Interactive Visual and Semantic Image Retrieval

Joost van de Weijer, Fahad Khan and Marc Masana Castrillo

1 Introduction

One direct consequence of recent advances in digital visual data generation and the direct availability of this information through the World-Wide Web, is a urgent demand for efficient image retrieval systems. The disclosure of the content of these millions of photos available on the internet is of great importance. The objective of image retrieval is to allow users to efficiently browse through this abundance of images. Due to the non-expert nature of the majority of the internet users, such systems should be user friendly, and therefore avoid complex user interfaces.

Traditionally, two sources of information are exploited in the description of images on the web. The first approach, called text-based image retrieval, describes images by a set of labels or keywords [1]. These labels can be automatically extracted from for example the image name (e.g. 'car.jpg' would provide information about the presence of a car in the image), or alternatively from the webpage text surrounding the image. Another, more expensive way would be to manually label images with a set of keywords. Shortcomings of the text-based approach to image retrieval are obvious: many objects in the scene will not be labeled, words suffer from the confusions in case of synonyms or homonyms, and words often fall short in describing the esthetics, composition and color scheme of a scene. However, until recently many image retrieval systems, such as e.g. Google-image search, were exclusively text based.

A second approach to image description is called content-based image retrieval (CBIR). Here users are provided with feedback from an image-query purely based

Joost van de Weijer
Computer Vision Center, Barcelona, Spain, e-mail: joost@cvc.uab.es

Fahad Khan
Computer Vision Laboratory, Linköping University, Sweden e-mail: fahad@cvc.uab.es

Marc Masana Castrillo
Computer Vision Center, Barcelona, Spain, e-mail: marc.masana@cvc.uab.es

on the visual content of images. These methods are better able to describe the scene composition and color scheme of images. However, they suffer from the semantic gap, which is the gap between low-level image features and high level semantics of the image [2]. Features which are popular in such systems range from global color description [3], to texture descriptions [4], to precise shape descriptions [5]. Due to their different nature, CBIR and text-based image retrieval were found to be complementary [1].

Given the complexity of the image retrieval problem, researchers have acknowledged that user feedback should be an integral part of any image retrieval system [2, 6]. Therefore, relevant feedback mechanisms have been a popular research subject in image retrieval. Users are asked to provide the system with some form of feedback, for example by selecting images which match or do not match the target image. The system then reorders the images given the user feedback. Interactive image retrieval provides a way to approach the inherent ambiguities which exist in image retrieval. Furthermore, it allows adapting the results to be user-dependent. Important is that these systems should operate in real-time which excludes the use of complex learning algorithms. From user studies we know that adding interactive feedback significantly improves the efficiency of retrieval systems [6].

In recent years object recognition and scene categorization have made significant advances [7, 8], especially due to the usage of machine learning techniques in combination with a local feature description of images. The combination of highly discriminative features [9], and the bag-of-words framework have resulted in significant progress [7, 10]. Also the introduction of standard benchmark data sets, such as the VOC PASCAL challenge [11], have further contributed to fast developments in the field of object recognition. One could say that these advances have significantly reduced the semantic gap, and state-of-the-art is currently able to automatically label images with semantically labels.

In the image retrieval system, described in this chapter, we investigate the usage of recent developments in object recognition to bridge the semantic gap. We are especially interested to investigate how such high-level information can improve interactive image retrieval. We will apply a bag-of-word based image representation method to automatically classify images in a number of categories. These additional labels are then applied to improve the image retrieval system. Next to these high-level semantic labels, we also apply a low-level image description to describe the composition and color scheme of the scene. Both descriptions are incorporated in a user feedback image retrieval setting. In conclusion, the novelty of our prototype for image retrieval can be summarized as follows:

- Apply bag-of-word based image classification to bridge the semantic gap by automatically labeling images with a set of semantic labels.
- Improve user feedback by allowing the user to select images to resemble the target image according to semantic or esthetic (color composition) content.

The main objective is to show that automatic labeling of images with semantic labels can improve image retrieval results.

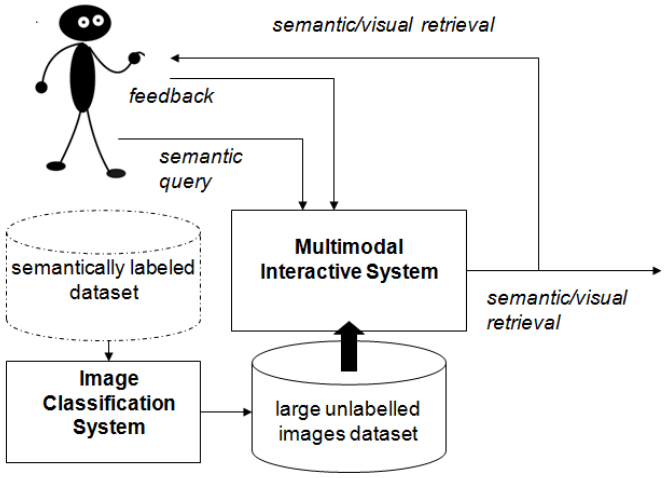

Fig. 1 Overview of image retrieval prototype combining both, semantic and visual queries. Adaptation of general model for multimodal iterative systems ([12]).

This chapter is organized as follows. In Section 2 an overview of our approach is given. In Section 3 the details of the both the semantic and the visual image representation are discussed. In section 4 the technical details and the user interface are discussed. In Section 5 a demonstration of the image retrieval system is given, and Section 6 finishes with concluding remarks.

2 Interactive Visual and Semantic Image Retrieval

A typical user of an image retrieval system is looking for images to use in a presentation, a report, or his webpage. Examples of images could be ”A city-scene during the night”, or ”A living room in retro style”. Communicating the desired image to other humans already can be a difficult task. To facilitate communicating these desires into queries for a computer, we differentiate between two sources of communication: semantic queries in the form of text, and visual queries in the form of images. Text queries, typically allow the user to communicate objects or buildings which should be present in the scene such as ”car”, or ”town” (e.g. Google image search is known to be mainly text-based). Visual queries allow the user to steer the composition, color arrangement and general atmosphere (e.g. cold or warm) of the query.

Due to the inherent ambiguity in the initial query (e.g. different users could envisage different images but use the same query) user feedback will be crucial for successfully navigation of the system. The propose system is given in Figure 1. The user will initialize the system with a text query. Based on an image classification

Fig. 2 Example of combination of visual and semantic query. The semantic query indicates that the user wants a ”horse” to be present in the image, whereas the visual query suggest that the horse should be situated in a green outdoor setting.

system a number of relevant images to the query will be presented to the user in the form of a ranked list. At this time, the user can precise his query by choosing visually and semantically relevant images. Furthermore, the user can leverage the importance of the text and visual query. Based on the combined query the system will re-rank the images and present them to the users. This loop can be repeated until the user is satisfied with the returned results. An example of a combination of visual and semantic query is provided in Figure 2.

In the following we give a more precise overview of our approach. Each image is defined by a visual $\mathbf{d}_v$ and a semantic descriptor $\mathbf{d}_s$ according to

$$\mathbf{d} = [\mathbf{d}_v, \mathbf{d}_s]\,. \tag{1}$$

The semantic query is coded by the vector $\mathbf{q}_s^0$ which is 1 for the classes which are indicated in the semantic query and zero otherwise. Initially the system returns a ranked list according to the following distance equation for all images (indexed by i)

$$\varepsilon_0^i = \mathbf{F}\left(\mathbf{q}_\mathrm{s}^0, \mathbf{d}_\mathrm{s}^\mathrm{i}\right) \tag{2}$$

where $\mathbf{F}$ is a distance function. Throughout this chapter we will use the following distance measure

$$F(\mathbf{a}, \mathbf{b}) = \frac{\sum_i (a_i - \min(a_i, b_i))}{\sum_i a_i} \tag{3}$$

where a_i denotes the element i of vector $\mathbf{a}$. Note that this distance measure is equal to histogram intersection in case of normalized vectors a and b (as we will see this is the case for the visual descriptors $\mathbf{d_v}$). However, it can also be used for unnormalized vectors, which the semantic descriptors $\mathbf{d}_s$ will turn out to be.

Next the user can improve his query by selecting relevant images. The selected images are contained in the set D_r. Given the selected relevant images the user is provided with new results based on the following distance measure

$$\varepsilon^i = \lambda\left(\mathbf{F}\left(\mathbf{q}_\mathrm{s}^0, \mathbf{d}_\mathrm{s}^\mathrm{i}\right) + \frac{1}{|D_r|}\sum_{j\in D_r}\mathbf{F}\left(\mathbf{d}_\mathrm{v}^\mathrm{j}, \mathbf{d}_\mathrm{v}^\mathrm{i}\right)\right) + (1-\lambda)\left(\frac{1}{|D_r|}\sum_{j\in D_r}\mathbf{F}\left(\mathbf{d}_\mathrm{s}^\mathrm{j}, \mathbf{d}_\mathrm{s}^\mathrm{i}\right)\right) \tag{4}$$

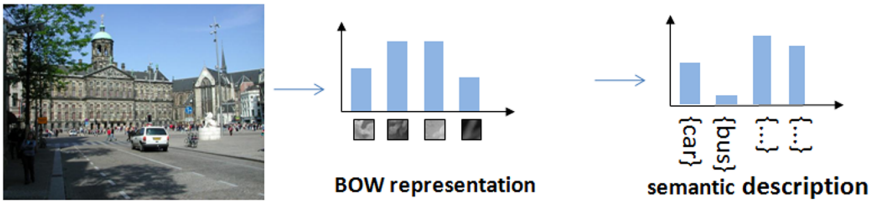

Fig. 3 We propose to use image classification methods based on bag-of-words to automatically label the image with semantic information, which are then applied for semantic image retrieval.

where the parameter λ allows to leverage the relative influence of both cues and will be set by the user. Compared with Eq. 2 this equation has two additional parts, corresponding to the semantic and the visual distance to the relevant image set D_r. In the following section we explain how the visual description $\mathbf{d}_v$ and the semantic description $\mathbf{d}_s$ are computed. The images with lowest distance ε^i to the query are presented to the user for further evaluation.

3 Image Representations

In this section we shortly describe the two image representation methods which are used to describe the semantic and the visual content of the image.

3.1 Semantic Image Representation

In recent years object recognition has advanced significantly. As a direct consequence the semantic gap which exists between low-level image features and high-level semantic content of the images has been narrowed. The main idea is to use image classification methods to automatically label the image with semantically relevant labels (see Figure 3). It is important to note that our approach differs from existing bag-of-word based image retrieval methods (e.g. [13]), in that these methods do not transform the histogram into semantic classes. In this section, we shortly describe our approach to semantic image representation. In particular, we will discus in detail how we combined several cues, in particular shape and color, into a single image representation. More details on our bag-of-words implementation can be found in [14, 15].

The bag-of-words approach which represents an image as a histogram of local features is currently the most successful approach for object and scene recognition [10, 9, 7, 8]. The approach works by constructing a visual vocabulary of local features after which a histogram is built by counting the occurrences of each visual

Method	Cue Binding	Cue Weighting	Scalability
Early Fusion	Yes	No	Yes
Late Fusion	No	Yes	Yes
Color Attention	Yes	Yes	No
Portmanteau	Yes	Yes	Yes

Table 1 Overview of properties for several methods to combine multiple cues into the bag-of-word framework. Only Portmanteau vocabularies combine all three desirable properties. See text for discussion of table.

word in an image. The histogram is then used to train a classifier. Consequently, given a test image the classifier is used to predict the category label of the image.

Introducing multi-modality, i.e. multiple cues, in bag-of-words image representations is an active field of research. Existing approaches used to combine color and shape information often provide below-expected results on a wide range of object categories. The inferior results obtained might be attributed to the way color is incorporated. Traditionally, there exist two approaches to combining color and shape features. The first approach, termed early fusion, combines color and shape features locally before the vocabulary construction stage. Therefore this representation has the *cue binding* property, meaning that the cue information is combined at the same location in the image. The second approach, called late fusion, combines the two visual cues after the vocabulary construction stage. In late fusion, separate visual vocabularies are constructed for color and shape and the two representations are then concatenated to construct the image representation. This representation lacks cue-binding, but possesses *cue weighting*, meaning that the relative weight of the cues can be balanced.

Recently, a method for combining multiple features, called color attention [16], has been introduced. This method combines both cue binding and cue weighting into a single representation. However, a disadvantage of this representation is that it does not possess *scalability* with the number of categories. Therefore, this method is not suitable for large class problems. An overview of the properties of the several methods to combine various cues in bag-of-words is given in Table 1. In the following, we will shortly describe the Portmanteau representation which we apply in our prototype [17]. This representation combines the desired properties cue binding, cue weighting and scalability.

A straightforward method to obtain the binding property is by considering a product vocabulary that contains a new word for every combination of shape and color terms. Assume that $S = \{s_1, s_2, ..., s_M\}$ and $C = \{c_1, c_2, ..., c_N\}$ represent the visual shape and color vocabularies, respectively. Then the product vocabulary is given by

$$\begin{aligned} W &= \{w_1, w_2, ..., w_T\} \\ &= \{\{s_i, c_j\} \mid 1 \leq i \leq M, 1 \leq j \leq N\}, \end{aligned} \tag{5}$$

where $T = M \times N$. A drawback of the product vocabularies is that they result in a very large vocabulary size. As result this yields inefficient image representation,

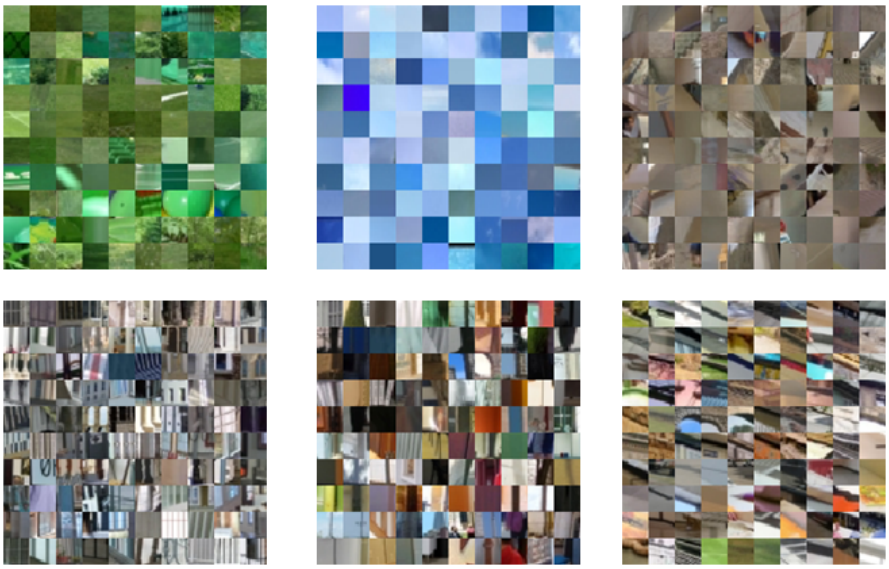

Fig. 4 Example of Portmanteau vocabulary: six different clusters are shown for the SUN data set, where every cluster is represented by 100 randomly sampled patches which are assigned to the cluster. Some clusters show constancy over color, whereas others are more constant over shape.

and in addition it is often difficult to obtain sufficient training data to prevent overtraining. Because of these drawbacks, compound product vocabularies have, not been pursued in literature. However, in recent years, several algorithms have been proposed which compress large vocabularies into small ones [18, 19]. Portmanteau vocabularies [17] are constructed by applying these algorithms to reduce the size of the product vocabularies. As a results we obtain a compact, multi-cue image representation. The algorithm joins words which have similar discriminative power over the set of classes in the image categorization problem. An example of the Portmanteau vocabulary for the SUN data set [20] is shown in Fig.4

In conclusion, we use the Portmanteau image representation in our prototype because it combines multiple cues, namely color and shape, it is compact, and it was shown to obtain state-of-the-art results. Having the Portmanteau representation of the images we learn a SVM classifier with intersection kernel to label the images with the probabilities over a set of class labels. These probabilities constitute the semantic description $\mathbf{d}_s$ of the image. When we compare different semantic descriptors with Eq. 3 we see that this has the desired property that the presence of objects which are not in the query does not increase the distance. However, the absence of objects which are in the query does increase the distance.

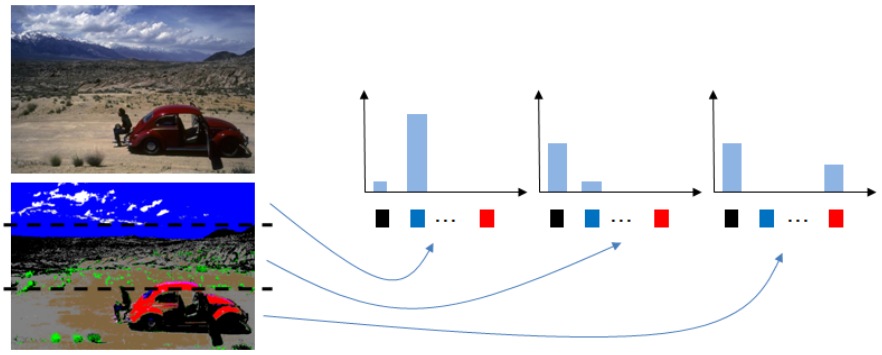

Fig. 5 The bottom image shows the color name assignment for the input image (superimposed lines indicate the three parts which are used to construct the final representation). For the visual representation of the image we concatenate the color histogram, over the eleven basic color terms of the English language, for the bottom, middle and top of the image.

3.2 Visual Image Representation

Here we describe the visual image representation which is applied in the prototype. The aim of the visual representation is to capture both the color sensation and the composition of the image.

For the color description of the image we use color names. Color names are linguistic terms which humans use to communicate colors, such as 'red', 'green' and 'blue'. We use the eleven basic color names of the English language, which are black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow. The mapping from RGB values to a probability over color names was learned from Google images (for a detailed description see [21][22]).

Color names have the advantages that they are intuitively understandable to humans and they provide a very compact color description of an image. Furthermore color names possess a certain degree of photometric invariance, since many different shades of green are all captured by the single color name 'green'. In addition, color names also describe the achromatic content of image, by using the color names 'black', 'grey' and 'white'. This information is normally lost when working with photometric invariants such as *hue*, and normalized *RGB*. Because of these properties, color names were found to be excellent color descriptors [22][23].

In addition, we use a weak composition descriptor image, by computing separate histograms over the color names for the bottom, the middle and the top of the image. This is similar to the spatial pyramids of Lazebnik [24] but was found to obtain better results. An overview of our visual image representation is given in Fig. 5. The final representation $\mathbf{d}_v$ is only 33 bins, i.e. a concatenation of the colors in the bottom, middle, and top image represented in the eleven color names.

4 Image Retrieval Application

In this section we provide the technical details of our system, and explain the user interface of our system.

4.1 Technical Implementation

The main challenges when implementing large scale image retrieval are user interaction and reaction speed. In our case, we need to provide visual feedback in the form of preferred images. Also, the user should be able to balance the strength of both semantic and visual queries which will result in a refined new query which can be repeated for further improvement.

To provide our system with these features, it is accessed with a web browser which is implemented in HTML. In addition to HTML we use PHP for the interactive functionality of the webpage. Furthermore, a combination of javascripts, AJAX and JSON are used to interact and perform computations on the client side. Finally, CSS is used to modify the style to make it more intuitive and pleasant for the user. PHP and HTML are interchangeable within the webpage, allowing the user to interact with the queries.

Each database is implemented into two PHP files. The first one contains the main part of the code to generate the webpage and relevant image ordering. The second one is a module file that contains functions to generate the remaining part of the website, mostly user interaction, but is not involved with the relevant image calculation.

Calculating distances for all images each time a query is performed, consumes a lot of time and slows the system down. To avoid this, pre-calculated distance values are stored which relieves the server side from laborious calculations. Each image has different distance values to all the other images in the database stored. Then, when dealing with a group of relevant images the user has provided, these stored values are loaded resulting in a fast response time. The system has been tested to function with large datasets up to 40,000 images. The system can be tested at the following website: `www.cat.uab.cat/Software/Image_Retrieval/index.php`.

4.2 User Interface

An example of the user interface of our system is given in Figure 6. The user can select a semantic category, in the example 'bicycle' has been selected. The user can further decide the number of images which should be returned (set to 15 in the example). In addition the user can select images which are considered relevant for the query. In the example the user has already selected three bike images on a grass background. Finally, the user can select the relevance of the semantic content

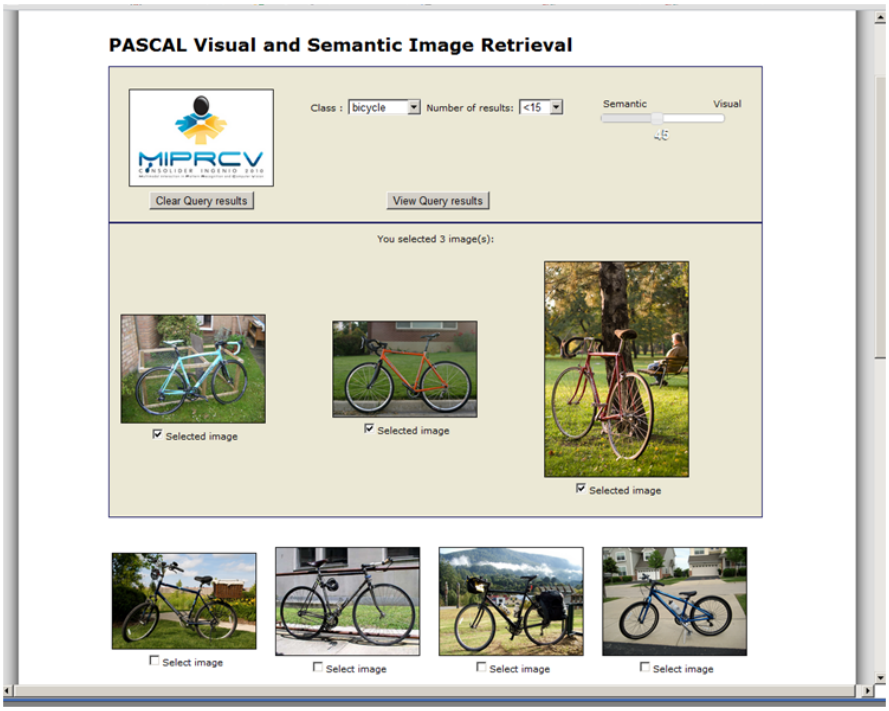

Fig. 6 Interface of the image retrieval system. See text for further explanation.

versus the visual content with a slider in the right top of the interface. Based on these inputs the system will return the most relevant images, four of which are given at the bottom of the example. In the experimental section we will investigate if the semantic image description leads to an improved image retrieval system.

5 Demonstration and Experiment Results

In the introduction we pointed out that the main objective of our image retrieval application is twofold: 1. Apply bag-of-word based image classification to bridge the semantic gap by automatically labeling images with a set of semantic labels, 2. Improve user feedback by allowing the user to select images to resemble the target image according to semantic or esthetic (color composition) content. In this section, we provide two experiments to evaluate these objectives.

To test our image retrieval system we use two large datasets. The PASCAL VOC 2009 dataset consists of 13704 images. The images are divided into 20 different object categories. In our experiment we train on 3473 images and test the retrieval system on 3581 images of the validation set. The SUN dataset consists of 39700 im-

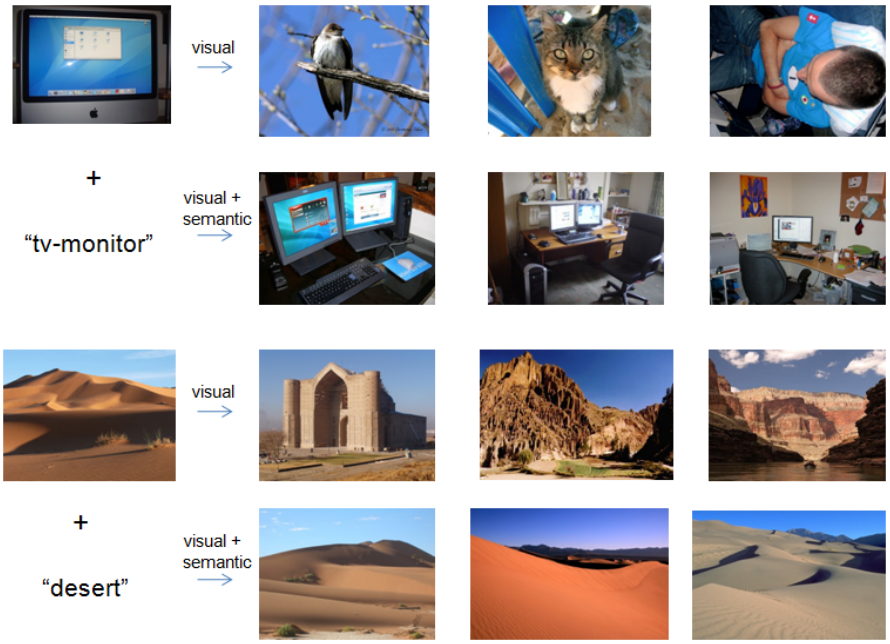

Fig. 7 Retrieval results for two different images. Top image from VOC PASCAL and bottom image from SUN data set. For both images the query is performed twice once only based on visual descriptor, and once on the combined visual and semantic descriptor. Note how the semantic description helps to improve the query results.

ages of 397 different scene categories. The dataset is divided into 19850 training and 19850 test images. Both datasets are difficult owing to large amount of variations both within an object category and across different object categories.

5.1 Semantic Image Description

In the first experiment we aim to evaluate if the semantic image description improves the overall image retrieval results. Two example retrievals which illustrate the importance of the semantic description are provided in Figure 7. To quantify the improvement we performed a small user study. Users were given a target image together with six retrieved images. The six images contained contained two images which were similar only in a visual sense, two in a semantic sense and two which are similar in both visual and semantic description. The images are randomly presented and the user is unaware which algorithm is related to which image. Next, the user is asked to select the image from the six which is most similar to the target image.

Data Set	Visual	Semantic	Visual+Semantic
VOC PASCAL 2009	17%	27%	56%
SUN	24%	34%	42%

Table 2 User preference for 'visual','semantic' or 'visual+semantic' description of images when asked which image is most similar to a target image. Results are provided in percentage of times the user selected the description.

In Table 2 the results of the experiment are summarized. The results are based on a total of ten test person which provided ten preferences each. The visual and semantic description is significantly more often selected than the results returned by visual only. In 56% of the queries on PASCAL and in 42% of the queries on SUN the combined description was preferred. This clearly shows the importance of the semantic description for image retrieval.

5.2 Interactive Visual and Semantic Retrieval

In the second experiment we desire to establish if the semantic description within the interface provided by Figure 1 is beneficial. To evaluate the user interface we designed a user experiment to measure the speed with which a user finds the desired image. Users are asked to find a given target image with the image retrieval system. The test is performed on the PASCAL data set. We compare the retrieval system with only visual image description (V-system) to the system with both visual and semantic information (VS-system). As an evaluation measure we compare the number of target images which were found within X rounds of interaction. Since in the VS-system the user can also balance the relative weight of semantic and visual information, we allow more rounds of interaction to the V-system. We choose X to be ten for the V-system and five for the VS-system.

The results of the experiment are presented in Table 3. Eight subjects have performed the experiment (five searches for the V-system and five for the VS-system). The same random set of images was evaluated by both systems. The results show that about double the amount of images were found by the VS-system, indicating that the additional semantic information does significantly improve the retrieval system. The fact that only eleven out of 40 images queries were found within five interactive rounds reveals that the user interaction can still be improved significantly. Users identified that they would have appreciate an additional feature which allows users to indicate wether the selected image is relevant for its semantic content or for its color and composition.

Data Set	Visual	Visual+Semantic
VOC PASCAL 2009	5	11

Table 3 Number of target images which were found by the system within X rounds of interaction for V-system and VS-system. The parameter X was set to ten for the V-system and to five for the VS-system.

6 Conclusions

In this chapter we have investigated the usage of image classification methods to bridge the semantic gap. We apply image classification to automatically label images with semantic terms. These terms are then used to facilitate image retrieval. Users can start their query with a semantic term, and subsequently improve the query by selecting relevant images. Queries can be considered relevant with respect to their visual or semantic content. The user interface allows users to leverage between these two cues. Initial results are promising and show that semantic queries improve retrieval quality.

As future work we see incorporating automatic semantic labeling in more developed retrieval systems such as RISE [25]. In this case terms which are contributed by the automatic labeling can be handled similarly as terms extracted from the surrounding webpage of images or the filename. Another extension in which we are interested is further improving the semantic labeling by using object detectors [26]. This would allow users to further specify which part of the image they consider relevant for the query. In conclusion, we expect that in the near future object recognition techniques will be an integral part of most image retrieval systems. The gained semantic descriptions of images will improve the quality of the retrieval system. Furthermore, knowledge of the location of the semantic content in the image will open up new ways of improved user feedback.

Acknowledgements We acknowledge the help and advice of Marc Paz in the implementation of this project. We also thank Carles Sánchez for his initial implementation of the system. We acknowledge the Spanish Research Program Consolider-Ingenio 2010: MIPRCV (CSD200700018); and Spanish project TIN2009-14173. Joost van de Weijer acknowledges support of Ramon y Cajal fellowship.

References

1. Ferecatu, M., Boujemaa, N., Crucianu, M.: Semantic interactive image retrieval combining visual and conceptual content description. ACM Multimedia Systems **13** (2008) 309–322
2. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval: the end of the early years. IEEE Trans. on pattern analysis and machine intelligence **22** (2000) 1349–1380
3. Gevers, T., Smeulders, A.: Color based object recognition. Pattern Recognition **32** (1999) 453–464

4. Fernandez, S.A., Salvatella, A., Vanrell, M., Otazu, X.: Low dimensional and comprehensive color texture description. Computer Vision and Image Understanding **116** (2012) 54–67
5. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)
6. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. Multimedia Syst. **8** (2003) 536–544
7. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE conference on Computer Vision and Patter Recognition. Volume 2. (2003) 264–271
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence **27** (2005) 1615–1630
9. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60** (2004) 91–110
10. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 530–534
11. Everingham, M., Gool, L.V., Williams, C.K.I., J.Winn, Zisserman, A.: The pascal visual object classes challenge 2007 results. (2007)
12. Toselli, A., Vidal, E., Casacuberta, F.: Multimodal Interactive Pattern Recognition and Applications. Springer (2011)
13. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. CVPR '06, IEEE Computer Society (2006) 2161–2168
14. Rojas-Vigo, D., Khan, F.S., van de Weijer, J., Gevers, T.: The impact of color on bag-of-words based object recognition. In: Int. Conference on Pattern Recognition (ICPR). (2010)
15. Elfiky, N., Khan, F.S., van de Weijer, J., Gonzalez, J.: Discriminative compact pyramids for object and scene recognition. Pattern Recognition (PR) **45** (2012) 1627–1636
16. Khan, F.S., van de Weijer, J., Vanrell, M.: Modulating shape features by color attention for object recognition. International Journal of Computer Vision (IJCV) **98** (2012) 49–64
17. Khan, F., Van de Weijer, J., Bagdanov, A., Vanrell, M.: Portmanteau vocabularies for multi-cue image representation. In: Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS 2011). (2011)
18. Fulkerson, B., Vedaldi, A., Soatto, S.: Localizing objects with smart dictionaries. In: European Conference on Computer Vision. (2008)
19. Dhillon, I., Mallela, S., Kumar, R.: A divisive information-theoretic feature clustering algorithm for text classification. Journal of Machine Learning Research (JMLR) **3** (2003) 1265–1287
20. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: IEEE conference on Computer Vision and Patter Recognition. (2010)
21. van de Weijer, J., Schmid, C.: Applying color names to image description. In: IEEE International Conference on Image Processing (ICIP), San Antonio, USA (2007)
22. van de Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing **18** (2009) 1512–1524
23. Khan, F., Anwer, R., van de Weijer, J., Bagdanov, A., Vanrell, M., Lopez, A.: Color attributes for object detection. In: IEEE conference on Computer Vision and Patter Recognition. (2012)
24. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE conference on Computer Vision and Patter Recognition. (2006) 2169–2178
25. Leiva, L.A., Villegas, M., Paredes, R.: Query refinement suggestion in multimodal interactive image retrieval. In: Proceedings of the 13th International Conference on Multimodal Interaction (ICMI). (2011) 311–314
26. Felzenszwalb, P.F., McAllester, D.A., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Computer Vision and Pattern Recognition. (2008)