MirBot: A multimodal interactive image retrieval system

Antonio Pertusa, Antonio-Javier Gallego, and Marisa Bernabeu

DLSI, University of Alicante
{pertusa,jgallego,mbernabeu}@dlsi.ua.es
http://www.dlsi.ua.es

Abstract. This study presents a multimodal interactive image retrieval system for smartphones (MirBot). The application is designed as a collaborative game where users can categorize photographs according to the WordNet hierarchy. After taking a picture, the region of interest of the target can be selected, and the image information is sent with a set of metadata to a server in order to classify the object. The user can validate the category proposed by the system to improve future queries. The result is a labeled database with a structure similar to ImageNet, but with contents selected by the users, fully marked with regions of interest, and with novel metadata that can be useful to constrain the search space in a future work. The MirBot app is freely available on the Apple app store.

Keywords: Image retrieval, multimodality, interactive labeling

1 Introduction

Content-based image retrieval systems have recently received a great deal of attention [1–3]. Searching for similar images in large datasets is a challenging task [4].

Large datasets [5, 6] are a critical resource for developing large scale image search algorithms. However, their dimensionality is a challenge for robust retrieval. In order to reduce the search space, a multimodal approach can be considered complementing the image information with metadata. Some methods have successfully combined image descriptors with textual information [7], and with features such as camera metadata [8].

One of the main contributions of this study is the inclusion of reverse geocoding information and novel metadata collected from smartphone sensors in order to constrain the search space in a multimodal scenario. These metadata, described in Sec. 2.2, can be used to reduce the number of images to be analyzed and to improve the search accuracy. For instance, if a user takes a photograph of an elephant, it is more likely that it will be in a zoo rather than on a beach, that the angle respect to the horizontal will be close to 90 degrees, and that the flash will be (hopefully) off.

Gathering labeled data is typically a tedious and extremely costly task. To overcome this problem, and similarly to [9, 10], the proposed application is designed as an entertainment game where users can validate the system response in

order to improve the results for future queries. Therefore, users feedback allows the database to grow with new labeled images continuously.

The result is a freely available research database organized according to the nouns of the WordNet [11] ontology, like ImageNet [5]. The main advantage of this categorization is its semantic structure, which prevents label ambiguity. WordNet *synsets* (synonym sets) are unique identifiers for meaningful semantic concepts. Each synset is linked to a definition, but it can be related to different words, e.g., *cellphone* and *mobile phone* share the same synset and definition.

The architecture of the system is presented in Fig. 1. After taking a photograph, the user can select the region of interest (ROI) of a target object. The image information within the ROI is sent to the server along with a set of metadata. Then, a ranking of similar images is calculated in the server, and the result is given to the user for its validation. The validated instance is finally added to the dataset for future queries.

Currently, the database is much more modest than ImageNet, which contains about 10 million images. However, the new images are stored with their associated metadata, within regions of interest, and they are theoretically gathered with minimum occlusions and plain backgrounds. This is a dynamic collaborative system that allows fast photo labeling and upload from smartphones, and which is continuously growing with the help of the users.

This study begins describing the features (Sec. 2) used to categorize the images, following by the classification stage (Sec. 3), and the user interaction interface (Sec. 4). Finally, the conclusions are addressed in Sec. 5.

2 Feature extraction

Local features and color histograms are extracted from the image within the ROI to classify the sample. This process is performed in the server side. Besides, a set of metadata is collected from the smartphone in order to allow the application of multimodal techniques in a future work, although only visual information is currently used for retrieval.

2.1 Image features

Local feature histograms. Most content-based image retrieval methods rely on local invariant descriptors [12–14]. The image descriptors obtained from a query image can be matched with the dataset prototypes using a nearest neighbour search technique. In order to improve the efficiency for large image datasets, the bag of features (BOF) image representation was introduced in [15]. This representation quantizes descriptors into visual words with the k-means algorithm. Then, an image can be represented by a frequency histogram of visual words obtained by assigning each descriptor to its closest visual word.

In the presented system, the TOP-SURF [16] toolkit is used to obtain an histogram of local descriptors for each image. This method calculates the Speed-Up Robust Features (SURF [13]) interest points, clustering them into a bag of features.

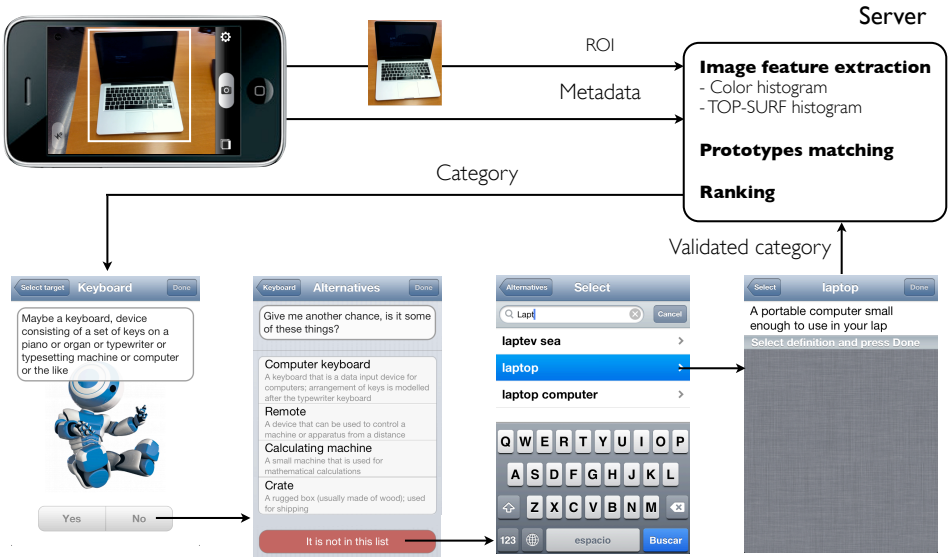

Fig. 1. Architecture of the presented method. This example corresponds to the longest interaction sequence.

The SURF algorithm uses a Hessian matrix-based measure for the detection of interest points and a distribution of Haar wavelet responses within the neighbourhood as descriptors [16]. These descriptors are clustered [17] in the TOP-SURF toolkit to yield a dictionary of visual words. The visual words that do not occur very frequently are emphasized using a tf-idf [18] weighting technique. Finally, the TOP-SURF image descriptor consists of a tf-idf histogram obtained by selecting the highest score visual words (the top visual words).

In the present method, a generic visual word dictionary with 100,000 words[1] is used as a basis to calculate the tf-idf histograms. Each of these histograms contains the top 100 visual words for that image.

Color histograms. SURF features do not consider color, which can be relevant to categorize certain objects [19]. In order to improve the classification results, SURF features are complemented with weighted color histograms in the presented method.

These histograms are computed in the YCbCr space. The histogram value of each color is weighted using a two-dimensional Gaussian function.

This weighting function allows to give less relevance to the colors that appear on the edges of the image and more weight to the middle, which is where the objects to recognize are located, as their ROI are already marked. The two-dimensional Gaussian function is defined as:

[1] Available at `http://press.liacs.nl/researchdownloads/topsurf/`

$$f(x,y) = Ae^{-\left(\frac{(x-x_o)^2}{2\sigma_x^2}+\frac{(y-y_o)^2}{2\sigma_y^2}\right)} \quad (1)$$

where A is the height of the curve's peak, x_o, y_o are the center position of the peak, and σ_x, σ_y define the width of the bell shape. In this study, the following values have been used: $A = 1$ to define the function between 0 and 1, x_o, y_o as the image center, and the width of the bell is the fifth of the width and the height of the image respectively.

2.2 Metadata

Device metadata. For each image, 29 metadata are obtained from the smartphone sensors. As shown in Tab. 1, these metadata correspond to the device information, to geolocation data, and to the sensor (accelerometer, gyroscope, and network) output values.

Metadata	Example value	Description
osversion	6.0	Operating system version
model	iPhone	Device model
timestamp_picture	2012-10-27 10:10:05	Timestamp
reliablelocation	YES	Location acquired with adequate precision
lat	38.387	Latitude
lng	-0.511615	Longitude
altitude	99.7184	Altitude
name	Universidad de Alicante	Main name of the feature
locality	San Vicente del Raspeig	Locality
sublocality		Sublocality
pc	03690	Postal Code
admin_area	Comunidad Valenciana	Name of administrative division of level 1
throroughfare		Name of administrative division of level 2
subthroroughfare		Name of administrative division of level 3
country	ES	Country code (ISO-639-1)
numaoi	2	Number of close areas of interest
closestaoi	Eps I	Name of the closest area of interest
horizontalerror	65	Approximate horizontal error in meters
verticalerror	10	Approximate altitude error in meters
wifi	YES	The data was sent using a WIFI network
yaw	0.192139	Rotation (yaw)
pitch	0.677016	Rotation (pitch)
roll	0.0703392	Rotation (roll)
angle	38.7902	Angle with respect to horizontal in degrees
accelerationx	-0.0283216	Acceleration (x)
accelerationy	0.0819144	Acceleration (y)
accelerationz	0.0515201	Acceleration (z)
globalacceleration	0.161756	Global acceleration
flash	OFF	Flash enabled

Table 1. Metadata obtained from the smartphone

Gisgraphy features. In addition, given a latitude and a longitude, reverse geocoding is also performed in the server with Gisgraphy[2], which uses the Geo-

[2] `http://www.gisgraphy.com/`

Metadata	Example value	Description
feature_id	6255088	Unique id to identify the feature
name	Universidad de Alicante	Name of the feature
adm1	60	Code for administrative division of level 1
adm2	A	Code for administrative division of level 2
amd3	03014	Code for administrative division of level 3
adm4		Code for administrative division of level 4
adm1_name	Comunitat Valenciana	Name of administrative division of level 1
adm2_name	Provincia de Alicante	Name of administrative division of level 2
adm3_name	Alicante	Name of administrative division of level 3
adm4_name		Name of administrative division of level 4
feature_class	S	The feature class
feature_code	UNIV	The feature code
country_code	ES	ISO 3166 country code
population	0	How many people lives in that place
elevation	110	Elevation in meters
gtopo30	91	Average elevation of 30'x30' area in meters
distance	282.141	Distance to the feature in meters

Table 2. Metadata obtained from reverse geocoding (Gisgraphy) in the server side.

Names geographical database. This allows to obtain valuable data such as the feature class and code[3] that provide information about the kind of place.

For each query image, the data of the closest feature (point of interest) is selected. The list of the 17 Gisgraphy features can be seen in Tab. 2. Some of these data such as `name` or `adm1` may seem redundant with respect to the geolocation data obtained from the smartphone, but their values differ. Here, the information is obtained from the Gisgraphy database, whereas the device features correspond to the Apple (for iOS) or Google Maps (for Android) geolocation data.

EXIF metadata The camera parameters of the photographs are also stored. The exchangeable image file format (EXIF) information [20] sent to the server includes 23 parameters such as the aperture value, brightness, ISO speed, white balance, etc.

3 Classification

Given a sample query, the category of the most likely image among the set of prototypes is given to the user for validation. The techniques used for classification are fast and incremental due to the real-time requirements of the proposed architecture.

The classification is performed searching for similar images from the dataset considering TOP-SURF and color histogram distances to the query image. Currently, the metadata are still not used in the classification, although they are stored.

[3] `http://gisgraphy.googlecode.com/svn-history/r14/trunk/gisgraphy/data/featureCodes.txt`

3.1 Image matching

To compare the TOP-SURF descriptors of two images a and b, the normalized cosine similarity d_t between their tf-idf histograms T and T' is calculated [16]:

$$d_t(a,b) = 1 - \frac{T \cdot T'}{|T||T'|} \tag{2}$$

Color histograms are compared using the Jensen-Shannon divergence (JSD) [21], defined as:

$$d_c(a,b) = \sum_{m=1}^{M} H_m log \frac{2H_m}{H_m + H'_m} + {H'}_m log \frac{2H'_m}{H'_m + H_m} \tag{3}$$

where H and H' are the color histograms of the images a and b respectively, and M is the histogram size.

3.2 Ranking

The TOP-SURF and color distances are combined to yield the final distance between two images:

$$d(a,b) = w \cdot d_t(a,b) + (1-w) \cdot d_c(a,b) \tag{4}$$

where w is a parameter to weight their contribution.

Given a query image a, its N nearest neighbour images from the set of prototypes are retrieved according to $d(a,b)$. Finally, the category of the first image in the ranking is given to the user for validation if its distance $d(a,b) < \theta$, where θ is a fixed threshold, or an "unknown object" message is yielded otherwise.

The system parameters are currently set to $M = 64$, $w = 0.8$, $N = 10$, and $\theta = 0.8$.

4 User interaction

This section describes the user interface in the MirBot app. As mentioned before, users can take a photograph with a smartphone and select the ROI with a finger. Then, the ROI image and its associated metadata are sent to the server.

A minimum number of SURF features $s >= 4$ in the image is required to perform the categorization, or else the user receives a message indicating that the selected region is empty. To enhance the users experience, the messages and the robot images that are shown are dependant on $d(a,b)$.

Besides the users' identity is completely anonymous, a random number is generated for each of them in order to track their uploaded contents. Before sending the data to the server, some parameters can be set by the users in order to constrain the search space. One of them allows a user to classify the target image considering only its previous images, or to classify it taking into account the images of all users (the whole dataset).

Other parameters of the application are the WordNet root categories (animal, object - artifact, food/drink, plant), that can be independently selected to reduce the search space.

If the user marks the server response as wrong, an alternative list with the categories of the first N images in the ranking (except by the first one, which was discarded by the user) is presented. If none of these categories correspond to the right one, a lemma (e.g., `key`) can be selected from the WordNet nouns, and then a definition related to this lemma (e.g., `Metal device shaped...`).

Besides setting the WordNet class, users can label their images (e.g. `my home keys`), and manage them (view, delete, etc.) in the app. They can also learn more about the lemmas with Wikipedia, and see the WordNet hierarchy for that class.

5 Conclusions and future work

A multimodal interactive system for smartphones image retrieval is presented in this study. The system is available for iOS devices[4] [22], and allows a user to interactively categorize images according to the WordNet hierarchy.

The research result is a hierarchical multimodal database organized according to the WordNet ontology. This dataset contains labeled objects within a ROI, with minimum occlusions, and typically with plain backgrounds, as the images are specific for this task and not downloaded from the Internet. The dataset is unbalanced, as the most common objects appear more frequently as long as it is user-driven.

A set of metadata that complements the image information allows the research community to apply different multimodal techniques. For instance, if metadata is used in a preprocessing stage, some image prototypes could be filtered out. A third component relying on metadata distances could also be added to $d(a, b)$. An alternative to this scheme is to perform content-based image retrieval first, and then to filter the results using the metadata. The proposed dataset can also be analyzed using hierarchical classification methods.

Although some statistics such as the number of images and the success rate can be seen in the app interface, currently the database lacks of enough data to perform a rigorous evaluation. This is planned for a future work, and also to include the ImageNet dataset when the option for searching between all users' images is enabled. In order to work in real time with a large amount of prototypes, an inverted index of the TOP-SURF descriptors would be required.

The database can be freely requested for research purposes at [22]. A web interface for researchers has been developed, where the images and metadata can be explored, reviewed and freely downloaded.

Acknowledgment. This study was supported by the Consolider Ingenio 2010 program (MIPRCV, CSD2007-00018), the PASCAL2 Network of Excellence IST-2007-216886, and the Spanish CICyT TIN2009-14205-C04-C1.

[4] The Android version is in progress

References

1. Lew, M.S, Sebe, N., Djeraba, C., Jain, R.: Content-based Multimedia Information Retrieval: State of the Art and Challenges, ACM Trans. on Multimedia Computing, Communications, and Applications, 2(1), pp. 1-19 (2006)
2. Smeulders, A., Worring, M. , Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(12), pp. 1349–1380 (2000)
3. Datta, R., Joshi, D., Li, J., Wang, J. Z.: Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys., 40(2), pp. 1–60 (2008)
4. Jegou, H. , Douze, M. , Schmid, C.: Recent Advances in Large Scale Image Search, LNCS, Emerging Trends in Visual Computing, vol. 5416, pp. 305–326 (2009)
5. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. IEEE CVPR, pp. 248–255 (2009)
6. Torralba, A., Fergus, R., Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Non-parametric Object and Scene Recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, 30(11), 1958–1970 (2008)
7. Dinakaran, B., Annapurna, J., Kumar, C.A.: Interactive image retrieval using text and image content, Cybernetics and Information Technologies, 10(3), pp. 20–30 (2010)
8. Boutell, M, Luo, J.: Beyond pixels: Exploiting camera metadata for photo classification, 38(6), pp. 935–946, Pattern Recognition (2005)
9. Barrington, L. L., Turnbull, D. D., Lanckriet, G. G.: Game-powered machine learning, Proc National Academy of Science (PNAS), 109(17), pp. 6411–6416 (2012)
10. Von Ahn, L., Dabbish, L.: Labeling images with a computer game, In CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems, NY, USA: ACM Press, pp. 319–326 (2004)
11. Fellbaum, C.: WordNet: An Electronic Lexical Database, MIT Press (1998)
12. Lowe, D.: Distinctive image features from scale-invariant keypoints, IJCV 60(2), pp. 91–110 (2004)
13. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF), Computer Vision and Image Understanding, 110(3) pp. 346–359 (2008)
14. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV 60(1), pp. 63–86 (2004)
15. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos, ICCV, pp. 1470–1477 (2003)
16. Thomee, B., Bakker, E.M., Lew, M.S.: TOP-SURF: a visual words toolkit. In: Proc. of the 18th ACM Int. Conf. on Multimedia, pp. 1473–1476, Firenze, Italy (2010)
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching, In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2007)
18. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill (1983)
19. Jeong, S.: Histogram-Based Color Image Retrieval, Stanford University (2001)
20. Exchangeable image file format for digital still cameras: Exif Version 2.3. CIPA, `http://www.cipa.jp/english/hyoujunka/kikaku/pdf/DC-008-2010_E.pdf`
21. Lin, J.: Divergence measures based on the Shannon entropy, IEEE Trans. On Information Theory, Vol. 37, No. 1, pp. 145-150 (1991)
22. MirBot, `http://www.mirbot.com`