An Interactive System for Content-Based Image Retrieval

José Manuel Torres[1,3], Luís Paulo Reis[2,4], David Hutchison [5]
[1]University Fernando Pessoa, Porto, Portugal
[2]Faculty of Engineering of the University of Porto, Porto, Portugal
[3]INESC Porto - Institute for Systems and Computer Engineering, Porto, Portugal
[4]Artificial Intelligence and Computer Science Lab., Porto, Portugal
[5]Lancaster University, Lancaster, United Kingdom

ABSTRACT: The increasing size of existing digital image collections means that manual annotation of images is becoming more and more an infeasible process. Content-based image retrieval (CBIR) systems attempt to solve this problem by automating the process of image indexing. Nevertheless, users want to search images at a conceptual level, and not only in terms of colour, texture or shape. Semantic modeling and the semantic gap are thus one of the biggest challenges in image retrieval. A key requirement for developing future image retrieval systems is to explore the synergy between humans and computers. Relevance Feedback (RF) and region-based representations are two effective ways to improve early CBIR systems. This paper presents Visual Object Information Retrieval (VOIR) an integrated CBIR system adopting a region-based image retrieval (RBIR) paradigm and using RF along with some results which emphasize the effectiveness of its long-term learning (inter query learning) algorithms.

KEYWORDS: Interactive Content-Based Image Retrieval, Region-Based Relevance Feedback, Inter Query Learning.

1 INTRODUCTION

An image retrieval system is a computer-based system for browsing, searching and retrieving images from large image repositories. Content-based image retrieval (CBIR) is the application of computer vision and image processing to the image retrieval problem. The search makes use of the contents of the images themselves, rather than relying only on human-inputted metadata such as captions or keywords. From a user perspective, the ideal CBIR system should also involve semantic. The user would be able to perform a request like "find me pictures of fishes". This type of query is very difficult for computers because there are all types of fishes of different sizes and shapes and other animals like dolphins that resemble a lot fishes.

The main objective of a CBIR system is the satisfaction of the user needs for some type of visual information. The design and conception of an image retrieval system should, consequently, follow the guidelines offered by the correct observation of what the users really want from the system. In practice, there are three fundamental aspects to be taken into account that make this task difficult:

- The diversity of applications for digital images.
- The diversity of image users with different perspectives, turning the problem of requirement definition extremely complex.
- The limitation within current state of the art of science and technology to mimic the human capacity of image understanding.

A key requirement for developing future image retrieval systems is to explore the synergy between humans and computers. Relevance feedback (RF) is a technique that engages the user and the retrieval system in a process of symbiosis. Following the formulation of the initial query, for subsequent iterations of query refinement, the system presents a set of results and the user evaluates the results in order to refine the set of images retrieved to his or her satisfaction. In image retrieval systems, this technique can be extremely useful to reduce the adverse effects of the three aspects mentioned above.

This paper analyses the use of relevance feedback in image retrieval applied to short-term learning (intra query) and long-term learning (inter query), presenting VOIR (Torres, 2005) a prototype image retrieval system. The experiments and the results presented in this paper are focused on the long-term learning of concepts associated with images.

The rest of the paper is organised as follows, section 2 reviews some of the most relevant related work. Section 3 introduces the VOIR framework and its two-layer model for describing visual items. Section 4 presents the methodology and the experimental results obtained in the long-term learning assessment. The final section gives some concluding remarks.

2 RELATED WORK

The increasing size of existing digital image collections means that manual annotation of images is becoming more and more an infeasible process. CBIR systems attempt to solve this problem by automating the process of image indexing. Nevertheless, users want to search images at a conceptual level, and not only in terms of colour, texture or shape. Semantic modeling and the semantic gap are thus one of the biggest challenges in image retrieval (Ritendra, Dhiraj, Jia, & James, 2008), (Jie & Qi, 2008), (Xiaofei, Deng, & Jiawei, 2008).

A key requirement for developing future image retrieval systems is to explore the synergy between humans and computers. Relevance Feedback (RF) and region-based representations are two effective ways to improve CBIR systems (Zhi-Hua, Ke-Jia, & Hong-Bin, 2006), (Giorgio, 2007), (Pratikakis, Vanhamel, Sahli, Gatos, & Perantonis, 2006), (Rahman, Bhattacharya, & Desai, 2007), (Ko & Byun, I, 2005), (Fei, Qionghai, Wenli, & Er, 2008).

Relevance feedback is a technique that engages the user and the retrieval system in a process of symbiosis. Following the formulation of the initial query, for subsequent iterations of query refinement, the system presents a set of results and the user evaluates the results in order to refine the set of images retrieved to his or her satisfaction (Jing, Li, Zhang, & Zhang, 2004). As pointed out by several authors (Carson, Belongie, Greenspan, & Malik, 2002), (Mezaris, Kompatsiaris, & Strintzis, 2004), (Wang, Rui, & Sun, 2004), (Zhang & Zhang, 2004), the adoption of a region-based representation in a concept-based image retrieval presents obvious advantages since, typically, each image normally contains several distinct visual concepts or objects. If, additionally, the system presents the possibility of result refinement through RF techniques, then, the relevance feedback at the region-level of granularity allows a much better interaction paradigm increasing the accuracy of the information flowed from the user to the system.

The task of visual information description is mainly concerned with transforming user needs into a suitable form to properly support searching procedures in visual collections. Moreover, the selected image indexing attributes should be sufficiently discriminatory to allow images to be retrieved in an effective and efficient way. Ideally, the descriptive information associated with the images, in an image retrieval system, should be closely related with the way that end users, i.e., humans, interpret images.

One of the facts deduced from several user studies in image retrieval is, as stated by Eakins (2002), that most image queries are performed at the logical level, identifying meaningful semantic objects in images such as, for instance, chairs or fruits. Although low-level features such as colour, texture or shape, sometimes are implicit in the user queries, rarely those features are used directly in the query formulation.

There are several disadvantages in using manual textual annotations to describe images, such as the human effort required to annotate large amounts of visual information, the subjectivity of the operation and the inconsistency in the textual term assignment. Nevertheless, some of these drawbacks can be significantly reduced if:

- The type of annotation is denotative (Barthes, 1977). This factual and expressional description, made at the visual object level, tends to be more objective and unambiguous.
- A controlled vocabulary is used to reduce the inconsistency in the term assignment, to establish a structure and to suggest preferred terms during the process of annotation.

3 VOIR

3.1 *Conceptual Image Retrieval Framework*

The VOIR framework aims to be used in conceptual image retrieval. It assumes that the target images of the user are fundamentally associated with concepts, such as, cars, chairs or airplanes. Each concept is represented by a textual term from a textual thesaurus, i.e., a hierarchic controlled vocabulary.

A region-based approach is used for representation, query and retrieval of images. It is assumed that the images were already segmented into regions before being indexed. During the indexing operation, each region is uniquely associated with a feature vector, f_i, representing low-level features such as colour, texture and shape. During query formulation, the user chooses textual terms from the thesaurus representing the desired concepts, and then selects, for each term, one of the visual regions already associated with the term to be used as the example during the content-based query.

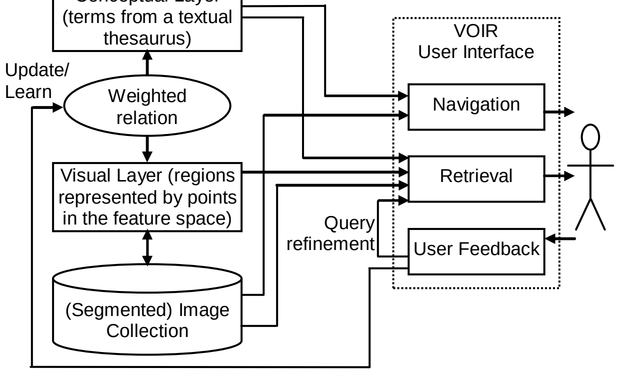
Conceptual Layer (terms from a textual thesaurus)
Update/ Learn
Weighted relation
Visual Layer (regions represented by points in the feature space)
(Segmented) Image Collection
VOIR User Interface
Navigation
Retrieval
Query refinement
User Feedback

Figure 1. Overview of the VOIR

Low-level features and conventional distance functions, usually, are not sufficient to support the correct discrimination of conceptual similarity between distinct visual regions.

VOIR framework implements a two-layer model separating conceptual categories at the upper layer from the visual layer composed by the low-level feature points. The visual layer is partitioned into visual categories, V_j. Each conceptual category, C_i, can be related with several visual categories. Each visual category is composed of several regions. The regions sharing the same visual category are conceptually and visually similar. The use of a textual thesaurus reduces inconsistency in term assignment and provides a knowledge structure that can be explored during the searching process.

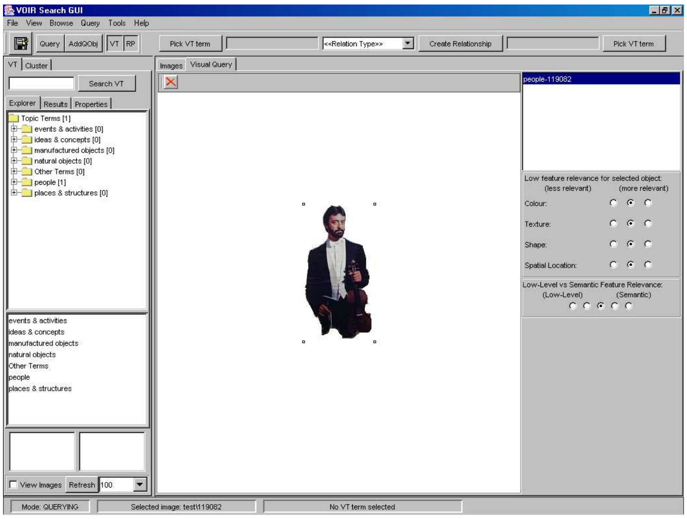

Figure 2. Snapshot of VOIR GUI interface for query composition

3.2 *Region-based Relevance Feedback*

The region-based relevance feedback information provided by the user supports refinement of the results and, additionally, is used to improve the behaviour of the image retrieval system in subsequent sessions. In the latter situation, the system is said to be evolving over time since it is learning the correct associations between terms and regions.

In each query session, the system implements a relevance feedback mechanism that attempts to move the query point towards the good points and away from the bad points. It also attempts to reweigh the query so as to increase the weight of the more discriminating features. These two methods have been used elsewhere (Rui, Huang, Ortega, & Mehrotra, 1998). The novelty of our approach is that, instead of limiting the number of query points to just one, it can expand the query by using additional query points in the feature space that are related with the same conceptual category.

When a new relevant example f_i is indicated by the user, a Boolean function will indicate if the designated point belongs to the same visual category of the evaluated visual item f_j or not. If true, the new point will be considered as one more positive point of the evaluated item. If false, this point will be considered as the seed of another visual category to be added to the current query.

The current implementation of the mentioned function, essentially compares distances $D_{ji} = distance(f_j, f_i)$ and $D_{jk} = distance(f_j, f_k)$ where $f_k \in \mathrm{F_K}$ the set of all visual items whose category C_k is different of the category C_i of point f_i. Basically the query expansion is done if $(D_{ji} / D_{jk}) > thr$, where thr is a predefined threshold level.

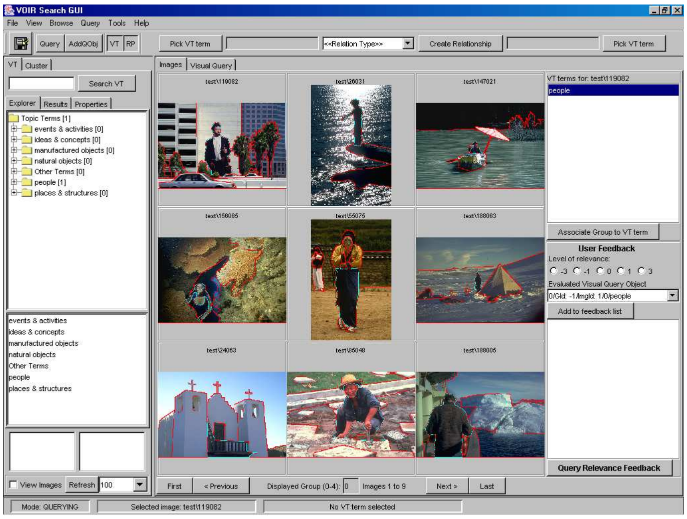

Figure 3. Snapshot of VOIR GUI interface for result set display and relevance feedback

3.3 *Learning term-region associations*

The association between terms and regions is characterized by having a normalized degree of confidence d_conf where the attribute $d\_conf \in [0, 100]$. This association is of fundamental importance since it constitutes the outcome of the process of concept learning. It can be done manually or automatically. In the first case d_conf is set to its maximum value (100), in the second case it will be defined or updated algorithmically.

The critical evaluation of the image results by the user during query sessions is used to create or update the existing associations. The outcome of this is that the system gradually learns associations between visual regions and labels from the textual thesaurus. The more the system learns, the more accurate and faster are the subsequent query sessions.

In the implementation used to carry out the experiments, the visual categories, used in the concept learning process, were defined off-line using a clustering algorithm that took low-level features extracted from each region as its input data. The automatic updating of the associations between term and visual item is done periodically after the query sessions or following new manually added associations. The updating process affects all the visual items that belong to the same visual category as the visual item whose situation was changed either because was explicitly associated with a keyword or because was evaluated during a query iteration.

4 EXPERIMENTS

Although there are actually diverse image datasets annotated in electronic format, virtually all are “per image”, i.e., the annotated words are associated with the whole image and the images in the collection are not segmented. In fact, due to the manual effort required to annotate large segmented image collections, indexing each region separately (Duygulu, 2003), the available collections are, typically, not very large.

The collection used was a database containing “ground-truth”, human-authored image segmentations made available for research use (Martin, Fowlkes, Tal, & Malik, 2001). It is composed of 300 images from the Corel dataset all labeled according to diverse categories such as animals, plants, people or landscape earth features. The total number of image segments is around 3100 representing an average of approximately 10 regions per image. The number of different keywords used in the categorization was 327, and each image has 4 or 5 different keywords associated.

The experiments were conducted using the VOIR system using, as textual thesaurus, the Australian Pictorial Thesaurus (Kingscote, 2003). From each segmented region, during the indexing process, are automatically extracted a collection of numerical properties.

The low-level descriptors and correspondent similarity measures used for the relevance feedback process were:

- L*a*b* Color Histogram (180-bin); histogram intersection.
- Edge Histogram descriptor adapted from the correspondent MPEG-7 descriptor (Manjunath, Salembier, & Sikora, 2002) (80-bin); histogram absolute difference.
- Shape descriptors vector composed by: proportion of image covered by the region, circularity, principal axis, six first invariants of the region central moments (9-dimension vector); euclidean distance.

To perform the clustering of existing regions, used in the process of learning term-region associations, a feature vector of dimension 13 is used. For each region the following features are computed:

- RGB color space: mean and standard deviation of each component (6 values);
- L*a*b* color space: mean and standard deviation of each component (6 values);
- Size: region or group relative size (1 value).

For the clustering algorithm, the value preselected for parameter k was 400, i.e. the number of obtained clusters is pre-defined as being 400. This value was chosen according to the image collection used in the tests. The total number of textual terms from the test collection was 327 and most of those terms were associated with just one region. The (term, region) relation is of type many-to-many, since it is possible for one term to be associated with several regions and, conversely, a region can be associated with several thesaurus terms.

4.1 *Experimental Procedure*

The frequency of the textual terms in the collection is diverse and, for instance, while the term *water* appears in 65 images (TF=65), about 240 terms occur in just one image. Due to this fact, the 37 most frequent single terms were grouped in five intervals (T1_I1 to T1_I5 from Table 1) having distinct frequencies. The same approach was applied to pairs and triplets of terms (Table 2). This has given origin to three classes of queries: queries with one (CL1), two (CL2) and three (CL3) terms.

Table 1. The 37 most frequent terms divided in 5 intervals.

Interval	Terms	Images (N)
T1_I1 (people, sky, water, trees, grasses, rocks)	6	N ≥ 30
T1_I2 (birds, clouds, buildings, landscapes,	5	30 > N ≥ 15
T1_I3 (cats, mountains, boats)	3	15 > N ≥ 12
T1_I4 (horses, flowers, bears, roads, mammals,	6	12 > N ≥ 9
T1_I5 (pyramids, tigers, churches, fish, …)	17	9 > N ≥ 6

For query class CL1, two distinct terms from each of the intervals T1_I1 to T1_I5 were selected. For each of the 10 terms selected, one region was randomly chosen as the query region (instant 1 in Table 3). In the query experiments for the class CL2 was considered one pair from each of the first five intervals T23_I1 to T23_I5 (instants 2-6). For the class CL3 were considered three triplets from the interval T23_I5 (instants 7-9).

Table 2. Frequency distribution for pairs and triplets of terms.

Interval	Pairs of terms	Triplets of terms	Images (N)
T23_I1	4	-	18 ≥ N ≥ 15
T23_I2	3	-	15 > N ≥ 12
T23_I3	2	-	12 > N ≥ 9
T23_I4	17	-	9 > N ≥ 6
T23_I5	66	12	6 > N ≥ 3

The kind of experiments carried out intent to measure fundamentally the following three aspects: the performance of the image retrieval system during a specific interaction session with one user (A1); the impact of the relevance feedback into the quality of the results presented during a specific interaction session with one user (A2); and the accuracy of the long-term (inter query) learned term-region associations (A3). Due to space constrains, this paper is focused in presenting the results obtained for aspect A3. The first two aspects where covered elsewhere (Torres, Hutchison, & Reis, 2007) .

An automatic evaluation system has been arranged to simulate a real user willing to cooperate with the system, i.e., giving to the system the maximum amount possible of positive feedback with respect to the first NR=30 image results delivered by VOIR in each iteration. Given the queries selected to perform the experiments, the evaluation framework, for each query iteration and the returned result set, selects the relevant regions within that result set and feeds that information back to the VOIR system. The process is done automatically after the first formulation of the query.

4.2 *Results*

Table 3 summarizes the results of the long-term learning along the experiment for *d_conf >= 30* (threshold of 30) for the 15 terms of the collection that were used in the several test queries. The column corresponding to instant 9 gives a snapshot of the state of knowledge of the system after that period of usage defined by the whole test.

For Table 3, *occur* means the real number of associations between that term and regions, *pred* the number of predicted associations between terms and regions, and *tp* (true positives) the number of correctly predicted associations by the learning algorithm.

From the 15 terms, only the first 10 were used in the queries with one term, executed before instant 1. Consequently, the other 5 terms have values of zero for instant 1. The term “bridges”, for instance, is used solely in the last query and consequently, only in instant 9 has prediction values different of zero. The term “water” was the one used in more queries (6 queries). The term “trees” used in 5 queries, “people” was used in 4 queries, “birds” and “boats” were used in 2 queries, and the terms “bears”, “bridges”, “clouds”, “flowers”, “hats”, “horses”, “mountains” “pyramids”, “sky” and “tigers” were used just in 1 query.

The obtained results demonstrate the accuracy of the inter query learning procedure. The difference value (*fp = pred - tp*), representing the *false positives*, is, most of the times, zero or almost zero. Additionally, the term-region associations evolve monotonically and steadily, confirming the effectiveness of the long term learning process.

Table 3. Evolution of the association terms-regions during execution of queries with one, two and three terms .

Term	Occur	Instant									
		1	2	3	4	5	6	7	8	9	
water	115	45	61	61	76	76	87	87	92	92	pred
		45	61	61	76	76	87	87	92	92	tp
trees	125	38	38	58	58	58	58	79	80	83	pred
		38	38	58	58	58	58	78	79	82	tp
clouds	28	9	9	9	9	9	9	9	9	9	pred
		9	9	9	9	9	9	9	9	9	tp
birds	59	24	24	24	24	24	35	35	35	35	pred
		24	24	24	24	24	35	35	35	35	tp
flowers	36	14	14	14	14	14	14	14	14	14	pred
		14	14	14	14	14	14	14	14	14	tp
boats	28	13	13	13	26	26	26	26	26	26	pred
		13	13	13	26	26	26	26	26	26	tp
horses	20	9	9	9	9	9	9	9	9	9	pred
		9	9	9	9	9	9	9	9	9	tp
bears	22	7	7	7	7	7	7	7	7	7	pred
		7	7	7	7	7	7	7	7	7	tp
pyramids	14	8	8	8	8	8	8	8	8	8	pred
		8	8	8	8	8	8	8	8	8	tp
tigers	8	6	6	6	6	6	6	6	6	6	pred
		6	6	6	6	6	6	6	6	6	tp
people	199	0	43	91	91	131	131	131	154	154	pred
		0	43	90	90	130	130	130	153	153	tp
hats	12	0	0	0	0	6	6	6	6	6	pred
		0	0	0	0	6	6	6	6	6	tp
mountains	25	0	0	0	0	0	0	12	12	12	pred
		0	0	0	0	0	0	12	12	12	tp
sky	100	0	0	0	0	0	0	35	35	35	pred
		0	0	0	0	0	0	35	35	35	tp
bridges	9	0	0	0	0	0	0	0	0	4	pred
		0	0	0	0	0	0	0	0	4	tp

5 CONCLUSIONS

This paper presents VOIR an image retrieval system based on relevance feedback and long-term concept learning. User interaction allows the refinement of the current search results.

Moreover, the interaction information is used in order to build a conceptual description of the segmented image collection. This description is updated across sessions and combined with the content-based similarity. The reported experiments, focused on the long-term learning aspect, and the quality assessment bear out the efficacy of the proposed method.

6 REFERENCES

Barthes, R. (1977). Rhetoric of the Image. In R.Barthes (Ed.), Image, music, text / trans. by Stephen Heath (pp. 32-51). London: Fontana.

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using Expectation-Maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 1026-1038.

Duygulu, P. (2003). Translating images to words : A novel approach for object recognition. PhD Middle East Technical University, Dept. of Computer Engineering.

Eakins, J. P. (2002). Towards intelligent image retrieval. Pattern Recognition, 35, 3-14.

Fei, L., Qionghai, D., Wenli, X., & Er, G. (2008). Multilabel Neighborhood Propagation for Region-Based Image Retrieval. Multimedia, IEEE Transactions on, 10, 1592-1604.

Giorgio, G. (2007). A nearest-neighbor approach to relevance feedback in content based image retrieval. In.

Jie, Y. & Qi, T. (2008). Semantic Subspace Projection and Its Applications in Image Retrieval. Circuits and Systems for Video Technology, IEEE Transactions on, 18, 544-548.

Jing, F., Li, M., Zhang, H. J., & Zhang, B. (2004). Relevance Feedback in Region-Based Image Retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 14, 672-681.

Kingscote, A. (2003). The Australian Pictorial Thesaurus 2 years on. In DC-ANZ Metadata Conference Australian National University, Canberra.

Ko, B. & Byun, H. M., I (2005). FRIP: a region-based image retrieval tool using automatic image segmentation and stepwise Boolean AND matching. 7, 105-113.

Manjunath, B. S., Salembier, P., & Sikora, T. (2002). Introduction to MPEG-7, Multimedia Content Description Interface. John Wiley & Sons Ltd.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its applications to evaluating segmentation algorithms and measuring ecological statistics. In Proc. IEEE 8th Int. Conf. Computer Vision (pp. 416-423). Vancouver, Canada.

Mezaris, V., Kompatsiaris, I., & Strintzis, M. G. (2004). Region-based Image Retrieval using an Object Ontology and Relevance Feedback. EURASIP Journal on Applied Signal Processing, 2004, 886-901.

Pratikakis, I., Vanhamel, I., Sahli, H., Gatos, B., & Perantonis, S. J. (2006). Unsupervised watershed-driven region-based image retrieval. Vision, Image and Signal Processing, IEE Proceedings -, 153, 313-322.

Rahman, M., Bhattacharya, P., & Desai, B. C. (2007). A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques With Relevance Feedback. Information Technology in Biomedicine, IEEE Transactions on, 11, 58-69.

Ritendra, D., Dhiraj, J., Jia, L., & James, Z. W. (2008). Image retrieval: Ideas, influences, and trends of the new age. Computing Surveys (CSUR), 40.

Rui, Y., Huang, T. S., Ortega, M., & Mehrotra, S. (1998). Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 8, 644-655.

Torres, J. M. (2005). Visual Information Retrieval through Interactive Multimedia Queries. PhD Lancaster University - UK.

Torres, J. M., Hutchison, D., & Reis, L. P. (2007). Semantic Image Retrieval Using Region-Based Relevance Feedback. Lecture Notes in Computer Science, LNCS 4398, 193-207.

Wang, T., Rui, Y., & Sun, J.-G. (2004). Constraint Based Region Matching for Image Retrieval. International Journal of Computer Vision, 56, 37-45.

Xiaofei, H., Deng, C., & Jiawei, H. (2008). Learning a Maximum Margin Subspace for Image Retrieval. Knowledge and Data Engineering, IEEE Transactions on, 20, 189-201.

Zhang, R. & Zhang, Z. (2004). Hidden Semantic Concept Discovery in Region Based Image Retrieval. In Proc. of the 2004 IEEE Conf. on Computer Vision and Pattern Recognition (pp. II-996-II-1001).

Zhi-Hua, Z., Ke-Jia, C., & Hong-Bin, D. (2006). Enhancing relevance feedback in image retrieval using unlabeled data. Transactions on Information Systems (TOIS), 24.