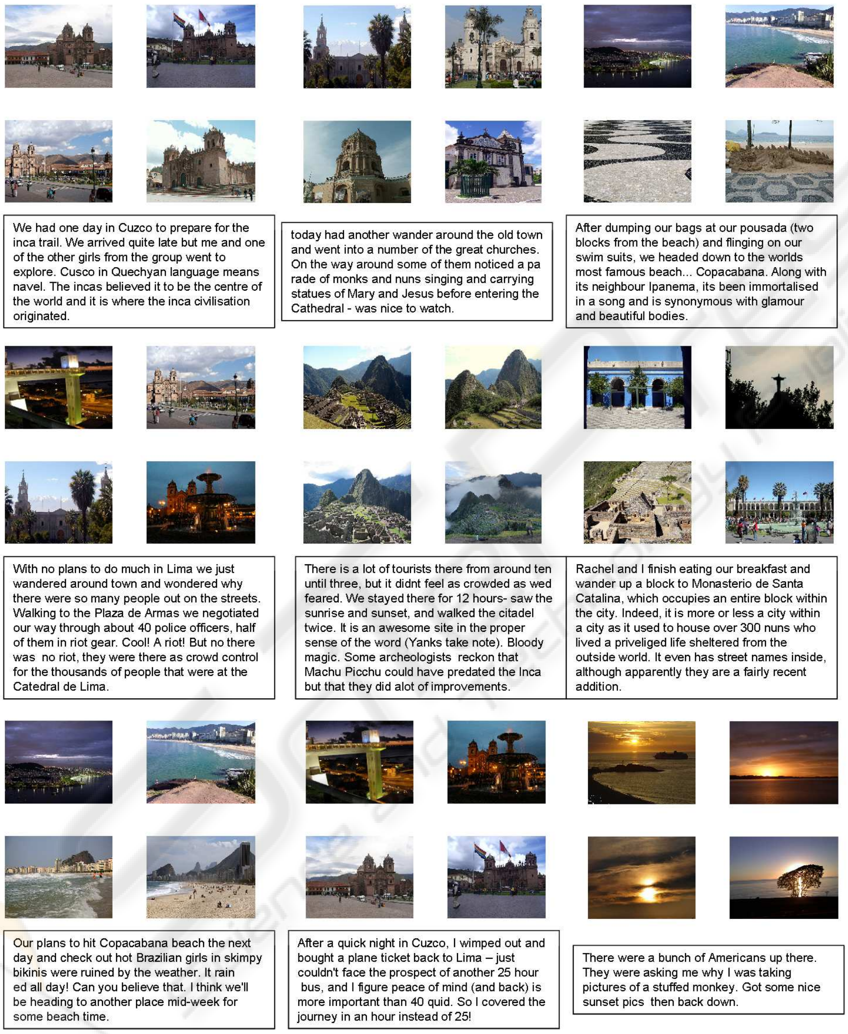


Fig. 5. Metadata (keywords) extracted for a query image and the TOP 4 neighbour images.

SCITEPRESS Science and Technology Publications



We had one day in Cuzco to prepare for the inca trail. We arrived quite late but me and one of the other girls from the group went to explore. Cusco in Quechyan language means navel. The incas believed it to be the centre of the world and it is where the inca civilisation originated.

today had another wander around the old town and went into a number of the great churches. On the way around some of them noticed a pa rade of monks and nuns singing and carrying statues of Mary and Jesus before entering the Cathedral - was nice to watch.

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

With no plans to do much in Lima we just wandered around town and wondered why there were so many people out on the streets. Walking to the Plaza de Armas we negotiated our way through about 40 police officers, half of them in riot gear. Cool! A riot! But no there was no riot, they were there as crowd control for the thousands of people that were at the Catedral de Lima.

There is a lot of tourists there from around ten until three, but it didnt feel as crowded as wed feared. We stayed 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 alot of improvements.

Rachel and I finish eating our breakfast and wander up a block to Monasterio de Santa Catalina, which occupies an entire block within the city. Indeed, it is more or less a city within a city as it used to house over 300 nuns who lived a priveliged life sheltered from the outside world. It even has street names inside, although apparently they are a fairly recent addition.

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

After a quick night in Cuzco, I wimped out and bought a plane ticket back to Lima – just couldn't face the prospect of another 25 hour bus, and I figure peace of mind (and back) is more important than 40 quid. So I covered the journey in an hour instead of 25!

There were a bunch of Americans up there. They were asking me why I was taking pictures of a stuffed monkey. Got some nice sunset pics then back down.

Fig. 6. Top 4 images to be associated with a blog's paragraph.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS