

Multilingual and Content Based Access to Flickr Images

Adrian Popescu

Laboratoire d'Intégration des Systèmes et des Technologies
Commissariat à l'Énergie Atomique
Fontenay aux Roses, France
adrian.popescu@cea.fr

Ioannis Kanellos

Computer Science Department
Telecom-Bretagne
Brest, France
ioannis.kanellos@telecom-bretagne.eu

Abstract: This paper outlines the *MLFLICKR* system, which is a multilingual query platform over *FLICKR*. In section I, we introduce the argument that deals with multilingual, conceptual driven image research. In section II, we present some related work; in Section III, we detail the problem targeted in this work, to focus, in Section IV, on a description of the image search architecture we developed. Before concluding, Section V presents some experiments and some results aiming at validating the approach we suggest.

Keywords: Multilingual image retrieval, automatic query translation, multilingual interface, *FLICKR*, *CBIR*

I. INTRODUCTION

Image queries constitute a hefty chunk of the total volume of information demands on the Web. There exist of course a considerable number of dedicated applications dealing with particular aspects of such queries but they generally suffer from a series of shortcomings, most of times inherent to semantic limitations. Nowadays, the market reformulates a recurrent demand of accessing pictures techniques better suited to users' profiles, needs and objectives. Although the total number of pictures on the Web widely exceeds several billions, the query space is still unequally covered and becomes often unbearable for languages other than English. The annoying consequence of such a state of affairs is that a large number of queries have no or very few answers.

This is not the sole problem: the ambiguity of the demanded concepts as query entries remains another hard and still open question in image retrieval. A significant number of initiatives in research and development concerns at present the inclusion of an automatic translation procedure in the search architecture in order to cope with the mentioned difficulties. On the other hand, and in spite of an important effort dedicated to the development of low-level content based image retrieval (*CBIR*) techniques, picture search mainly relies on indexing techniques founded on textual data. The idea we would like to promote is that content based techniques introduced as a complement of keyword search can be of a great help in large scale applications. In this paper, we present *MLFLICKR*, a new service for accessing *FLICKR* pictures, combining automatic query translation and content based techniques for accessing images.

The third important question is the linguistic isolation: queries are generally monolingual and fail systematically to integrate indexing material already at disposal in other languages.

Such a R&D space is indeed determined by three general and recurrent problems: i) image data massiveness ii) linguistic ambiguity and iii) multilingual indexing. It seems yet wide for setting up targeted services improving the already existing solutions by convenient architectures. It may also join, as we shall show, *CBIR* techniques in order to refine the overall accuracy and recovering performances.

II. RELATED WORK

Founded on research of the Turing Centre of the University of Washington, *PanImages*¹ introduces a multilingual image retrieval framework based on the alignment of translation dictionaries for over 100 languages. But the system proposed can process only mono-term queries and if several translations between the two languages are found, the user has to choose the one she/he prefers. The system is especially useful for enriching image sets associated to languages other than English. It is worth to notice that there is no image processing included in the *PanImages* architecture.

The automatic text translation received a lot of attention both in the research and in the industrial world. Existing systems, like *Google Language Tools* [2] or *Systran* [9] focus on the translation of full text. In this paper, we follow a different direction insofar as we concentrate our attention on the translation of series of keywords. This adaptation seems to us necessary in order to cope with the way people query Internet search engines.

[10] and [6] review a large number of research works in content based image retrieval (*CBIR*). In spite of the attention received by *CBIR*, these techniques are not currently integrated in Web search engines. The impossibility to translate the human notion of image similarity (highly subjective, context sensible, objective dependant and thoroughly centered on concepts) to a machine-based model (which is mainly perceptual) explains perhaps the epistemological barriers underlying such common situation. The confinement of the search space using prior textual information [8], [3] constitutes an intere-

¹ <http://www.panimages.org/>

sting alternative in order to conciliate the two types of similarity. A second problem with CBIR is the difficulty to scale up the systems to volumes of data such as the Internet image corpus. To our knowledge, the largest scale content based search application is Cortina [7], which works on 11 millions images. This state of affairs is to be compared to the 2 billions images in *FLICKR*.

III. PROBLEM STATEMENT

This paper looks precisely for an alternative solution to a recurrent problem, that is better formulated when split into two complementary questions:

(i) How can we reliably translate a keyword query in order to enrich the picture representing the request? (ii) How can the keyword-based search and the CBIR be aggregated in order to exploit the advantages offered by each approach?

In other words, the system has to cope with the following situations: first, given a keyword query (KQ), present a set of representative pictures (PS) as returned by the search engine; second, given an image query (IQ), select among the other images associated to KQ the closest neighbors from a visual point of view. Both situations give rise to problems that are far from being trivial. We set out to show that quite satisfactory results can be obtained if the approach is well tuned to the application domain.

Our proposition is called *MLFLICKR* (for *Multi-Lingual FLICKR*). It is outlined the following section.

IV. MLFLICKR

A. Architectural Overview

In *Figure 1*, we propose an overview of the architecture of *MLFLICKR*.

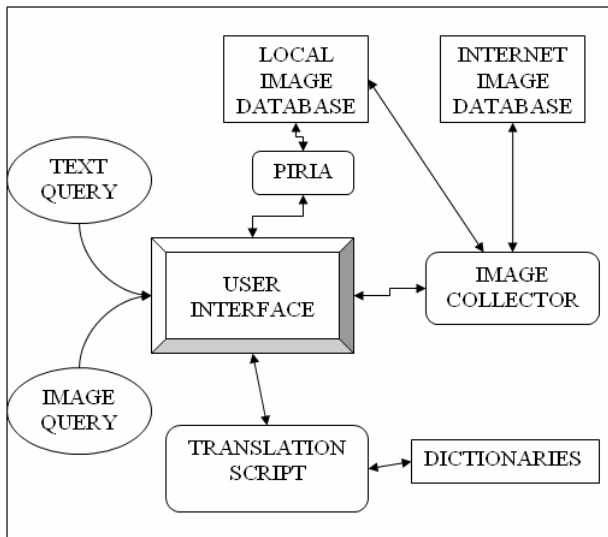


Figure 1: The architecture of *MLFLICKR*. The data sources are represented as rectangles; the active parts of the application as rounded rectangles; the user interface as a rounded square and the user interaction modalities as ellipses. The arrows indicate the interactions between different modules.

Hereafter, we give a brief description of the way the system functions. The user inputs a keyword query, which is sent to the translation script. This script employs two data sources (*WordReference* and *Wikipedia*) in order to propose an automatic translation of the query. Next, both the original and the translated version of the query are used by the image collector. This script checks firstly if the query is already uttered and a set of corresponding pictures is already locally stored. If so, these images are displayed; otherwise, i.e. if the system deals with a new query, the query and its translation are launched in *FLICKR*. When one of the displayed images is clicked, a CBIR process is launched in order to find visually close images among the items representing the same query.

The following subsections include a more detailed presentation of the system components. A more interactive illustration of the system functioning is available via a demonstration video².

B. *FLICKR*

FLICKR is seemingly the most successful photo sharing platform on the Web. Currently, the image corpus in this resource contains over 2 billion images with associated textual annotations. The annotation process is in no way constrained and the users can associate any terms to the images. Consequently, *FLICKR* pictures can be described by a rich set of words, in any language. The representation of different languages in *FLICKR* varies, with a preponderance of English. While the words attached to an image vary, insofar as different languages vary, their content is understood in a quite similar way, regardless the selected language by the user (*dog* (English), *chien* (French) and *cane* (Italian) describe similar visual entities).

C. The Local Image Database

We created a cache containing the image sets associated to the queries that were already formulated. If a query is not new, the image collection script does not go on the Internet to find images but rather recovers them from local, already stored images. The cache is not essential to the system; it is only an ergonomic expedient that simply allows the system to process queries faster.

D. Dictionaries

The query translation in our case is based on the use of *WordReference*³, a classical translation dictionary and that of *Wikipedia*. The first data source is fitted for translating common terms (like *dog* or *duck*), while the second is useful when translating proper names (for example *Parigi*, the Italian version of *Paris*). Currently the system works on three languages, English, French and Italian, with English as pivot language. This last choice is determined by the fact that English is by far better represented than other languages in *FLICKR* (and on the Web in general). The approach can be easily extended to other languages, provided that a translation dictionary between that language and English is available.

² <http://moromete.net/flickml.avi>

³ <http://www.wordreference.com/>

E. Translation Script

Textual queries are expressed as series of keywords. We found appropriate to propose a term-by-term translation procedure which includes a processing of composed terms. Current translation systems, like *Google Language Tools*, do not deal with composed terms like *Irish setter*. This term will be translated as *poseur irlandais* in French while it should be translated as *setter irlandais*. If a query is composed of n words, we first check if the term composed of the n words is a composed term and if so, we propose that translation as correct. If the n -words expression is not a composed term, we then check all combinations of $n-1$ words and, if no composed terms are found, the process is iterated until all 1-word terms are translated. Currently, the script translated queries up to 4 terms, which compose a large majority of Web queries [4].

A second important feature of our translation method is the translation of named entities, which is performed using *Wikipedia*. Classical dictionaries provide a reduced coverage of these types of terms while *Wikipedia* is richer and thus more adapted for the task. As the system has no prior information about the nature of terms, the translation is launched in both dictionaries.

Finally, if a term is polysemic or has different translations in *WordReference* and *Wikipedia*, we propose to retain that translation that appears more frequently on the Web. For example, *dog* is translated as *cane* when using *WordReference* and as *Canis familiaris* when using *Wikipedia*; but it is *cane* that will be retained as Italian translation of *dog*. Polysemous terms are also a recurrent problem; it appears regularly when the user employs secondary senses of terms. Such cases are not currently processed. We simply suppose that the first sense of a translated term appears more frequently than the others; it is thus preferred. Nevertheless, in the future, we plan to propose alternative translations in order to better treat term ambiguity.

F. Image Collector

This element is a Perl script using the *FLICKR* API in order to collect images for a given query. As we already mentioned it, the Image Collector exploits the results of the translation procedure and, if the user inputs a query such as *setter irlandais*, the *MLFLICKR* will equally search for images for *irish setter* allowing a drastic enrichment of the answers set (going, for this particular example, from 3 to over 2600 answers).

The translation procedure also enables the automatic reformulation of short queries (which are likely to be ambiguous) and propose the answers containing both the original term and its translation.

We illustrate the multilingual disambiguation with a query with *loup* (*wolf* in French). In Figure 1a and 1b, we present the best ranked results for this query when only querying *FLICKR* with *loup*, respectively with *loup+wolf*.

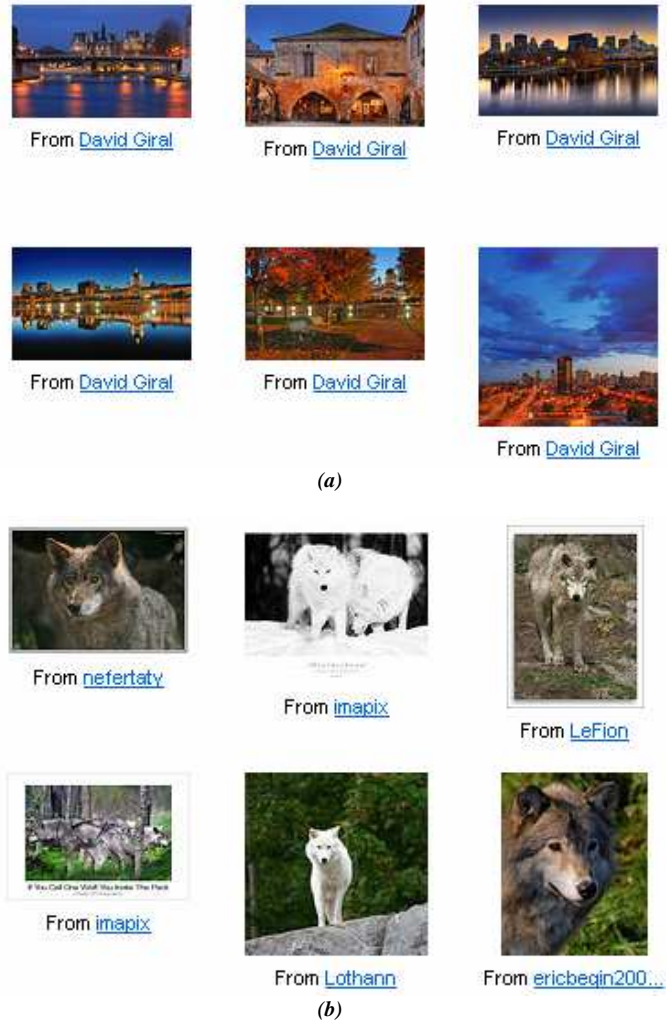


Figure 2a, 2b: *FLICKR* answers for a query with respectively *loup* (a) and *loup+wolf* (b).

The images in Figure 2b provide a better representation of *loup* (in its primary sense, i.e. that of *wolf*) compared to Figure 2a. This improvement is a direct consequence of the automatic multilingual disambiguation of the initial query.

G. PIRIA

For the content based counterpart of our exploration, we used *PIRIA* [5], a tool that performs low-level image indexing and retrieval. *PIRIA* proposes moreover several types of content indexing. In this paper, we chose to perform a low-level indexing based on a combination of the texture and the color features of the images.

V. EXPERIMENTS AND RESULTS

We performed three types of experiments in order to validate the image retrieval approach we implemented in *MLFLICKR*. Firstly, we assessed the enrichment of the answers set using an automatic translation. Secondly, we evaluated the quality of our translation procedure compared to *Google Language Tools*. Finally, we compared the results of the constrai-

ned CBIR to the performances of *Alipr*⁴, a visual search engine that also works on *FLICKR* photos.

A. Results Set Enrichment

We performed a comparative study of the queries having no answer for 200 synonymous queries of different complexity in English (En), French (Fr) and Italian (It). The following table gives a synthetic view of the obtained results:

Table 1: Comparison of *FLICKR* results for English, French and Italian

Number of answers/terms	1 term	2 terms	3 terms	4 terms
English	0/50	1/50	3/50	11/50
French	0/50	2/50	20/50	36/50
Italian	0/50	3/50	23/50	37/50

The results in Table 1 clearly show that English queries are more frequently answered; this is especially true for complex queries (including 3 or even 4 terms). When launching queries with 4 terms, there are only 11 requests with no answers in English; there are respectively 36 and 37 such queries for French and Italian. We already presented the example of *setter irlandais* (French) and *Irish setter* (English), a query pointing to the same concept which is depicted by 3 items in the first form and by over 2600 items in the second. The interest of the translation procedure is obvious for the enrichment of the answers sets.

B. Translation Procedure

The term-by-term translation method developed in this paper was confronted with the results provided by *Google Language Tools*. We carried out a comparative evaluation of the precision associated to the two translation procedures for a series of 200 queries containing one to four terms. The results are illustrated in Table 2. A translation is considered as correct if each element in the query is well translated.

Table 2: Precision of the translation procedures in *MLFLICKR* and *Google Language Tools*. (En, Fr and It stand for English, French and Italian.)

Precision/ Translation	En to Fr	Fr to En	En to It	It to En
<i>Google Language Tools</i>	71%	67%	59%	66,5%
<i>MLFLICKR</i>	92,5%	89%	92,5%	94%

The results in Table 2 show that our translation procedure clearly outperforms that in *Google Language Tools* for the given application. Typically, the precision is improved with over 20% in all translation cases. The average precision of the translation in *MLFLICKR* is around 90%, which can be

considered satisfactory for a completely automated method. The errors that appear usually correspond to polysemous terms.

C. Content Based Image Retrieval

CBIR systems are rather inefficient for accessing pictures in a large scale image corpus like *FLICKR*. Even for much smaller volumes of data (like, for instance, *Alipr*), the answers usually fail to be both conceptually and visually similar to the query image and to correspond to the users' expectancies. An experiment that compares our CBIR method and the one employed in *Alipr* on a number of 20 images corresponding to textual queries of different complexity (10 for 1 term queries and 10 for requests formed of 2 or 3 terms) is carried out.

Alipr was chosen because, similarly to *MLFLICKR*, the visually indexed image database is obtained from *FLICKR* and the type, quality and content of the pictures are quite comparable. The same query image is launched in *Alipr* and in *MLFLICKR* and a user is asked to count the number of similar answers.

Table 3: Performance of the CBIR function

Precision/ Terms per query	1 term	2 and 3 terms	Average
<i>Alipr</i>	18%	2%	10%
<i>MLFLICKR</i>	40%	31%	35.5%

The comparison in Table 3 demonstrates that the restriction of the query space using prior information introduced by the user improves significantly the precision of the content based search (which goes from 10% in *Alipr* to 35.5% in *MLFLICKR*). The difference is more important for complex queries (2% against 31%) than for those containing only one term (18% against 40%). Precision is not the only cue: an interesting measure that accounts for the quality of the CBIR process is the number of query images for which at least one similar answer is found. For *Alipr*, at least one similar result was judged similar in 7 cases out of 20. The corresponding ratio for *MLFLICKR* is 18 out of 20.

⁴ <http://alipr.com/>

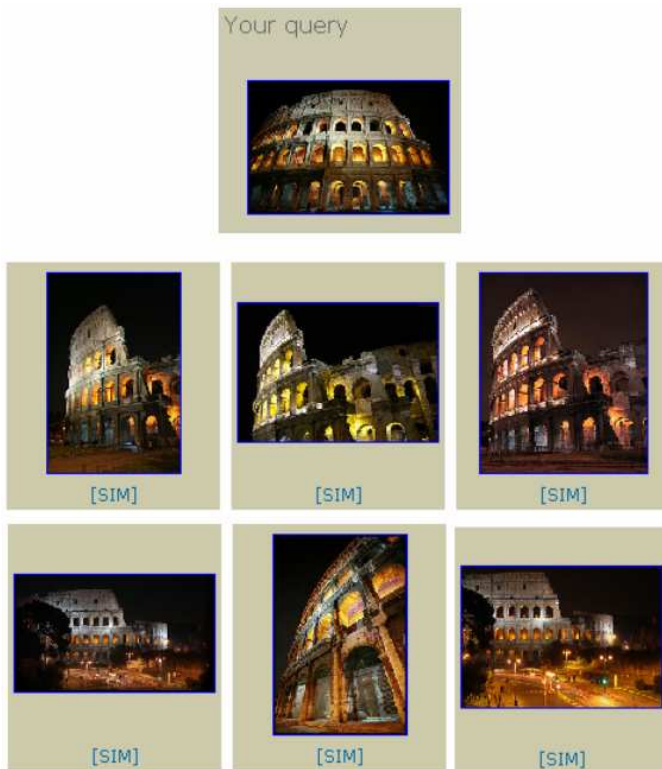


Figure 3: An example of CBIR query

In Figure 3, we present the results of a content-based search for an image of the *Coliseum*. The closest images are both conceptually and visually close to the query.

The comparison of the textually-constrained CBIR and that of a pure CBIR shows that, when systems implementing the two methods are evaluated by a user, the constrained version of CBIR is by far more efficient than the pure version of the process. The only limiting assumption implicated by the use of textual information is that if a user wants to see an image representing a given term, she/he would prefer, of course, to have answers conceptually coherent with her/his query. A typical use case is presented when the user wishes to have at disposal an image of, let us say, the Eiffel Tower and she/he also wishes to see afterwards visually resembling images to it coming up from the database. Clearly, the probability to consider as similar other images of the Eiffel Tower is greater than that of making the same assumption representing the *Umayyad Mosque* or, worse, a *tiger*.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced *MLFLICKR*, a service over *FLICKR* enabling an automatic reformulation of query in other languages in order to disambiguate them and mainly to enrich the answers set. In addition, we showed how to add a simple but efficient content based search function in *MLFLICKR* allowing an access to *FLICKR* photos that combines visual and conceptual features. Our system can process queries up to 4 terms long. It is easy to generalize the translation algorithm for

longer queries. For the moment, three languages are supported; but the translation procedure is not language dependent and it allows a straightforward integration of other languages.

The translation and the content based search were compared to competitive baseline systems. The results indicate that important improvements are obtained in both cases. The translation in *MLFLICKR* seems well adapted to the given task; the translation of a series of one or more keywords outperforms, in average, *Google Language Tools* by over 20%. As for the constrained CBIR, the results we obtained in this work furnish yet another proof that the human notion of similarity is primarily based on conceptual resemblance and only secondarily on perceptual features of the image, like its color or its texture.

In the future, we plan to improve the translation procedure in order to better process ambiguous terms. Currently, when translating, we only retain the main sense of a term; we would like to suggest other translation means to the users making easier for them to choose the sense of the term they desire and get relevant results to the chosen sense.

A second work direction we currently envisage concerns the integration of other languages in this framework. A complete use of *WordReference* would add two more languages, Spanish and Portuguese, to the already existing English, French and Italian (always with English as a pivot language). As proven by systems like *PanImages*, it is possible to integrate a large number of languages in multilingual image retrieval frameworks. We similarly are intended to considerably increase the number of languages processed in the service *MLFLICKR*.

REFERENCES

- [1] O. Etzioni, K. Reiter, S. Soderland, and M. Sammer, "Lexical translation with application to image search on the Web", in *Proceedings of the MT Summit XI*, 2007.
- [2] Google Language Tools - http://www.google.com/language_tools?hl=en
- [3] A. Popescu, C. Millet, and P.-A. Moëllic "Ontology Driven Content Based Image Retrieval", in *Proc. of ACM CIVR*, 2007.
- [4] J. Jansen, A. Goodrum and A. Spink, "Searching for multimedia: analysis of audio, video and image Web queries", in *World Wide Web Journal* 3(4), 2000.
- [5] M. Joint, P. A. Moëllic, P. Hède, and P. Adam, "PIRIA: A general tool for indexing, search and retrieval of multimedia content", in *Proc. of SPIE Image processing: algorithms and systems*, 2004, pp. 116-125.
- [6] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics", *Pattern Recognition* 40(1), 2007.
- [7] T. Quack, U. Monich, L. Thiele, and B. S. Manjunath, "Cortina: A System for Largescale, Content-based Web Image Retrieval", in *Proc. of ACM Multimedia*, 2004.
- [8] J. R. Smith and S. F. Chang, "Visually searching the Web for Content", *IEEE Multimedia*, 4(3), pp. 12-20.
- [9] Systran - <http://www.systran.fr/>
- [10] R. Veltkamp and M. Tanase, "Content based image retrieval systems: a survey", *Technical Report*, University of Utrecht, 2000.